Practical Data Analysis with JMP®

Third Edition

Robert H. Carver

Student Solutions

sas.com/books

The correct bibliographic citation for this manual is as follows: Carver, Robert. 2019. *Practical Data Analysis with JMP®, Third Edition*. Cary, NC: SAS Institute Inc.

Practical Data Analysis with JMP®, Third Edition

ISBN 978-1-64295-614-6 (Hardcover)
ISBN 978-1-64295-610-8 (Paperback)
ISBN 978-1-64295-611-5 (Web PDF)
ISBN 978-1-64295-612-2 (EPUB)
ISBN 978-1-64295-613-9 (Kindle)

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

October 2019

Chapter 2: Solutions to Application Scenarios

Scenario 2

- Quantity of cement (component 1), expressed as kg in a m^3 mixture.
- Quantity of Superplasticizer (component 5), expressed as kg in a m^3 mixture.
- Quantity of Fine Aggregate (component 7), expressed as kg in a m^3 mixture.

Scenario 4

NHANES does not contain experimental data because the experimenters are not manipulating any of the variables and there was no random assignment of treatments. The data was not obtained through a designed experiment but through observation.

Scenario 6

This data table contains monthly stock values and volume from the Nikkei 225 Index, between December 31, 2013 through 1 December 31, 2018. Data were collected by observation on the first day of each month. The date column is continuous because it is a chronological variable. Open, High, Low, Close, Adj Close, Volume, and change% are all Continuous columns containing numeric measurements. Open is the index's opening price. High represents the high price for the day. Low is the low price for that day. Close is the closing price for that day. Volume is the number of shares exchanged during the day. change% is how much the index changed from open to close.

Scenario 8

This table contains observational data from the World Health Organization (WHO) regarding tobacco use, cardiovascular disease and cancer rates. **Code** is a nominal variable uniquely identifying each nation. **Country** is a nominal variable that provides the name of the country relating to the data. **Region** is also a nominal variable indicating the region where the country is located in. **TobaccoUse** is a continuous variable observed describing the prevalence of tobacco use in that country. Female and Male are both continuous variables that were found observationally which describe the prevalence of tobacco use for both genders. **CVmort** is the mortality rate from cardiovascular disease for this country and **CancerMort** is the cancer mortality rate for this country. Both are continuous.

Scenario 10

The columns are as follows:

marst : marital status (nominal). Respondent's marital status, one of six levels

empstat: employment status (nominal). Respondent's employment status, one of five levels.

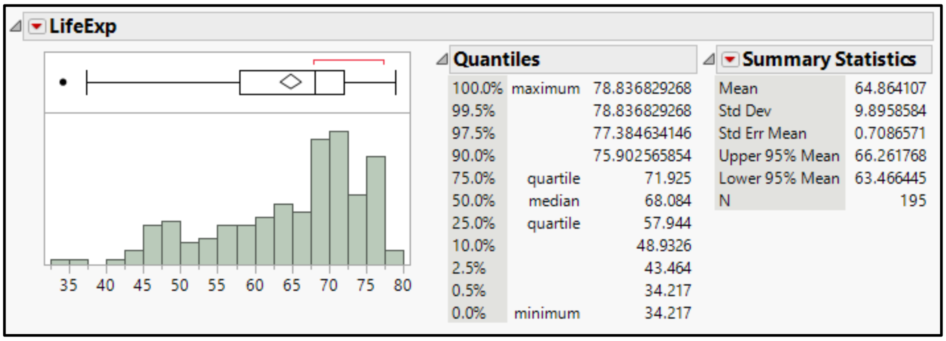
sleeping minutes spent sleeping each day (continuous).

telff minutes spent on the telephone with family and friends each day (continuous).

Chapter 3: Solutions to Application Scenarios

Scenario 2

a. This histogram from 1990 has a shape that is skewed to the left, has a mean of approximately 65, and a spread described by a range from 34 to 79 years. It has a peak near 72 and a lesser peak just above 75. There appears to be one outlier with a life expectancy of only 34.2 years in 1990.



Quantiles

100.0%	maximum	78.836829268
99.5%		78.836829268
97.5%		77.384634146
90.0%		75.902565854
75.0%	quartile	71.925
50.0%	median	68.084
25.0%	quartile	57.944
10.0%		48.9326
2.5%		43.464
0.5%		34.217
0.0%	minimum	34.217

Summary Statistics

Mean	64.864107
Std Dev	9.8958584
Std Err Mean	0.7086571
Upper 95% Mean	66.261768
Lower 95% Mean	63.466445
N	195

b. The five-number summary from 1990 has a minimum of 34.2, a 25% quartile of 57.9, a 75% quartile of 71.9, and a maximum of 78.8 with a median of 68.1.

The 2015 data has a minimum of 51.5, a 25% quartile of 66.4, a 75% quartile of 77.5, and a maximum of 84.3 with a median of 71.96. Clearly, every statistic from the five-number summary has increased indicating life expectancy has gone up across the entirety of the distribution. Both distributions are strongly left-skewed. Over the 25-year time period, the minimum life expectancy increased by approximately 15 years and the maximum by 5 years.

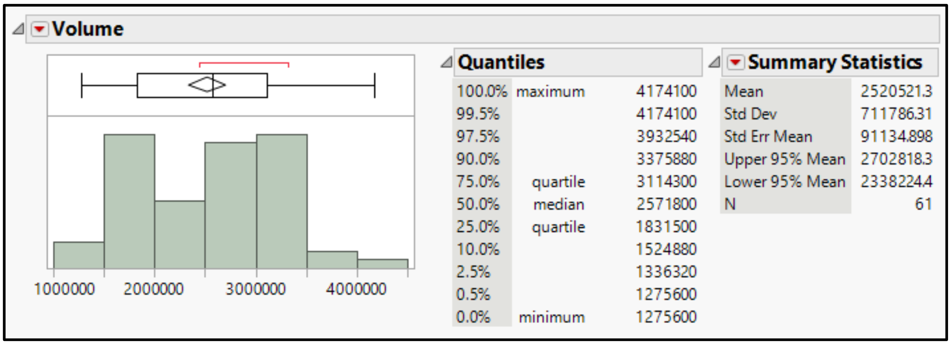
c. The standard deviation is 9.9 years in the 1990 data compared to 7.9 in the 2015 data.

This suggests that variability in life expectancy has decreased across countries.

d. Similar to the 2015 distribution, the mean is less than the median in 1990, which is indicative of a left-skewed distribution.

Scenario 4

a. Volume has a moderately symmetrical and bimodel distribution. It ranges from 1,275,600 to 4,174,100 shares with a median of 2.57 million shares and a mean of 2.52 million shares. There are no outliers.



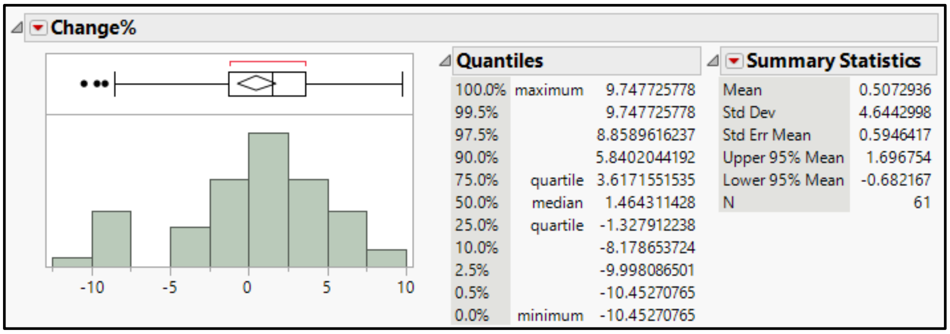
Quantiles

100.0%	maximum	4174100
99.5%		4174100
97.5%		3932540
90.0%		3375880
75.0%	quartile	3114300
50.0%	median	2571800
25.0%	quartile	1831500
10.0%		1524880
2.5%		1336320
0.5%		1275600
0.0%	minimum	1275600

Summary Statistics

Mean	2520521.3
Std Dev	711786.31
Std Err Mean	91134.898
Upper 95% Mean	2702818.3
Lower 95% Mean	2338224.4
N	61

b. Change% has a mildly right-skewed, unimodel shape ranging from -10.45% to 9.74%. Its center can be described by the mean of .5073 and a median of 1.4631. There are three outlying months at the low end.



Change%

Quantiles

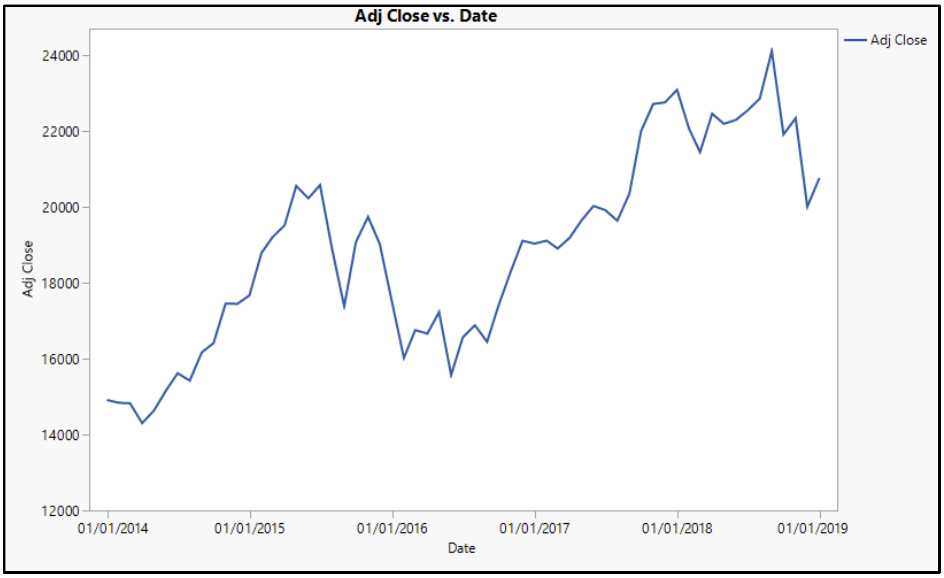
100.0%	maximum	9.747725778
99.5%		9.747725778
97.5%		8.8589616237
90.0%		5.8402044192
75.0%	quartile	3.6171551535
50.0%	median	1.464311428
25.0%	quartile	-1.327912238
10.0%		-8.178653724
2.5%		-9.998086501
0.5%		-10.45270765
0.0%	minimum	-10.45270765

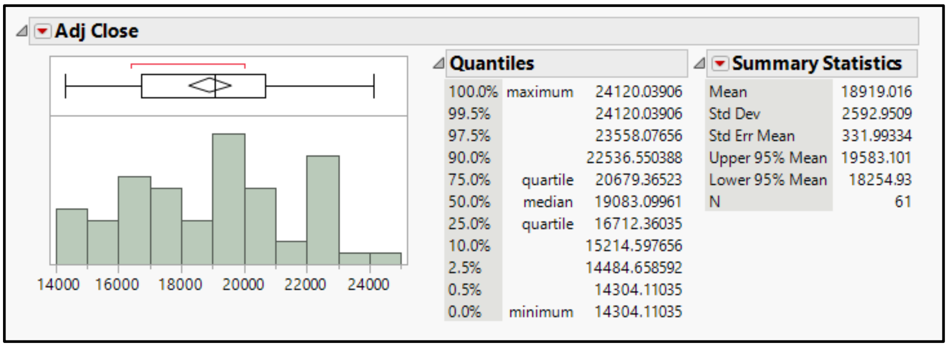
Summary Statistics

Mean	0.5072936
Std Dev	4.6442998
Std Err Mean	0.5946417
Upper 95% Mean	1.696754
Lower 95% Mean	-0.682167
N	61

c. The Nikkei declines somewhere between 25% and 50% of months. By clicking on all histogram bars to the left of 0, we find 24 of 61 rows selected, representing approximately 39% of months.

d. Both graphs clearly show the range of the Adjusted Close variable. The up-and-down growth over time is clear in the line chart, but not in the histogram, where there is no time element. On the other hand, the multiple peaks that are so evident in the histogram are invisible in the line graph.





Adj Close

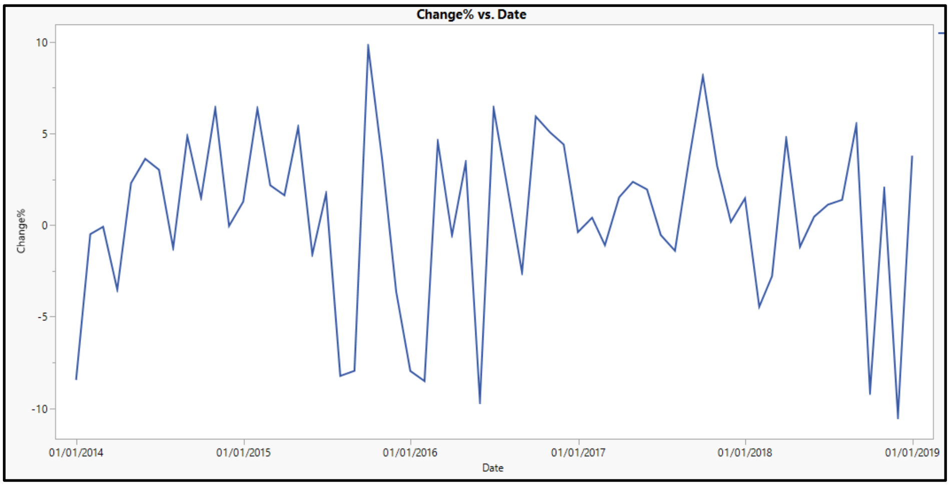
Quantiles

100.0%	maximum	24120.03906
99.5%		24120.03906
97.5%		23558.07656
90.0%		22536.550388
75.0%	quartile	20679.36523
50.0%	median	19083.09961
25.0%	quartile	16712.36035
10.0%		15214.597656
2.5%		14484.658592
0.5%		14304.11035
0.0%	minimum	14304.11035

Summary Statistics

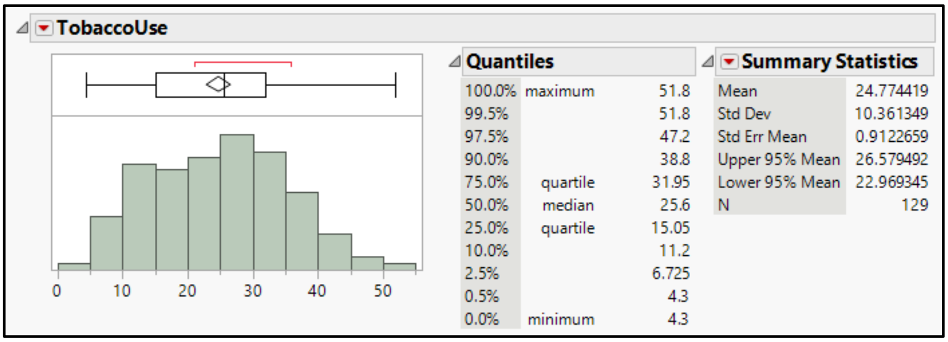
Mean	18919.016
Std Dev	2592.9509
Std Err Mean	331.99334
Upper 95% Mean	19583.101
Lower 95% Mean	18254.93
N	61

e. This line graph shows fluctuation without any obvious pattern. The monthly percentage change seems to vary at random from month to month, typically remaining between – 5% and +5%. There is no obvious growth over the five years, in contrast to the closing index value.



Scenario 6

a. **TobaccoUse** is nearly symmetrical with a mean of 24.77 and median of 25.6. It ranges from 4.3 to 51.8, with no outliers.



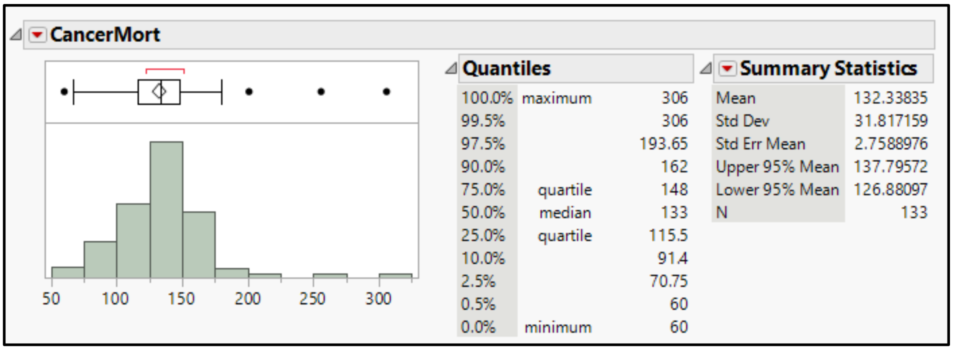
Quantiles

100.0%	maximum	51.8
99.5%		51.8
97.5%		47.2
90.0%		38.8
75.0%	quartile	31.95
50.0%	median	25.6
25.0%	quartile	15.05
10.0%		11.2
2.5%		6.725
0.5%		4.3
0.0%	minimum	4.3

Summary Statistics

Mean	24.774419
Std Dev	10.361349
Std Err Mean	0.9122659
Upper 95% Mean	26.579492
Lower 95% Mean	22.969345
N	129

b. **CancerMort** is mildly skewed to the right with a mean of 132.3 and median of 133. It ranges from 60 to 306, and there is one low-end outlier plus 3 outlying countries to the right.



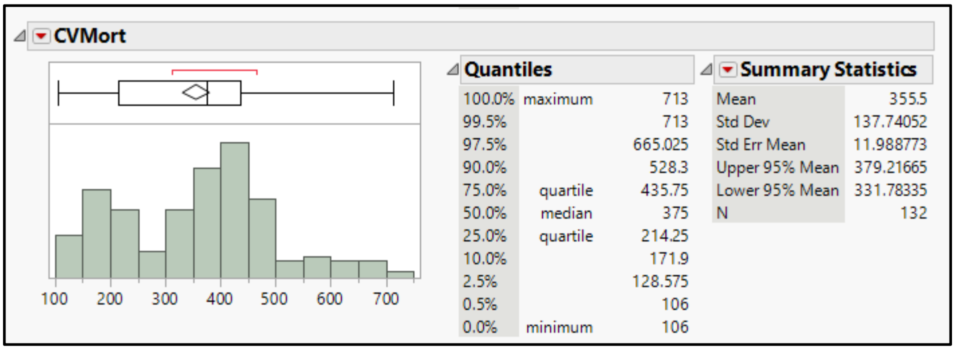
Quantiles

100.0%	maximum	306
99.5%		306
97.5%		193.65
90.0%		162
75.0%	quartile	148
50.0%	median	133
25.0%	quartile	115.5
10.0%		91.4
2.5%		70.75
0.5%		60
0.0%	minimum	60

Summary Statistics

Mean	132.33835
Std Dev	31.817159
Std Err Mean	2.7588976
Upper 95% Mean	137.79572
Lower 95% Mean	126.88097
N	133

c. **CVMort** has two peaks near 150 and 400. It is skewed to the right. It has a mean of 355.5 and a median of 375. It ranges from 106 to 713, with no outliers.



CVMort

Quantiles

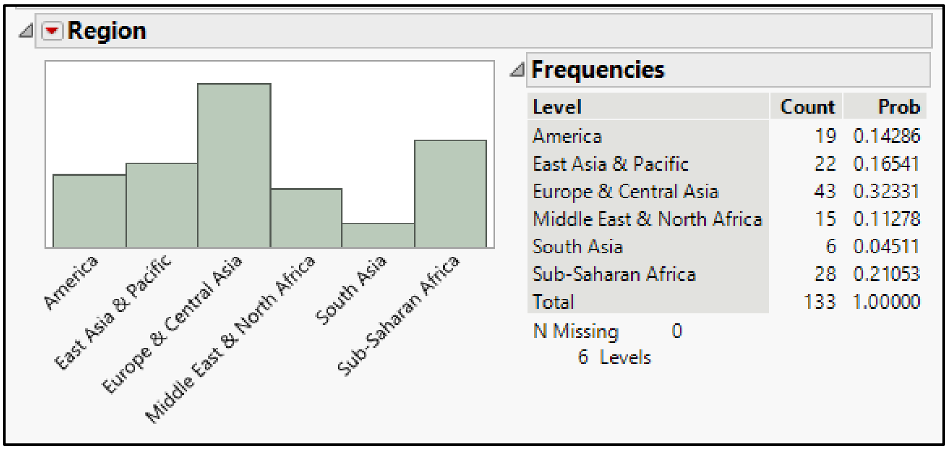
100.0%	maximum	713
99.5%		713
97.5%		665.025
90.0%		528.3
75.0%	quartile	435.75
50.0%	median	375
25.0%	quartile	214.25
10.0%		171.9
2.5%		128.575
0.5%		106
0.0%	minimum	106

Summary Statistics

Mean	355.5
Std Dev	137.74052
Std Err Mean	11.988773
Upper 95% Mean	379.21665
Lower 95% Mean	331.78335
N	132

d. Overall, **TobaccoUse** is more uniform than **CancerMort** and **CVMort**. **CancerMort** has one peak and **CVMort** has two peaks. **CVMort** has the largest range and **TobaccoUse** has the smallest range. **TobaccoUse** is the most symmetrical of the three, while **CancerMort** and **CVMort** are both skewed right.

e. Europe & Central Asia and Sub-Saharan Africa have the highest count of countries in this data table. South Asia has the lowest count and America, East Asia & Pacific and Middle East & North Africa all fall in the middle.



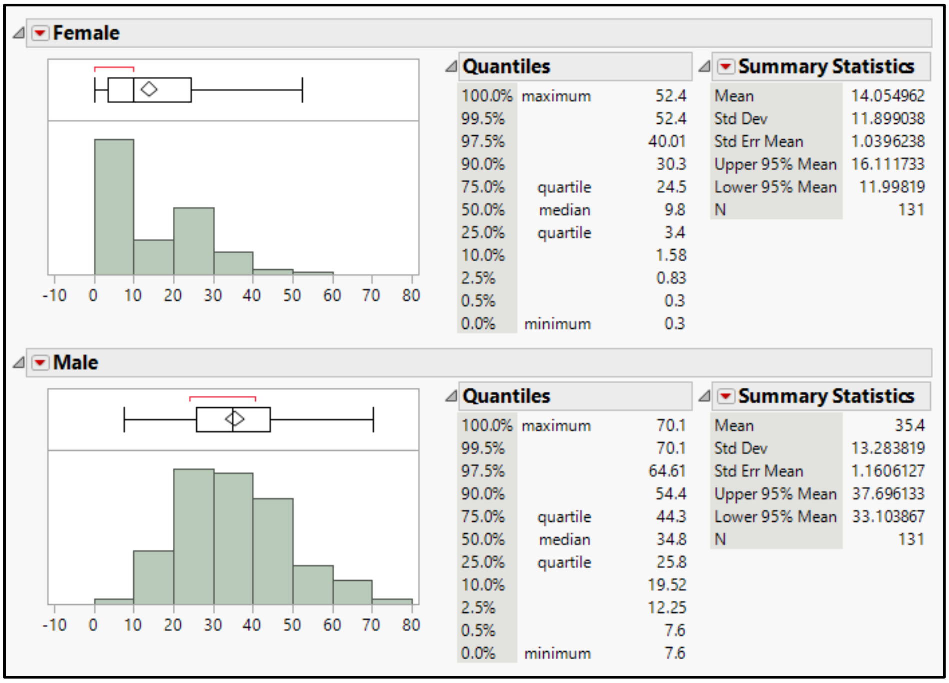
Region

Frequencies

Level	Count	Prob
America	19	0.14286
East Asia & Pacific	22	0.16541
Europe & Central Asia	43	0.32331
Middle East & North Africa	15	0.11278
South Asia	6	0.04511
Sub-Saharan Africa	28	0.21053
Total	133	1.00000

N Missing 0
6 Levels

f. Women generally use less Tobacco than men do. The center for the male distribution is approximately 35 compared to around 10 for women. The distribution for women is strongly right-skewed, with many clustering between 0 and 10. In contrast, the values for men are symmetric and more widely varied.



Female

Quantiles

100.0%	maximum	52.4
99.5%		52.4
97.5%		40.01
90.0%		30.3
75.0%	quartile	24.5
50.0%	median	9.8
25.0%	quartile	3.4
10.0%		1.58
2.5%		0.83
0.5%		0.3
0.0%	minimum	0.3

Summary Statistics

Mean	14.054962
Std Dev	11.899038
Std Err Mean	1.0396238
Upper 95% Mean	16.111733
Lower 95% Mean	11.99819
N	131

Male

Quantiles

100.0%	maximum	70.1
99.5%		70.1
97.5%		64.61
90.0%		54.4
75.0%	quartile	44.3
50.0%	median	34.8
25.0%	quartile	25.8
10.0%		19.52
2.5%		12.25
0.5%		7.6
0.0%	minimum	7.6

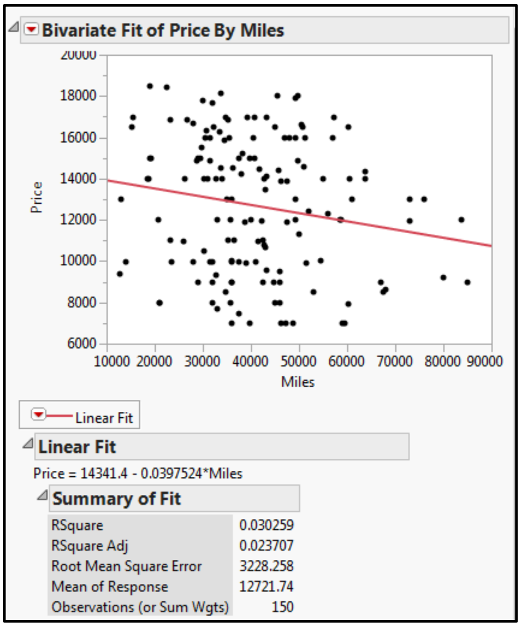
Summary Statistics

Mean	35.4
Std Dev	13.283819
Std Err Mean	1.1606127
Upper 95% Mean	37.696133
Lower 95% Mean	33.103867
N	131

Chapter 4: Solutions to Application Scenarios

Scenario 2

a.



Bivariate Fit of Price By Miles

Linear Fit

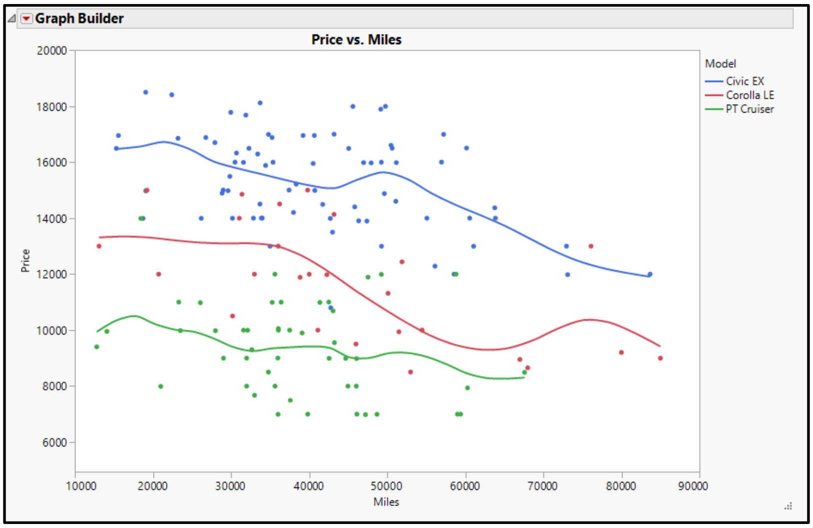
Price = 14341.4 - 0.0397524*Miles

Summary of Fit

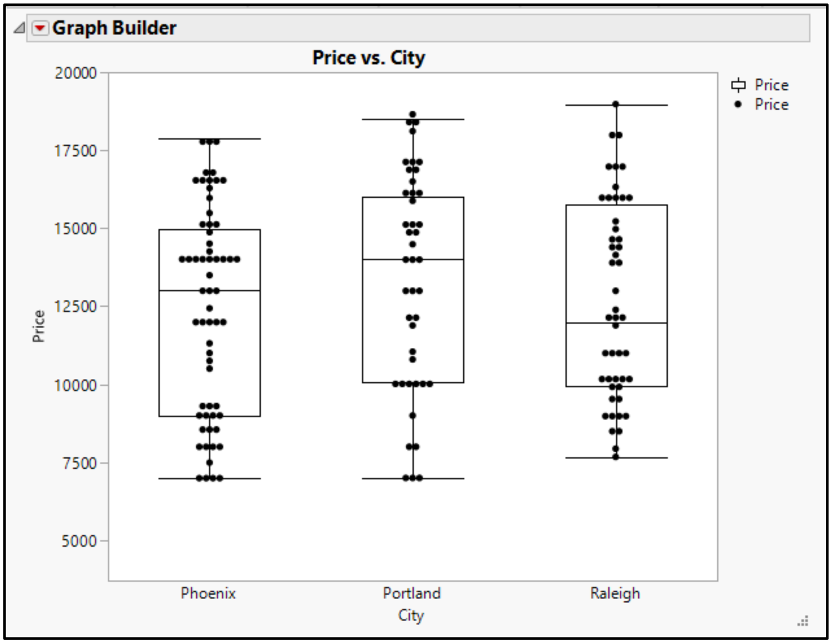
RSquare	0.030259
RSquare Adj	0.023707
Root Mean Square Error	3228.258
Mean of Response	12721.74
Observations (or Sum Wgts)	150

The plot, equation and Rsquare are shown above. It is not obvious that there is a linear pattern at all. The correlation coefficient is –0.17395. There is a weak negative relationship between mileage and price: the higher the mileage, the lower the price.

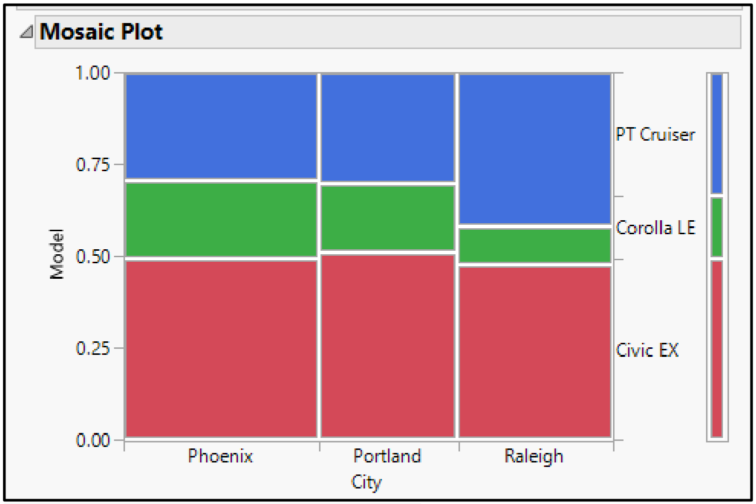
b. Below is one graph with car model in the Overlay zone of the graph building. For all models the relationship appears to be negative (downward sloping), but least so for the PT Cruiser: additional miles do not reduce prices as much as for the other models.

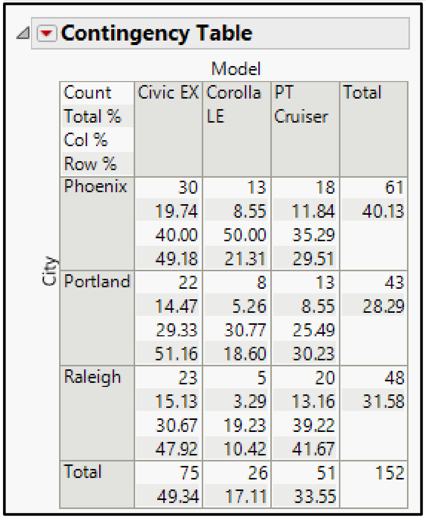


c. The distribution of price across the three cities is similar. The box plot shows similar middle 50% with varying medians, with Portland having the highest median price. They also have very similar spreads.



d. The models are not equally favored across cities. For example Corollas are more popular in Phoenix than the other two cities. The third line in each cell of the contingency table says what percentage of that certain model are located in each city.



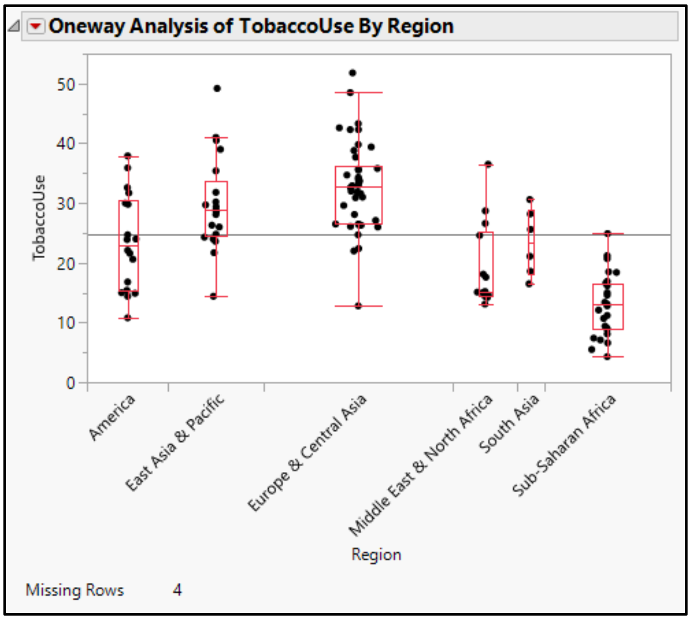


Contingency Table

		Model			
	Count Total % Col % Row %	Civic EX	Corolla LE	PT Cruiser	Total
City	Phoenix	30 19.74 40.00 49.18	13 8.55 50.00 21.31	18 11.84 35.29 29.51	61 40.13
	Portland	22 14.47 29.33 51.16	8 5.26 30.77 18.60	13 8.55 25.49 30.23	43 28.29
	Raleigh	23 15.13 30.67 47.92	5 3.29 19.23 10.42	20 13.16 39.22 41.67	48 31.58
	Total	75 49.34	26 17.11	51 33.55	152

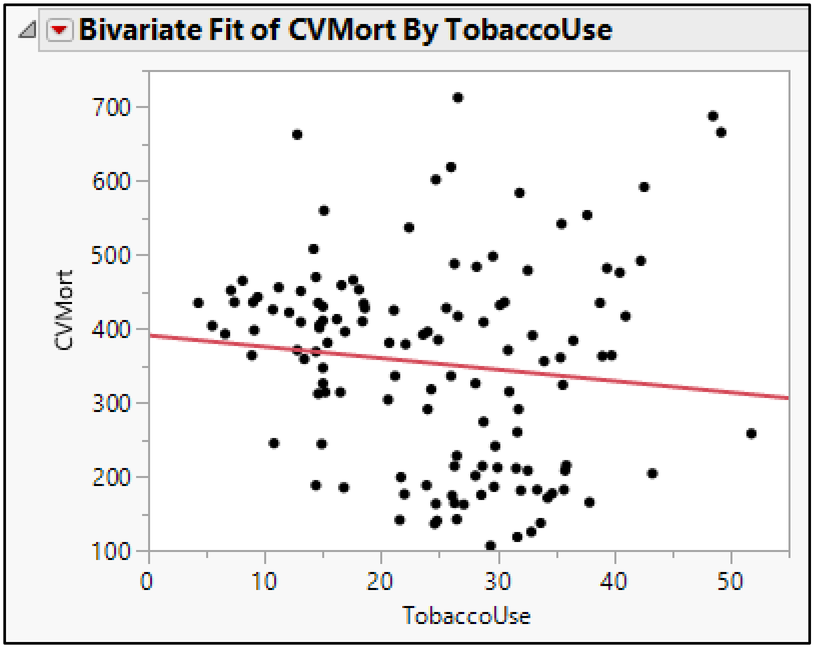
Scenario 4

a. Tobacco is most heavily used in Europe & Central Asia and to a lesser extent in East Asia and the Pacific. There is a moderate use in the Middle East and North Africa as well as the Americas while Sub-Saharan Africa has the lowest tobacco use.

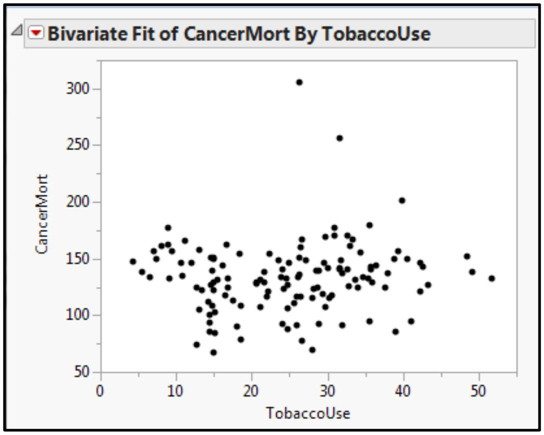


Missing Rows 4

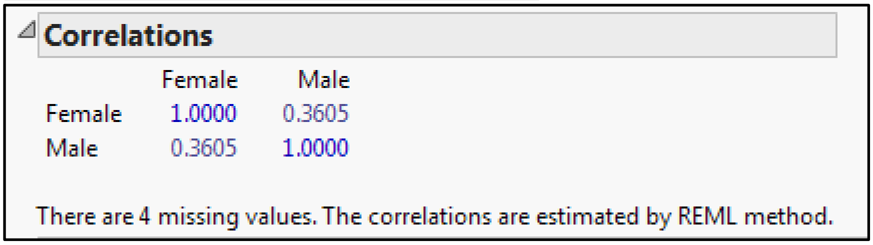
b. There does not seem to be any strong linear relationship to the two variables in this sample.



c. Here again, we find scant evidence of a relationship.



d.



Correlations

	Female	Male
Female	1.0000	0.3605
Male	0.3605	1.0000

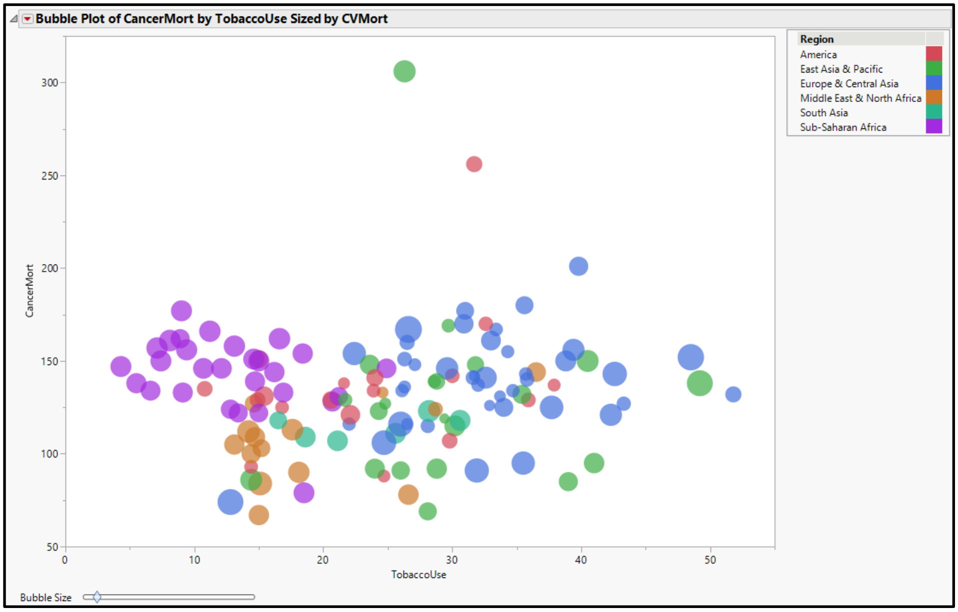
There are 4 missing values. The correlations are estimated by REML method.

The correlation of .3605 is very weak between male and female tobacco use.

e. The bubble plot below indicates that the relationship between tobacco use and cancer mortality may vary by region.

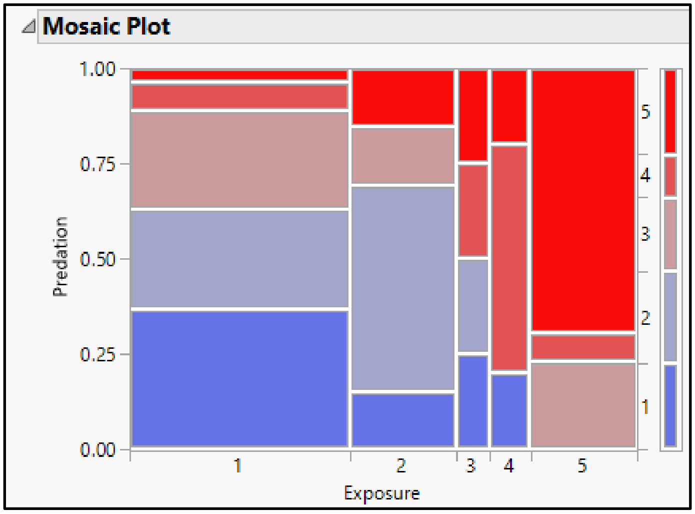
We see the sub-Saharan African nations clustered together on the left side, showing little or no relationship between tobacco use and cancer mortality. The nations of Europe and central Asia may display a weak positive relationship, as do those in the East Asia and Pacific region. In the latter group, Mongolia is a clear outlier with an exceptionally high rate of cancer mortality.

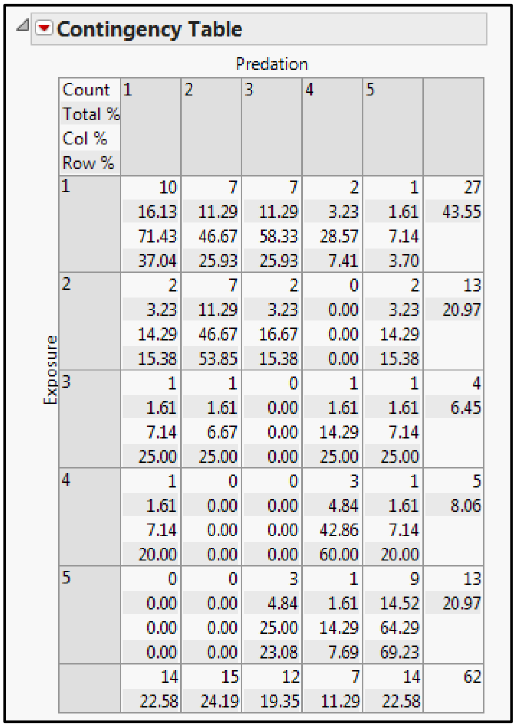
The bubble sizes (CV mortality) do not shed much light.



Scenario 6

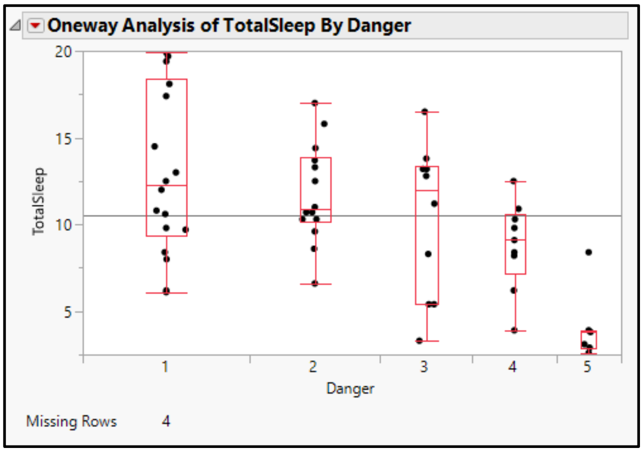
a. Animals with lower exposure values seem to have lower predation ratings. Conversely, creatures with higher exposure values also had higher predation ratings.




Contingency Table

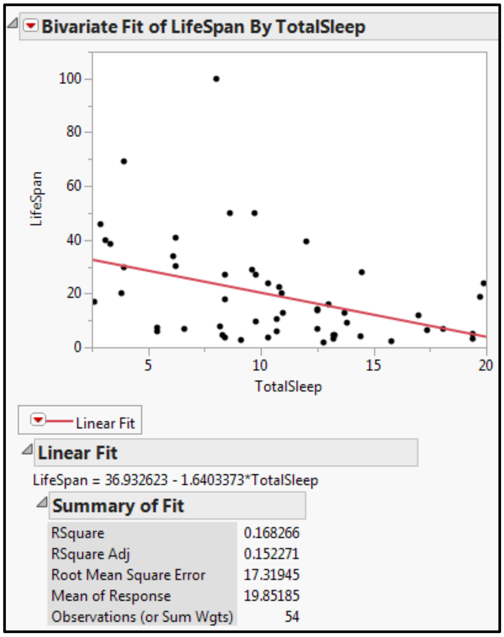
		Predation					
	Count Total % Col % Row %	1	2	3	4	5	
Exposure	1	10 16.13 71.43 37.04	7 11.29 46.67 25.93	7 11.29 58.33 25.93	2 3.23 28.57 7.41	1 1.61 7.14 3.70	27 43.55
	2	2 3.23 14.29 15.38	7 11.29 46.67 53.85	2 3.23 16.67 15.38	0 0.00 0.00 0.00	2 3.23 14.29 15.38	13 20.97
	3	1 1.61 7.14 25.00	1 1.61 6.67 25.00	0 0.00 0.00 0.00	1 1.61 14.29 25.00	1 1.61 7.14 25.00	4 6.45
	4	1 1.61 7.14 20.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 4.84 42.86 60.00	1 1.61 7.14 20.00	5 8.06
	5	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 4.84 25.00 23.08	1 1.61 14.29 7.69	9 14.52 64.29 69.23	13 20.97
		14 22.58	15 24.19	12 19.35	7 11.29	14 22.58	62

b. Generally, animals with lower scores on the danger index slept more, while those who had higher danger values slept for fewer hours.



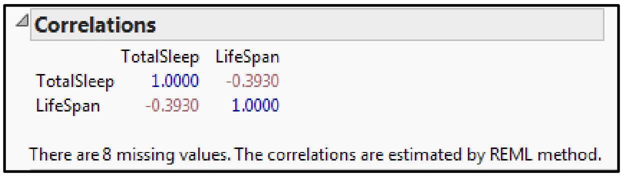
```
Missing Rows    4
```

c. There is evidence of a weak negative relationship between lifespan and total sleep. Longer-lived animals tend to get or need less sleep (humans are the outlier in this graph).



```
Linear Fit
LifeSpan = 36.932623 - 1.6403373*TotalSleep
Summary of Fit
RSquare                          0.168266
RSquare Adj                      0.152271
Root Mean Square Error           17.31945
Mean of Response                 19.85185
Observations (or Sum Wgts)             54
```

d.

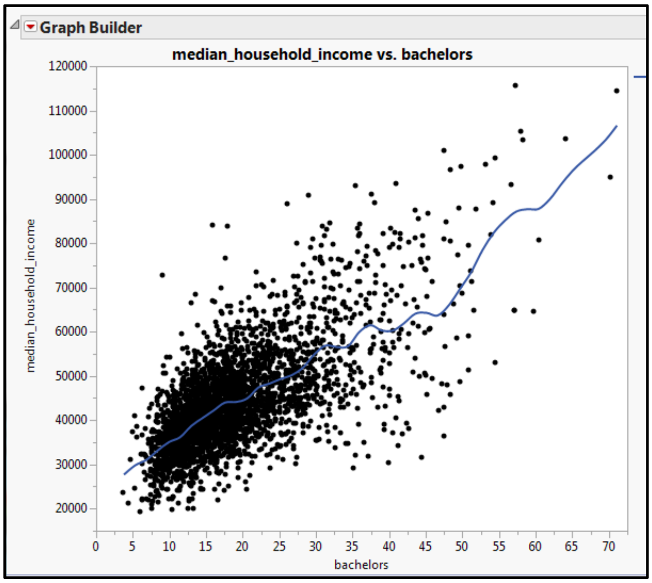


```
Correlations
              TotalSleep   LifeSpan
TotalSleep        1.0000    -0.3930
LifeSpan         -0.3930     1.0000

There are 8 missing values. The correlations are estimated by REML method.
```

There is a weak negative correlation between total sleep time and life span, confirming what we saw in the graph. In other words, mammals with long life spans may tend to sleep less than other mammals.

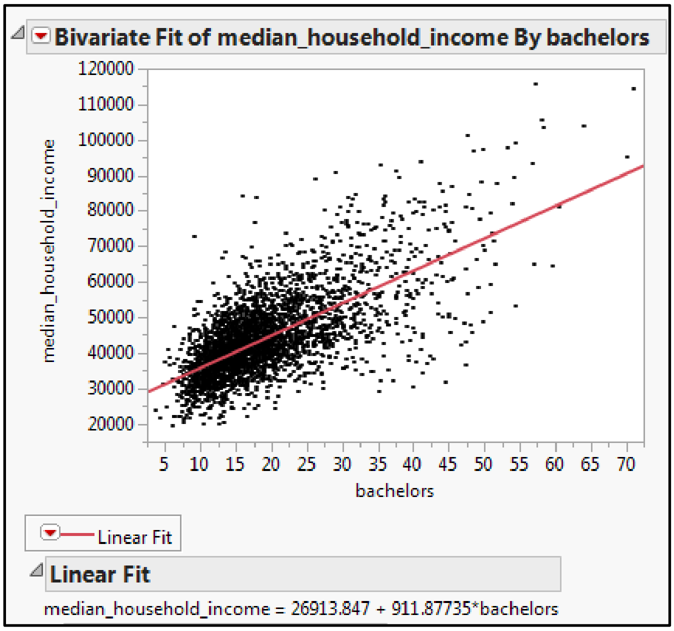
Scenario 8

a.



Using Graph Builder to investigate this relationship we find a positive but inconsistent relationship between income and percentage of population with a bachelor's degree. There is a clear upward pattern with a lot of scatter, indicating that a moderate linear relationship.

b.

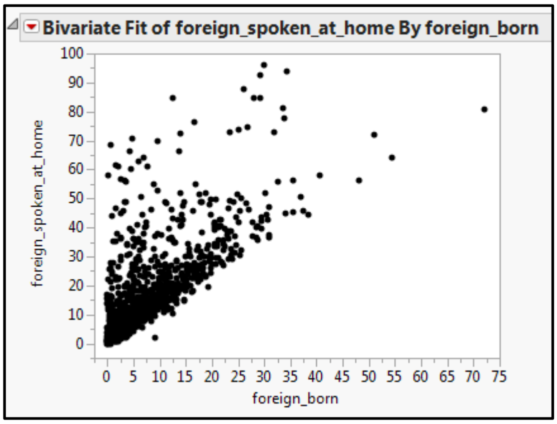


Linear Fit

median_household_income = 26913.847 + 911.87735*bachelors

Using the Fit Y by X platform, we obtain the results shown here. The linear fit is shown below the scatter plot. Substituting 25 for bachelors, we get median_household_income = 26913.847 + 911.87735* 25 = $ 49, 710.78.

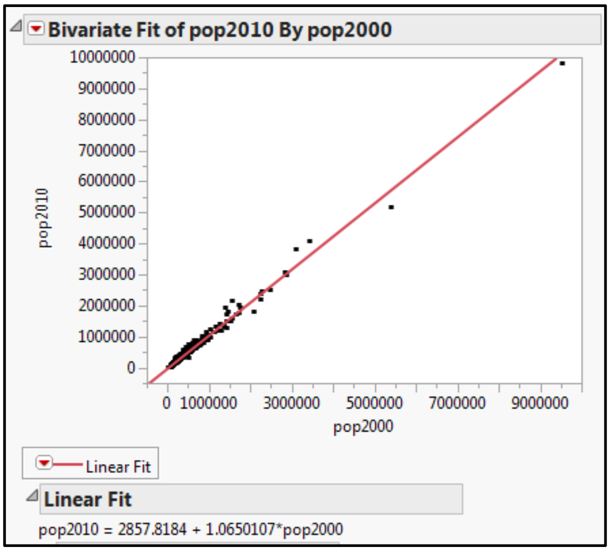
c.



There are very few counties lying below the 45-degree diagonal line, indicating that the percentage of homes where a foreign language is spoken almost always exceeds the percentage of homes with a foreign-born member. This makes sense, assuming that homes with no foreign-born members would be less inclined to speak a foreign language.

d. The correlation is 0.7911; there is a moderately strong association between the percentage number of households where a foreign language is spoken and the percentage of households with a foreign-born member.

e.



pop2010 = 2857.8184 + 1.0650107*pop2000

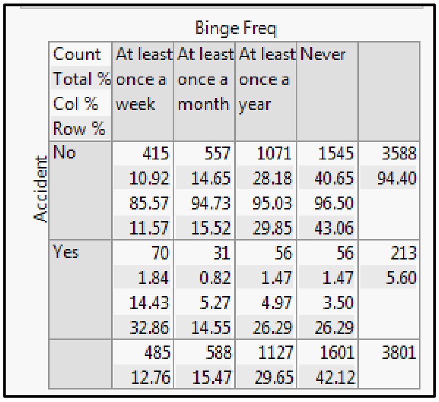
The slope of the line is approximately 1.065, indicating that on average, the population of US counties grew by 6.5% from 2000 to 2010.

f. Cook County lost population between 2000 and 2010; hence it's growth rate was substantially less than + 6.5%.

Chapter 6: Solutions to Application Scenarios

Scenario 2

For the questions that follow, we can use this contingency table:



		Binge Freq				
Accident	Count Total % Col % Row %	At least once a week	At least once a month	At least once a year	Never	
	No	415	557	1071	1545	3588
		10.92	14.65	28.18	40.65	94.40
		85.57	94.73	95.03	96.50	
		11.57	15.52	29.85	43.06	
	Yes	70	31	56	56	213
		1.84	0.82	1.47	1.47	5.60
		14.43	5.27	4.97	3.50	
		32.86	14.55	26.29	26.29	
		485	588	1127	1601	3801
		12.76	15.47	29.65	42.12	

a. *Pr(Binge at least once a week)* = 0.1276.

b. *Pr(Never binge)* = 0.4212.

c. *Pr(Accident)* = 0.0560.

d. *Pr(Accident or binge at least once a week)* = *Pr(Accident)* + *Pr(at least once a week)* – *Pr(Accident and binge at least once a week)* = 0.0560 + 0.1276 – .0184 = 0.1552.

e. *Pr(Accident | binge at least once a week)* = 0.1443.

f. *Pr(Binge at least once a week| Accident)* = 0.3286.

g. No. Comparing the results in parts a and f or parts c and e should lead to the conclusion that because the relevant marginal probabilities do not equal the corresponding conditionals, the events are not independent.

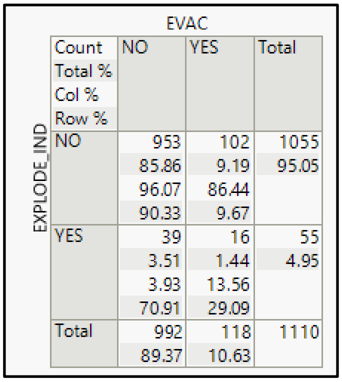
Scenario 4

a. *Pr(Equipment failure)*= 0.31456

b. *Pr(ignited)* = 0.11352.

c. *Pr(Evacuation)* = 0.10631.

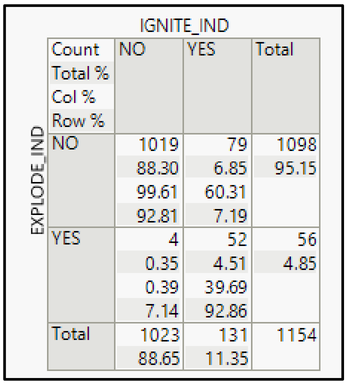
d. Use this contingency table:



		EVAC		
EXPLODE_IND	Count Total % Col % Row %	NO	YES	Total
	NO	953	102	1055
		85.86	9.19	95.05
		96.07	86.44	
		90.33	9.67	
	YES	39	16	55
		3.51	1.44	4.95
		3.93	13.56	
		70.91	29.09	
	Total	992	118	1110
		89.37	10.63	

Pr(Evacuation | Explosion) = 0.2909 (row %)

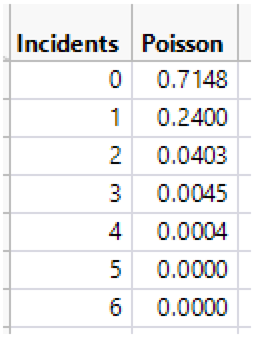
e.



		IGNITE_IND		
	Count Total % Col % Row %	NO	YES	Total
EXPLODE_IND	NO	1019 88.30 99.61 92.81	79 6.85 60.31 7.19	1098 95.15
	YES	4 0.35 0.39 7.14	52 4.51 39.69 92.86	56 4.85
	Total	1023 88.65	131 11.35	1154

Pr(Ignition or Explosion) = Pr(Ignition) + Pr(Explosion) – Pr(Ignition and Explosion) = 0.1135 + 0.0485 – 0.0451 = 0.1169.

f. Here is a table of computed Poisson probabilities:



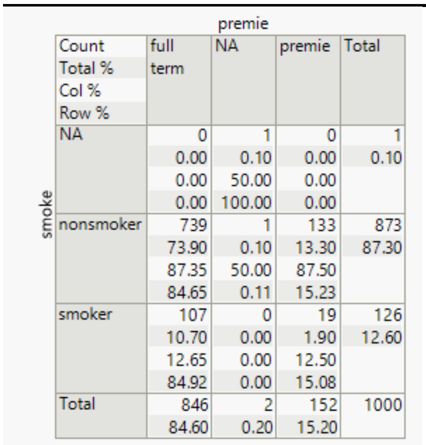
Incidents	Poisson
0	0.7148
1	0.2400
2	0.0403
3	0.0045
4	0.0004
5	0.0000
6	0.0000

In the data we observed 1 incident 24% of the time, which matches the theoretical probability. We observed 2 incidents 3% of the time, which is slightly less than the theoretical probability of 0.0403, and both the model and the observed data show no 5-incident observations.

The model fits the data quite well.

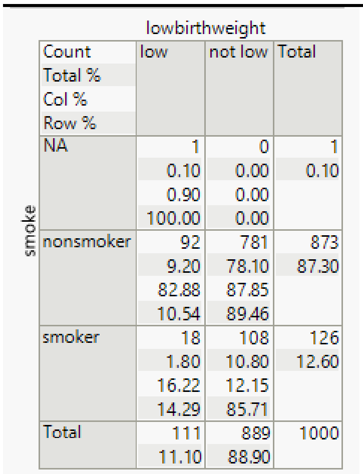
Scenario 6

a. *Pr(smoker AND premie)* = 0.019



		premie			
	Count Total % Col % Row %	full term	NA	premie	Total
smoke	NA	0 0.00 0.00 0.00	1 0.10 50.00 100.00	0 0.00 0.00 0.00	1 0.10
	nonsmoker	739 73.90 87.35 84.65	1 0.10 50.00 0.11	133 13.30 87.50 15.23	873 87.30
	smoker	107 10.70 12.65 84.92	0 0.00 0.00 0.00	19 1.90 12.50 15.08	126 12.60
	Total	846 84.60	2 0.20	152 15.20	1000

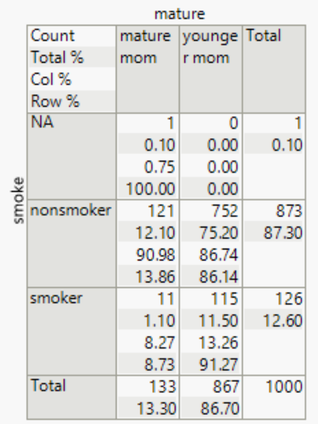
b. *Pr(smokers AND low birth weight)* = 0.018.



lowbirthweight

	Count Total % Col % Row %	low	not low	Total
smoke	NA	1 0.10 0.90 100.00	0 0.00 0.00 0.00	1 0.10
	nonsmoker	92 9.20 82.88 10.54	781 78.10 87.85 89.46	873 87.30
	smoker	18 1.80 16.22 14.29	108 10.80 12.15 85.71	126 12.60
	Total	111 11.10	889 88.90	1000

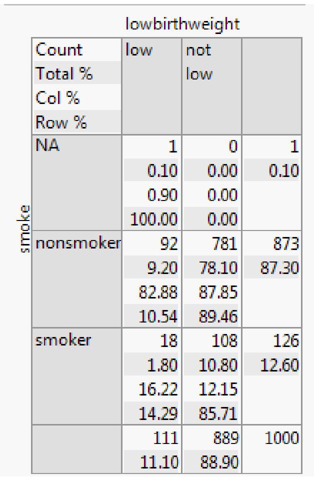
c. *Pr(mature AND smoker)* = 0.011.



mature

	Count Total % Col % Row %	mature mom	younger mom	Total
smoke	NA	1 0.10 0.75 100.00	0 0.00 0.00 0.00	1 0.10
	nonsmoker	121 12.10 90.98 13.86	752 75.20 86.74 86.14	873 87.30
	smoker	11 1.10 8.27 8.73	115 11.50 13.26 91.27	126 12.60
	Total	133 13.30	867 86.70	1000

d.



lowbirthweight

	Count Total % Col % Row %	low	not low	
smoke	NA	1 0.10 0.90 100.00	0 0.00 0.00 0.00	1 0.10
	nonsmoker	92 9.20 82.88 10.54	781 78.10 87.85 89.46	873 87.30
	smoker	18 1.80 16.22 14.29	108 10.80 12.15 85.71	126 12.60
		111 11.10	889 88.90	1000

We know that if *Pr(A|B)=Pr(A)*, events A and B are independent. Here, for example, we know that *Pr(low birthweight)* = 0.111.

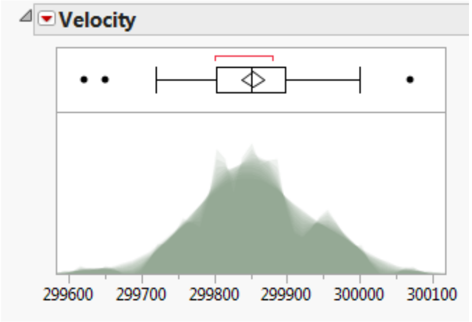
We can also see in the table that *Pr(low birthweight|smoker)* = 0.1429.

Because 0.1110 ≠ 0.1429, we conclude that low birthweight and smoker are not independent.

Chapter 7: Solutions to Application Scenarios

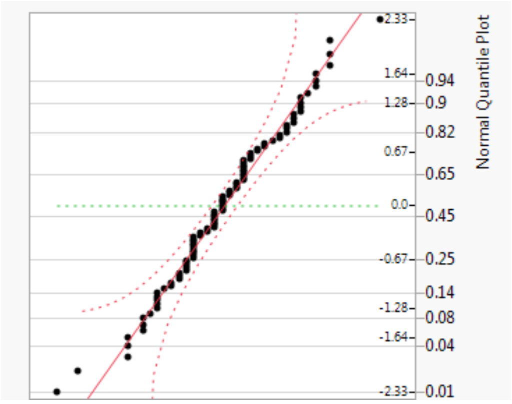
Scenario 2

a.



In the shadowgram to the left we see a generally symmetric distribution that seems to be mound-shaped with a peak near 299,850 km/sec. There may be some indication of a secondary peak at approximately 299,950 km/sec., but the overall impression is that the distribution might be well-described by the normal model.

b.



In the normal quantile plot the points closely follow the 45-degree diagonal line – further suggesting that suitability of the normal model.

c. The data set provides some support for the assumption. Michelson's various measurements of the speed of light seem to vary according to an approximate normal distribution.

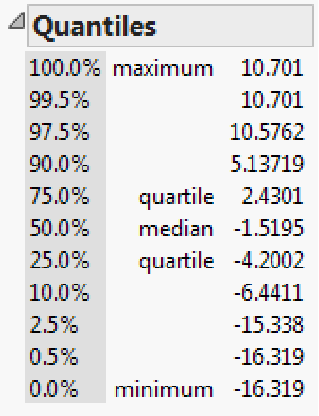
Scenario 4

a. Student answers will vary. Most will likely choose the weekly change column corresponding to the Hang Seng or Tel Aviv market index. In these graphs, the points track most closely to the diagonal line, but in the other graphs they do not.

b. Student answers will vary here as well. The FTSE and Madrid (IGBM) weekly changes have normal quantile plots that deviate most from the diagonal line.

c. The mean and standard deviation of the changes in Hang Seng are –1.102065 and 5.242892. For a normal distribution with that mean and standard deviation,
Pr (X <0) =0.5832, or approximately 0.58.

d.



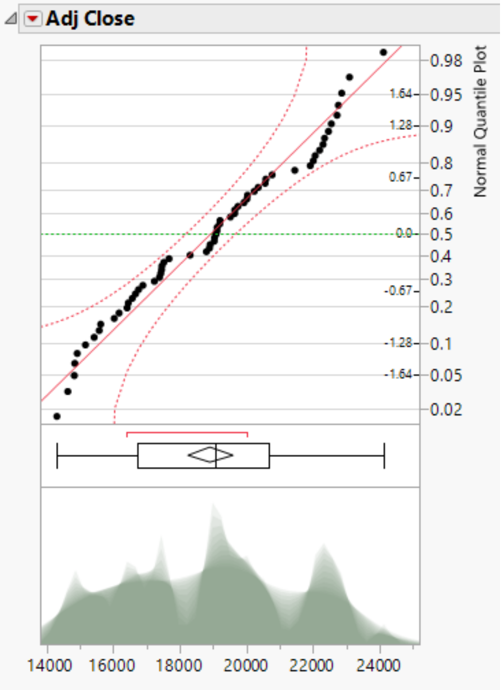
Quantiles

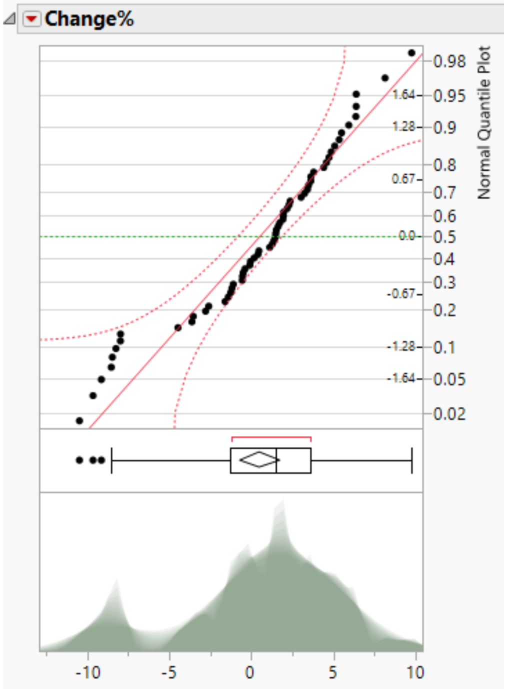
100.0%	maximum	10.701
99.5%		10.701
97.5%		10.5762
90.0%		5.13719
75.0%	quartile	2.4301
50.0%	median	-1.5195
25.0%	quartile	-4.2002
10.0%		-6.4411
2.5%		-15.338
0.5%		-16.319
0.0%	minimum	-16.319

Looking at the table of quantiles (left), we see that the 75th percentile is at 2.43% and the 50th percentile is at –1.5195%. We know therefore that the Hang Seng index lost value somewhere between 50% and 75% of the time. This is consistent with the result in part c.

Scenario 6

These graphs can be used to respond to parts a and b.

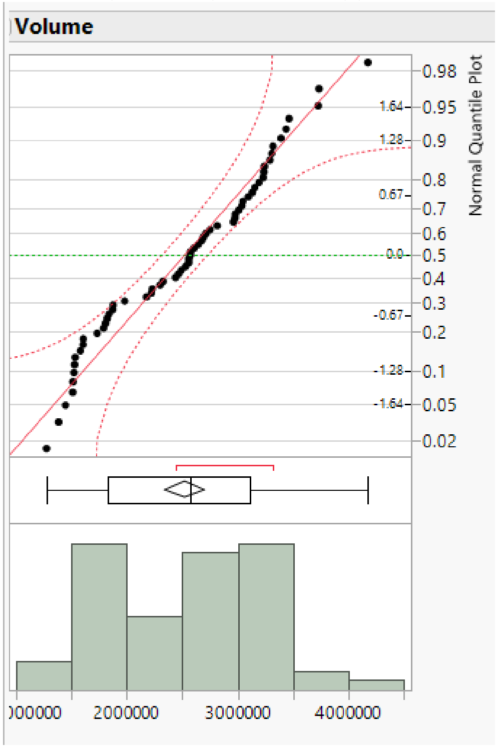




a. Adjusted closing values are relatively symmetric but multi-modal, with three or four peaks. The median of the distribution is close to 19,000 and it ranges from approximately 14,000 to 24,000. In contrast, the %change column is distinctly bimodal and left-skewed, with a major peak near 2% and a minor peak near –8%. Most of the distribution lies between –10 % and +10 %.

b. NOTE: This question should not be assigned. The Close and Adjusted Close columns are identical, so the response to part a also applies here.

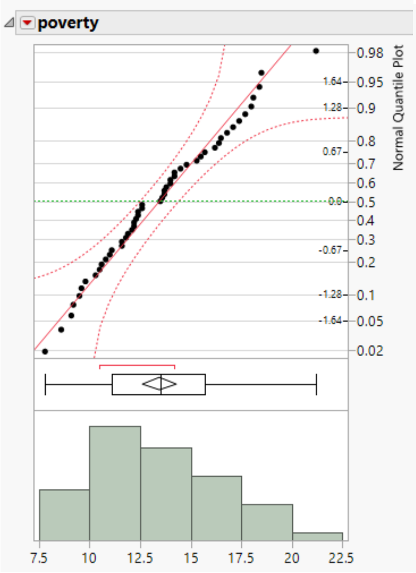
c.

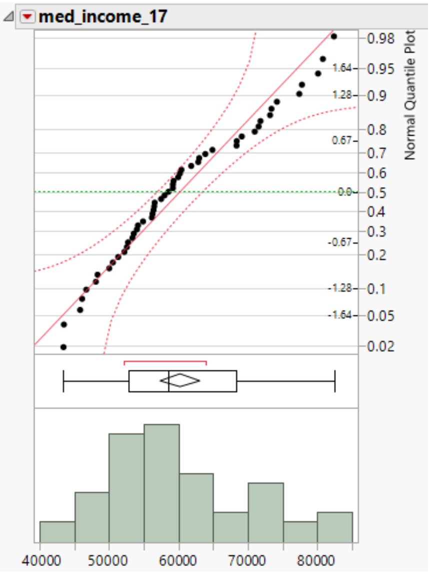


The volume column is bimodal and right skewed. The normal quantile plot do not track well along the diagonal line, and therefore a normal model would not be appropriate.

Scenario 8

a. Use Analyze > Distribution to obtain histograms, then red triangle Normal Quantile plot to produce these two plots:



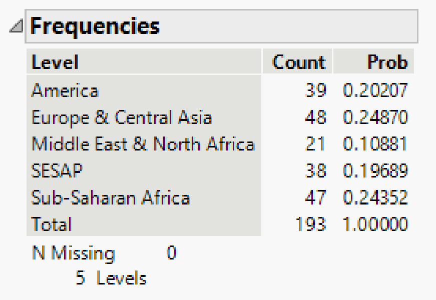


b. Both distributions are unimodal and right-skewed. Both normal quantile plots show departures from the normal model, with the poverty data being more nearly normal than the income data.

Chapter 8: Solutions to Application Scenarios

Scenario 2

a.



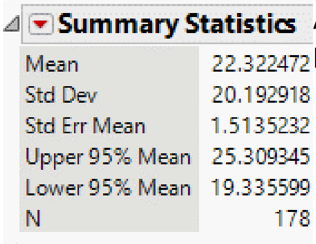
Frequencies

Level	Count	Prob
America	39	0.20207
Europe & Central Asia	48	0.24870
Middle East & North Africa	21	0.10881
SESAP	38	0.19689
Sub-Saharan Africa	47	0.24352
Total	193	1.00000

N Missing 0
5 Levels

The proportion of countries in Sub-Saharan Africa is 0.24352.

b.

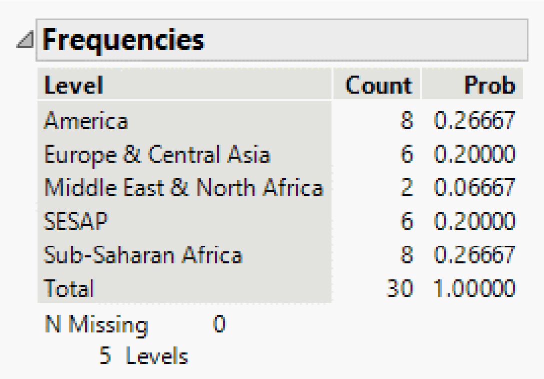


Summary Statistics

Mean	22.322472
Std Dev	20.192918
Std Err Mean	1.5135232
Upper 95% Mean	25.309345
Lower 95% Mean	19.335599
N	178

As shown to the right, the mean is 22.322 deaths per 1,000 live births; the standard deviation is 20.193.

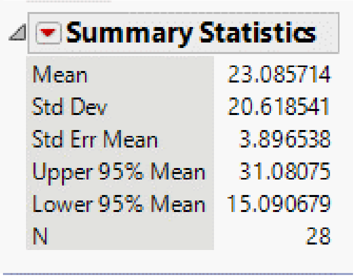
c.



Frequencies

Level	Count	Prob
America	8	0.26667
Europe & Central Asia	6	0.20000
Middle East & North Africa	2	0.06667
SESAP	6	0.20000
Sub-Saharan Africa	8	0.26667
Total	30	1.00000

N Missing 0
5 Levels



Summary Statistics

Mean	23.085714
Std Dev	20.618541
Std Err Mean	3.896538
Upper 95% Mean	31.08075
Lower 95% Mean	15.090679
N	28

Student answers will vary due to random sampling. Above we find the results of one random sample—8 of the 30 countries are in Sub-Saharan Africa (26.7%), which is slightly higher than the proportion in the full list.

The mean mortality rate in the sample is 23.09 (note that in this sample only 28 of 30 countries reported an infant mortality rate). In general students' results will not match the population values shown in parts a & b due to sampling variation.

Scenario 4

a. Student responses will vary. In general, the sampling distribution will be bell-shaped and symmetrical, centered very near 15 with an overall standard error (std. deviation of the sample means) approximately equal to 0.10 and ranging from about 14.7 to 15.3.

b. Student responses will again vary. In general, the sampling distribution will be bell-shaped and symmetrical, centered very near 15 with an overall standard error (std. deviation of the sample means) approximately equal to 0.20 and ranging from about 14.4 to 15.6.

c. Student responses will again vary. In general, the sampling distribution will be bell-shaped and symmetrical, centered very near 15 with an overall standard error (std. deviation of the sample means) approximately equal to 0.40 and ranging from about 13.8 to 16.2.

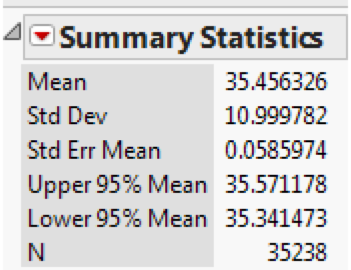
d. Student responses will again vary. In general, thanks to the Central Limit Theorem the sampling distribution will be bell-shaped and symmetrical, centered very near 15 with an overall standard error (std. deviation of the sample means) approximately again equal to 0.10 and ranging from about 14.7 to 15.3.

e. The results will be very similar to parts a and d though each student may have slightly different numerical results.

f. Reducing the sample size gradually increases the standard error of the sampling distribution (i.e. increases the variability across samples). Populations with relatively large standard deviations generate samples with comparatively large sampling variation. With samples this large (*n* = 1000) the shape of the parent population has no appreciable effect on the center, shape or spread of the sampling distribution.

Scenario 6

a.

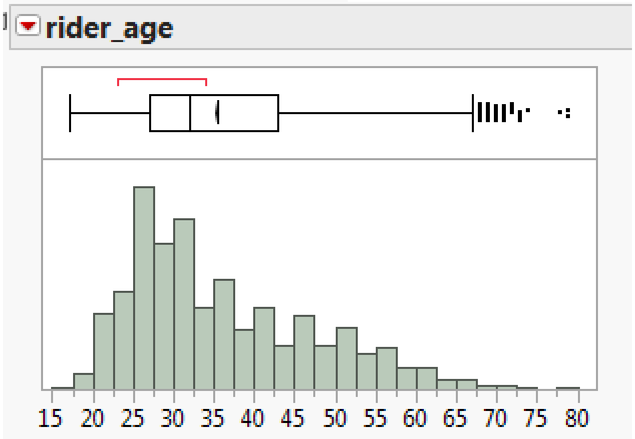


Summary Statistics

Mean	35.456326
Std Dev	10.999782
Std Err Mean	0.0585974
Upper 95% Mean	35.571178
Lower 95% Mean	35.341473
N	35238

The mean rider age is 35.46 years.

b.

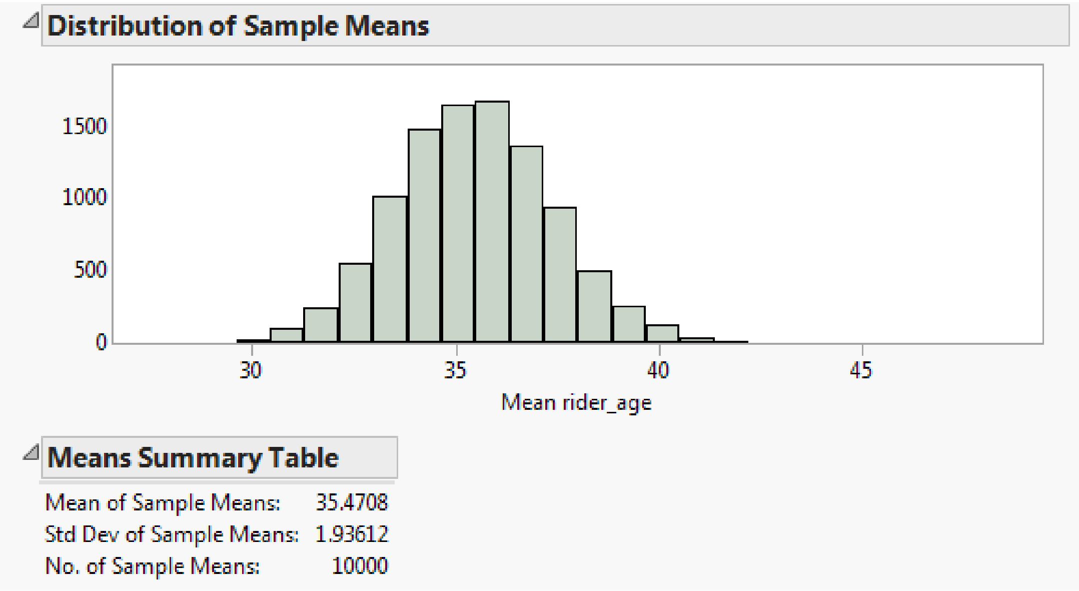


The distribution is unimodal, strongly skewed to the right, with a relatively small number of outliers.

c. Using the CLT, we'd expect the sampling distribution of the sample mean to approach an approximately normal distribution as the sample size, *n*, grows large. The mean of the distribution should be 35.46 years with a standard error equal to approximately 11/(sqrt(n)).

d. Each student will obtain a different result, reported in part e. It is important to base the simulation on the Hubway data.

e. Here are the results of **one** such simulation, rescaled for clarity:



Distribution of Sample Means

Means Summary Table

Mean of Sample Means:	35.4708
Std Dev of Sample Means:	1.93612
No. of Sample Means:	10000

The sampling distribution is symmetric and unimodal, with a mean at 35.47 years and a standard error of 1.936. Note that in part c the CLT would have predicted a mean of 35.46 and a standard error of 11/√50 = 1.56

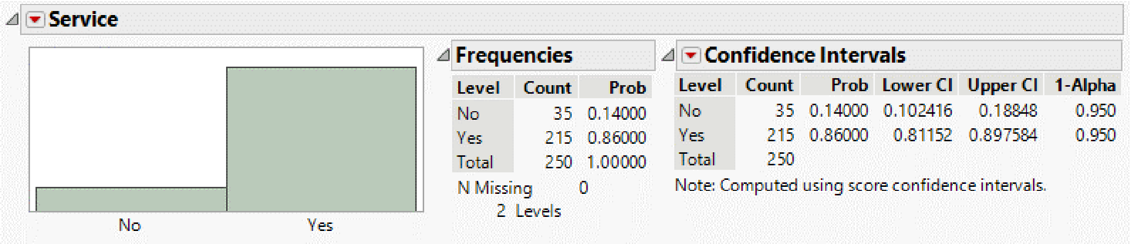
f. Samples with a mean > 35 years are quite common, because 35 is near the center of the sampling distribution.
Samples with means less than 30 are quite rare, because 30 lies in the far-left tail.
Samples with means more than 45 would be rarer still, because 45 is extremely far from the center of the distribution.

Chapter 10: Solutions to Application Scenarios

Scenario 2

a. Yes. We have a random sample of adequate size without exceeding 10% of the population. Because this is a simple random sample, we can safely assume that respondents are independent.

b. NOTE: You must create a small data table with 2 columns: “Service” and “Frequency”



Frequencies

Level	Count	Prob
No	35	0.14000
Yes	215	0.86000
Total	250	1.00000

N Missing 0
2 Levels

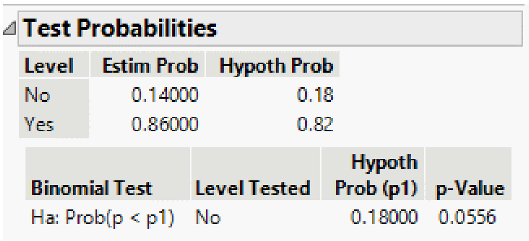
Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
No	35	0.14000	0.102416	0.18848	0.950
Yes	215	0.86000	0.81152	0.897584	0.950
Total	250				

Note: Computed using score confidence intervals.

We are 95% confident that the proportion of homes without Internet service is between 0.102 and 0.188.

c.



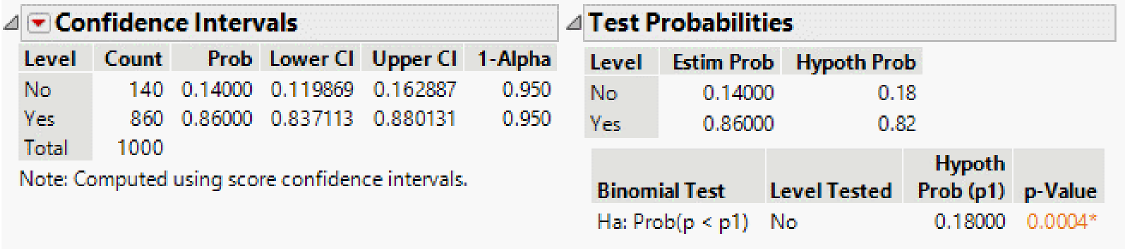
Test Probabilities

Level	Estim Prob	Hypoth Prob
No	0.14000	0.18
Yes	0.86000	0.82

Binomial Test	Level Tested	Hypoth Prob (p1)	p-Value
Ha: Prob(p < p1)	No	0.18000	0.0556

With a p-Value of 0.0556, this sample falls short of statistical significance, assuming the customary 5% significance level. The sample does not provide sufficient evidence to conclude that the rate is currently below 18%.

d.



Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
No	140	0.14000	0.119869	0.162887	0.950
Yes	860	0.86000	0.837113	0.880131	0.950
Total	1000				

Note: Computed using score confidence intervals.

Test Probabilities

Level	Estim Prob	Hypoth Prob
No	0.14000	0.18
Yes	0.86000	0.82

Binomial Test	Level Tested	Hypoth Prob (p1)	p-Value
Ha: Prob(p < p1)	No	0.18000	0.0004*

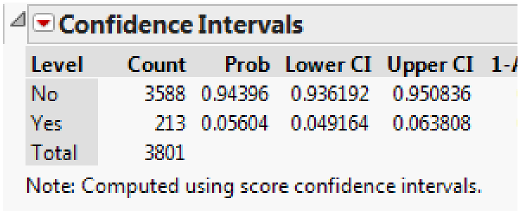
Now the confidence interval is narrower—from 0.12 to 0.16—and we would **reject** the null hypothesis and conclude that the current proportion of homes without Internet service is less than 0.18.

e. A larger sample with the very same proportion provides more precision in the confidence interval (i.e. a narrower interval) and enhances the statistical significance of the test result.

Scenario 4

a. Yes. We have a random sample of independent respondents and of sufficient size to draw inferences.

b.



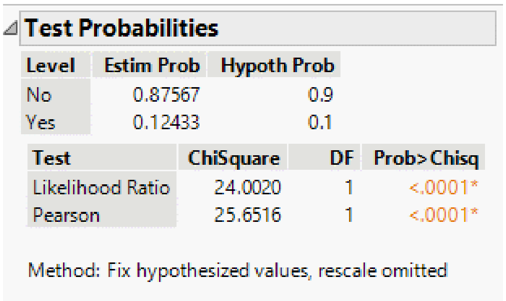
Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-/
No	3588	0.94396	0.936192	0.950836	
Yes	213	0.05604	0.049164	0.063808	
Total	3801				

Note: Computed using score confidence intervals.

We can be 95% confident that between 4.9% and 6.4% of all individuals in 18-39 age range have been in accidents after drinking.

c.



Test Probabilities

Level	Estim Prob	Hypoth Prob
No	0.87567	0.9
Yes	0.12433	0.1

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	24.0020	1	<.0001*
Pearson	25.6516	1	<.0001*

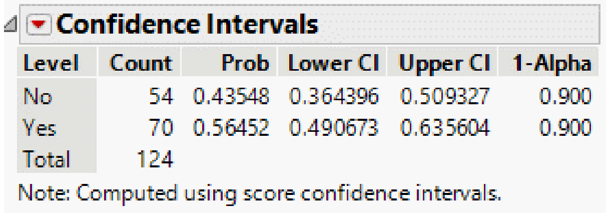
Method: Fix hypothesized values, rescale omitted

(For this question, it is simplest to create a small summary table). Create a Because of the question's wording, a two-tailed test is most appropriate here. By default, JMP produces a ChiSquare test (covered in Chapter 12), which is equivalent to the one-sample, two-sided test for a proportion in this case. Rely on the *P*-value to make the conclusion.

Based on this random sample, we can confidently conclude that it is *not* credible to conclude that 10% of the population binge drinks at least once per week. If anything, this sample suggests a higher population proportion.

Scenario 6

a.



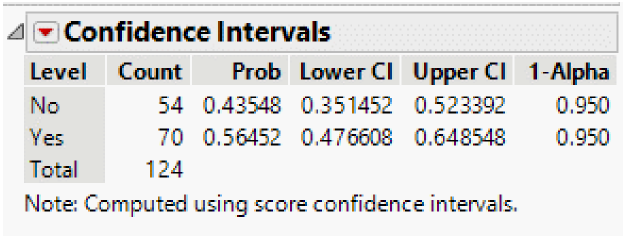
Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
No	54	0.43548	0.364396	0.509327	0.900
Yes	70	0.56452	0.490673	0.635604	0.900
Total	124				

Note: Computed using score confidence intervals.

We can be 90% confident that the proportion of trading days on which McDonald’s stock increases is somewhere between 0.491 and 0.636.

b.



Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
No	54	0.43548	0.351452	0.523392	0.950
Yes	70	0.56452	0.476608	0.648548	0.950
Total	124				

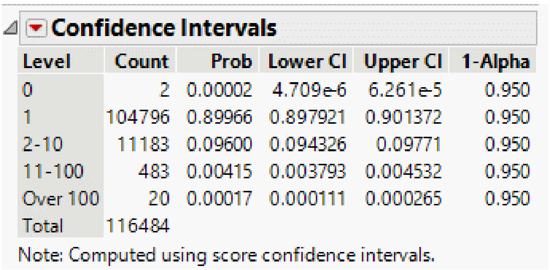
Note: Computed using score confidence intervals.

This interval is a bit wider that the earlier one: both are constructed around the point estimate of 0.56452, but the 95% interval reaches from 0.3515 to 0.4766. Here again, the higher confidence level requires a larger margin of error and hence a wider interval.

Scenario 8

a. We should proceed with caution. First, we would need to know more before assuming that the nature and frequency of bird strikes reported on one day is independent of other reports. It also depends on which variables we examine. Even though we have a large data table, many columns have a considerable amount of missing data, so that we may not have enough observations of some variables.

b.



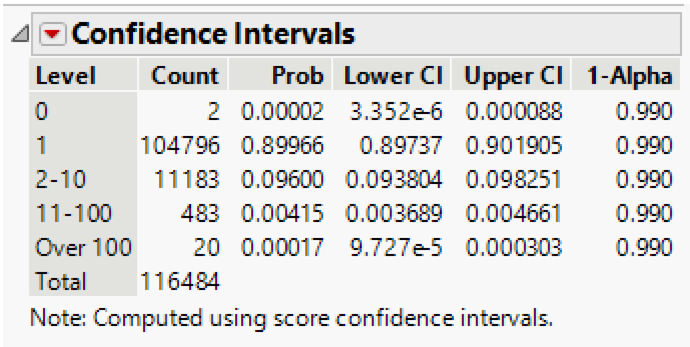
Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
0	2	0.00002	4.709e-6	6.261e-5	0.950
1	104796	0.89966	0.897921	0.901372	0.950
2-10	11183	0.09600	0.094326	0.09771	0.950
11-100	483	0.00415	0.003793	0.004532	0.950
Over 100	20	0.00017	0.000111	0.000265	0.950
Total	116484				

Note: Computed using score confidence intervals.

We can be 95% confident that, out of all instances where there is a bird strike, a single bird is struck somewhere between 89.8% and 90.1% of the time.

c.



Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
0	2	0.00002	3.352e-6	0.000088	0.990
1	104796	0.89966	0.89737	0.901905	0.990
2-10	11183	0.09600	0.093804	0.098251	0.990
11-100	483	0.00415	0.003689	0.004661	0.990
Over 100	20	0.00017	9.727e-5	0.000303	0.990
Total	116484				

Note: Computed using score confidence intervals.

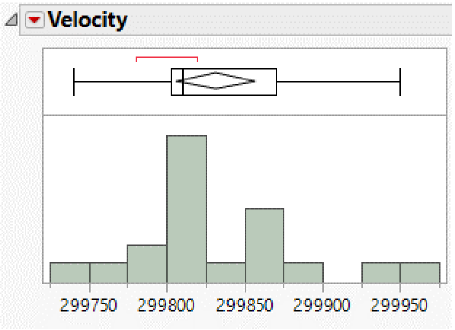
We can be 99% confident that, out of all instances where there is a bird strike, a single bird is struck somewhere between 89.7% and 90.2% of the time. The 99% CI is very slightly wider than the 95% CI.

d. First, recall that this data set only contains reported episodes when some kind of wildlife was struck. We’ve just seen in parts c and d that 90% is within both confidence intervals, so yes—it is plausible that the population proportion is 90%.

Chapter 11: Solutions to Application Scenarios

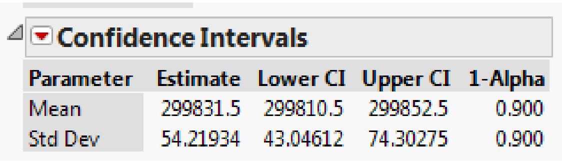
Scenario 2

a.



Yes. We do not know the population σ so we will use the t-distribution. Because the sample is small (n = 20) we want to see if the sample data suggest that the population is roughly normal in shape. The histogram and normal quantile plots indicate mild skewness but no serious indication of non-normality.

b.



Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	299831.5	299810.5	299852.5	0.900
Std Dev	54.21934	43.04612	74.30275	0.900

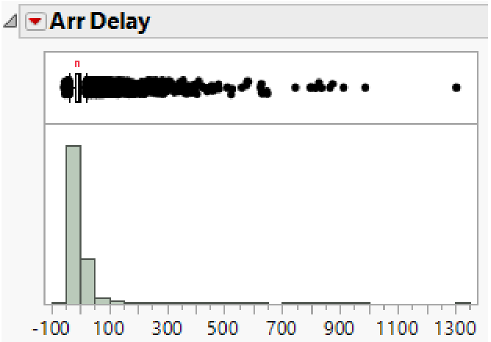
Based on this sample data, we can be 90% confident that the speed of light is between 299,810.5 and 299,852.5 km. per second.

c. From the confidence interval in part b we can see that Michelson would probably have (erroneously) concluded that the value 300,000 kps is not credible. The two-tailed hypothesis test yields a P-value < 0.0001 and a test statistic equal to –13.898; Michelson would have rejected a null hypothesis that the constant speed of light is 300,000 kps.

d. Student answers may vary, but assuming a significance level of 0.05 and a two-sample test, if the null value were approximately 299,857 Michelson would not have rejected the null hypothesis.

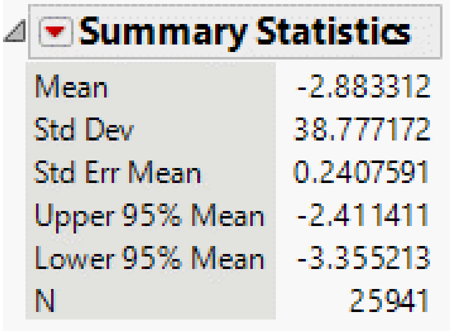
Scenario 4

a.



Yes. This is a highly skewed distribution, but because the sample is so large (n = 25,941) we can rely on the Central Limit Theorem to proceed. We do not know the population σ so we will use the t-distribution.

b.



Summary Statistics

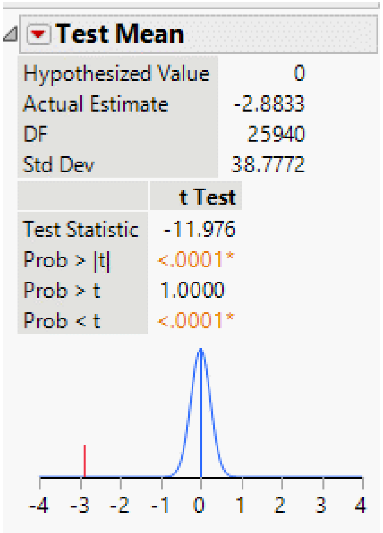
Mean	-2.883312
Std Dev	38.777172
Std Err Mean	0.2407591
Upper 95% Mean	-2.411411
Lower 95% Mean	-3.355213
N	25941

We can be 95% confident that the mean flight "delay" was not a delay at all, somewhere between 2.41 and 3.35 minutes early on average.

c. NOTE: There is a typographical error in the earliest printings of this book. The question should refer to Atlanta rather than to Chicago.

No. The interval is an estimate of the population mean, not the range of individual values. The interval provides an estimate of the location of the population mean acknowledging the uncertainty that arises from using a sample.

d.



Test Mean

Hypothesized Value	0
Actual Estimate	-2.8833
DF	25940
Std Dev	38.7772

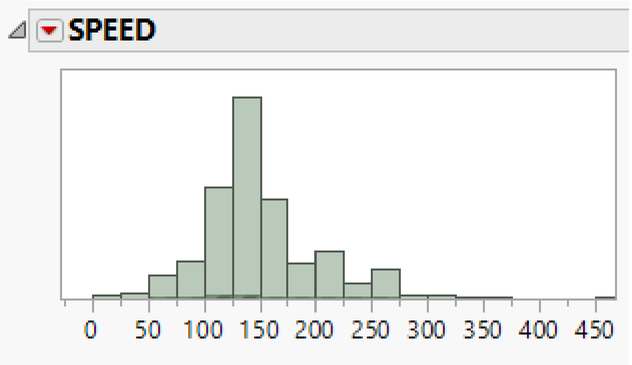
	t Test
Test Statistic	-11.976
Prob > |t|	<.0001*
Prob > t	1.0000
Prob < t	<.0001*

For this test we see that the reported p-value is < 0.0001. Because this is less than any conventional alpha, we reject the null hypothesis and we conclude that there is compelling evidence to conclude that the mean is less than 0 minutes. In other words, we are convinced that flights to Atlanta do tend to arrive ahead of schedule.

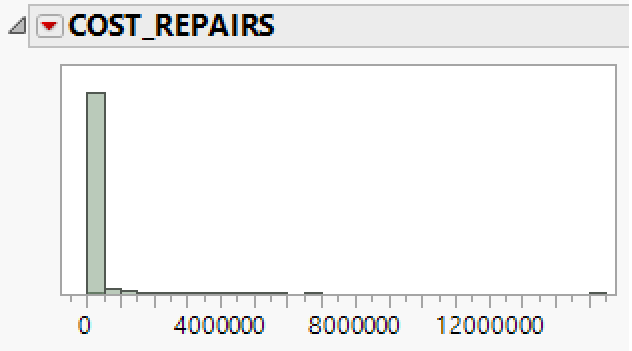
e. NOTE: In the animator tool, you will need to drag the horizontal axis to reveal the blue line at -2.88. Grab the square at the top of the blue line and drag it towards 2.find the blue line If the true population mean actually = -2 minutes, the power of this test would be approximately 1. In other words, if the reality were that the mean flight is 2 minutes early, this test would surely detect that flights arrive early.

Scenario 6

a.

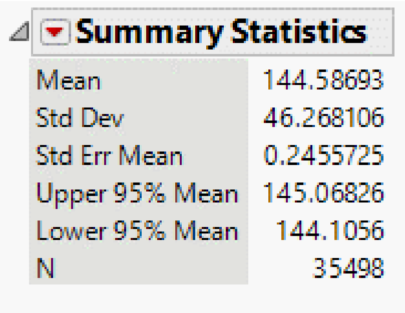


The speed column does seem to satisfy the conditions: it is moderately symmetric, and the sample is very large (n = 35,498) so we can rely on the Central Limit Theorem to proceed. We do not know the population σ so we will use the t-distribution.



The Cost of Repairs column is a smaller sample (n = 1,885) and very strongly skewed. Even with the CLT, we should proceed with caution.

b.

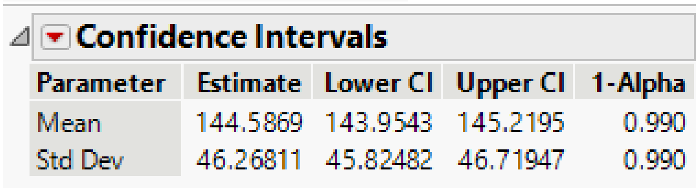


Summary Statistics

Mean	144.58693
Std Dev	46.268106
Std Err Mean	0.2455725
Upper 95% Mean	145.06826
Lower 95% Mean	144.1056
N	35498

We can be 95% confident that the mean flight speed at impact is between 144.1 and 145.1 MPH.

c.



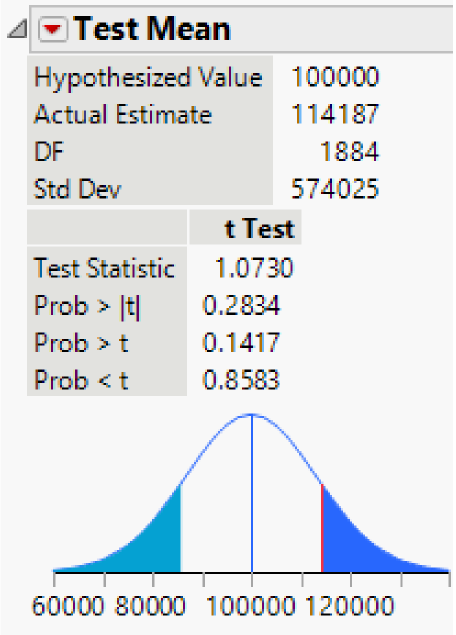
Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	144.5869	143.9543	145.2195	0.990
Std Dev	46.26811	45.82482	46.71947	0.990

At the 99% confidence level, we can be 99% confident that the mean flight speed at impact is between 143.9 and 145.2 MPH.

d. When we increase the confidence level, we widen the interval.

e.



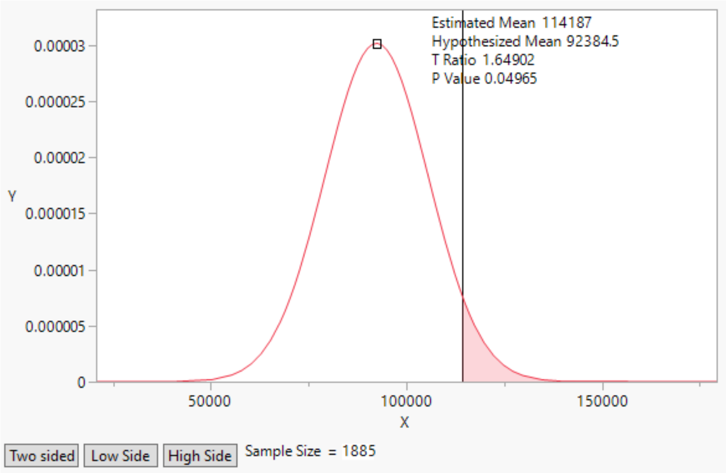
Test Mean

Hypothesized Value	100000
Actual Estimate	114187
DF	1884
Std Dev	574025

	t Test
Test Statistic	1.0730
Prob > |t|	0.2834
Prob > t	0.1417
Prob < t	0.8583

The test results indicate that this sample does not provide convincing evidence to reject the null hypothesis, yielding an upper-tail P-value of 0.1417. The sample is, as noted, very right-skewed, but if anything that would overstate the population mean. Even with a sample mean of $114,187 we should not conclude that the mean cost exceeds 100,000.

f.

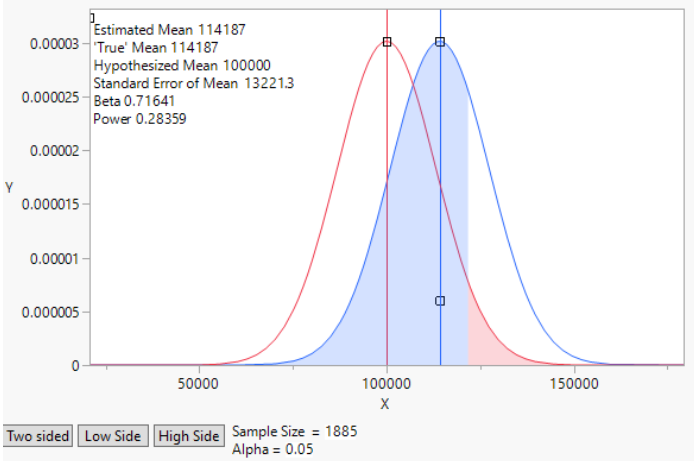


Estimated Mean 114187
Hypothesized Mean 92384.5
T Ratio 1.64902
P Value 0.04965

Sample Size = 1885

Student answers will vary slightly, but if the hypothesized mean were less than approximately $923,845, the upper-tailed p-value would fall below 0.05.

g.



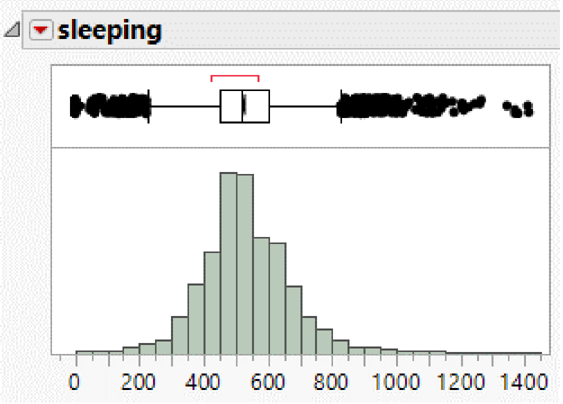
Estimated Mean 114187
'True' Mean 114187
Hypothesized Mean 100000
Standard Error of Mean 13221.3
Beta 0.71641
Power 0.28359

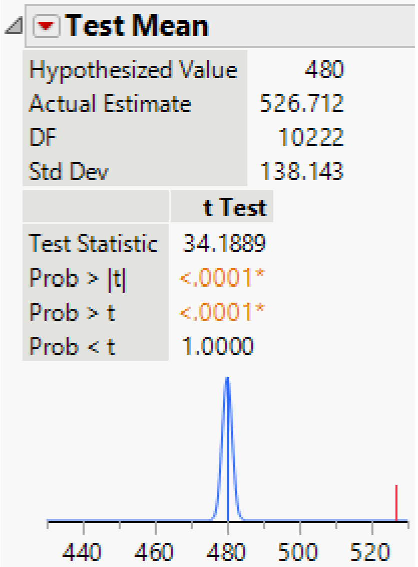
Sample Size = 1885
Alpha = 0.05

Student answers will vary, but should compare the current test to alternatives where the true mean is considerably larger than 100,000. As a starting point, if the true mean matched the sample mean, power = 0.28. At an indicator of a powerful test, if the true mean were approximately $138,712, power would equal 0.90.

Scenario 8

a.





Test Mean

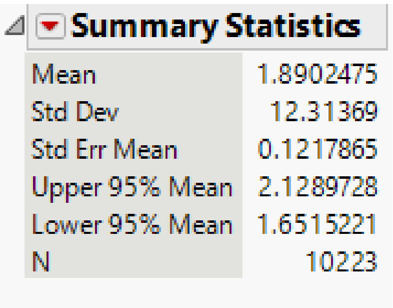
Hypothesized Value	480
Actual Estimate	526.712
DF	10222
Std Dev	138.143

	t Test
Test Statistic	34.1889
Prob > |t|	<.0001*
Prob > t	<.0001*
Prob < t	1.0000

We have a quite symmetric distribution and a very large sample, so we are firm ground in conducting a t-test.

As the test report shows, this sample provides convincing evidence that in 2017, US adults slept more than 480 minutes per night, on average.

b.



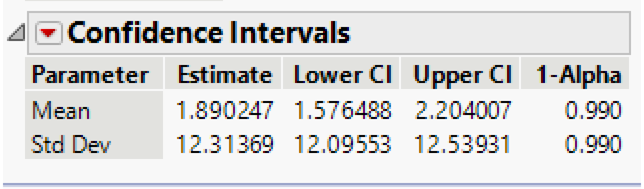
Summary Statistics

Mean	1.8902475
Std Dev	12.31369
Std Err Mean	0.1217865
Upper 95% Mean	2.1289728
Lower 95% Mean	1.6515221
N	10223

This distribution is quite skewed, but we do have a large sample and can rely on the Central Limit Theorem.

We can be 95% confident that the mean time devoted to reading personal emails was between 1.65 and 2.13 minutes per day.

c.



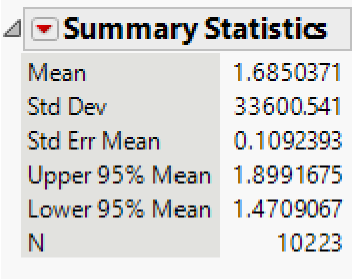
Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	1.890247	1.576488	2.204007	0.990
Std Dev	12.31369	12.09553	12.53931	0.990

We can be 99% confident that the mean time spent on reading personal emails in 2017 was between 1.58 and 2.20 minutes. Both intervals are centered on the sample mean, but the 99% interval is wider than the 95% interval. In general, increasing the confidence level leads to a wider interval.

d. No. The confidence interval estimates the population mean, not the range of individual behaviors.

e.



Summary Statistics

Mean	1.6850371
Std Dev	33600.541
Std Err Mean	0.1092393
Upper 95% Mean	1.8991675
Lower 95% Mean	1.4709067
N	10223

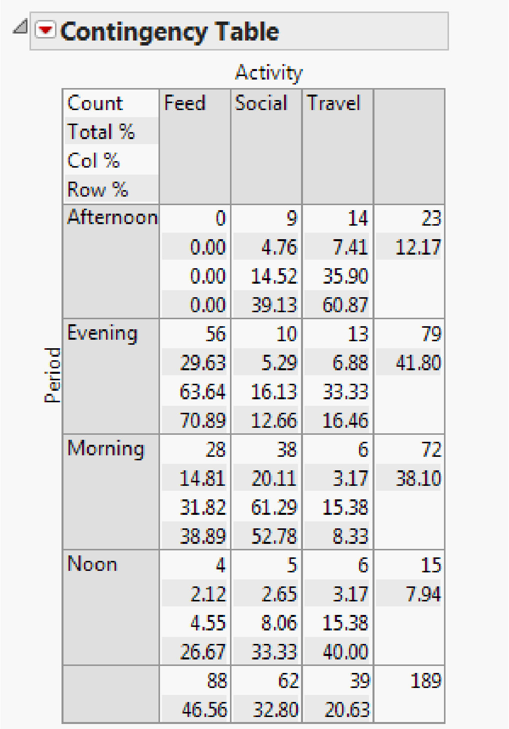
By adjusting for sampling weight, the sample mean is smaller (1.69 minutes vs. 1.89 minutes). The weighted confidence interval is (1.47, 1.90), as compared to the unweighted result of (1.65, 2.13) minutes.

The unweighted (and less accurate) estimate is higher than the weighted one; we would tend to overestimate the amount of time people spent on email.

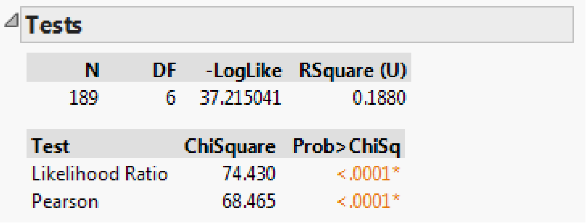
Chapter 12: Solutions to Application Scenarios

Scenario 2

a.



Period	Count Total % Col % Row %	Feed	Social	Travel	
	Afternoon	0	9	14	23
		0.00	4.76	7.41	12.17
		0.00	14.52	35.90	
		0.00	39.13	60.87	
	Evening	56	10	13	79
		29.63	5.29	6.88	41.80
		63.64	16.13	33.33	
		70.89	12.66	16.46	
	Morning	28	38	6	72
		14.81	20.11	3.17	38.10
		31.82	61.29	15.38	
		38.89	52.78	8.33	
	Noon	4	5	6	15
		2.12	2.65	3.17	7.94
		4.55	8.06	15.38	
		26.67	33.33	40.00	
		88	62	39	189
		46.56	32.80	20.63	

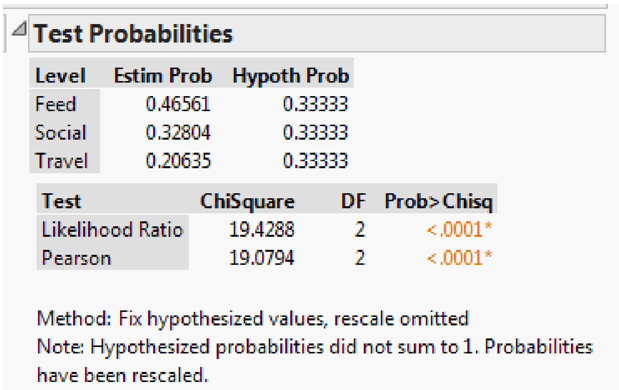


N	DF	-LogLike	RSquare (U)
189	6	37.215041	0.1880

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	74.430	<.0001*
Pearson	68.465	<.0001*

Because there are some cells with very small counts and expected counts, we should use caution making inferences from the ChiSquare test. However, we can note that the evidence points towards rejection of the null hypothesis of independence and we can also note (for example) that dolphins were regularly observed feeding in the morning and evening, but rarely if ever at other times.

b.



Level	Estim Prob	Hypoth Prob
Feed	0.46561	0.33333
Social	0.32804	0.33333
Travel	0.20635	0.33333

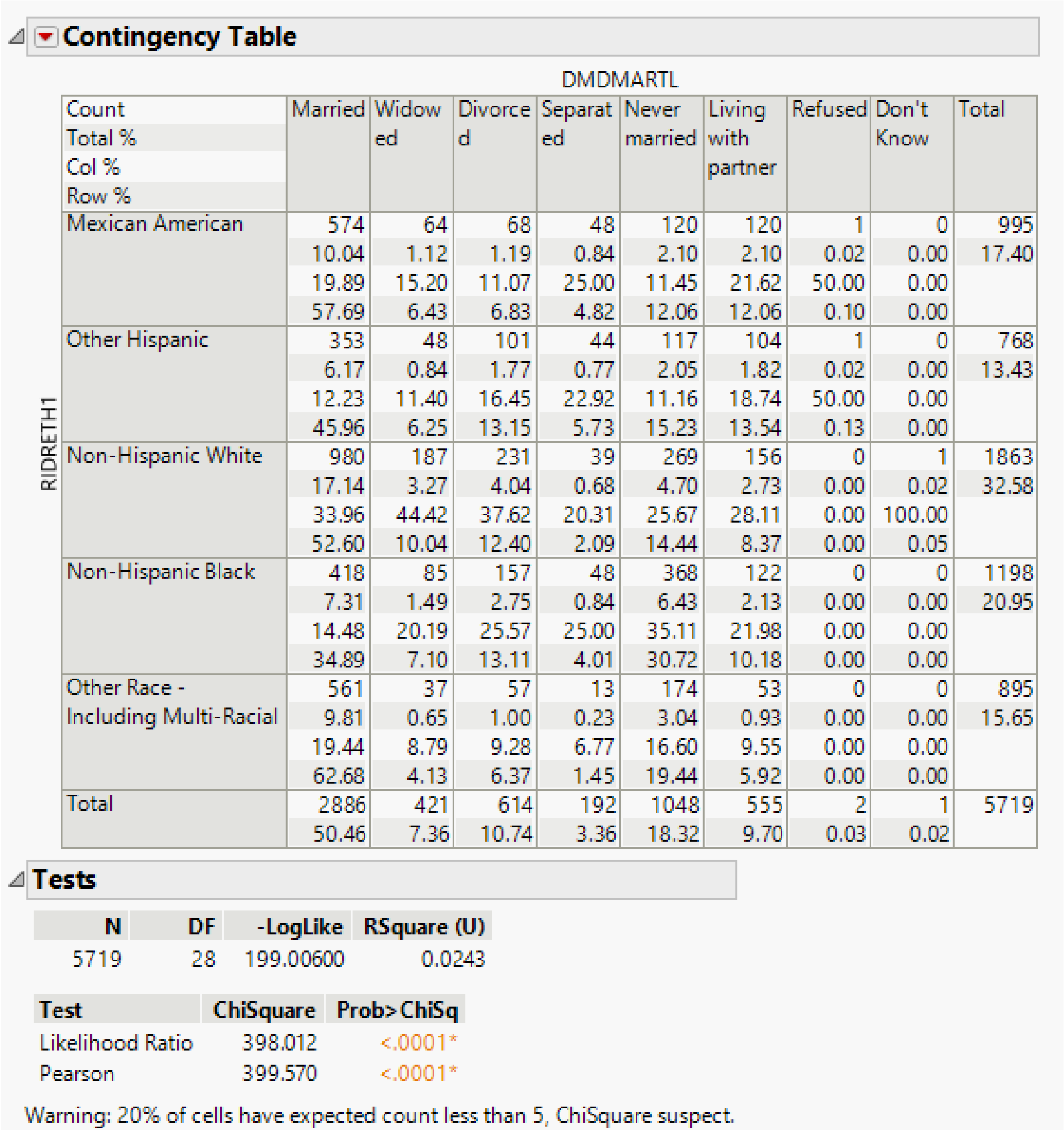
Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	19.4288	2	<.0001*
Pearson	19.0794	2	<.0001*

Method: Fix hypothesized values, rescale omitted
Note: Hypothesized probabilities did not sum to 1. Probabilities have been rescaled.

No. At the 0.05 level of significance we reject the null hypothesis of equal probabilities.

Scenario 4

a.



Contingency Table

DMDMARTL

	Count Total % Col % Row %	Married	Widowed	Divorced	Separated	Never married	Living with partner	Refused	Don't Know	Total
RIDRETH1	Mexican American	574 10.04 19.89 57.69	64 1.12 15.20 6.43	68 1.19 11.07 6.83	48 0.84 25.00 4.82	120 2.10 11.45 12.06	120 2.10 21.62 12.06	1 0.02 50.00 0.10	0 0.00 0.00 0.00	995 17.40
	Other Hispanic	353 6.17 12.23 45.96	48 0.84 11.40 6.25	101 1.77 16.45 13.15	44 0.77 22.92 5.73	117 2.05 11.16 15.23	104 1.82 18.74 13.54	1 0.02 50.00 0.13	0 0.00 0.00 0.00	768 13.43
	Non-Hispanic White	980 17.14 33.96 52.60	187 3.27 44.42 10.04	231 4.04 37.62 12.40	39 0.68 20.31 2.09	269 4.70 25.67 14.44	156 2.73 28.11 8.37	0 0.00 0.00 0.00	1 0.02 100.00 0.05	1863 32.58
	Non-Hispanic Black	418 7.31 14.48 34.89	85 1.49 20.19 7.10	157 2.75 25.57 13.11	48 0.84 25.00 4.01	368 6.43 35.11 30.72	122 2.13 21.98 10.18	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1198 20.95
	Other Race - Including Multi-Racial	561 9.81 19.44 62.68	37 0.65 8.79 4.13	57 1.00 9.28 6.37	13 0.23 6.77 1.45	174 3.04 16.60 19.44	53 0.93 9.55 5.92	0 0.00 0.00 0.00	0 0.00 0.00 0.00	895 15.65
	Total	2886 50.46	421 7.36	614 10.74	192 3.36	1048 18.32	555 9.70	2 0.03	1 0.02	5719

Tests

N	DF	-LogLike	RSquare (U)
5719	28	199.00600	0.0243

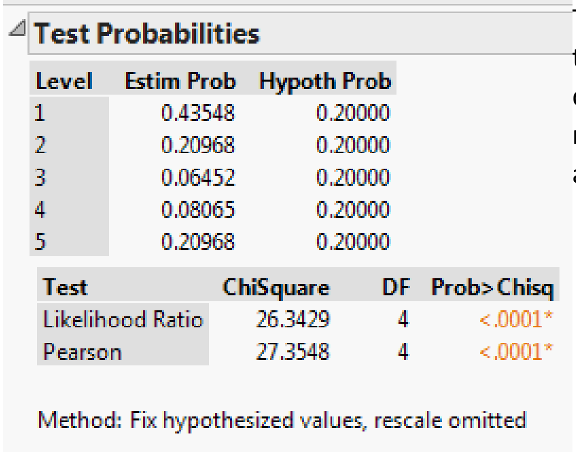
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	398.012	<.0001*
Pearson	399.570	<.0001*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Because there are a substantial proportion of cells with very small expected counts, we should use caution making inferences from the ChiSquare test. However, we can note that the evidence points toward rejecting the null hypothesis of independence. We might observe (for example) that married respondents were disproportionately non-Hispanic whites.

Scenario 6

a.



Test Probabilities

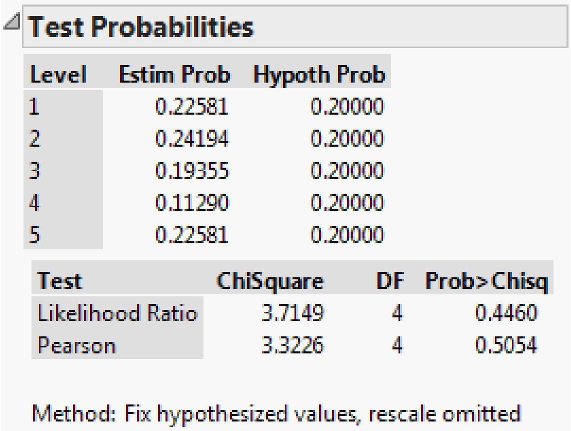
Level	Estim Prob	Hypoth Prob
1	0.43548	0.20000
2	0.20968	0.20000
3	0.06452	0.20000
4	0.08065	0.20000
5	0.20968	0.20000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	26.3429	4	<.0001*
Pearson	27.3548	4	<.0001*

Method: Fix hypothesized values, rescale omitted

The Chi-Square goodness-of-fit test indicates that the five categories are not equally distributed across mammalian species. We reject the null hypothesis that all proportions are equal at 0.20.

b.



Test Probabilities

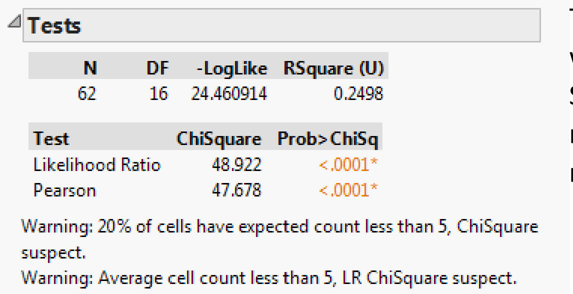
Level	Estim Prob	Hypoth Prob
1	0.22581	0.20000
2	0.24194	0.20000
3	0.19355	0.20000
4	0.11290	0.20000
5	0.22581	0.20000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	3.7149	4	0.4460
Pearson	3.3226	4	0.5054

Method: Fix hypothesized values, rescale omitted

In this case the Chi-Square goodness of fit test does *not* reject the null hypothesis of equal distribution. In other words, we should NOT conclude that species are unequally distributed across the predation index.

c.



Tests

N	DF	-LogLike	RSquare (U)
62	16	24.460914	0.2498

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	48.922	<.0001*
Pearson	47.678	<.0001*

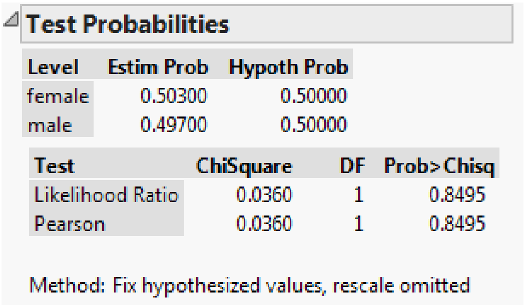
Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Warning: Average cell count less than 5, LR ChiSquare suspect.

The total sample size here leads to many cells with expected counts < 5, making the Chi-Square test unreliable. That said, the test results point in the direction of rejecting the null hypothesis.

Scenario 8

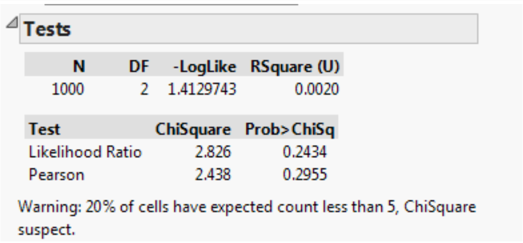
a.



```
Test Probabilities
Level    Estim Prob   Hypoth Prob
female     0.50300      0.50000
male       0.49700      0.50000

Test              ChiSquare   DF   Prob>Chisq
Likelihood Ratio     0.0360    1       0.8495
Pearson              0.0360    1       0.8495

Method: Fix hypothesized values, rescale omitted
```

According to the Chi-Square test there is not sufficient evidence to reject a null hypothesis that mothers are equally likely to give birth to a male as a female baby.

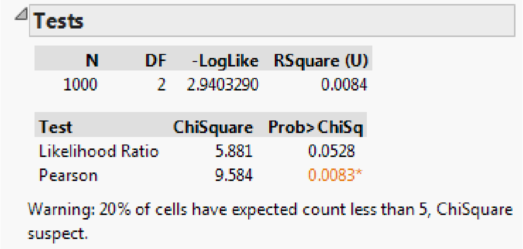
b.



```
Tests
   N    DF    -LogLike   RSquare (U)
1000     2   1.4129743        0.0020

Test              ChiSquare   Prob>ChiSq
Likelihood Ratio      2.826       0.2434
Pearson               2.438       0.2955
Warning: 20% of cells have expected count less than 5, ChiSquare
suspect.
```

We should be reluctant to draw inferences about this question because of the high number of cells with counts less than 5. At any rate, there does not seem to be sufficient evidence to reject a null hypothesis that they are independent.

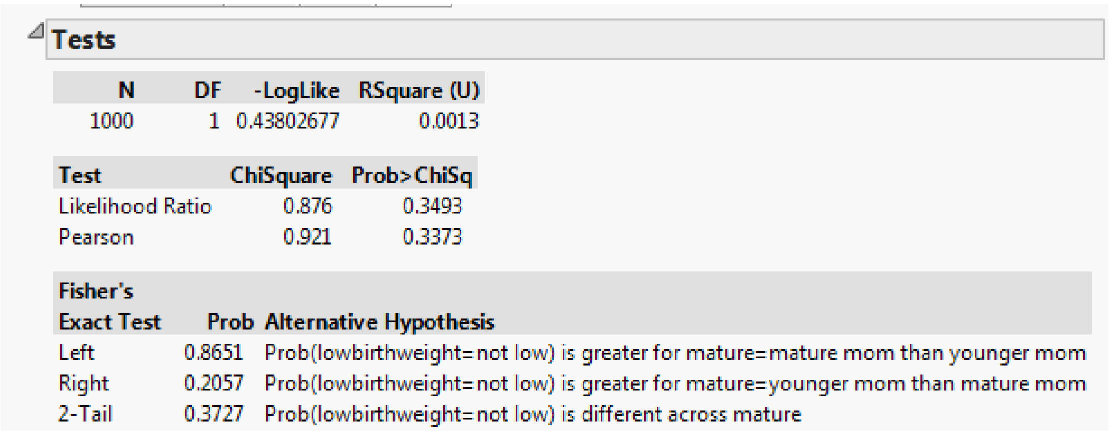
c.



```
Tests
   N    DF    -LogLike   RSquare (U)
1000     2   2.9403290        0.0084

Test              ChiSquare   Prob>ChiSq
Likelihood Ratio      5.881       0.0528
Pearson               9.584       0.0083*
Warning: 20% of cells have expected count less than 5, ChiSquare
suspect.
```

We should be reluctant to draw inferences about this question because of the high number of cells with counts less than 5. That said, Pearson's test does indicate sufficient evidence to reject a null hypothesis that they are independent. It would be wise to obtain a larger sample before drawing a conclusion.

d.



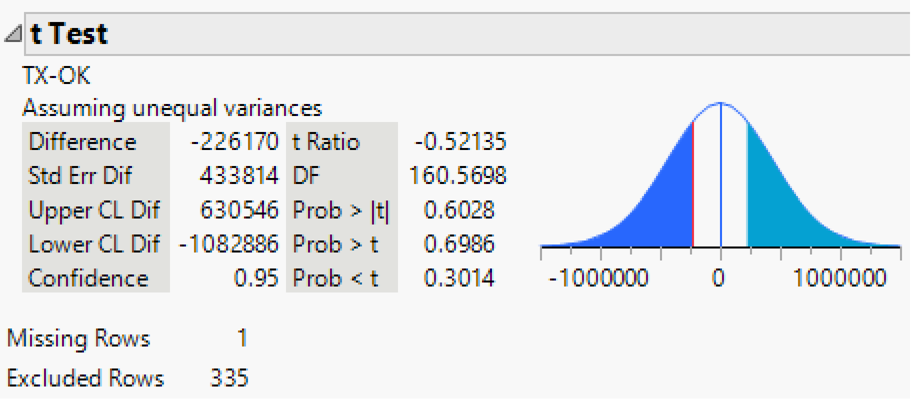
```
Tests
   N    DF    -LogLike   RSquare (U)
1000     1  0.43802677        0.0013

Test              ChiSquare   Prob>ChiSq
Likelihood Ratio      0.876       0.3493
Pearson               0.921       0.3373

Fisher's
Exact Test     Prob   Alternative Hypothesis
Left         0.8651   Prob(lowbirthweight=not low) is greater for mature=mature mom than younger mom
Right        0.2057   Prob(lowbirthweight=not low) is greater for mature=younger mom than mature mom
2-Tail       0.3727   Prob(lowbirthweight=not low) is different across mature
```

According to the Chi-Square test and Fisher's Exact test, there is not sufficient evidence to reject a null hypothesis that they are independent.

Chapter 13: Solutions to Application Scenarios

Scenario 2

a.



```
t Test
TX-OK
Assuming unequal variances
Difference     -226170  t Ratio    -0.52135
Std Err Dif     433814  DF         160.5698
Upper CL Dif    630546  Prob > |t|   0.6028
Lower CL Dif  -1082886  Prob > t     0.6986
Confidence        0.95  Prob < t     0.3014

Missing Rows        1
Excluded Rows     335
```

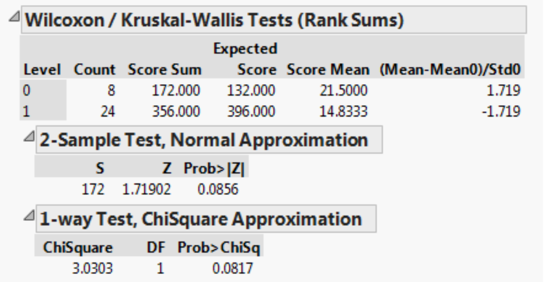
We should first note that the distribution of property damage costs is highly skewed in both states. The samples are moderately large, so the Central Limit Theorem may apply. The computed 95% interval is between --$ 1,082,886 and $630,546.

b. Using the output shown in part a, we see no strong evidence of a difference. We fail to reject the null hypothesis of no difference in costs between the two states.

Scenario 4

a. Student answers will differ. We have only 8 individuals without PD, and for the baseline pitch and jitter, the distributions appear bimodal with few observations in the "center"; shimmer may be normally distributed for non-PD observations. Among individuals with PD (n = 24) the distributions tend to be skewed. As such, with non-normal distributions and small samples, this sample does not satisfy the conditions for the use of the t-test.

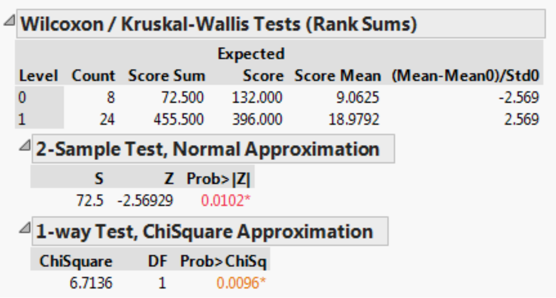
b.



```
Wilcoxon / Kruskal-Wallis Tests (Rank Sums)
                          Expected
Level  Count  Score Sum      Score  Score Mean  (Mean-Mean0)/Std0
0          8    172.000    132.000     21.5000              1.719
1         24    356.000    396.000     14.8333             -1.719

2-Sample Test, Normal Approximation
      S        Z  Prob>|Z|
    172  1.71902    0.0856

1-way Test, ChiSquare Approximation
ChiSquare    DF  Prob>ChiSq
   3.0303     1      0.0817
```

Based on the Wilcoxon test (assuming a significance level of α = 0.05) we fail to reject the null hypothesis that the mean fundamental frequency is equal for both groups. There is no significant difference in this sample data.

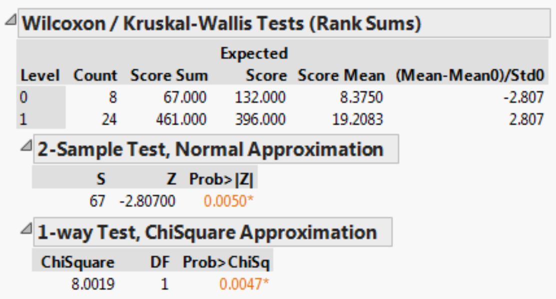
c.



```
Wilcoxon / Kruskal-Wallis Tests (Rank Sums)
                               Expected
Level  Count  Score Sum     Score  Score Mean  (Mean-Mean0)/Std0
0          8     72.500   132.000      9.0625             -2.569
1         24    455.500   396.000     18.9792              2.569

  2-Sample Test, Normal Approximation
          S         Z  Prob>|Z|
       72.5  -2.56929   0.0102*

  1-way Test, ChiSquare Approximation
   ChiSquare   DF  Prob>ChiSq
      6.7136    1     0.0096*
```

Based on the Wilcoxon test (assuming a significance level of $\alpha = 0.05$) we reject the null hypothesis that the mean jitter measurement is equal for both groups. There is a statistically significant difference in this sample data.

d.

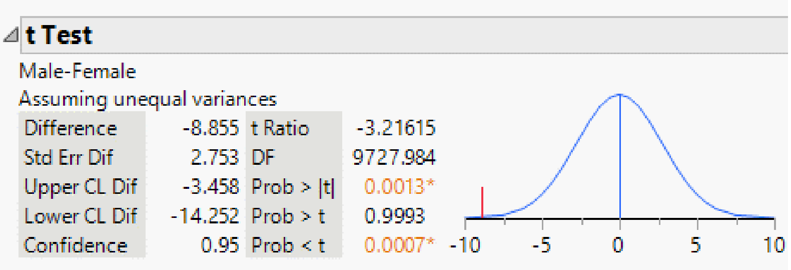


```
Wilcoxon / Kruskal-Wallis Tests (Rank Sums)
                               Expected
Level  Count  Score Sum     Score  Score Mean  (Mean-Mean0)/Std0
0          8     67.000   132.000      8.3750             -2.807
1         24    461.000   396.000     19.2083              2.807

  2-Sample Test, Normal Approximation
          S         Z  Prob>|Z|
         67  -2.80700   0.0050*

  1-way Test, ChiSquare Approximation
   ChiSquare   DF  Prob>ChiSq
      8.0019    1     0.0047*
```

Based on the Wilcoxon test (assuming a significance level of $\alpha = 0.05$) we reject the null hypothesis that the mean shimmer measurement is equal for both groups. There is a statistically significant difference in this sample data.

Scenario 6

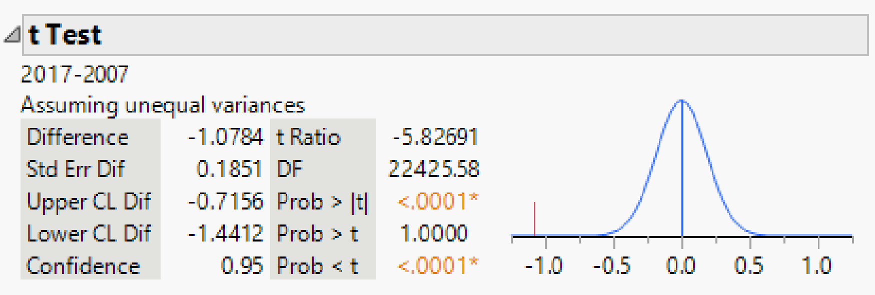
a.



```
t Test
Male-Female
Assuming unequal variances
Difference     -8.855  t Ratio    -3.21615
Std Err Dif     2.753  DF         9727.984
Upper CL Dif   -3.458  Prob > |t|   0.0013*
Lower CL Dif  -14.252  Prob > t     0.9993
Confidence       0.95  Prob < t     0.0007*
```

Using just the 2017 data, we find symmetric distributions in two large subsamples. We estimate with 95% confidence that females reported sleeping between 3.46 and 14.25 minutes more than males.

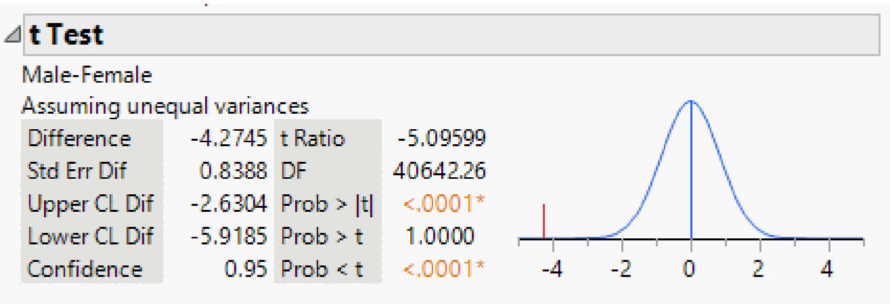
b.



```
t Test
2017-2007
Assuming unequal variances
Difference    -1.0784  t Ratio    -5.82691
Std Err Dif    0.1851  DF         22425.58
Upper CL Dif  -0.7156  Prob > |t|   <.0001*
Lower CL Dif  -1.4412  Prob > t     1.0000
Confidence       0.95  Prob < t     <.0001*
```

We can safely draw inferences because despite the skewed distributions, the samples are large enough to rely on the Central Limit Theorem. We can infer that people spent more time on email in 2007 than in 2017: we estimate with 95% confidence that the mean time devoted to email was somewhere between 0.72 and 1.44 minutes longer per day in 2007 than in 2017.

c.



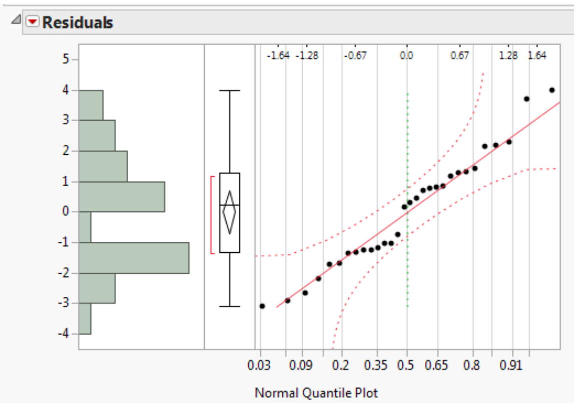
```
t Test
Male-Female
Assuming unequal variances
Difference     -4.2745  t Ratio    -5.09599
Std Err Dif     0.8388  DF         40642.26
Upper CL Dif   -2.6304  Prob > |t|  <.0001*
Lower CL Dif   -5.9185  Prob > t    1.0000
Confidence        0.95  Prob < t    <.0001*
```

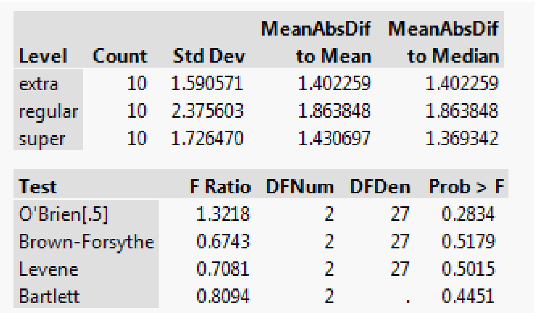
Combining all of the data from both years, we can conclude with 95% confidence that men spend, on average, 2.6 to 5.9 fewer minutes per day socializing than do women.

Chapter 14: Solutions to Application Scenarios

Scenario 2

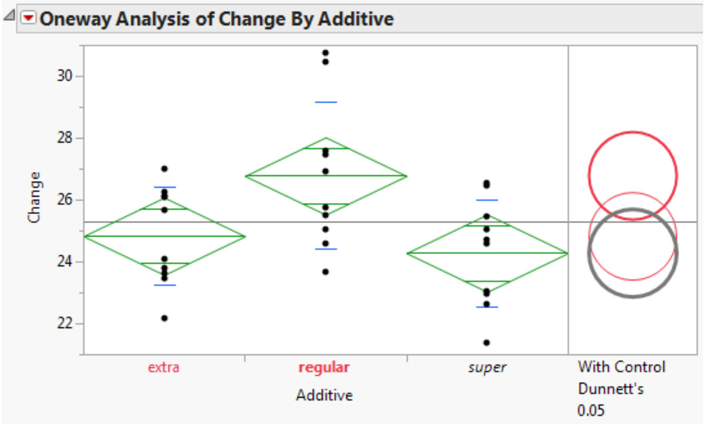
a. We see no evidence that the ANOVA assumptions have been violated; variances across the three groups appear to be equal and residuals are approximately normal. The F Ratio of 4.6275 and corresponding P-value of 0.0187 indicate that we should reject the null hypothesis of equal means; there is compelling evidence that the different additives lead to different mean changes.

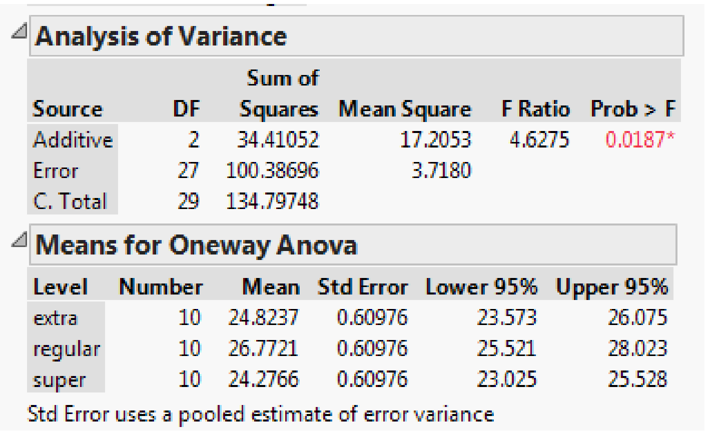




Level	Count	Std Dev	MeanAbsDif to Mean	MeanAbsDif to Median
extra	10	1.590571	1.402259	1.402259
regular	10	2.375603	1.863848	1.863848
super	10	1.726470	1.430697	1.369342

Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	1.3218	2	27	0.2834
Brown-Forsythe	0.6743	2	27	0.5179
Levene	0.7081	2	27	0.5015
Bartlett	0.8094	2	.	0.4451





Analysis of Variance

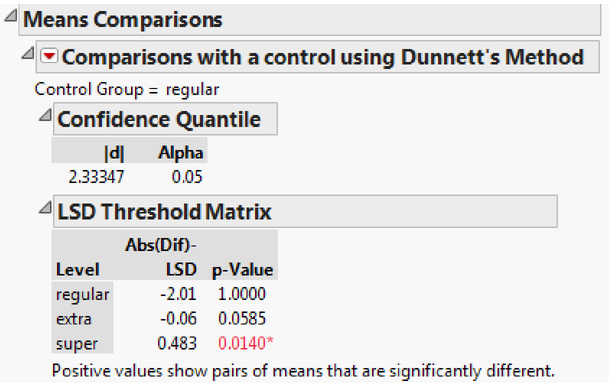
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Additive	2	34.41052	17.2053	4.6275	0.0187*
Error	27	100.38696	3.7180		
C. Total	29	134.79748			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
extra	10	24.8237	0.60976	23.573	26.075
regular	10	26.7721	0.60976	25.521	28.023
super	10	24.2766	0.60976	23.025	25.528

Std Error uses a pooled estimate of error variance

b.



Means Comparisons

Comparisons with a control using Dunnett's Method

Control Group = regular

Confidence Quantile

|d|	Alpha
2.33347	0.05

LSD Threshold Matrix

Level	Abs(Dif)-LSD	p-Value
regular	-2.01	1.0000
extra	-0.06	0.0585
super	0.483	0.0140*

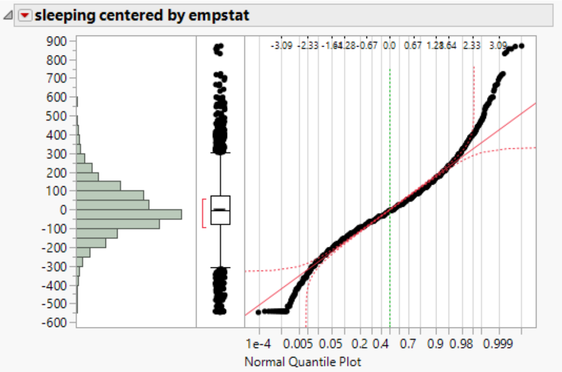
Positive values show pairs of means that are significantly different.

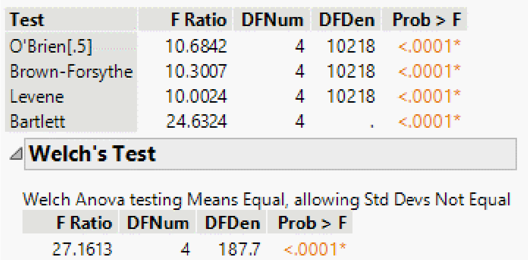
Because we have a control group, we should use Dunnett's method to compare the means.

c. We find that there is a significant improvement in insulation with the "super" additive—the temperature change is smallest with that additive. The company should switch from regular to super.

Scenario 4

a.





Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	10.6842	4	10218	<.0001*
Brown-Forsythe	10.3007	4	10218	<.0001*
Levene	10.0024	4	10218	<.0001*
Bartlett	24.6324	4	.	<.0001*

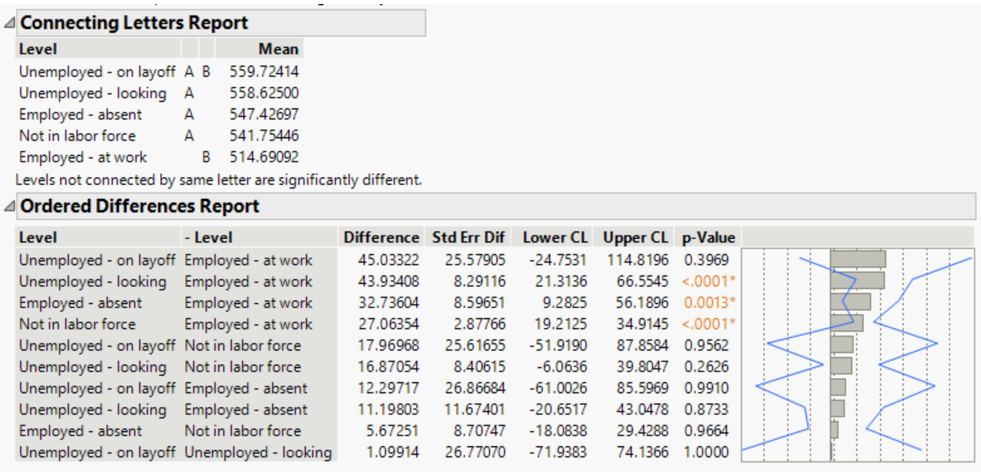
Welch's Test

Welch Anova testing Means Equal, allowing Std Devs Not Equal

F Ratio	DFNum	DFDen	Prob > F
27.1613	4	187.7	<.0001*

As usual we start by evaluating assumptions. We have a very large sample, so the Central Limit Theorem applies. The residuals are unimodal and symmetric but depart from the normal model in the tails). We also see evidence that the variances are unequal. In practice, because of the very large sample it is not surprising that we find significant differences.

Both Welch's test and the standard ANOVA results strongly indicate that there are significant differences in group means.



Connecting Letters Report

Level			Mean
Unemployed - on layoff	A	B	559.72414
Unemployed - looking	A		558.62500
Employed - absent	A		547.42697
Not in labor force	A		541.75446
Employed - at work		B	514.69092

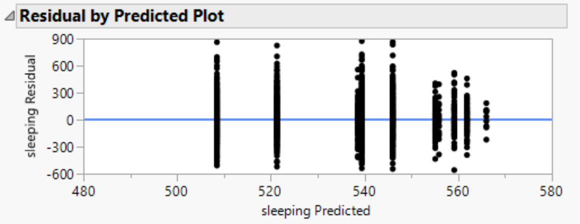
Levels not connected by same letter are significantly different.

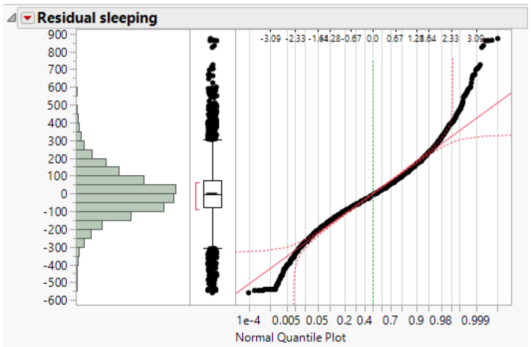
Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
Unemployed - on layoff	Employed - at work	45.03322	25.57905	-24.7531	114.8196	0.3969
Unemployed - looking	Employed - at work	43.93408	8.29116	21.3136	66.5545	<.0001*
Employed - absent	Employed - at work	32.73604	8.59651	9.2825	56.1896	0.0013*
Not in labor force	Employed - at work	27.06354	2.87766	19.2125	34.9145	<.0001*
Unemployed - on layoff	Not in labor force	17.96968	25.61655	-51.9190	87.8584	0.9562
Unemployed - looking	Not in labor force	16.87054	8.40615	-6.0636	39.8047	0.2626
Unemployed - on layoff	Employed - absent	12.29717	26.86684	-61.0026	85.5969	0.9910
Unemployed - looking	Employed - absent	11.19803	11.67401	-20.6517	43.0478	0.8733
Employed - absent	Not in labor force	5.67251	8.70747	-18.0838	29.4288	0.9664
Unemployed - on layoff	Unemployed - looking	1.09914	26.77070	-71.9383	74.1366	1.0000

There is no control group here. Tukey's HSD indicates that employed people at work get the least sleep and unemployed people on layoff report the most. All others are indistinguishable from one another.

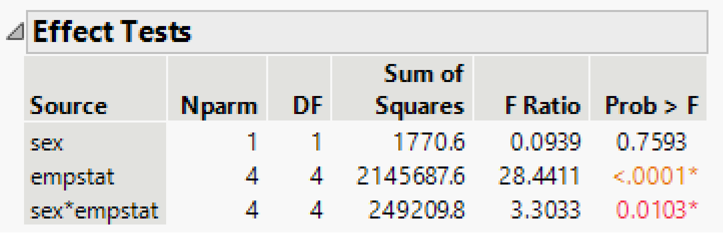
b.





We start again by evaluating assumptions. The residual by Predicted Plot, shown above with X axis rescaled for clarity, seems to indicate non-constant variance. We have a very large sample, so the Central Limit Theorem applies and we need not be overly concerned with normality (above we see the residuals are unimodal and symmetric, but depart from the normal model in the tails).

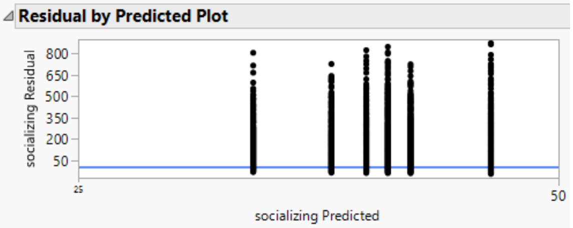
Adding the second variable (sex) to the model does not improve it much.
As we can see in the Effects Tests, sex has no main effect, but there is a significant interaction effect. The effect of employment status on an individual's sleeping patterns is different for men and women.



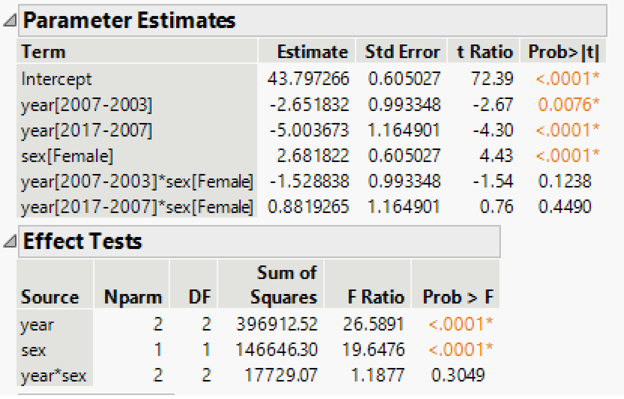
Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
sex	1	1	1770.6	0.0939	0.7593
empstat	4	4	2145687.6	28.4411	<.0001*
sex*empstat	4	4	249209.8	3.3033	0.0103*

c.



The rescaled residual by predicted plot appears to show near-constant variance, but strongly right-skewed residuals. here again, we have large subgroup sizes so normality is not a major issue.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	43.797266	0.605027	72.39	<.0001*
year[2007-2003]	-2.651832	0.993348	-2.67	0.0076*
year[2017-2007]	-5.003673	1.164901	-4.30	<.0001*
sex[Female]	2.681822	0.605027	4.43	<.0001*
year[2007-2003]*sex[Female]	-1.528838	0.993348	-1.54	0.1238
year[2017-2007]*sex[Female]	0.8819265	1.164901	0.76	0.4490

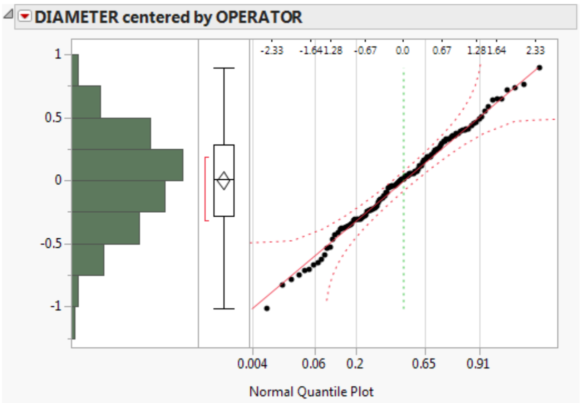
Effect Tests

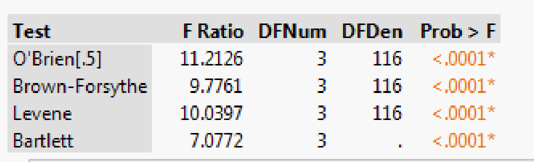
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
year	2	2	396912.52	26.5891	<.0001*
sex	1	1	146646.30	19.6476	<.0001*
year*sex	2	2	17729.07	1.1877	0.3049

The estimates and effect tests indicate there are significant main effects but no interaction between year and sex. Overall, people seem to be socializing less each year of the survey, and women socialize more than men. The dropoff in time spent socializing was nearly twice as large from 2007 to 2017 as it was between 2003 and 2007.

Scenario 6

a.



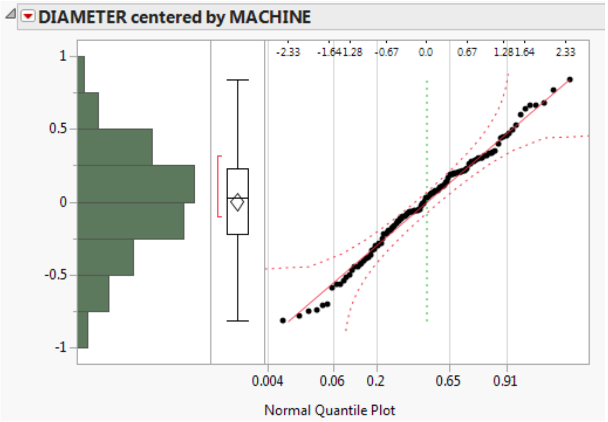


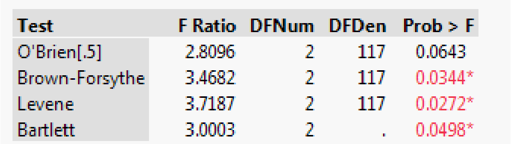
Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	11.2126	3	116	<.0001*
Brown-Forsythe	9.7761	3	116	<.0001*
Levene	10.0397	3	116	<.0001*
Bartlett	7.0772	3	.	<.0001*

We start by examining assumptions. The residuals appear to be normally distributed (the sample sizes are large enough to rely on the Central Limit Theorem in this case), but the subsamples appear not to share a common variance.

Both Welch's test and the conventional ANOVA find no significant differences among group means.

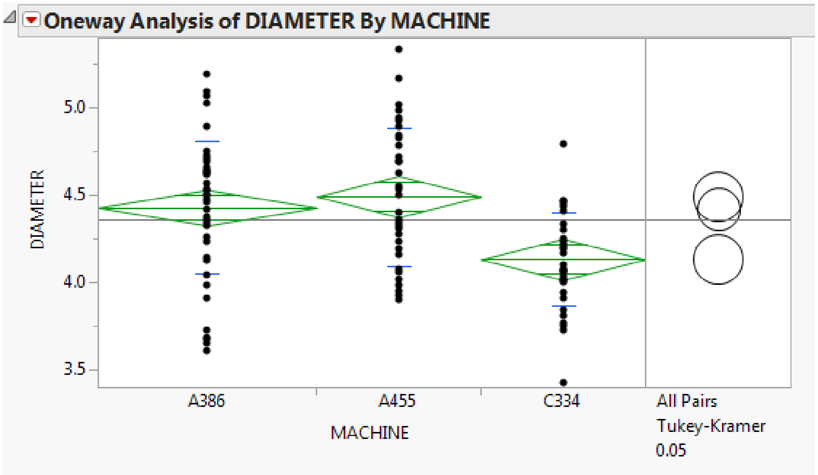
b.



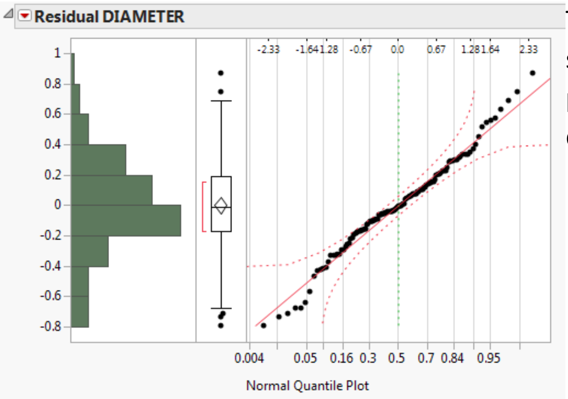


Test	F Ratio	DFNum	DFDen	Prob > F
O'Brien[.5]	2.8096	2	117	0.0643
Brown-Forsythe	3.4682	2	117	0.0344*
Levene	3.7187	2	117	0.0272*
Bartlett	3.0003	2	.	0.0498*

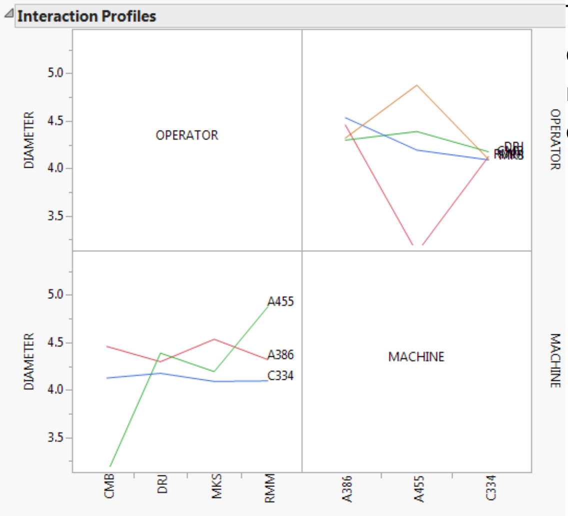
In this analysis the assumption of normality is satisfied; the tests for equal variances are not all in agreement so we may question that assumption. Both Welch's test and the ANOVA indicate a significant difference in mean diameters for at least one machine. Tukey's HSD finds that machine C334 has lower mean diameters than the other machines.



c. The assumption of normality does appear to be satisfied; visual inspection of residuals vs. predicted values does not reveal any obvious differences in group variances.



The interaction plots indicate interaction effects between operator and machine, making it difficult to interpret the main effects of machine and operator separately.

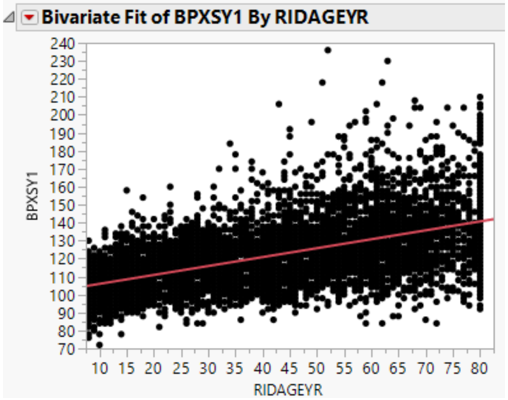


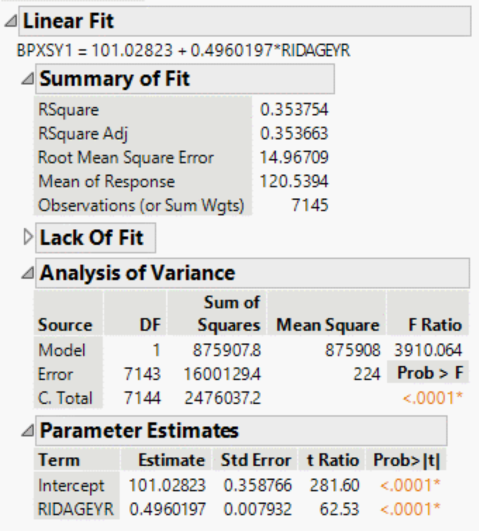
d. The interaction plot is a bit difficult to read because the Operator initials are superimposed on one another. The profiler makes it easier to see that the *extent* to which machines produce tubing of differing widths varies by operator. Thus, for example, when Operator RMM is involved, machine A455 regularly makes the widest diameters; otherwise it does not. RRM's tubing diameters appear to vary widely by machine, whereas DRJ's do not.

Chapter 15: Solutions to Application Scenarios

Scenario 2

a.





Linear Fit

BPXSY1 = 101.02823 + 0.4960197*RIDAGEYR

Summary of Fit

RSquare	0.353754
RSquare Adj	0.353663
Root Mean Square Error	14.96709
Mean of Response	120.5394
Observations (or Sum Wgts)	7145

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	875907.8	875908	3910.064
Error	7143	1600129.4	224	Prob > F
C. Total	7144	2476037.2		<.0001*

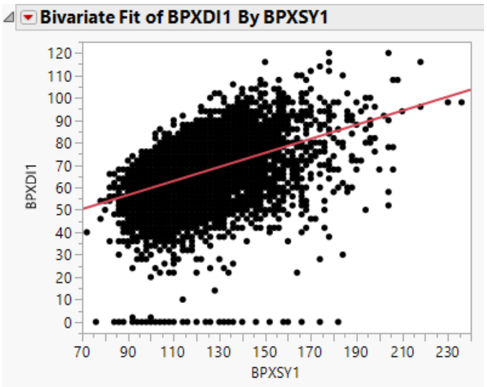
Parameter Estimates

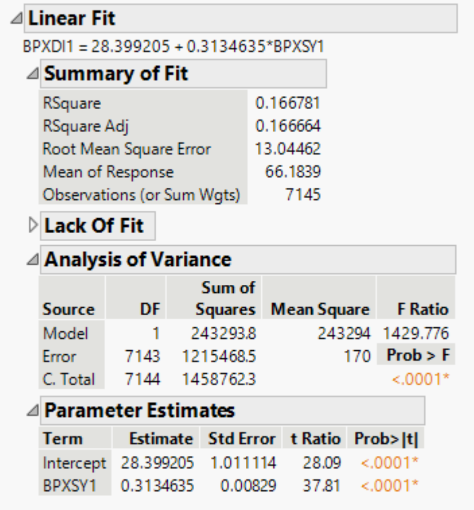
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	101.02823	0.358766	281.60	<.0001*
RIDAGEYR	0.4960197	0.007932	62.53	<.0001*

In this regression we find a weak ($R^2 = 0.35$) but highly significant positive relationship. Subjects who differ in age by 1 year tend to have, on average, systolic BP that is approximately 0.496 points higher per year. This is not a strong relationship because age accounts for less than one-third of the variation in systolic BP.

b. NOTE: The question does not specify which column should be treated as Y and which as X. Because systolic pressure is the pressure of blood leaving the heart, and diastolic is the pressure of returning blood, it makes sense to use Diastolic as Y. Students who reverse the columns will see the same R^2 and significance levels.

NOTE ALSO: If we use the entire data table (as shown here) we find a horizontal row of points corresponding to respondents for whom we have systolic readings, but diastolic readings of 0.





Linear Fit

BPXDI1 = 28.399205 + 0.3134635*BPXSY1

Summary of Fit

RSquare	0.166781
RSquare Adj	0.166664
Root Mean Square Error	13.04462
Mean of Response	66.1839
Observations (or Sum Wgts)	7145

Lack Of Fit

Analysis of Variance

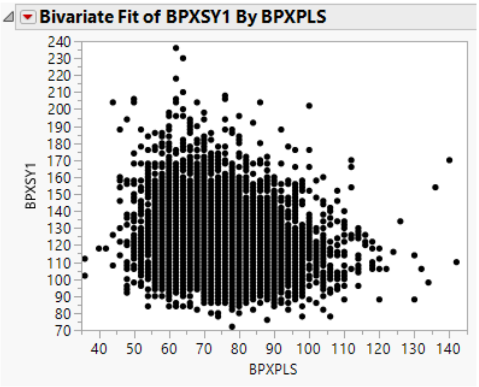
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	243293.8	243294	1429.776
Error	7143	1215468.5	170	Prob > F
C. Total	7144	1458762.3		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	28.399205	1.011114	28.09	<.0001*
BPXSY1	0.3134635	0.00829	37.81	<.0001*

Here we find a significant but weak (R^2 = 0.17) positive relationship. For each additional 1 point of systolic BP, diastolic increases by 0.313 points. If we exclude the 0 diastolic points, R^2 increases only slightly to 0.2 and the slope barely changes, to 0.316.

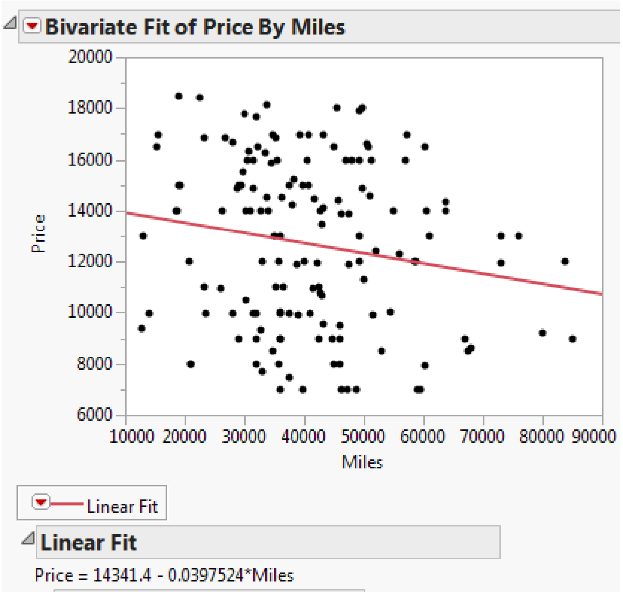
c.



The scatterplot to the left shows little or no relationship between pulse and systolic BP. If anything, there may be a very weak negative relationship here, contrary to the suspicion expressed in the question.

Scenario 4

a.



Linear Fit

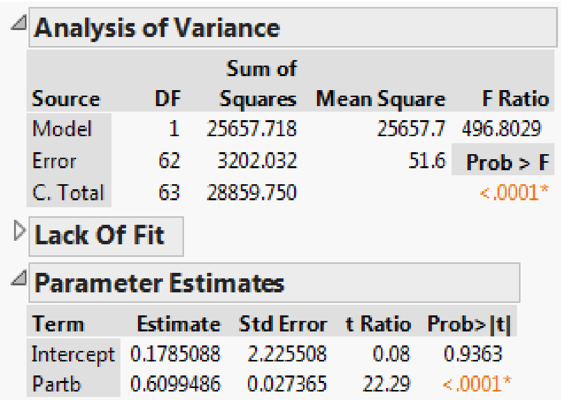
Price = 14341.4 - 0.0397524*Miles

The equation appears beneath the graph, and R^2 = 0.03.

This regression shows there is a weak, significant negative relationship between mileage and price for used cars. The further a car has been driven, on average the lower the price (about 4 cents per mile, on average). However, there is considerable scatter around the line.

Scenario 6

a.



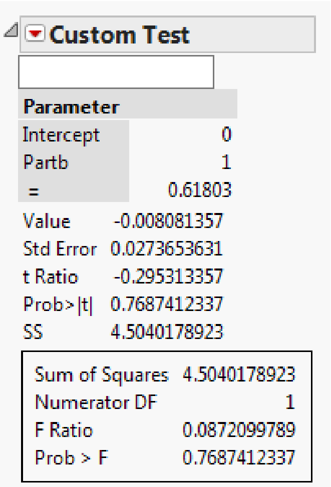
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	25657.718	25657.7	496.8029
Error	62	3202.032	51.6	Prob > F
C. Total	63	28859.750		<.0001*

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.1785088	2.225508	0.08	0.9363
Partb	0.6099486	0.027365	22.29	<.0001*



Custom Test

Parameter	
Intercept	0
Partb	1
=	0.61803

Value	-0.008081357
Std Error	0.0273653631
t Ratio	-0.295313357
Prob>|t|	0.7687412337
SS	4.5040178923

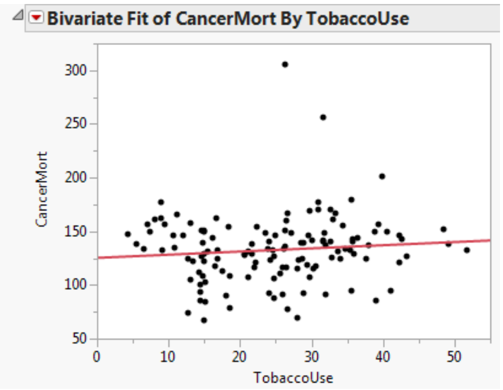
Sum of Squares	4.5040178923
Numerator DF	1
F Ratio	0.0872099789
Prob > F	0.7687412337

Using the Haydn data, we find a similar story to the one we saw with Mozart. We again find the Golden Mean model plausible.

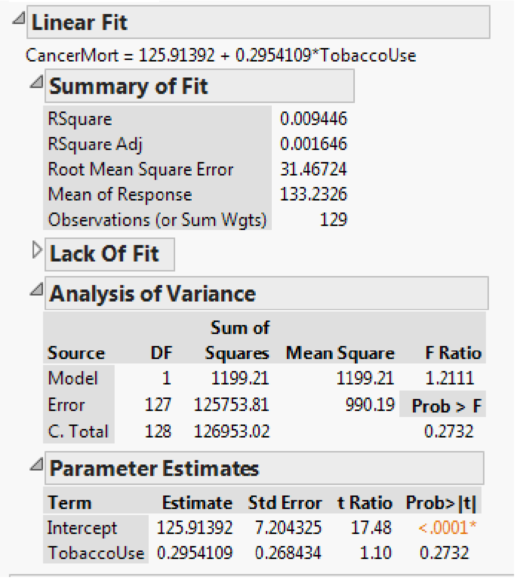
b. Here, the R^2 value (not shown) is .889; with the Mozart data R^2 was .938 which is slightly better. In both cases the linear model fits the data very well.

Scenario 8

a.



Bivariate Fit of CancerMort By TobaccoUse



Linear Fit

CancerMort = 125.91392 + 0.2954109*TobaccoUse

Summary of Fit

RSquare	0.009446
RSquare Adj	0.001646
Root Mean Square Error	31.46724
Mean of Response	133.2326
Observations (or Sum Wgts)	129

Lack Of Fit

Analysis of Variance

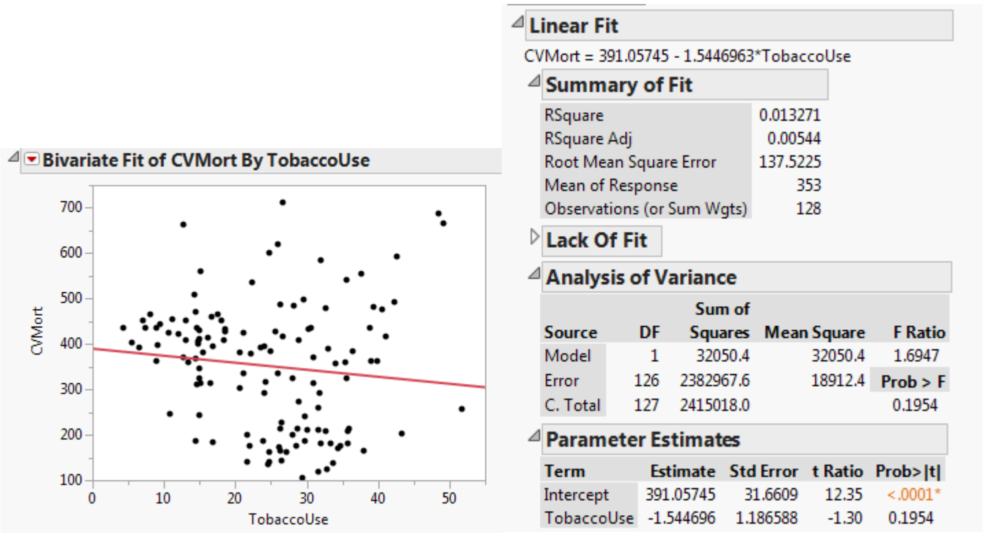
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1199.21	1199.21	1.2111
Error	127	125753.81	990.19	Prob > F
C. Total	128	126953.02		0.2732

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	125.91392	7.204325	17.48	<.0001*
TobaccoUse	0.2954109	0.268434	1.10	0.2732

We find a non-significant relationship here – Tobacco Use is not a useful predictor of cancer deaths in a country.

b.



Linear Fit

CVMort = 391.05745 - 1.5446963*TobaccoUse

Summary of Fit

RSquare	0.013271
RSquare Adj	0.00544
Root Mean Square Error	137.5225
Mean of Response	353
Observations (or Sum Wgts)	128

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	32050.4	32050.4	1.6947
Error	126	2382967.6	18912.4	Prob > F
C. Total	127	2415018.0		0.1954

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	391.05745	31.6609	12.35	<.0001*
TobaccoUse	-1.544696	1.186588	-1.30	0.1954

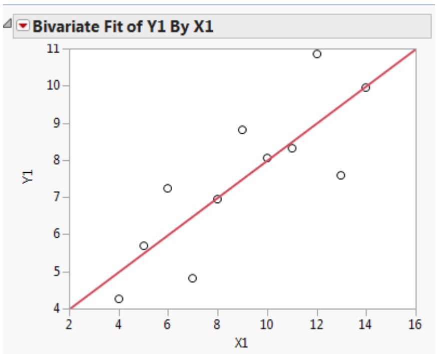
This is also a non-significant relationship. Tobacco Use does not predict cardiovascular mortality rate.

c. The aggregate prevalence of tobacco use obscures the fine distinctions in the amount and length of tobacco use in individuals. We'd really want to look at data at the individual level in order to determine the degree to which increased tobacco use influences the risks of death from cancer or from cardiovascular disease.

Scenario 10

a. There are slight differences, but when we round the major statistics, we find that all four models are nearly identical: $Y_i = 3 + 0.5\ X_i$. All R^2 (0.66) and p-values (0.0022 for the slope) are the same.

b.

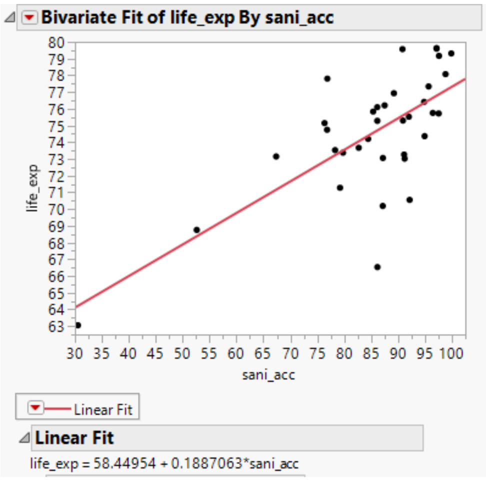


The linear model is an apt description of these points. There is a general linear trend with points scattering evenly above and below the line.

c. In the other three graphs, the points do not fall in a linear pattern at all. This illustrates a substantial risk in running a linear regression without first examining the data visually. (In JMP we *always* see a scatterplot of the points either prior to fitting a model or in conjunction with fitting a model).

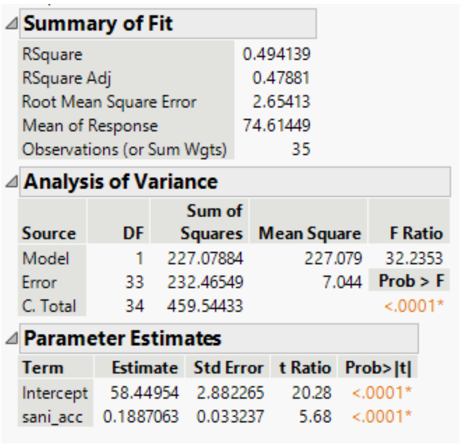
Scenario 12

a.



Linear Fit

life_exp = 58.44954 + 0.1887063*sani_acc



Summary of Fit

RSquare	0.494139
RSquare Adj	0.47881
Root Mean Square Error	2.65413
Mean of Response	74.61449
Observations (or Sum Wgts)	35

Analysis of Variance

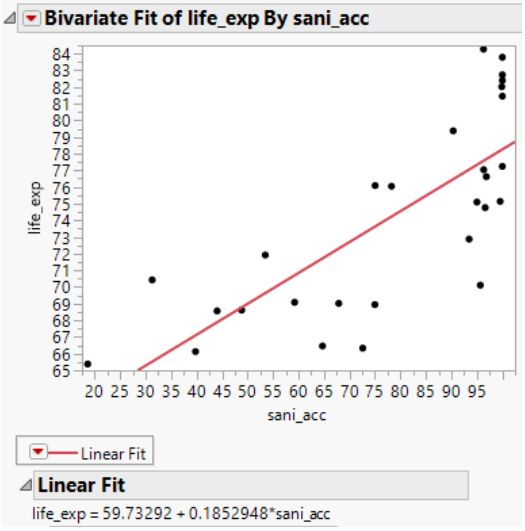
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	227.07884	227.079	32.2353
Error	33	232.46549	7.044	Prob > F
C. Total	34	459.54433		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	58.44954	2.882265	20.28	<.0001*
sani_acc	0.1887063	0.033237	5.68	<.0001*

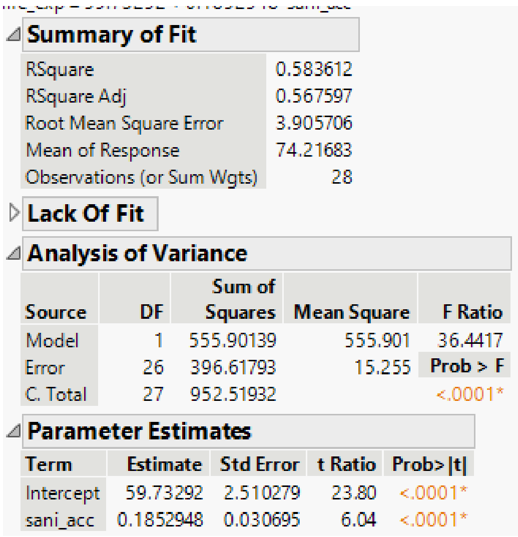
In Latin America & Caribbean Region, countries in which higher percentages of citizens have access to sanitation have greater life expectancies. The equation appears beneath the fitted line plot. The slope is significant at the 0.0001 level, and $R^2 = 0.49$.

b.



Linear Fit

life_exp = 59.73292 + 0.1852948*sani_acc



Summary of Fit

RSquare	0.583612
RSquare Adj	0.567597
Root Mean Square Error	3.905706
Mean of Response	74.21683
Observations (or Sum Wgts)	28

Lack Of Fit

Analysis of Variance

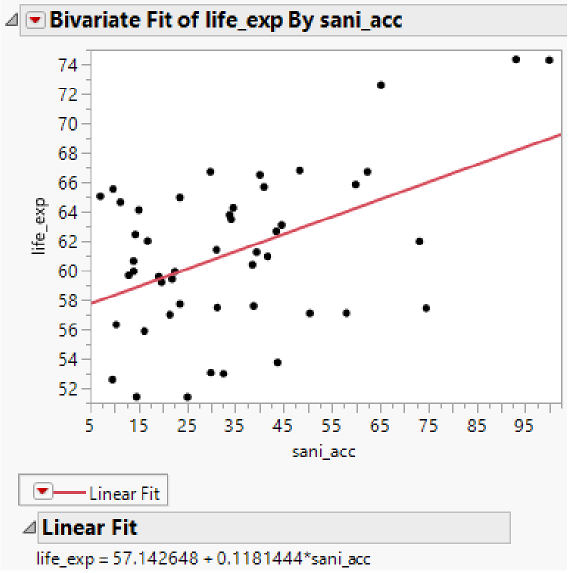
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	555.90139	555.901	36.4417
Error	26	396.61793	15.255	Prob > F
C. Total	27	952.51932		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	59.73292	2.510279	23.80	<.0001*
sani_acc	0.1852948	0.030695	6.04	<.0001*

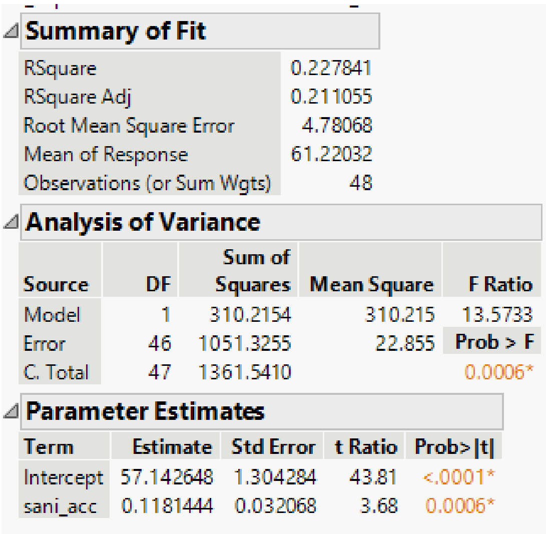
In East Asia & Pacific region, countries in which higher percentages of citizens have access to sanitation have greater life expectancies. The equation appears beneath the fitted line plot. The slope is significant at the 0.0001 level, and $R^2 = 0.58$.

c.



Linear Fit

life_exp = 57.142648 + 0.1181444*sani_acc



Summary of Fit

RSquare	0.227841
RSquare Adj	0.211055
Root Mean Square Error	4.78068
Mean of Response	61.22032
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	310.2154	310.215	13.5733
Error	46	1051.3255	22.855	Prob > F
C. Total	47	1361.5410		0.0006*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	57.142648	1.304284	43.81	<.0001*
sani_acc	0.1181444	0.032068	3.68	0.0006*

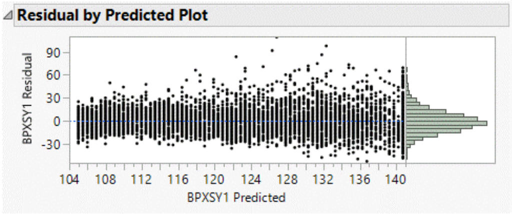
In Sub-Saharan Africa, countries in which higher percentages of citizens have access to sanitation have greater life expectancies. The equation appears beneath the fitted line plot. The slope is significant at the 0.0006 level, but R^2 only = 0.23, indicating a weak fit.

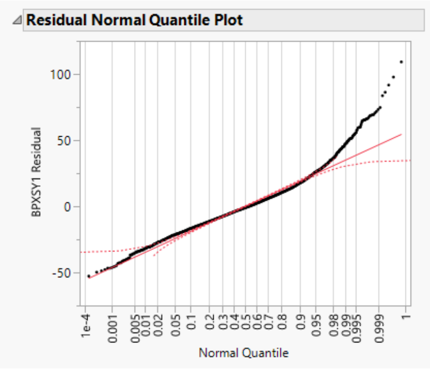
d. The three models are similar in that they provide significant results, with regional differences in slope and intercepts. All three have base life expectancies (intercept) between 57 and 59 years, and marginal increases in life expectancy between 0.12 and 0.19 years. Student answers will vary about reasons for the differences. They should note that the African sample is the largest, but least significant; hence the difference in significance should not be due to small sample size. Responses should refer to other factors influencing life expectancy, and differences within the regions.

Chapter 16: Solutions to Application Scenarios

Scenario 2

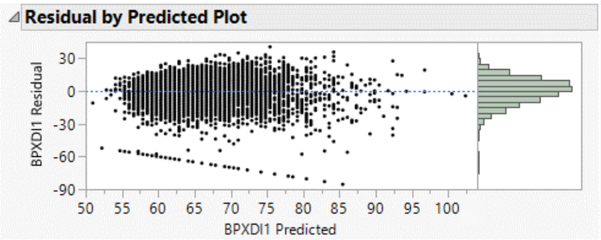
a.

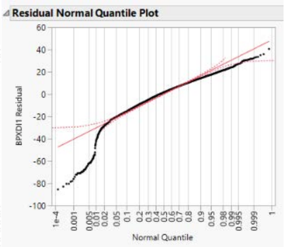




Once again we see the suggestion of heteroskedasticity in the Residual by Predicted Plot. The residuals are largely normal in shape, though somewhat right-skewed. We can probably use the model safely.

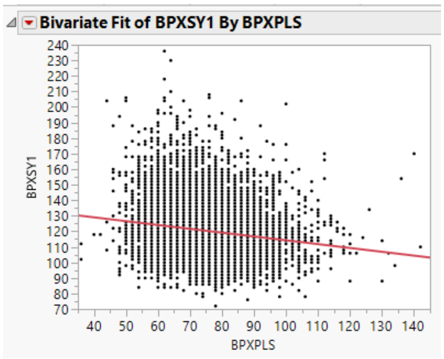
b.



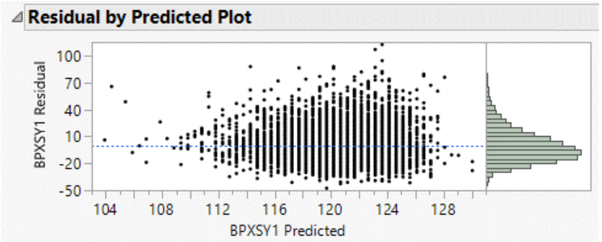


As in the prior chapter, we find a goodly number of observations for which diastolic pressure was reported as 0, but with varied systolic pressures. These appear in the left-hand plot as a downward sloping line. The other residuals for this regression appear to satisfy the assumptions of constant variance, but less so for normality. There is some indication that the variance increase moving from left to right, but the evidence is ambiguous. The residuals are distinctly left-skewed, and the sample size may not be large enough to overcome the skewness.

c.



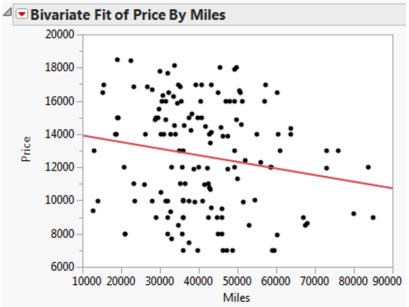
The scatterplot to the left shows little or no relationship between pulse and systolic BP. If anything, there may be a very weak negative relationship here, contrary to the suspicion expressed in the question.

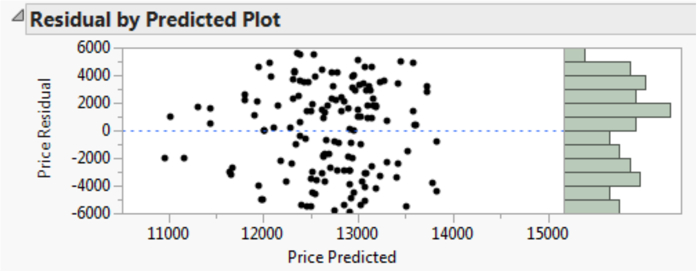


The residuals graphs cast doubt on both normality and constant variance.

Scenario 4

a.





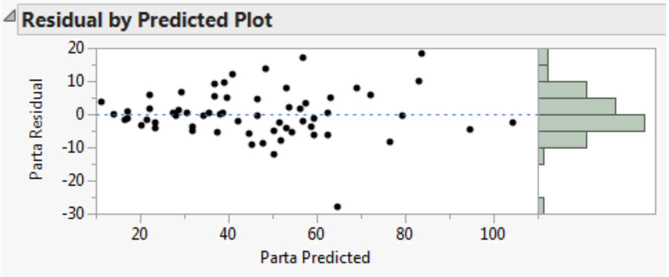
(Note: it is wise to adjust the horizontal axis on the residual by predicted plot to more clearly see the pattern.) The residuals are not normally distributed, there may be a problem with constant variance on the left side of the graph. The sample size may be large enough to rely on the Central Limit Theorem.

b. The 95% confidence interval for the marginal decrease in price associated with each additional mile driven is [– $ 0.076, – $ 0.003].

c. Student answers will vary. The prediction bands on this graph are quite wide, and even with rescaling the axes it is difficult to read predicted values of Y. A reasonable response would be that the price should fall between $6200 to $19,500.

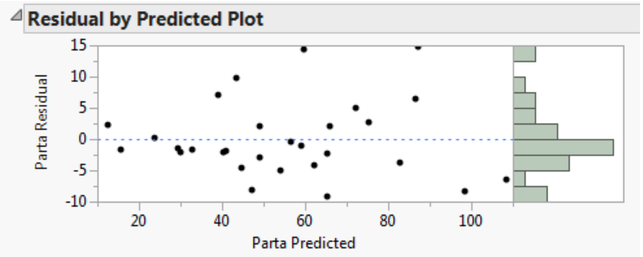
Scenario 6

a.



With the Haydn data, in the Residual vs. Partb plot we find a heteroskedastic pattern; the residual do deviate slightly from normality, but the distribution is single peaked, so inference is probably appropriate.

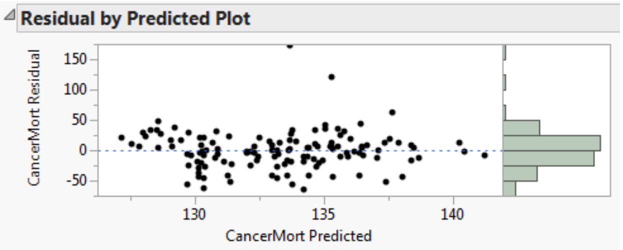
b.



With the Mozart data we also find heteroskedasticity and probable non-normality. Both issues present reasons not to interpret the regression results. With the relatively small Mozart sample, we cannot rely on the Central Limit Theorem with regard to the non-normality.

Scenario 8

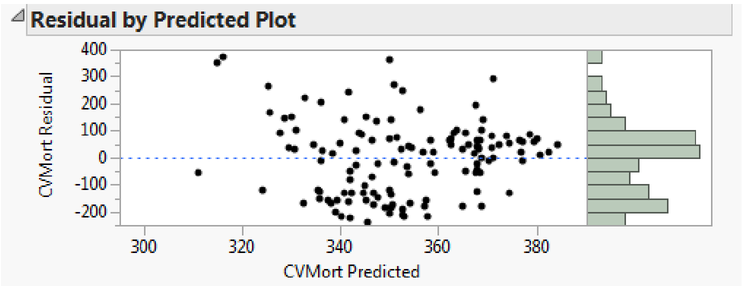
a.



(Note: it is wise to adjust the horizontal axis on the residual by predicted plot to more clearly see the pattern.)

Recall that we find a non-significant relationship here – Tobacco Use is not a useful predictor of cancer deaths in a country. The residuals seem to show more variability in the middle range of tobacco use (non-constant variance), and residuals are nearly normal, with a long upper tail but large sample size. This model is not useful for inference.

b.

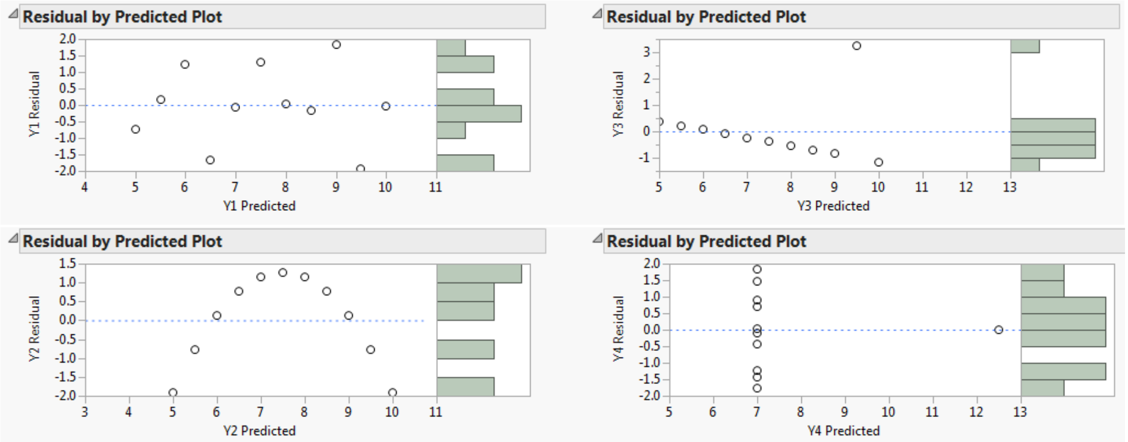


Again, adjust the horizontal axis for clarity. Recall that this is also a non-significant relationship. Tobacco Use does not predict cardiovascular mortality rate.

The residuals indicate some possible curvature (non-linearity) as well as heteroskedasticity. They are not very close to a normal distribution, though the large sample size would permit us to invoke the CLT. This model should not be put to use based on this sample.

Scenario 10

a.

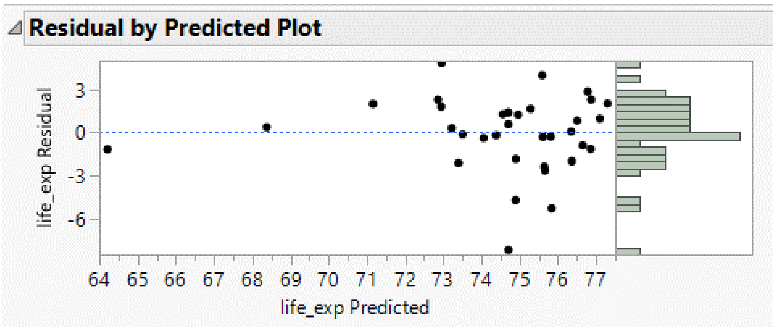


Above are the four plots of residuals vs. predicted. The residuals in the first regression are homoskedastic and approximately normal. The others indicate non-linearity and/or heteroskedasticity. Normality plots also indicate non-normal residuals in these small samples.

b. The four residual vs. X plots indicate that only the first model is suitable for interpretation and use.

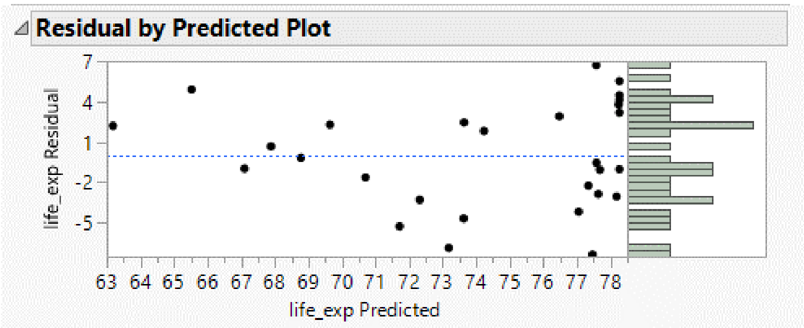
Scenario 12

a.



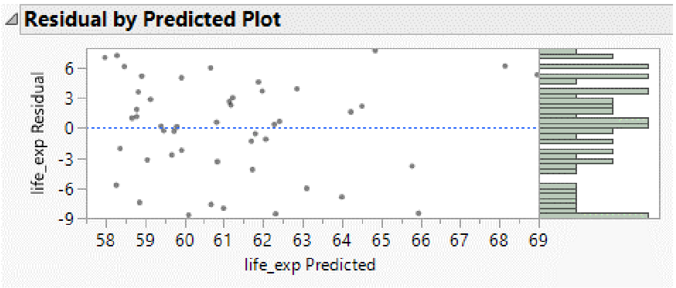
Latin America & Caribbean: These residuals are unimodal and somewhat symmetric. With a small sample it is difficult to determine non-constant variance, but the variance does seem to increase from left to right. Avoid inference.

b.



East Asia & Pacific: These residuals may be non-normal and may have non-constant variance, though with a small sample it is difficult to determine. Avoid inference.

c.



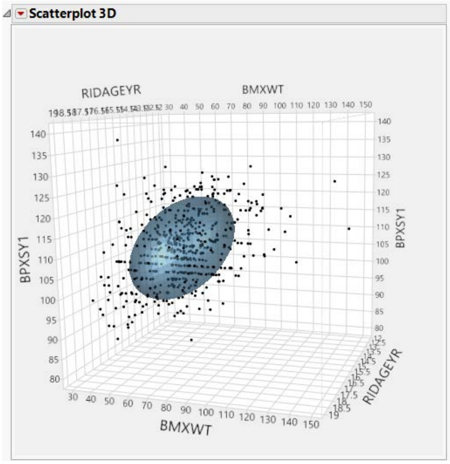
Sub-Saharan Africa: The residuals seem to show constant variance, but normality is questionable. Sample size is large enough to safely make inferences.

d. All three subsamples present some possible violations of the assumptions. Differences are likely due to other variables that are not yet part of the regression models.

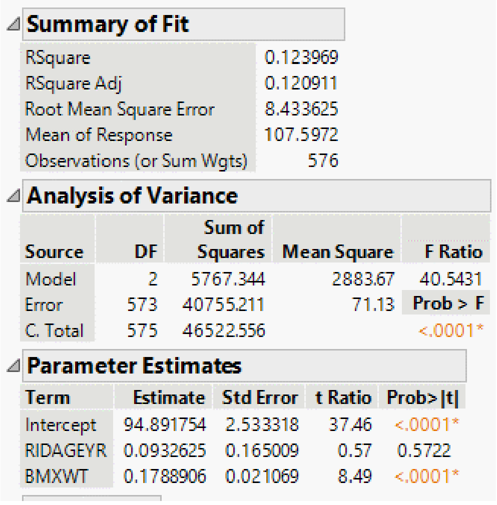
Chapter 18: Solutions to Application Scenarios

Scenario 2

a. Student answers will vary. One rotated scatterplot is shown here (including a density ellipsoid). We see a weak tendency for systolic BP to increase both as age and weight increase.



b.



Summary of Fit

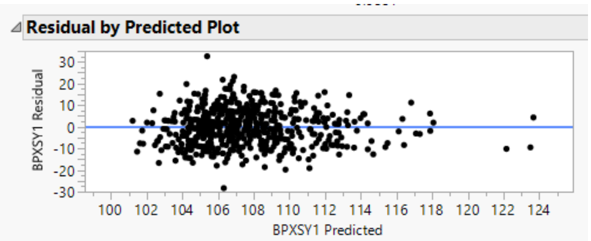
RSquare	0.123969
RSquare Adj	0.120911
Root Mean Square Error	8.433625
Mean of Response	107.5972
Observations (or Sum Wgts)	576

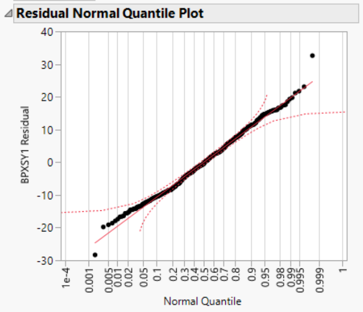
Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	2	5767.344	2883.67	40.5431
Error	573	40755.211	71.13	Prob > F
C. Total	575	46522.556		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	94.891754	2.533318	37.46	<.0001*
RIDAGEYR	0.0932625	0.165009	0.57	0.5722
BMXWT	0.1788906	0.021069	8.49	<.0001*

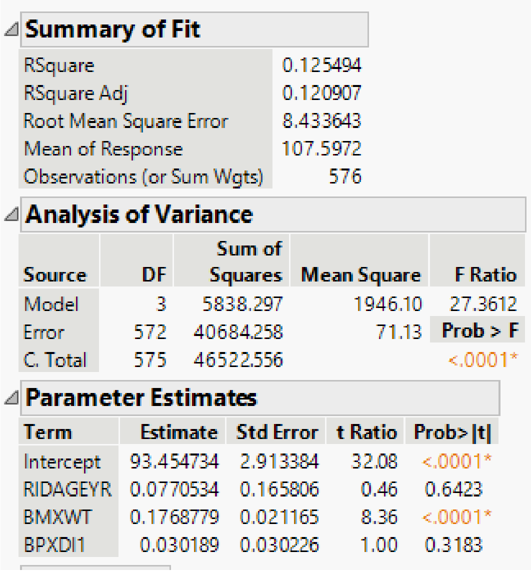




The residuals indicate a possible problem with heteroskedasticity; normality looks fine.

When we look at the regression results, we conclude that there is a significant positive relationship between systolic BP and weight, but that age has no significant effect once age is considered. The overall model fit is poor.

c.



Summary of Fit

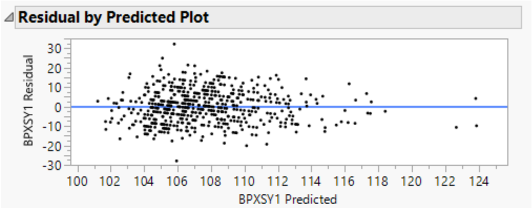
RSquare	0.125494
RSquare Adj	0.120907
Root Mean Square Error	8.433643
Mean of Response	107.5972
Observations (or Sum Wgts)	576

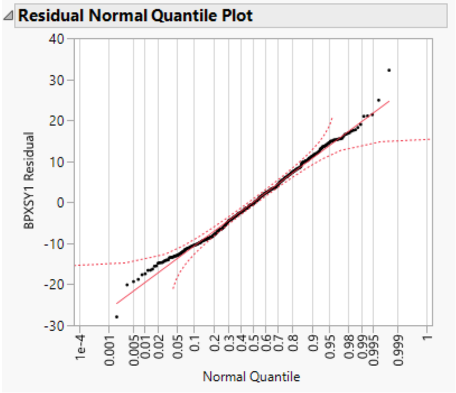
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	5838.297	1946.10	27.3612
Error	572	40684.258	71.13	Prob > F
C. Total	575	46522.556		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	93.454734	2.913384	32.08	<.0001*
RIDAGEYR	0.0770534	0.165806	0.46	0.6423
BMXWT	0.1768779	0.021165	8.36	<.0001*
BPXDI1	0.030189	0.030226	1.00	0.3183





Here again we find concerns about heteroskedasticity and normality; if we continue on to interpret the coefficient estimates, we see that the Diastolic BP adds no significant explanatory power to the model. The estimated value is not significantly different from zero, and the adjusted R^2 is slightly *less* than in the prior model using just 2 factors in the model. This model is no meaningful improvement over the prior one.

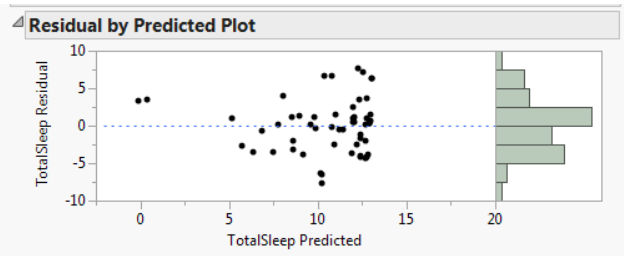
d. In the profiler, the line for systolic pressure is nearly flat. The mean systolic BP for 12-year old females is approximately the same as for 19-year old’s.

e. Student answers will vary, but after five splits the variables that emerge as informative are BMSXT, BMXBMI, PAQ679 , and BMXHT.

f. Student answers will vary. Among the promising table columns to include in a model are those identified by the Partition platform. The key in these responses is whether students accurately assess the residuals and the significance and properly interpret the meaning of parameter estimates.

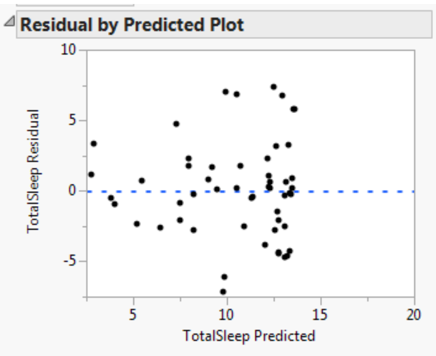
Scenario 4

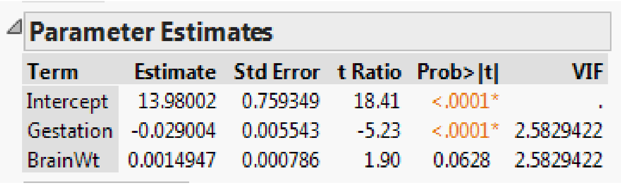
a.



When we estimate a simple linear model using gestation as the factor, we find a heteroskedastic pattern in which the variability of residuals diminishes as the Gestation period lengthens. Normality is not ideal, but the sample size is large enough to rely on the CLT. Given the non-constant variance, we should be reluctant to interpret or use the results of the regression.

b.





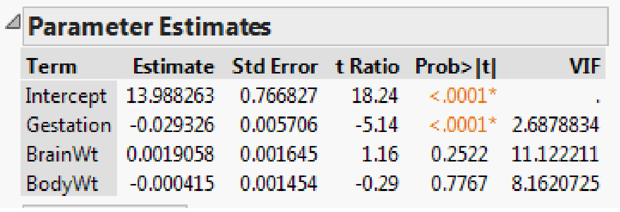
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	13.98002	0.759349	18.41	<.0001*	.
Gestation	-0.029004	0.005543	-5.23	<.0001*	2.5829422
BrainWt	0.0014947	0.000786	1.90	0.0628	2.5829422

With the addition of the BrainWeight variable, the residuals are still heteroskedastic suggesting caution in interpretation of the other results. The leverage plot (not shown) for BrainWt indicates a possible collinearity problem.

The Brain Weight variable is not significant at the customary 5% level, though the P-value is small (0.0628). This model is not a substantial improvement over the first model.

c.



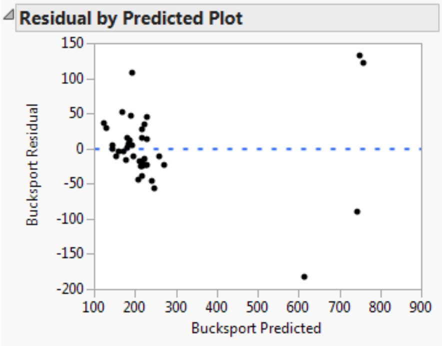
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	13.988263	0.766827	18.24	<.0001*	.
Gestation	-0.029326	0.005706	-5.14	<.0001*	2.6878834
BrainWt	0.0019058	0.001645	1.16	0.2522	11.122211
BodyWt	-0.000415	0.001454	-0.29	0.7767	8.1620725

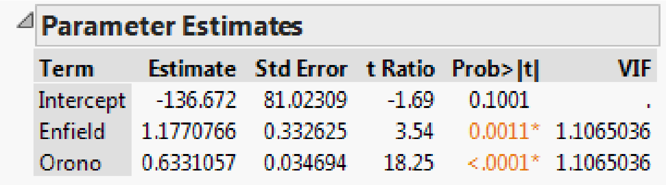
This model is not an improvement over the prior two. We still see heteroskedasticity in the plot of residuals vs. fitted values (not shown here). We see evidence of collinearity in the large VIF for BrainWt, and only the Gestation variable is statistically significant.

Scenario 6

a. Student models will vary. Here is one plausible result using the Enfield and Orono columns:



The residuals appear to have a non-constant variance, which raises a problem with using this model for prediction or estimation. The model adjusted R^2 is approximately 0.9 which indicates a very good fit. Both variables are statistically significant, and we see no real evidence of collinearity.



Parameter Estimates

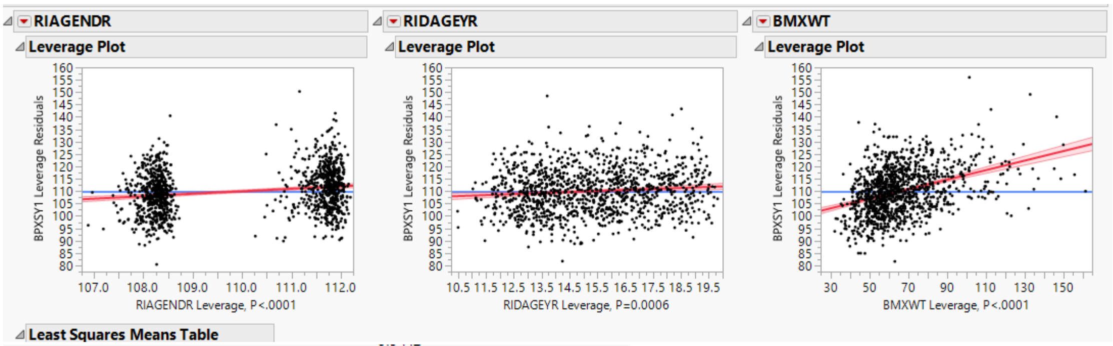
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-136.672	81.02309	-1.69	0.1001	.
Enfield	1.1770766	0.332625	3.54	0.0011*	1.1065036
Orono	0.6331057	0.034694	18.25	<.0001*	1.1065036

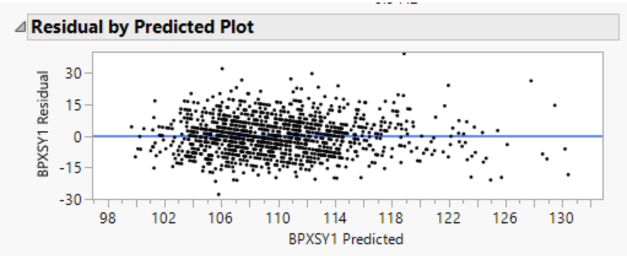
b. All of these communities have been exposed to the same state and national trends described in the question. Thus, the same factors that have led to reduced waste collections in one community also lead to reduced collections in another.

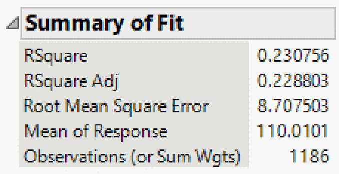
Chapter 19: Solutions to Application Scenarios

Scenario 2

a.

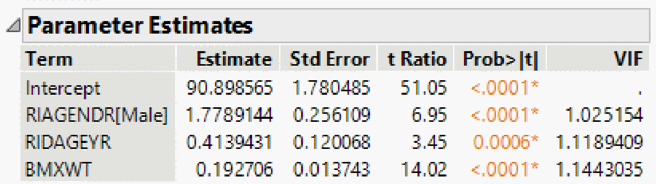






Summary of Fit

RSquare	0.230756
RSquare Adj	0.228803
Root Mean Square Error	8.707503
Mean of Response	110.0101
Observations (or Sum Wgts)	1186

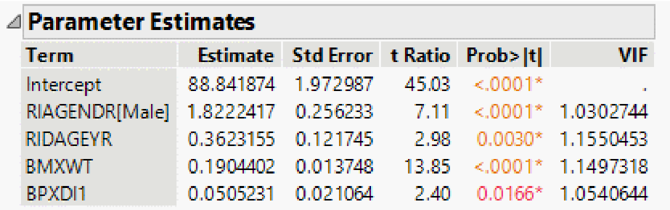


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	90.898565	1.780485	51.05	<.0001*	.
RIAGENDR[Male]	1.7789144	0.256109	6.95	<.0001*	1.025154
RIDAGEYR	0.4139431	0.120068	3.45	0.0006*	1.1189409
BMXWT	0.192706	0.013743	14.02	<.0001*	1.1443035

The leverage plots and VIFs indicate no major collinearity problems and residuals appear to have constant variance. We see that the model has rather poor fit, but all three variables are statistically significant and their signs are plausible.

b.

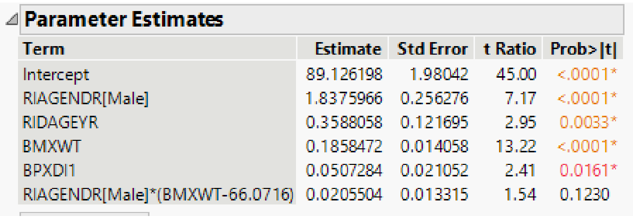


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	88.841874	1.972987	45.03	<.0001*	.
RIAGENDR[Male]	1.8222417	0.256233	7.11	<.0001*	1.0302744
RIDAGEYR	0.3623155	0.121745	2.98	0.0030*	1.1550453
BMXWT	0.1904402	0.013748	13.85	<.0001*	1.1497318
BPXDI1	0.0505231	0.021064	2.40	0.0166*	1.0540644

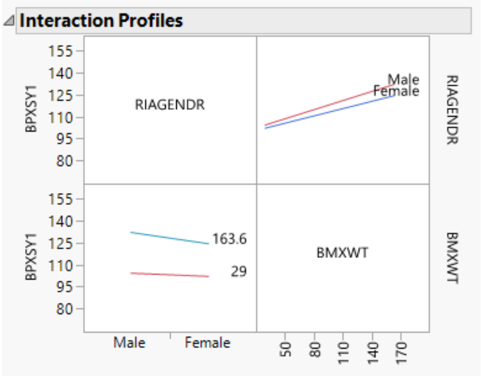
Adding the diastolic blood pressure measurement does help somewhat; it is statistically significant (as shown above), and residuals and leverage plots look fine. The summary of fit measures are improved very slightly in this model.

c.



Parameter Estimates

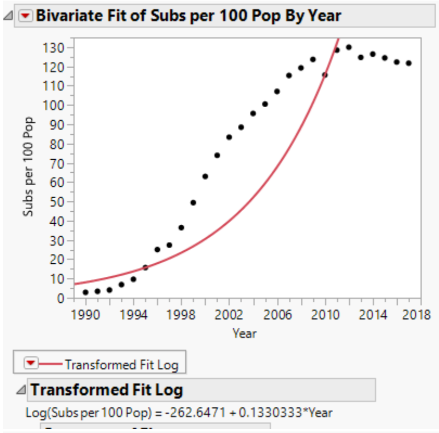
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	89.126198	1.98042	45.00	<.0001*
RIAGENDR[Male]	1.8375966	0.256276	7.17	<.0001*
RIDAGEYR	0.3588058	0.121695	2.95	0.0033*
BMXWT	0.1858472	0.014058	13.22	<.0001*
BPXDI1	0.0507284	0.021052	2.41	0.0161*
RIAGENDR[Male]*(BMXWT-66.0716)	0.0205504	0.013315	1.54	0.1230



There is no significant interaction between Gender and Weight. The interaction term does not add value to the model.

Scenario 4

a.

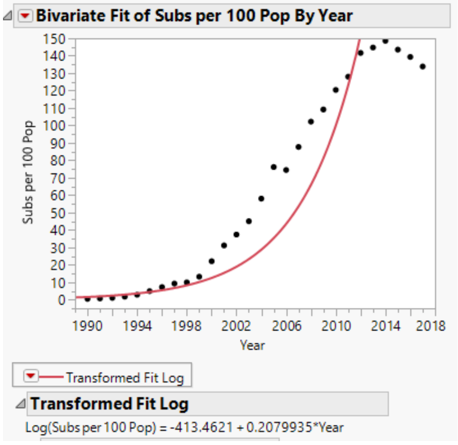


Transformed Fit Log

Log(Subs per 100 Pop) = -262.6471 + 0.1330333*Year

For Denmark, the log-linear estimated annual growth rate is $e^{0.133033}$–1 = 0.142 or 14.2% per year. The model clearly does not describe the pattern in the data.

b.

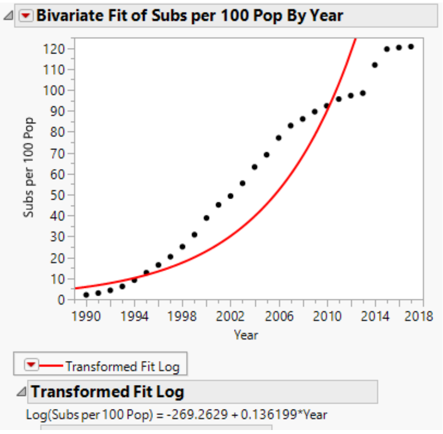


Transformed Fit Log

Log(Subs per 100 Pop) = -413.4621 + 0.2079935*Year

For Malaysia, the log-linear estimated annual growth rate is $e^{0. 2079935}$–1 = 0.231 or 23.1% per year. The model clearly does not describe the pattern in the data.

c.



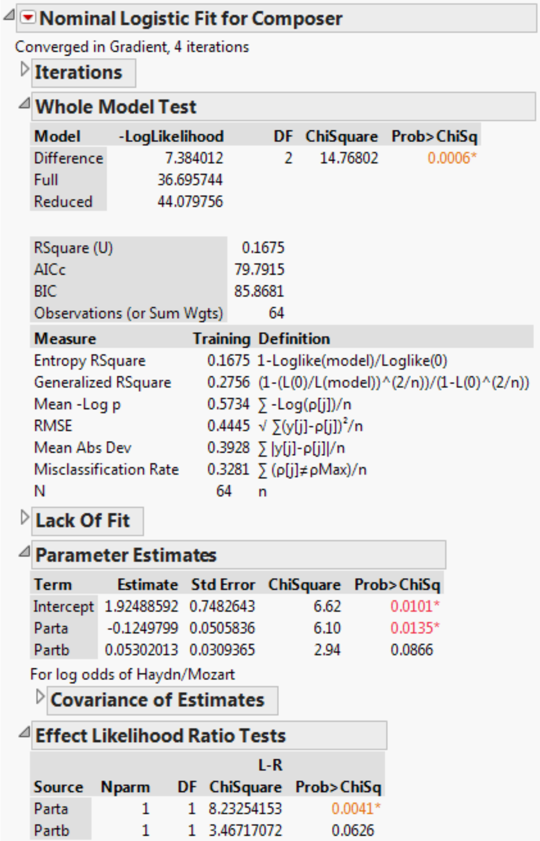
```
Transformed Fit Log
Log(Subs per 100 Pop) = -269.2629 + 0.136199*Year
```

For the U.S., the log-linear estimated annual growth rate is $e^{0.\ 136199}$–1 = 0.146 or 14.6% per year.

d. The log-linear model does not fit any of these countries particularly well, but may be useful in comparing the growth rates. In all of the countries, rapid growth was followed by a flattening out of the points in recent years. The US and Denmark had lowest growth rates, followed by Malaysia and Sierra Leone. Note that in all countries, the actual figures fall below the fitted line in the most recent observations.

Scenario 6

a.



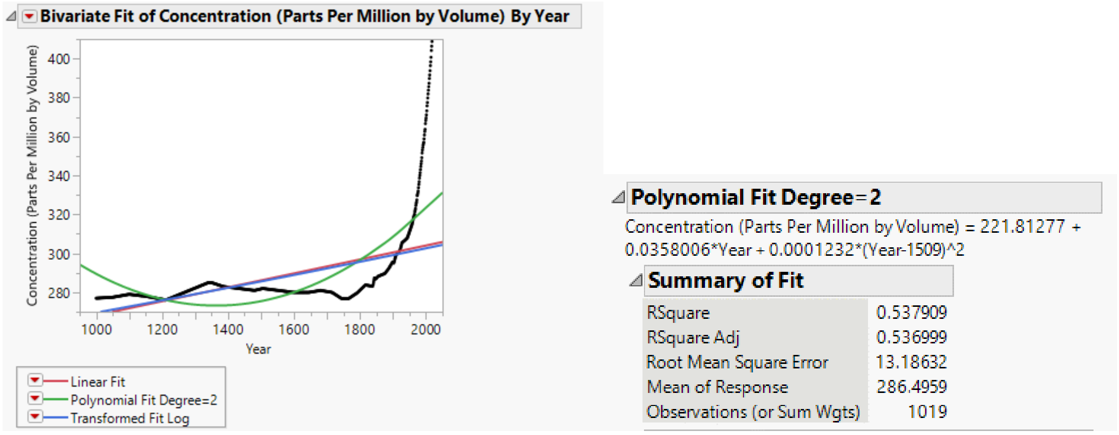
```
Nominal Logistic Fit for Composer
Converged in Gradient, 4 iterations
Iterations
Whole Model Test
Model        -LogLikelihood    DF   ChiSquare   Prob>ChiSq
Difference         7.384012     2    14.76802      0.0006*
Full              36.695744
Reduced           44.079756

RSquare (U)                  0.1675
AICc                        79.7915
BIC                         85.8681
Observations (or Sum Wgts)       64

Measure                  Training  Definition
Entropy RSquare            0.1675  1-Loglike(model)/Loglike(0)
Generalized RSquare        0.2756  (1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p                0.5734  ∑ -Log(ρ[j])/n
RMSE                       0.4445  √ ∑(y[j]-ρ[j])²/n
Mean Abs Dev               0.3928  ∑ |y[j]-ρ[j]|/n
Misclassification Rate     0.3281  ∑ (ρ[j]≠ρMax)/n
N                              64  n
Lack Of Fit
Parameter Estimates
Term        Estimate     Std Error   ChiSquare   Prob>ChiSq
Intercept   1.92488592   0.7482643        6.62      0.0101*
Parta       -0.1249799   0.0505836        6.10      0.0135*
Partb       0.05302013   0.0309365        2.94      0.0866
For log odds of Haydn/Mozart
Covariance of Estimates
Effect Likelihood Ratio Tests
                                L-R
Source   Nparm   DF    ChiSquare   Prob>ChiSq
Parta        1    1   8.23254153      0.0041*
Partb        1    1   3.46717072      0.0626
```

The results are to the left. We find that the whole model is significant with a rather poor fit, as measured by U. Other things being equal, the longer Part a is the lower the odds that it was composed by Haydn. Conversely, the longer Part b is (holding Part a constant) the higher the odds that it was composed by Haydn.

b. [note: to solve this problem, one needs to refer to outside sources about Logistic Regression]To decide which composer is more likely to have written a sonata with a 72-measure Parta and 112 measure Partb, we first substitute the values into the estimated equation: Logodds = 1.92488592 –0.1249799(72) +0.05302013(112) = –1.13541232. This is the log of the odds ratio for Haydn/Mozart, so the odds ratio is $e^{1.13541232}$ = 0.3213. Because the estimated ratio is well below 1, it is far more likely that Mozart would have composed such a sonata rather than Haydn.

Scenario 8

a. Here are the results for the linear, quadratic and log- linear fits: The linear and log-linear are nearly indistinguishable. None of the models fit particularly well, which visual inspection makes clear. The quadratic model has the best fit of the three, but it is weak.


Bivariate Fit of Concentration (Parts Per Million by Volume) By Year
Concentration (Parts Per Million by Volume)
400
380
360
340
320
300
280
1000
1200
1400
1600
1800
2000
Year
Linear Fit
Polynomial Fit Degree=2
Transformed Fit Log

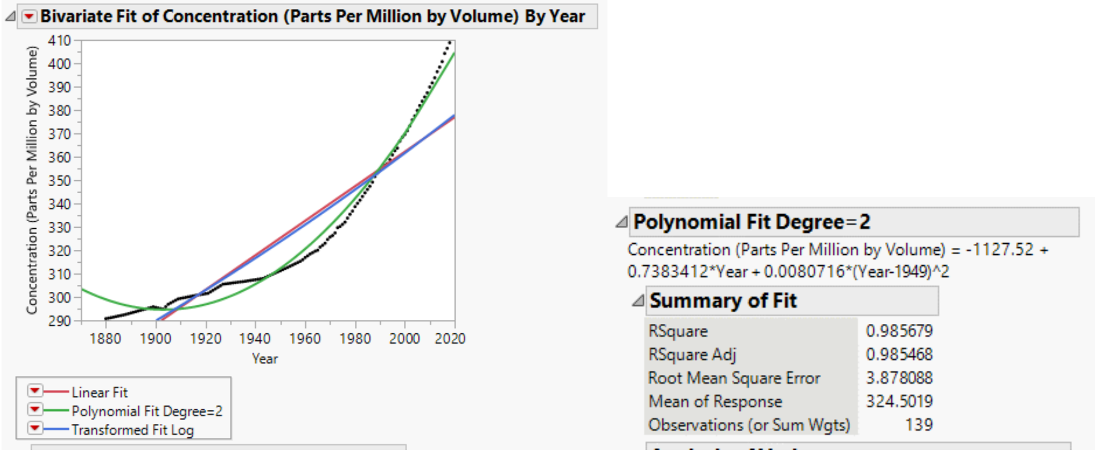
Polynomial Fit Degree=2

Concentration (Parts Per Million by Volume) = 221.81277 + 0.0358006*Year + 0.0001232*(Year-1509)^2

Summary of Fit

RSquare	0.537909
RSquare Adj	0.536999
Root Mean Square Error	13.18632
Mean of Response	286.4959
Observations (or Sum Wgts)	1019

b. The quadratic model still fits best, and the fit is considerably improved using the more recent data.



Polynomial Fit Degree=2

Concentration (Parts Per Million by Volume) = -1127.52 + 0.7383412*Year + 0.0080716*(Year-1949)^2

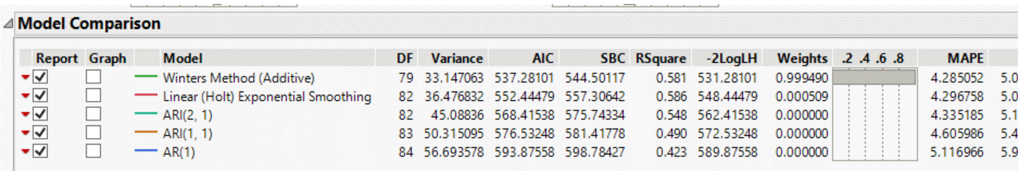
Summary of Fit

RSquare	0.985679
RSquare Adj	0.985468
Root Mean Square Error	3.878088
Mean of Response	324.5019
Observations (or Sum Wgts)	139

Chapter 20: Solutions to Application Scenarios

Scenario 2

a. Student answers will vary. Responses should note that Durables show a marked upward trend with likely seasonal component. Below are summary results for several reasonable approaches. Among the methods available through the Time Series platform, Winters Method outperforms the others according to the measures we have studied. The adjusted RSquare statistics for the regression-based models are superior to all of the Time Series models, as follows: Linear, (.667), Quadratic (.675), LogLinear (.671). However, the regression models do not capture seasonal shifts.



Model Comparison

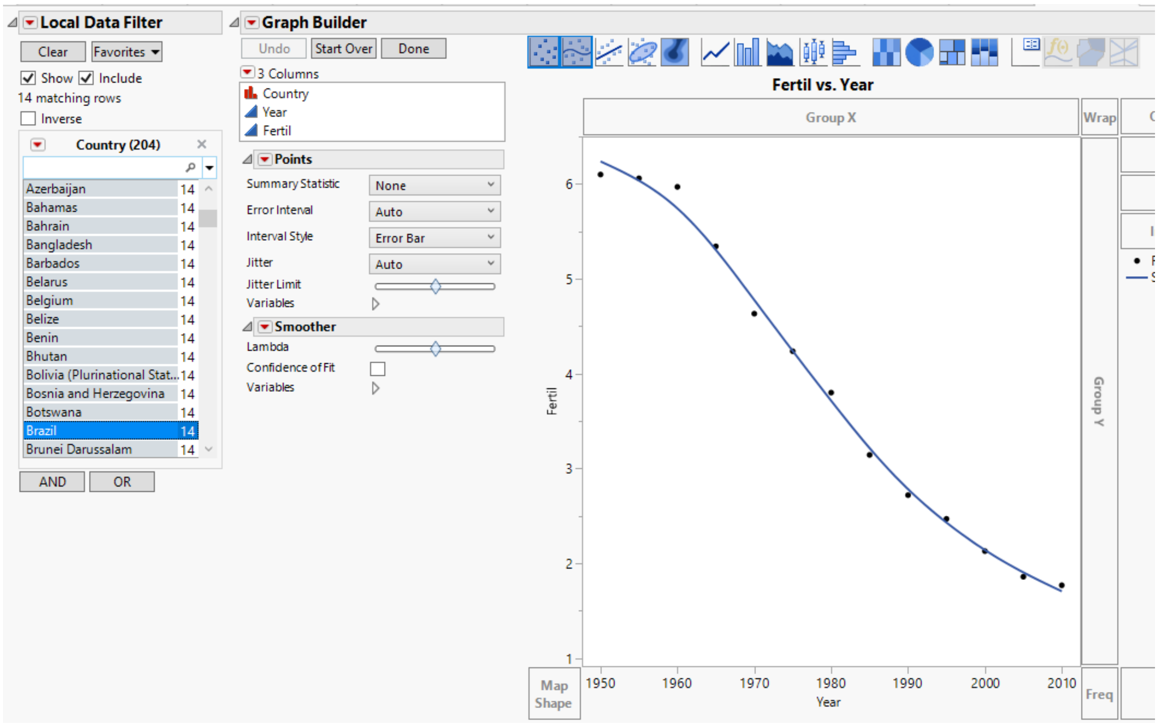
Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2 .4 .6 .8	MAPE	
☑	□	Winters Method (Additive)	79	33.147063	537.28101	544.50117	0.581	531.28101	0.999490		4.285052	5.0
☑	□	Linear (Holt) Exponential Smoothing	82	36.476832	552.44479	557.30642	0.586	548.44479	0.000509		4.296758	5.0
☑	□	ARI(2, 1)	82	45.08836	568.41538	575.74334	0.548	562.41538	0.000000		4.335185	5.1
☑	□	ARI(1, 1)	83	50.315095	576.53248	581.41778	0.490	572.53248	0.000000		4.605986	5.4
☑	□	AR(1)	84	56.693578	593.87558	598.78427	0.423	589.87558	0.000000		5.116966	5.9

b. Student answers will vary. This table summarizes the results for the top models cited in Part a. A good response will accurately report the predictions and compare them to the actual figures.

Period	Actual	Holt	Winters	AR(1,1)	AR(2,1)	Linear	Quadratic
84	131.9	131.2	133.7	127.4	126.6	131.6	129.9
85	127.1	131.5	131.7	129.0	129.3	131.9	130.1
86	133.4	131.5	129.0	130.0	128.7	132.2	130.3

Scenario 4

To illustrate the suggested starting approach, here is Graph Builder with a Local Data Filter active:



Local Data Filter

Clear | Favorites

☑ Show ☑ Include

14 matching rows

□ Inverse

Country (204)

Azerbaijan	14
Bahamas	14
Bahrain	14
Bangladesh	14
Barbados	14
Belarus	14
Belgium	14
Belize	14
Benin	14
Bhutan	14
Bolivia (Plurinational Stat...	14
Bosnia and Herzegovina	14
Botswana	14
Brazil	14
Brunei Darussalam	14

AND | OR

Graph Builder

Undo | Start Over | Done

3 Columns: Country, Year, Fertil

Points

Summary Statistic: None

Error Interval: Auto

Interval Style: Error Bar

Jitter: Auto

Jitter Limit

Variables

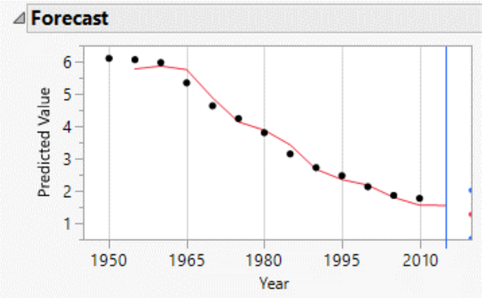
Smoother

Lambda

Confidence of Fit □

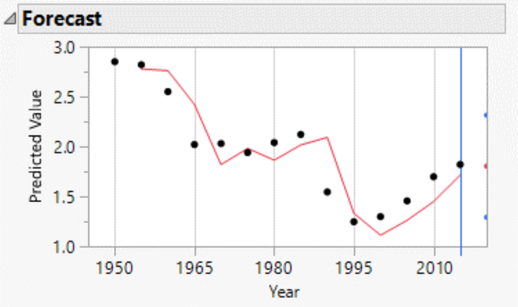
Variables

a. The fertility rate in Brazil has declined following an S-shaped curve:



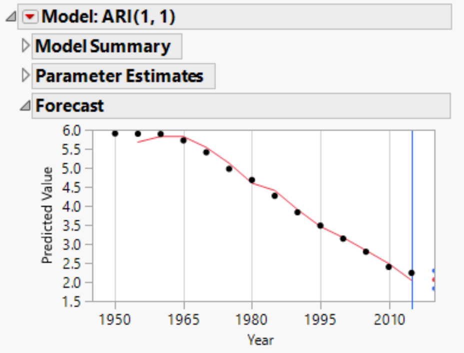
An AR(1,1) model fits moderately well, with relatively high RSquare (0.982), low variance (0.047) and MAPE and MAE of 4.76 % and 0.167 respectively.

b.



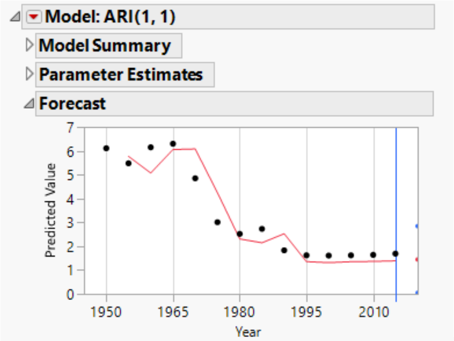
The decline in the Russian Federation fertility rate has been rather irregular, and will not be well-modeled by any of the regression methods. Simple exponential smoothing or AR(1,1) [shown above] models serve well.

c.



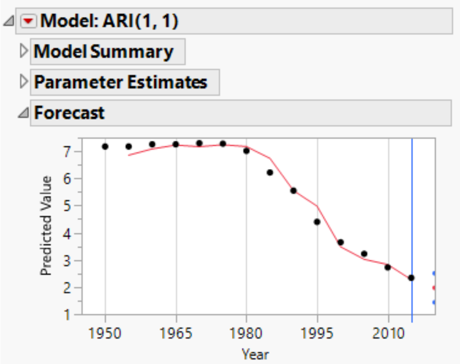
India’s decline is very regular, especially since 1960. Linear Exponential Smoothing (Holt’s method) and AR(1,1) models both fit extremely well.

d.



The decline in China’s fertility rate has been rather irregular and will not be well-modeled by any of the regression methods. An AR(1,1) model fits well.

e.

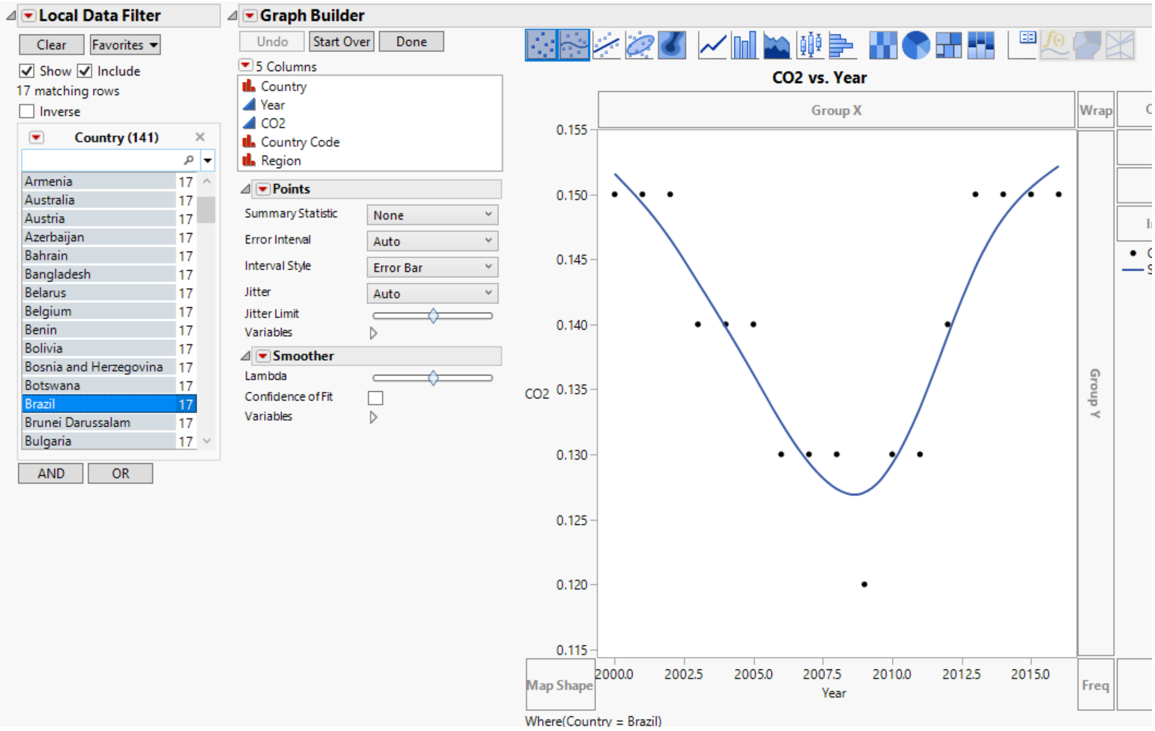


Saudi Arabia’s decline is very regular, especially since 1980. Linear Exponential Smoothing (Holt’s method) and AR(1,1) models both fit extremely well, with AR(1,1) fitting slightly better.

f. It is difficult to say with certainty. Simple Exponential smoothing estimates the rate in 2015 as 1.77, which is closer to the UN figure than any of the other models presented in the chapter. The AR(1,1) model, for example, produces a 2015 estimate of 1.55.

Scenario 6

To illustrate the suggested starting approach, here is Graph Builder with a Local Data Filter active:



Local Data Filter

☑ Show ☑ Include

17 matching rows

□ Inverse

Country (141)	
Armenia	17
Australia	17
Austria	17
Azerbaijan	17
Bahrain	17
Bangladesh	17
Belarus	17
Belgium	17
Benin	17
Bolivia	17
Bosnia and Herzegovina	17
Botswana	17
Brazil	17
Brunei Darussalam	17
Bulgaria	17

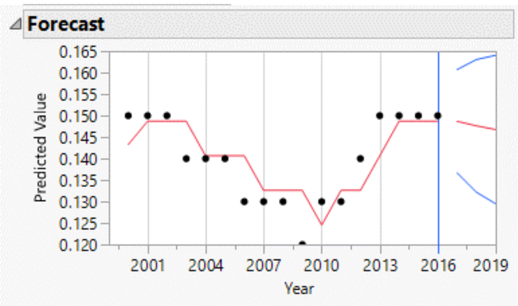
Graph Builder

5 Columns
- Country
- Year
- CO2
- Country Code
- Region

Points
- Summary Statistic: None
- Error Interval: Auto
- Interval Style: Error Bar
- Jitter: Auto
- Jitter Limit
- Variables

Smoother
- Lambda
- Confidence of Fit: □
- Variables

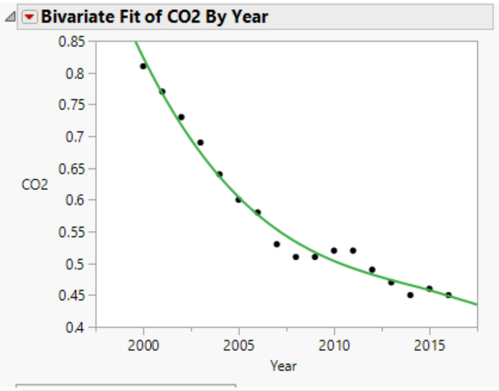
a.



CO2 emissions in Brazil fell at the start of the series but then have risen and have leveled off in most recent years.

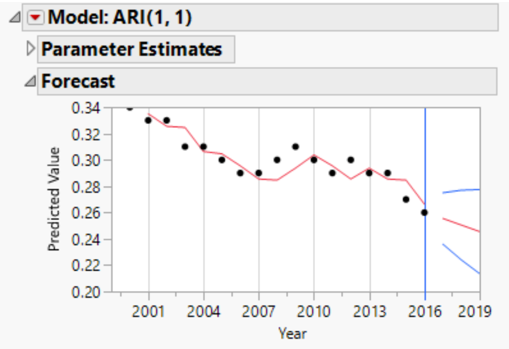
This series is difficult to model well, but an AR(1) model fits better than most.

b.



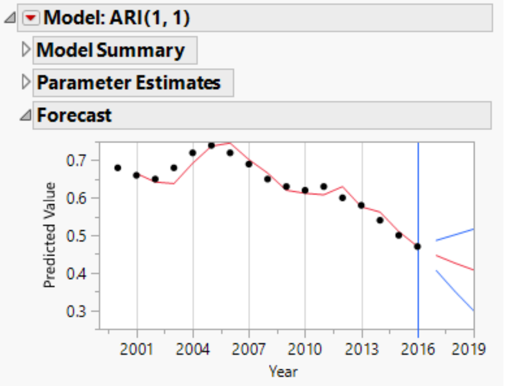
In the Russian Federation there has also been a steady decline, with an unusual jump in 2010 & 2011. A 3rd-degree polynomial (cubic; shown here) provides a moderately good fit, as does AR(1,1).

c.



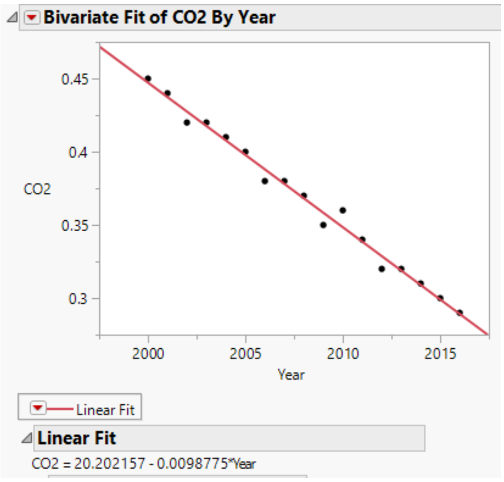
India's CO2 emissions have trended downward with several increases along the way. The AR(1,1) model fits better than alternatives and forecasts a continued downward trend.

d.



After some early increases, China's emissions have been decreasing. The AR(1,1) model performs well here.

e.



Linear Fit

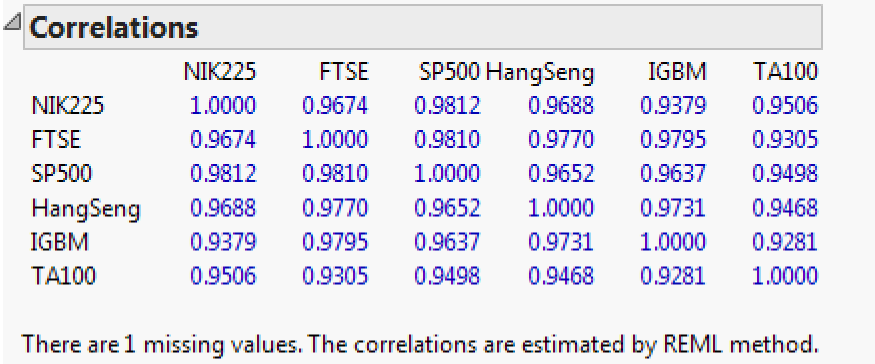
CO2 = 20.202157 - 0.0098775*Year

CO2 emissions in the US fell steadily over the period. A simple linear regression model fits better than the alternatives.

f. There is no single model that fits all of these series probably because the use of CO2-generating technologies varies considerably across these countries as does environmental public policy . Some are reducing emissions while others are making greater use of activities that emit CO2.

Scenario 8

a.



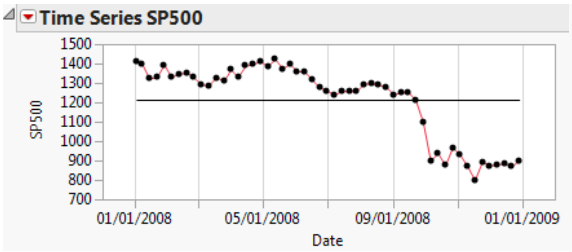
Correlations

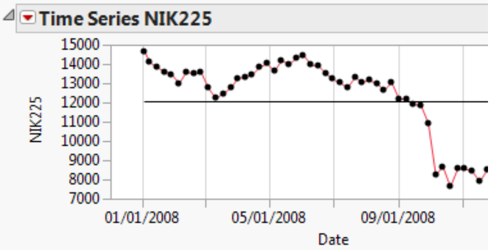
	NIK225	FTSE	SP500	HangSeng	IGBM	TA100
NIK225	1.0000	0.9674	0.9812	0.9688	0.9379	0.9506
FTSE	0.9674	1.0000	0.9810	0.9770	0.9795	0.9305
SP500	0.9812	0.9810	1.0000	0.9652	0.9637	0.9498
HangSeng	0.9688	0.9770	0.9652	1.0000	0.9731	0.9468
IGBM	0.9379	0.9795	0.9637	0.9731	1.0000	0.9281
TA100	0.9506	0.9305	0.9498	0.9468	0.9281	1.0000

There are 1 missing values. The correlations are estimated by REML method.

The Nikkei225 has the highest correlation with the S&P500 (0.9812) and the FTSE100 is close behind with *r* = 0.9810)

b. The models should be for the Nikkei and S&P. The two series are shown below.





For the S&P no model is perfect, AR(2,1) provides a comparably low variance, MAE, MAPE, and high RSqr.

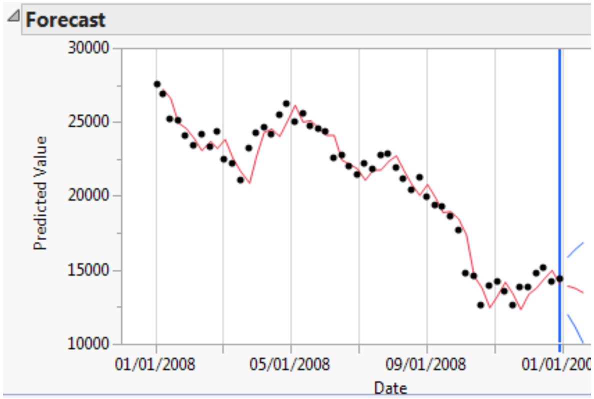
Much like the S&P series, the Nikkei is well-modeled with an AR(2,1) model.

c. Yes. Both markets are engaged in competition in the same global markets, and move very closely together as indicated by their very high correlation.

d. Student answers will vary. The AR(2,1) model works rather well for the HangSeng data. Based on that model, the forecasts are as follows:

01/05/2009	13972.2114
01/12/2009	13804.8764
01/19/2009	13510.4924

The graph to the right shows the extrapolation. We can be 95% confident that the index would lie within the blue confidence limits. Beyond that it is difficult to specify a confidence level in the point estimates, but they appear to be a reasonable extrapolation beyond the observed data.



Chapter 21: Solutions to Application Scenarios

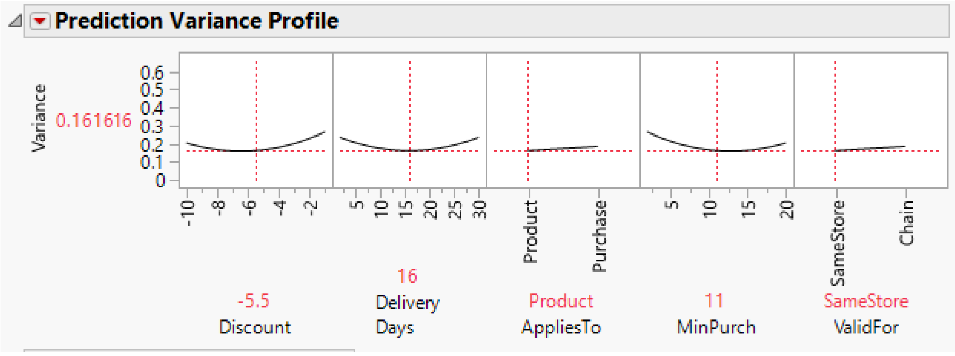
Scenario 2

a. this part just calls for entry of five factors.

b. The design has 18 runs

c. The Custom Design with all 2nd order interactions requires a minimum of 6 runs. With the addition of three more factors in this screening design, we need more runs.

d.



When we compare the variance prediction to that shown in Figure 21.13, we see that this design has a smaller initial variance prediction.

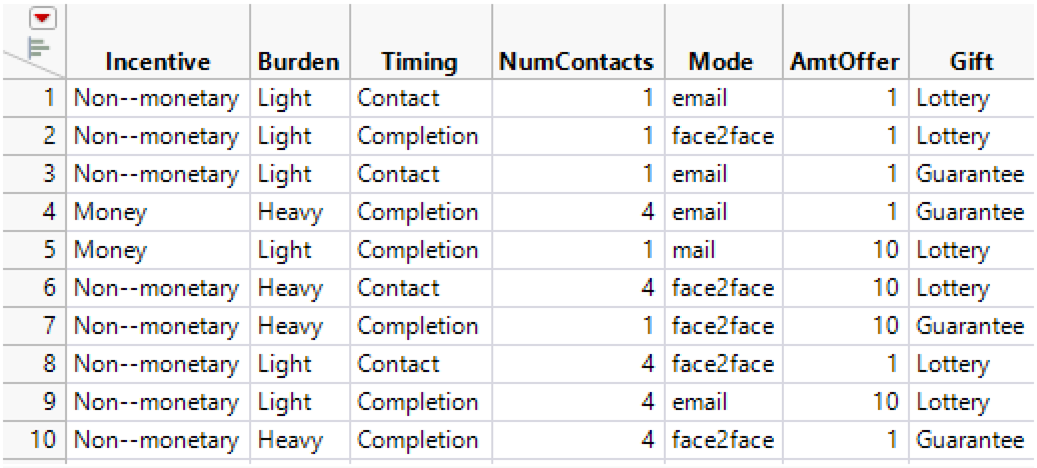
Scenario 4

a. Categorical factors: type of incentive, burden of survey, timing of incentive, survey mode, guarantee vs. lottery.
Continuous factors: number of contacts made, amount of money offered.
[some students might classify “burden” of survey as continuous.]

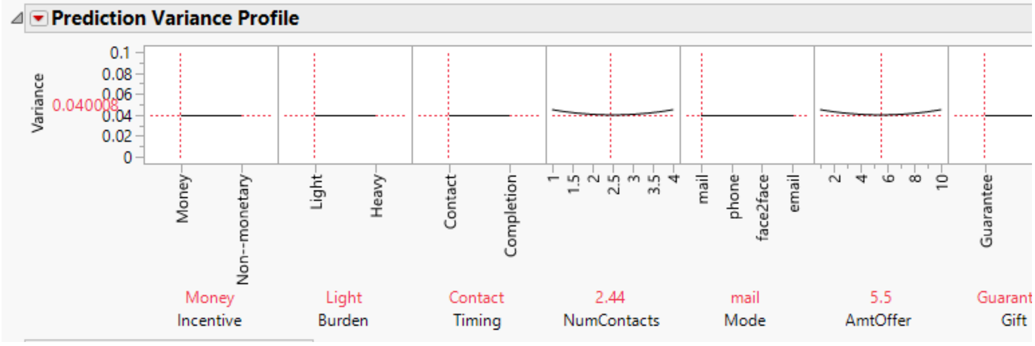
b. [Student answers will vary]
Type of incentive: monetary/ non-monetary. Might also include “none” as a control, or vary the specific non-monetary incentives.
Timing of incentive: as described, point of contact vs. completion of survey
Survey mode: mail, telephone, face-to-face, email.
Nature of gift: guarantee vs. entry into lottery

c. Assuming that we use minimal number of factor levels described in b, and two factor levels for the continuous factors, we would have five dichotomous categorical factors (four with 2 levels and one with 4 levels) and two continuous factors. After entering all factors and specifying main and interaction effects, the minimum number of runs is 43, and Default is 48 runs.

d. Here are the first 10 of the 200 rows:



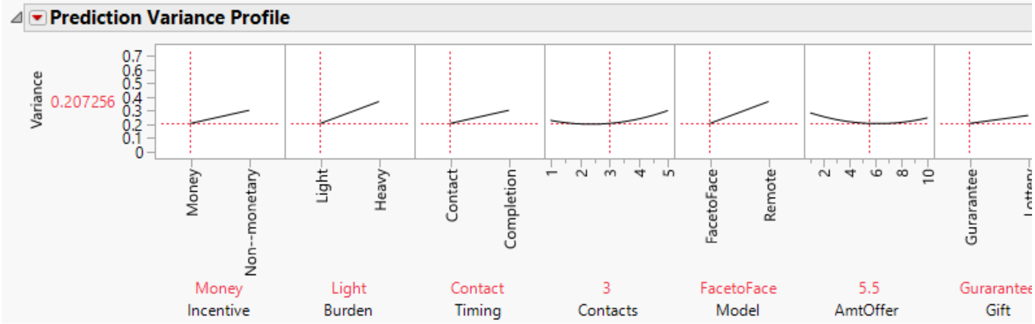
	Incentive	Burden	Timing	NumContacts	Mode	AmtOffer	Gift
1	Non--monetary	Light	Contact	1	email	1	Lottery
2	Non--monetary	Light	Completion	1	face2face	1	Lottery
3	Non--monetary	Light	Contact	1	email	1	Guarantee
4	Money	Heavy	Completion	4	email	1	Guarantee
5	Money	Light	Completion	1	mail	10	Lottery
6	Non--monetary	Heavy	Contact	4	face2face	10	Lottery
7	Non--monetary	Heavy	Completion	1	face2face	10	Guarantee
8	Non--monetary	Light	Contact	4	face2face	1	Lottery
9	Non--monetary	Light	Completion	4	email	10	Lottery
10	Non--monetary	Heavy	Completion	4	face2face	1	Guarantee



Prediction Variance Profile

The minimum prediction variance is approximately 0.04, and maximum is approximately 0.05 according to the Prediction Variance Profiler.

Without added runs, the DSD design calls for 22 runs by default. Here is the prediction variance profiler, showing settings that minimize prediction variance at 0.207for the DSD is here:

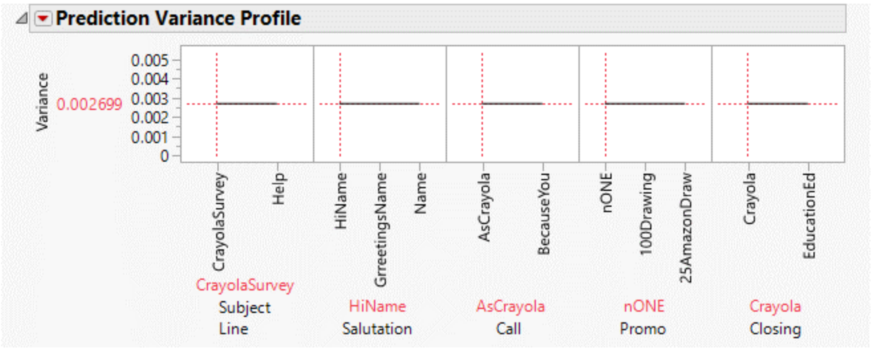


Prediction Variance Profile

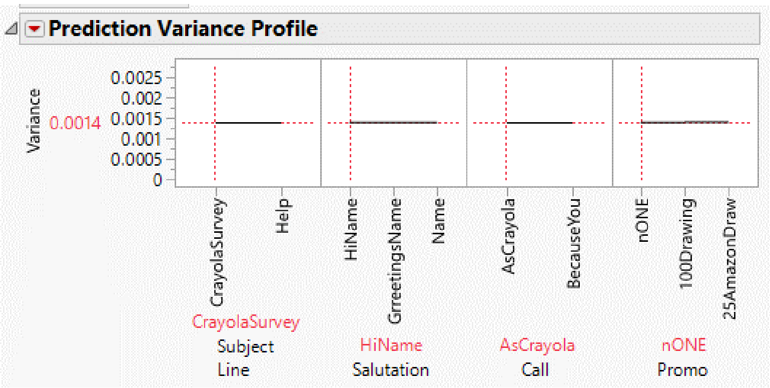
The maximum variance with 22 runs is 0.738.

Scenario 6

a. The minimum estimated variance is 0.0027



b. With the reduced design, the minimum variance decreases to 0.0014



c. [NOTE: This question draws the reader into unfamiliar territory. The Compare Designs report is extensive, but readers should be able to draw some conclusions.]

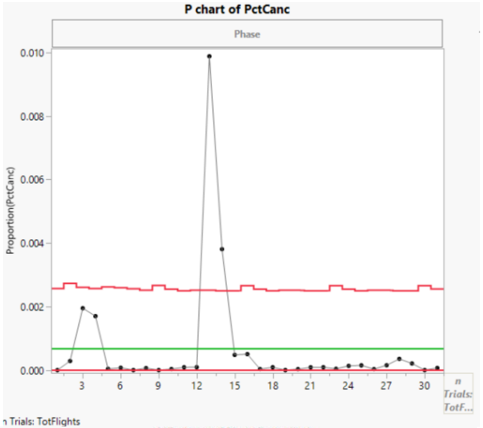
Key conclusions include:

- This platform compares main effects analysis in both designs, restricted to the four factors common to both. Interaction effects are dropped from the first design.
- Both designs have similar power.
- It becomes clear that the Prediction Variance for the 5000-run design is twice the size of the 10,000-run.
- The smaller design has approximately 70% of the estimation efficiency of the first design.

Chapter 22: Solutions to Application Scenarios

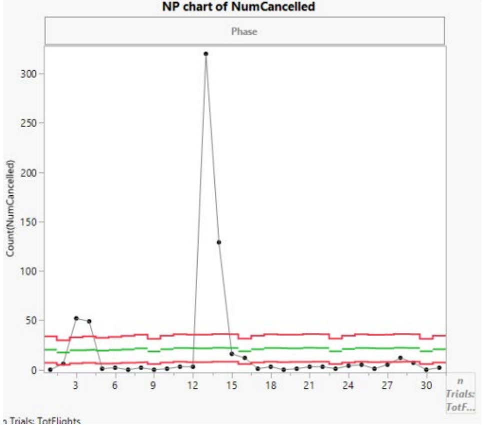
Scenario 2

a.



This process is out of control at two points. Because a day with 0 cancellations is desirable, we should not be concerned about dates with values below the LCL. However, the chart shows 2 dates well above the UCL. Presumably there was severe weather on those dates.

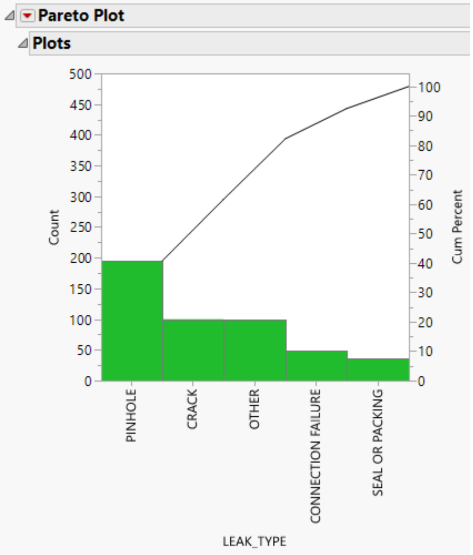
b.



This is essentially the same data as the previous chart again showing the process out of control at several points. Because a day with 0 cancellations is desirable, we should not be concerned about dates with values below the LCL. However, the chart shows 4 dates above the UCL.

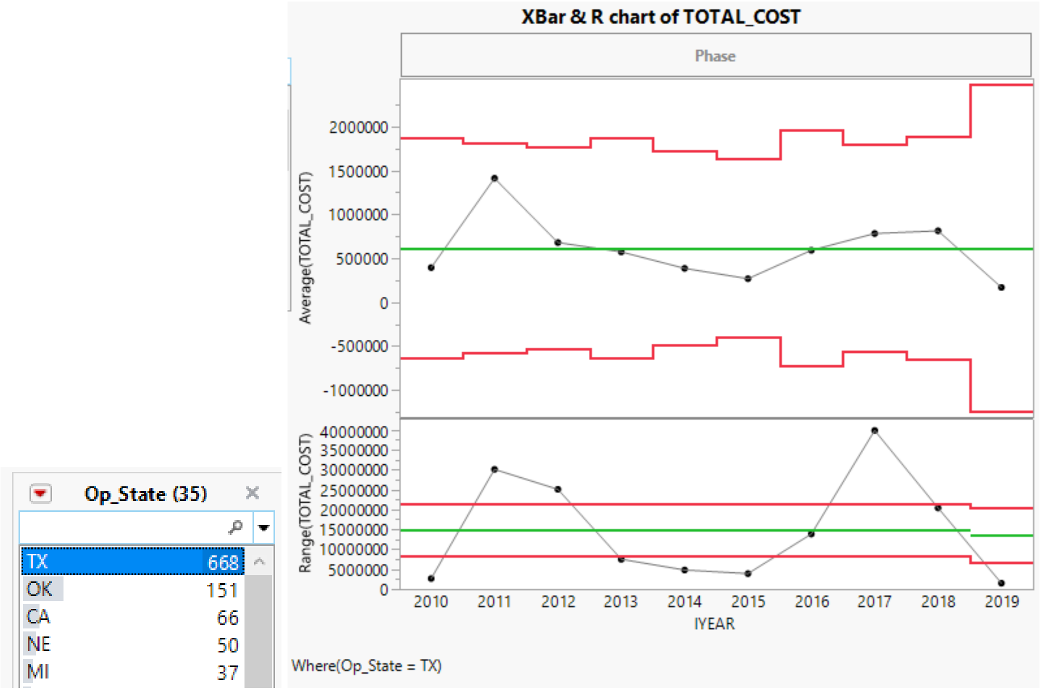
Scenario 4

a.



Pinholes and cracks account for approximately half of all leaks where the type is known. Connection failures and seal or packing issues are comparatively rare.

b.

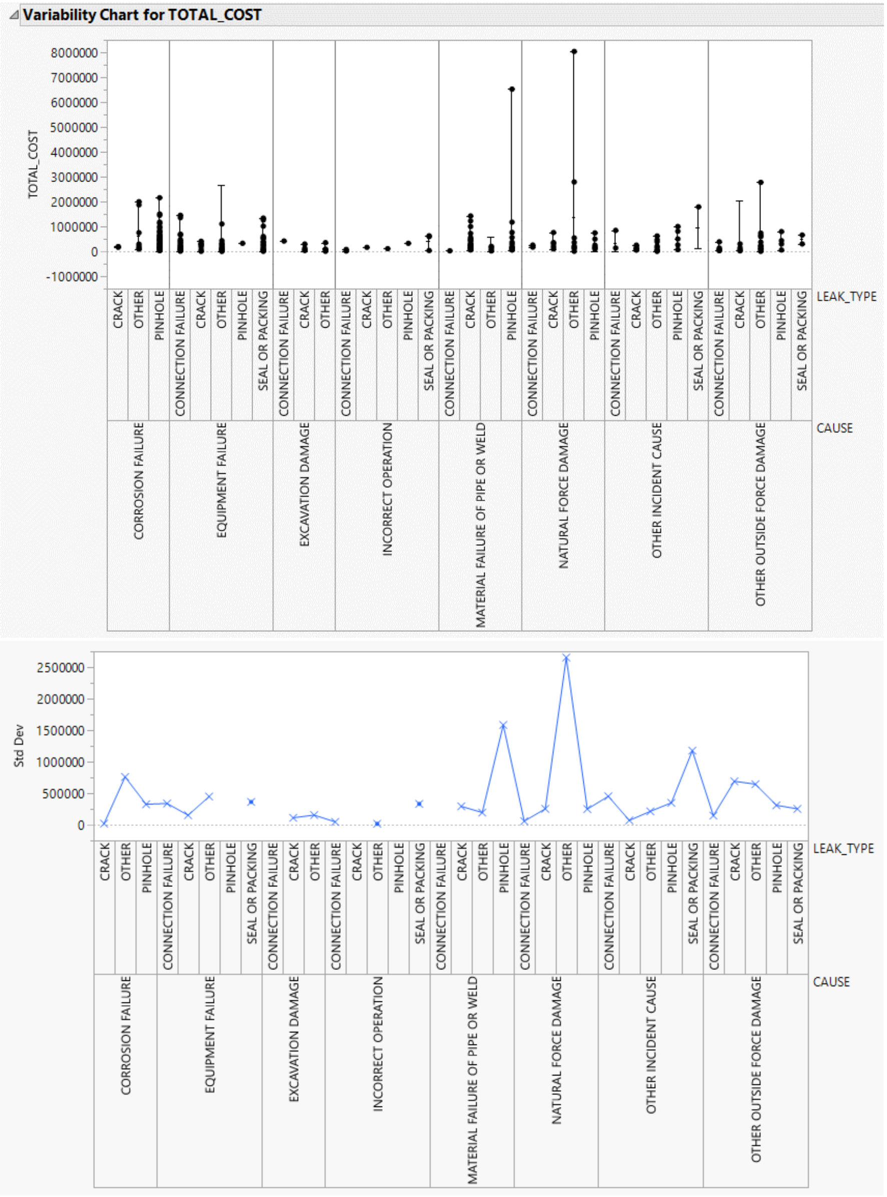


TX	668
OK	151
CA	66
NE	50
MI	37

The top four states each have unique situations.

- In Texas (shown), the range chart shows instability in variation, with wide fluctuations year to year. The mean of all years is over $600,000 and there are costly leaks each year with no extraordinary variation.
- Oklahoma has very low range variability in most years, with one large spike in 2016. The grand mean exceeds $783,000 due largely to extraordinary costs in 2016.
- With the exception of 2010, California appears to be under control, with near-zero costs in subsequent years. The grand mean is over $9 million including the disastrous year of 2010. Excluding 2010, the grand mean falls to $732,000.
- Nebraska, like Oklahoma, has low range variability in most years except for 2014. Otherwise, the process is relatively stable with a grand mean of just more than $262,000.

c.

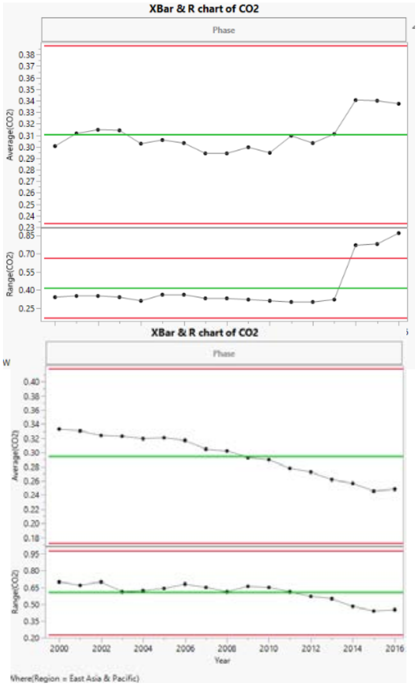


The standard deviation chart indicates that the greatest variabilty is present when the cause is either natural force damage or material failure of the pipe or welds. These causes are also associated with comparatively high costs. With natural force damage, the greatest variation and highest costs are associated with “Other” leak types. With material failures, pinhole leaks have the highest costs and variation.

Utility operators cannot do much about natural force damage. From Figure 22.11, we know that the most common causes are equipment failure and corrosion failure, and that we see here that interruptions from these causes also tend to be costly, it makes sense to prioritize prevention of such leaks.

Scenario 6

a.



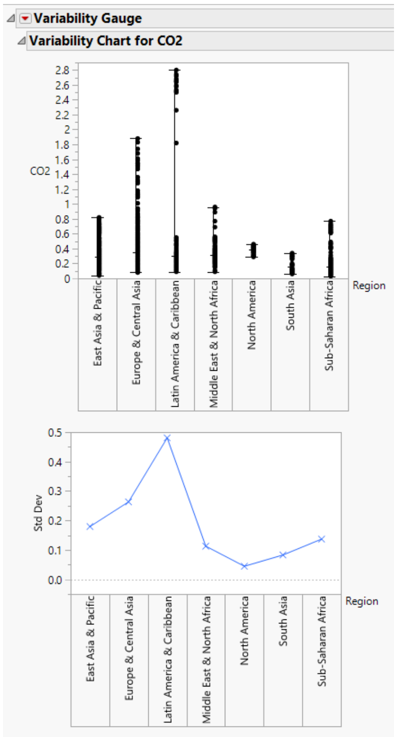
Emissions levels trends are notably different around the world, though the processes are in control most everywhere. In Latin America & Caribbean, South Asia and Sub-Saharan Africa, the levels have been stationary.

One exception is the Middle East and North Africa (shown to the left), where we see a level shift starting in 2014.

The other regions (East Asia & Pacific, Europe & Central Asia, and North America) have seen declining CO2 emissions. The one process out of statistical control is North America, due to steady declines.

b. The general message in this set of charts is the same as in Part a. The Xbar charts are all identical to what we saw previously. The S Chart for Europe & Central Asia now exhibits variability beyond the control limits.

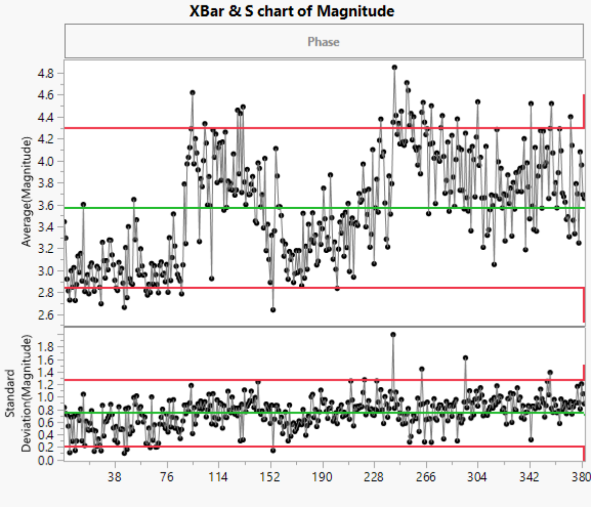
c.



The greatest variability is in Latin America and the Caribbean, followed by Europe & Central Asia. North America shows the least variability.

Scenario 8

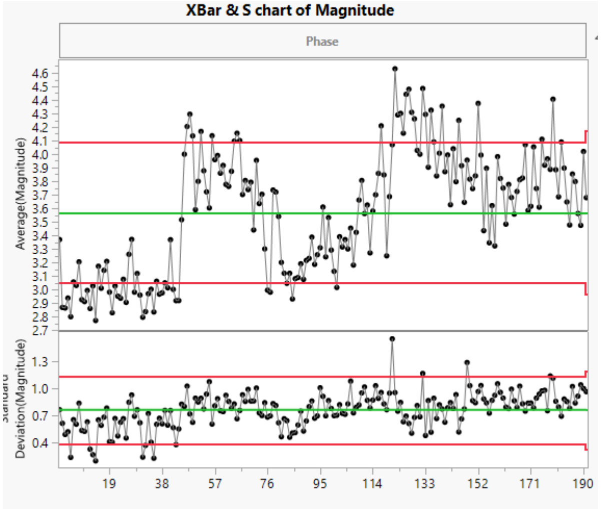
a.



Earthquakes are a natural process that are beyond human control, but we can note that the variability of the process standard deviation has increased over time, with several spikes beyond the upper control limit in the S-chart..

Although we should not interpret the means chart due to the instability of variability, the process appears to be non-stationary with both level shifts and a long-term upward drift.

b.



With a larger sample size there are naturally fewer sample means. The averages in both mean charts are the same, but otherwise the computed values are different. In this chart, the control limits for both graphs are closer to the mean than in the earlier chart. Again we see increasing oscillation in the sample standard deviations, and the general pattern of means is similar to the prior chart.