Paper SAS252-2014

The Desert and the Dunes: Finding Oases and Avoiding Mirages with SAS Visual Analytics

Nascif Abousalh-Neto, SAS Institute

ABSTRACT

Once upon a time, a writer compared a desert to a labyrinth. A desert has no walls or stairways, but you can still find yourself utterly lost in it. And oftentimes, when you think you found that oasis you were looking for, what you are really seeing is an illusion, a mirage.

Similarly, logical fallacies and misleading data patterns can easily deceive the unaware data explorer. In this paper, we discuss how they can be recognized and neutralized with the power of the SAS® Visual Analytics Explorer. Armed with this knowledge, you will be able to safely navigate the dunes to find true insights and avoid false conclusions.

INTRODUCTION

In this paper, we briefly explore a visual analytics methodology that allow us to find insights (oases) in a dataset; we describe the problems that can arise due to incorrect data preparation, mismatch of data structure, and mistakes during data analysis; and we deep dive in a few case studies that show how the SAS® Visual Analytics Explorer can be used to identify and prevent these problems to ensure the correct interpretation of the insights that were found.

FINDING OASES

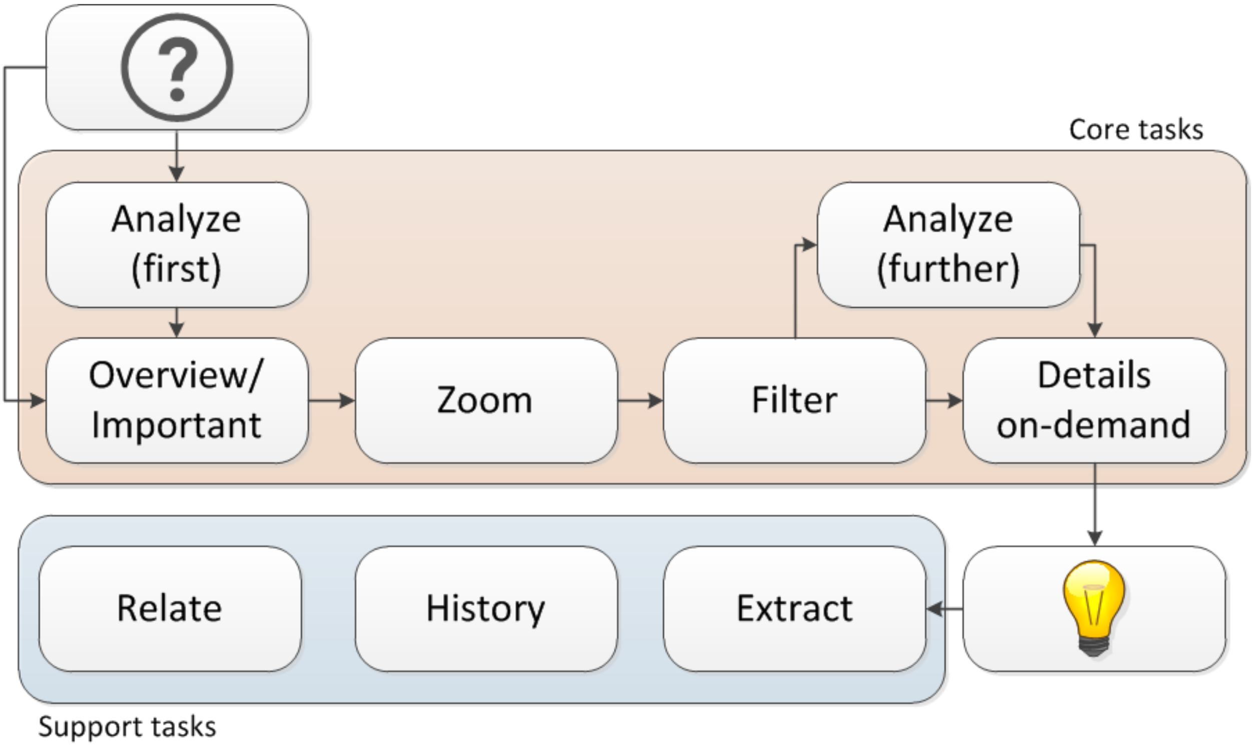


Figure 1 - Visual Analytics Mantra tasks and workflow

In a previous paper [Abousalh2013], we looked at how the visual analytics science has evolved over the last decades, distilling its multi-disciplinary lessons into a workflow called the Visual Analytics Mantra:

"Analyze first, show the important, zoom, filter and analyze further, details on demand."

The SAS® Visual Analytics Explorer, part of the SAS Visual Analytics suite, was designed from the ground up as an exploratory data analysis application enhanced with a number of advanced analytical tools. That makes the

SAS Visual Analytics Explorer perfectly suited to carry out the Visual Analytics Mantra tasks in a seamless workflow, as illustrated by Figure 1.

Implementing the visual analytics workflow with the SAS Visual Analytics Explorer is an effective way to find insights in your data. Next we will look at situations in which these insights turn out to be mirages.

AVOIDING MIRAGES

Tools like the SAS Visual Analytics Explorer make data analysis accessible to a larger audience. But the complexity inherent to the visual analytics discipline still requires a great deal of knowledge and awareness from the user in order to prepare the data, select the correct analysis, and interpret the results.

Unlike information processing tools such as network applications where input errors (for example, an invalid host address) result in clear error messages ("host not found"), most numerical methods used by data analysis, when presented with an invalid input, produce invalid results as well. It is up to the user to identify them as such. This phenomenon is typically called "garbage in, garbage out," and the best defense against it is to be careful, at all times, with the steps taken to prepare the data while being aware of the assumptions and limitations of the algorithms applied to it.

In the next sections, we break down some of the potential problems a user can find during different stages of the exploratory data analysis process: data preparation, (choice of) data structure, and data analysis.

DATA PREPARATION

The data preparation step requires special attention because problems introduced during this stage not only are difficult to trace but can also invalidate any insights found in later exploratory data analysis steps. The "garbage in, garbage out" issue is usually triggered at the data preparation stage.

Data preparation can go wrong in many different ways. Issues caused by incorrect data preparation include data elimination, data corruption, and data duplication.

Data Elimination

When the user is crafting the queries that prepare the data for loading into the SAS Visual Analytics system, a misplaced or too aggressive *where* clause might drop observations that would have been relevant to the analysis.

This issue is compounded by the survivorship bias, which causes invalid patterns to be identified in incomplete data. The problem is compounded if by chance the insights derived from the filtered data confirm the desired outcome. It is in our nature to accept positive results without further questioning, thanks to another human cognitive system "bug" called confirmation bias [McRaney2011].

Sometimes, data elimination is intentional—for example, to limit the amount of data that is loaded in the SAS Visual Analytics environment. In such cases, the user must be careful when generalizing later findings because the analyzed data subset might not be representative of the larger dataset.

Data Corruption

A common data preparation step is the conversion of the format used to store the original data into a format that is suitable for analysis in the SAS Visual Analytics system. The user must make sure that the data is not transformed in an unintended way during this conversion.

For example, users might want to analyze data contained in CSV files and Excel spreadsheets. These files are loaded by using the new SAS Visual Analytics self-service data import features. Since internally the SAS Visual Analytics system relies on PROC IMPORT to convert these files into a SAS dataset before loading them into memory, it is possible that critical data attributes, such as length and format, could be incorrectly assigned. This issue is more frequent with long categorical values and datasets that have a large number of rows. The end result would be the unintended truncation of some of the data values. [McQuown2005] describes a few workarounds for this issue that can be executed prior to importing the data.

Data Duplication

The SAS Visual Analytics system is designed to operate with analytical base tables. Therefore, it requires the de-normalization of the (potentially many) input tables in the customer data warehouse and their combination into a single table. This combination of multiple data sources, if not done properly, can introduce unintended data duplication. Two scenarios are particularly dangerous:

- A *fan trap* happens when multiple one-to-many related (fanned-out) tables are used in a SQL join, with attributes selected from the different joined tables.

- A *chasm trap* happens when there is a many-to-one-to-many join relationship between three tables, with attributes selected from the “many” end of the joined tables.

Both traps produce overstated or inflated (and therefore incorrect) results, referred to as Cartesian products. Once identified, this issue must be resolved—perhaps by splitting the query or pre-aggregating the data—before the data analysis process can proceed.

DATA STRUCTURE

An exploratory data analysis workflow typically requires that different questions are asked of the same data in order to get a complete understanding of it. Partial understanding can be a dangerous state as it lends a veneer of legitimacy to incomplete, and therefore potentially wrong, insights.

Different questions require different analytical procedures. Each procedure carries its own assumptions about the data structure it was designed to operate on [Sglavo2012]. The challenge in this step is making sure that the data under analysis is compatible with all these different assumptions, and if not, finding out how to make it so.

The most prominent data mart structure in data mining is the cross-sectional data. It is characterized by having a one-row-per-subject data structure. It is required by many statistical methods, in particular regression analysis and decision trees, two of the most useful analytical tools supported by the SAS Visual Analytics Explorer.

In the data mining sense, the term *subject* refers to data attributes that describe the subject identity, usually a categorical value or a numeric value that can be converted to a category. This data structure requires that the subject information is unique for each row considered for the analysis. Remaining attributes (the variables that describe what the subject did or how it behaved) are not under the same constraints. For example, in a cross-sectional dataset of a health study, the subject might be a patient identified by its unique social security number while remaining columns contain non-unique attributes such as blood pressure and temperature. The data structure underlying assumption is that each observation (each row under analysis) contains data from a different patient.

A simple way to validate this assumption by using the SAS Visual Analytics Explorer is to look at the frequency of the subject identifier categorical column. To make sure that the identifier is unique as expected, we just need to check the highest value and make sure that it is equal to 1. This check is trivial since the required frequency plot (conveniently sorted by descending value by default) can be created with a single drag-and-drop of the target category into the canvas while in auto-charting mode. We also want to make sure that this category contains no missing values.

Note that value repetition or non-uniqueness of the subject category is perfectly legitimate for other types of analysis and their corresponding data structures. For example, time series data—time-stamped data collected over time at a particular frequency—will likely have an observation for the same subject for all the "moments" when data was collected. While not suitable (without changes, see case study below) for regression analysis and decision trees, time series data is the required data structure for forecast and scenario analysis in the SAS Visual Analytics Explorer. Examples of time series data include web hits per hour, sales per month, and trades per weekday.

DATA ANALYSIS

After double-checking the data preparation steps and validating the data structure assumptions, we are ready to move into the fun stuff: data analysis. But here too a data explorer needs to be aware of pitfalls that might be lurking in the way to true insights.

One cornerstone of exploratory data analysis is the use of descriptive statistics to get a broader understanding of large datasets. Descriptive analytics accomplishes this goal by summarizing the data and removing detail. However, this summary has the potential of leading to incorrect conclusions. It is up to the user to know if the removed details are relevant to the question being asked or if they can be safely ignored. The application always provides *precise* results, but they might not be *accurate*. These two terms are not interchangeable in the context of data analysis. Precision is a measure of the exactitude with which a value, for example a summary statistic, can be expressed. Accuracy, on the other hand, is a measure of whether a figure is broadly consistent with the truth [Wheelan2013]. For example, saying that your home is “3.475 miles to the North” is a very precise statement, but it will be inaccurate if you are facing the wrong direction and your home is actually to the South.

The dichotomy between precision and accuracy poses a number of challenges to the data explorer, from matching the correct analysis to the question and data at hand to selecting the most relevant facet of the many and often conflicting aspects of a complex phenomenon. The main issue is not that you can lie with statistics, as Mark Twain once said, but that you might get insights from it that, while technically true, don’t further your understanding of the issue being analyzed.

In the next sections, we look into some of the basic analytical tools offered by the SAS Visual Analytics Explorer and discuss scenarios where they should be used with extra care.

Mean and Median

The summary statistics *mean* (the average of all values) and *median* (the central value of the sorted value set, or the average of the central pair in the case of an even number of measures) are common measures of the "middle" of a distribution. But while the first is vulnerable to outliers, the second is not. This crucial difference has severe implications on how they are affected by the underlying data and correlate with how well suited they are for answering a specific user question.

The current debate around wealth distribution in the USA provides a good example of how the use of averages can be misleading in cases where outliers can grossly sway the final result. The insights provided by looking at the average income trends are completely different once you isolate the top 1%, and even more so when you look at the top 0.01% [Thompson2014]. Similar results are found for other indicators such as life expectancy; not only the averages don't tell the whole story, but the false insight they provide is likely to be used to justify broad impact changes like increasing the retirement age [Fletcher2013].

That is not to say that the median is always a superior value and that it should be always preferred over the mean. The fact that it is not vulnerable to outliers can be detrimental as well. For example, when evaluating medical treatments for a fatal disease, a patient is told that the median increase in life expectancy provided by the treatment is just one year. But that hides the fact that a minority of the patients is completely cured by it. The fact that the mortality distribution is "right-skewed" blindsides the median. This scenario was described by the evolutionary biologist Stephen Jay Gould in his landmark paper, "The Median Isn't the Message" [Gould1975].

Regression Analysis

Regression analysis is likely the most important tool when it comes to finding meaningful associations between measures in large datasets. Because many scenarios in the social studies and health sciences don't lend themselves to controlled experiments, regression analysis of carefully selected measures in longitudinal data studies has been a key technology in unlocking new insights in these areas.

The problem with regression analysis is that it can provide precise answers to complex questions, but these answers might not be accurate. To avoid a misleading or even completely wrong insight, the data explorer must watch out for the following common regression analysis mistakes [Wheelan2013]:

Analyzing a nonlinear relationship

Linear regression tries to fit a straight line with the data. But a single coefficient (the slope of that line) cannot accurately reflect the relationship between the selected measures unless that relationship is linear or approximately constant for all observations. If that basic assumption is violated, you still get a coefficient, but it won't be meaningful.

Assuming that correlation is equal to causation

Although it is tempting to do so, we cannot infer a causation effect from a strong correlation found by regression analysis; it only means that the association is strong. For example, the rise in the number of Facebook users might have a strong correlation with global warming, but it probably didn't cause it.

Ignoring the possibility of reverse causality

Not only a strong association between A and B doesn't prove that A causes B, it also doesn't preclude the opposite (that B caused A). Causality may go in both directions and needs to be taken into account when selecting explanatory and dependent variables, something that the user only can do with his knowledge of the meaning of the variables in the data.

Omitting variables

When an explanatory variable is left out of the analysis (perhaps by mistake or lack of knowledge about the data), another correlated variable might pick up the effects of the omitted variable, leading to invalid or even incorrect conclusions about the associations in the data. This issue is also known as the omitted variable bias.

Using highly correlated explanatory variables

When two measures are highly correlated, adding both to the analysis is usually not a good practice. If the number of observations that satisfy one measure but not the other is just a small fraction of the dataset, no meaningful inference can be made. In this case, it might be better to create a new variable that can represent the combined effect of both variables and use that variable in the analysis instead.

Extrapolating beyond the data

The results of regression analysis, like any other form of statistical inference, are only valid for a population that is similar to the sample on which the analysis has been done. A user must keep that in mind when using data preparation or dynamic filters to reduce the original dataset. Any conclusions reached by using regression after that point must be considered in the context of the reduced dataset.

Selecting too many variables

One needs to be particularly wary of the “shotgun” approach to regression analysis. If you put enough variables in a regression analysis, one of them is bound to meet the threshold for statistical significance by chance. And thanks to the post-hoc fallacy [McRaney2011], a user might be able to build a theory that “explains” what is actually just an artifact of the analysis.

CASE STUDIES

In this section, we look at practical examples of how the SAS Visual Analytics Explorer can be used to identify and remediate some the issues described in the previous sections.

CASE STUDY 1: CORRELATION MATRIX AND TIME SERIES DATA

Let’s say we are exploring a table of economic indicators for different countries. One would usually start by creating a correlation matrix to identify potential relationships between these economic indicators (Figure 2).

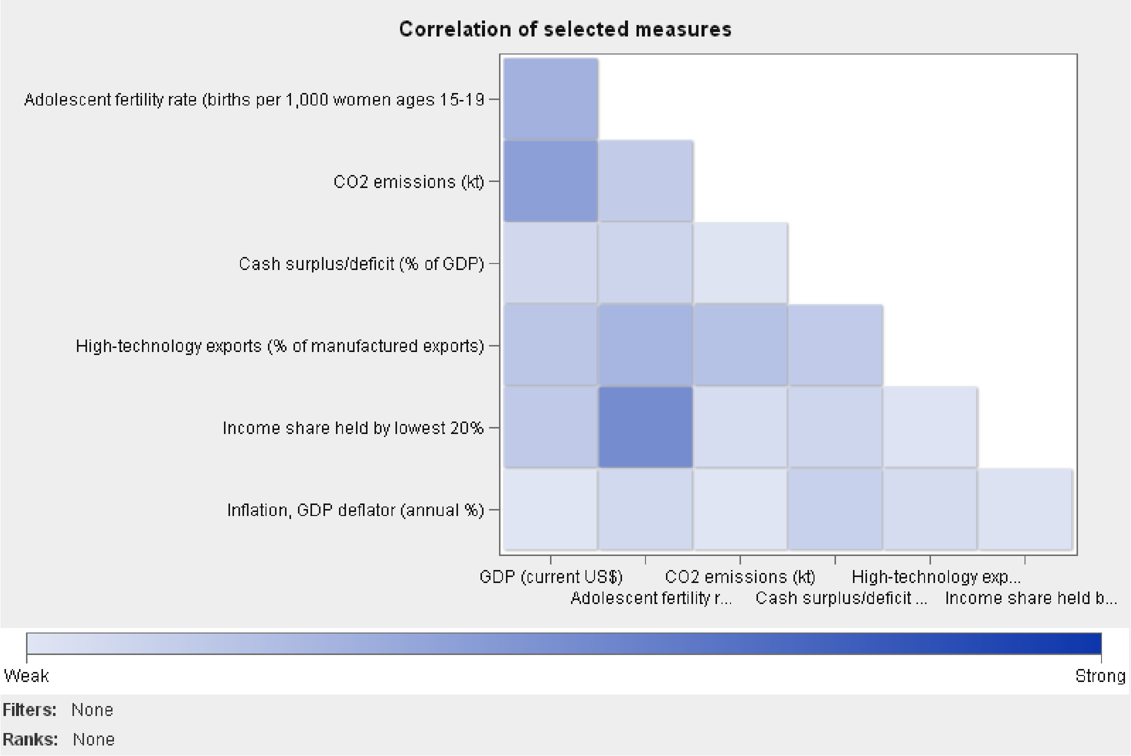
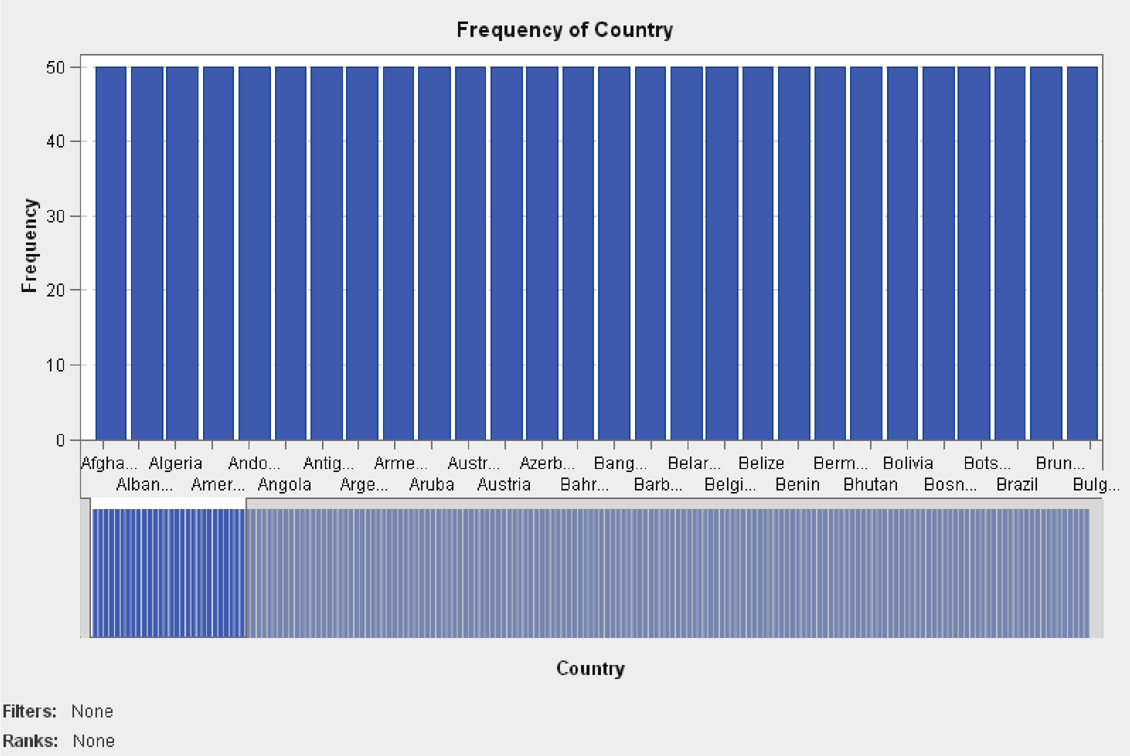


Figure 2 - correlation of countries economic indicators

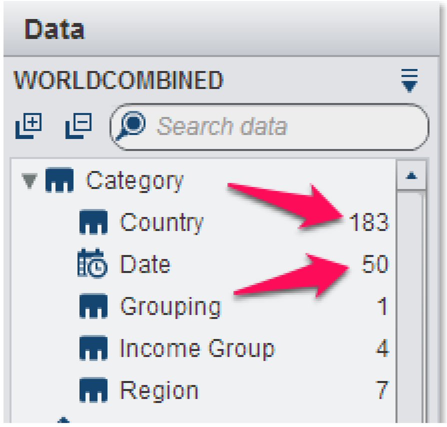
The SAS Visual Analytics Explorer correlation matrix visualization displays the correlation value between pairs of measures. These values are generated by using the Pearson product-moment correlation coefficient method, which assumes the one-row-per-subject data structure. That in turn requires that each subject (in this case, each country) is associated with a single observation or row under analysis. We can use a frequency plot to check this assumption. Figure 3 has the result, and it shows that the assumption does not hold—in fact, we have 50 observations for each country.



Filters: None
Ranks: None

Figure 3 - Frequency Plot for Country,

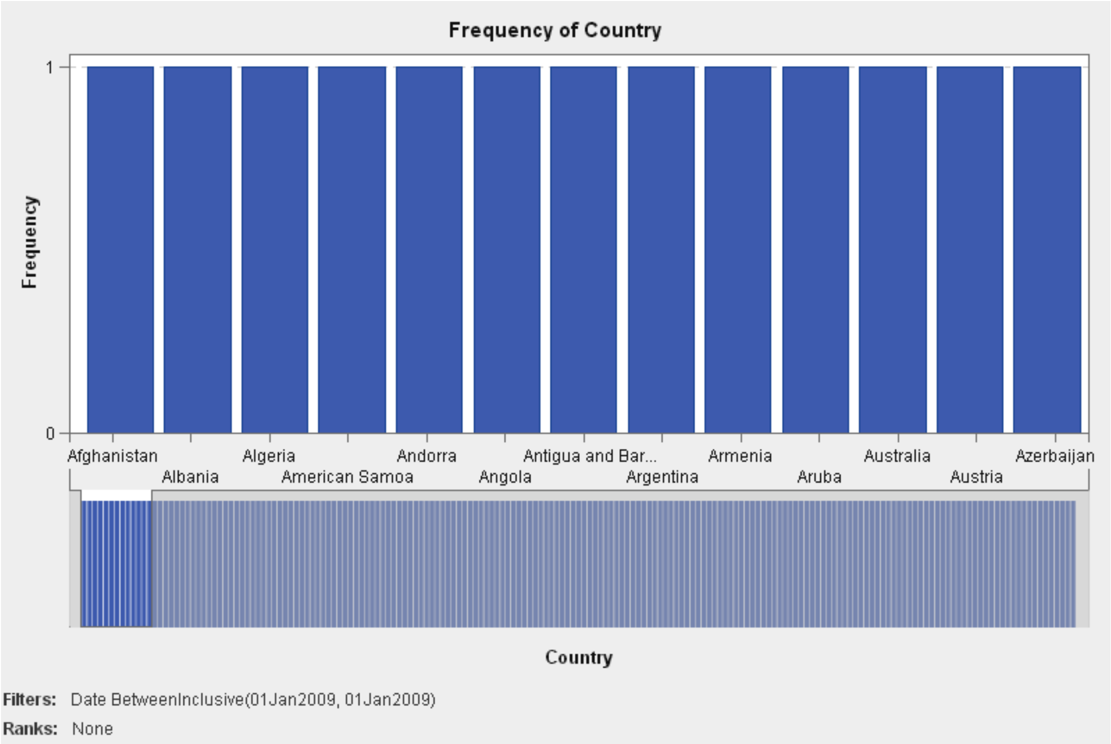
By looking at the SAS Visual Analytics Explorer Data panel, we can quickly identify the source of this multiplicity. The Date variable has 50 unique values (cardinality), which indicates that this dataset is actually an instance of time series data.



Category	
Country	183
Date	50
Grouping	1
Income Group	4
Region	7

Figure 4 - Data Panel

We can proceed with a correlation analysis if we can find a way to fulfill its data structure assumption. To do so, we need to collapse all observations in the time range we are interested on into a single value per subject. A generic approach for doing that is to calculate for each subject a sample correlation coefficient, where the sample values are the measurements taken during that time range. The sample correlation coefficient of two variables A and B is defined as the sample covariance of the two variables, divided by their sample standard deviation. For simplicity, we can take a different and more limited approach, using a data filter to isolate a single time value to make sure that only a single row for each country is selected. Figure 5 confirms that our uniqueness assumption for country observation is now valid, after the data was filtered by a particular year (2009).

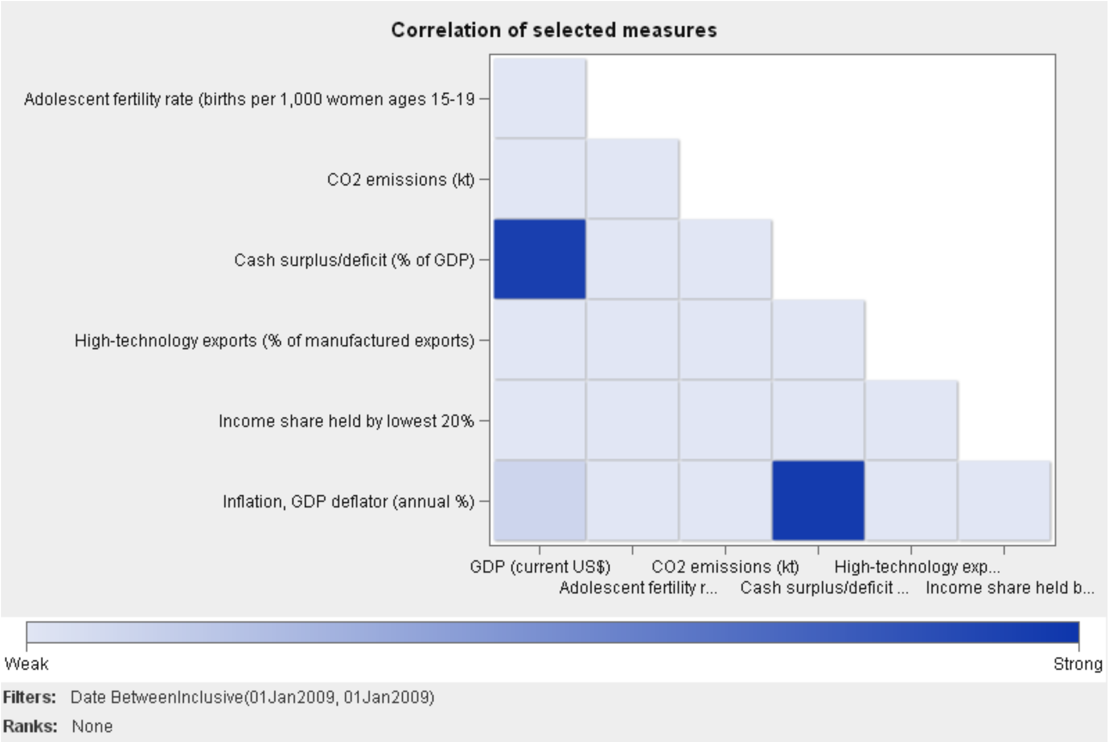


Filters: Date BetweenInclusive(01Jan2009, 01Jan2009)

Ranks: None

Figure 5 - Frequency Plot after Filter

After we verified that the data structure assumptions have been satisfied, we can safely proceed with our correlation matrix and look for strong correlations between economic indicators for a given year. Figure 6 shows the resulting visualization.


Correlation of selected measures
Adolescent fertility rate (births per 1,000 women ages 15-19
CO2 emissions (kt)
Cash surplus/deficit (% of GDP)
High-technology exports (% of manufactured exports)
Income share held by lowest 20%
Inflation, GDP deflator (annual %)
GDP (current US$)
Adolescent fertility r...
CO2 emissions (kt)
Cash surplus/deficit ...
High-technology exp...
Income share held b...
Weak
Strong

Filters: Date BetweenInclusive(01Jan2009, 01Jan2009)

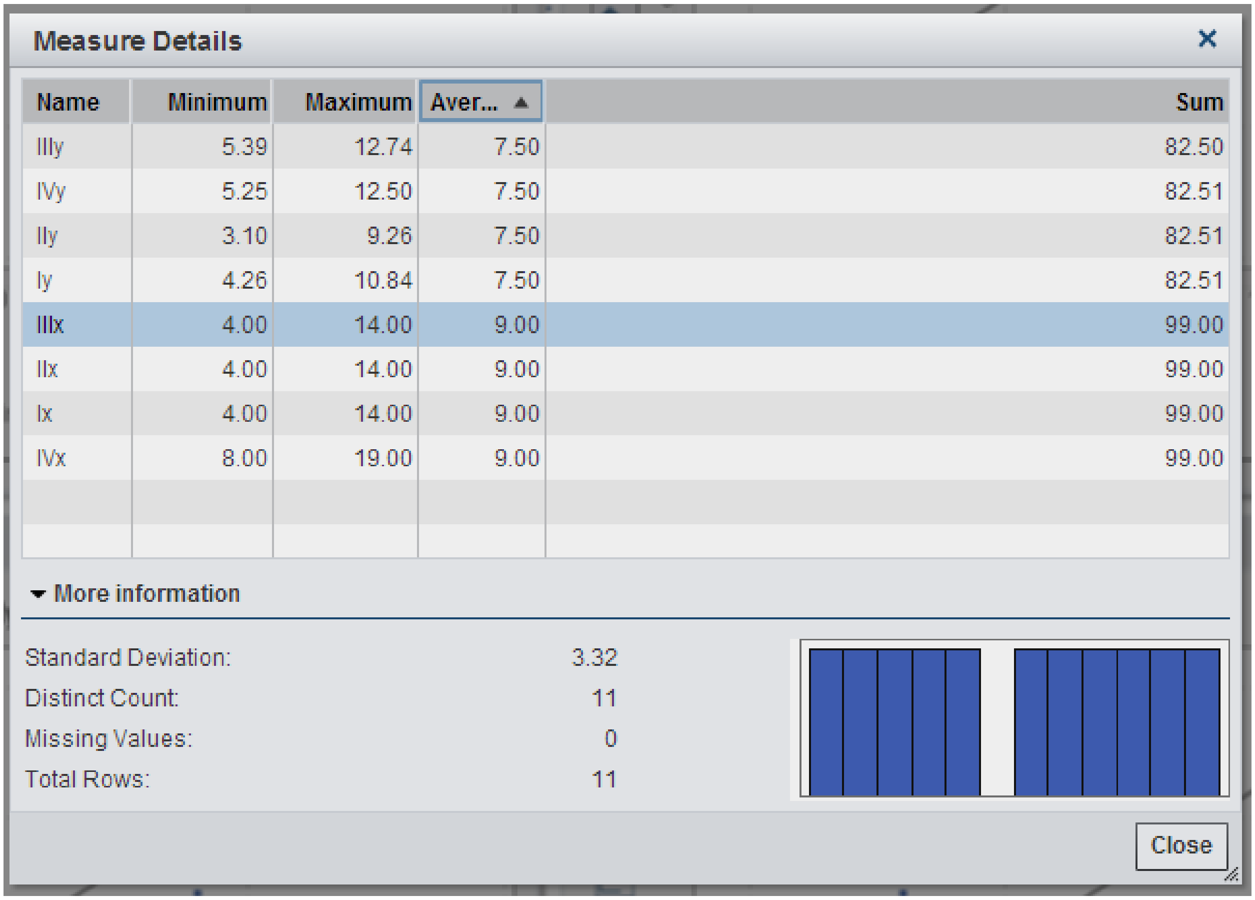
Ranks: None

Figure 6 - Correlation Matrix after Filter

CASE STUDY 2: ANSCOMBE'S QUARTET

In 1973, the statistician Francis J. Anscombe published four datasets (each with eleven x and y pairs of measures) designed to demonstrate the importance of graphing data and accounting for outliers when calculating descriptive statistics. The four datasets, famously known as the *Anscombe's Quartet,* have nearly identical statistical properties: same mean (for x and y), variance (for x and y), correlation between x and y, and linear regression line.

Some of these properties can be seen in the SAS Visual Analytics Explorer Measure Details display, illustrated by Figure 7.



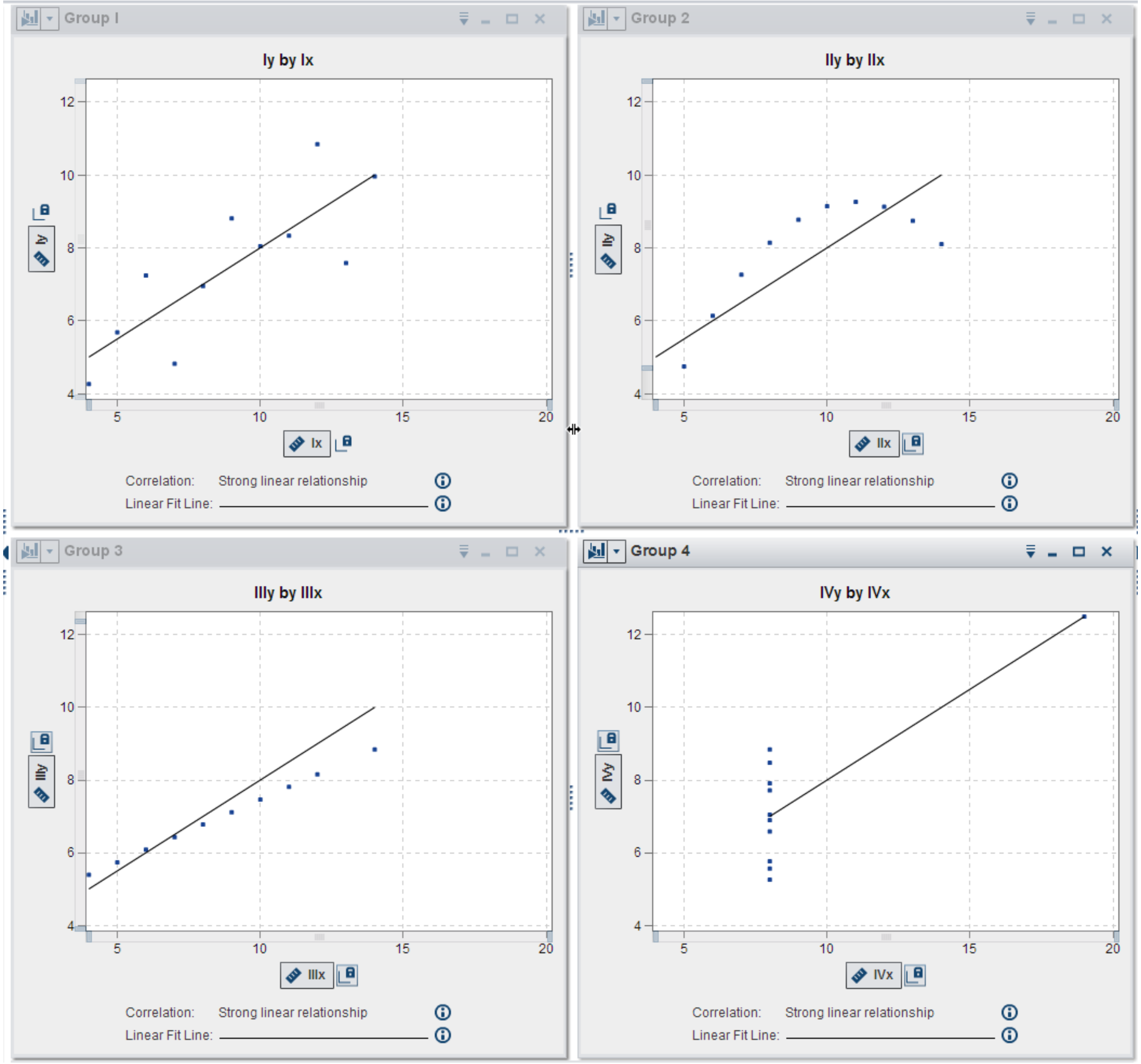
Name	Minimum	Maximum	Aver... ▲	Sum
IIIy	5.39	12.74	7.50	82.50
IVy	5.25	12.50	7.50	82.51
IIy	3.10	9.26	7.50	82.51
Iy	4.26	10.84	7.50	82.51
IIIx	4.00	14.00	9.00	99.00
IIx	4.00	14.00	9.00	99.00
Ix	4.00	14.00	9.00	99.00
IVx	8.00	19.00	9.00	99.00

Standard Deviation:	3.32
Distinct Count:	11
Missing Values:	0
Total Rows:	11

Figure 7 - Measure Details for Anscombe's Quartet

When visualized in scatter plots with a linear regression fit line overlay (Figure 8), the unique characteristics of each member of the quartet come to life. Moreover, the fact that some of the datasets violate the assumptions of a linear regression analysis becomes self-evident. This is a classic demonstration of the dangers of relying on statistic properties alone to understand the data and the importance of visualization during the exploratory data analysis process.

After equalizing the axes to allow for an effortless comparison between the visualizations, we can see that the dataset represented by Group 1 exhibits a straightforward linear relationship. Group 2, on the other hand, displays an approximately quadratic relationship between its variables. Group 3 displays a linear relationship but also an outlier that changes the course of the regression line and lowers the correlation value. Group 4 shows the power of an extreme outlier to influence the correlation coefficient calculation, producing a strong value even for a clearly non-linear relationship. Other than Group 1, all datasets violate some of the assumptions of the linear regression analysis.



Correlation: Strong linear relationship
Linear Fit Line:

Correlation: Strong linear relationship
Linear Fit Line:

Correlation: Strong linear relationship
Linear Fit Line:

Correlation: Strong linear relationship
Linear Fit Line:

Figure 8 - Anscombe's Quartet, Visualized

Figure 9 shows an alternative for the classic visualization of the Anscombe's quartet, using box plots instead of scatter plots. While we cannot see the linear regression fit line overlay (which is inaccurate in all but the first dataset), we can quickly grasp the influence of outliers and the asymmetry of some of the distributions as evidenced by the difference in their mean and median values.

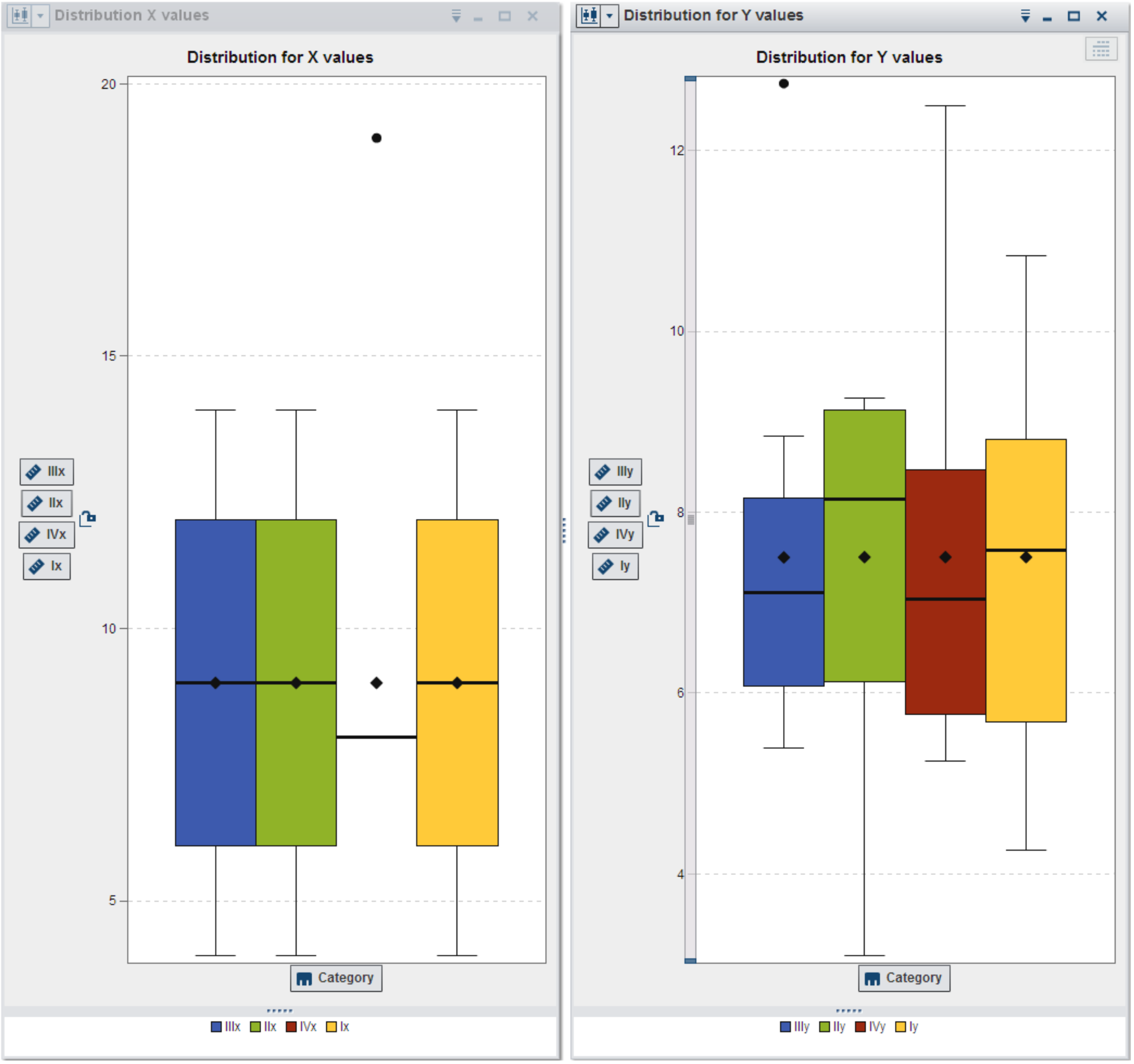


Figure 9 Boxplots for Anscombe's Quartet

CASE STUDY 3: SIMPSON’S PARADOX

Sometimes statistics results can be anything but intuitive. The Simpson's paradox is a mind-boggling example. It happens when the correlation calculated for a dataset changes direction if the dataset is split into groups. In other words, a positive correlation found when looking at the whole dataset not only disappears but is replaced with negative correlations when the data is partitioned. The Simpson's paradox illustrates the need to be skeptical of conclusions that rely on statistics that depend on how the data is aggregated [Bickel1975]. The Simpson's paradox is a type of omitted variable bias, in which one variable takes over the explanation of a missing or lurking variable and leads to an incorrect interpretation of an apparently obvious insight.

One of the best known examples of the Simpson’s paradox comes from a 1973 case in which UC Berkeley was sued for gender discrimination in graduate school admissions. An analysis of the admission data shows that males had a higher chance of being admitted that females (Figure 10). It was easy to jump to the conclusion that the university had a bias towards male admission.

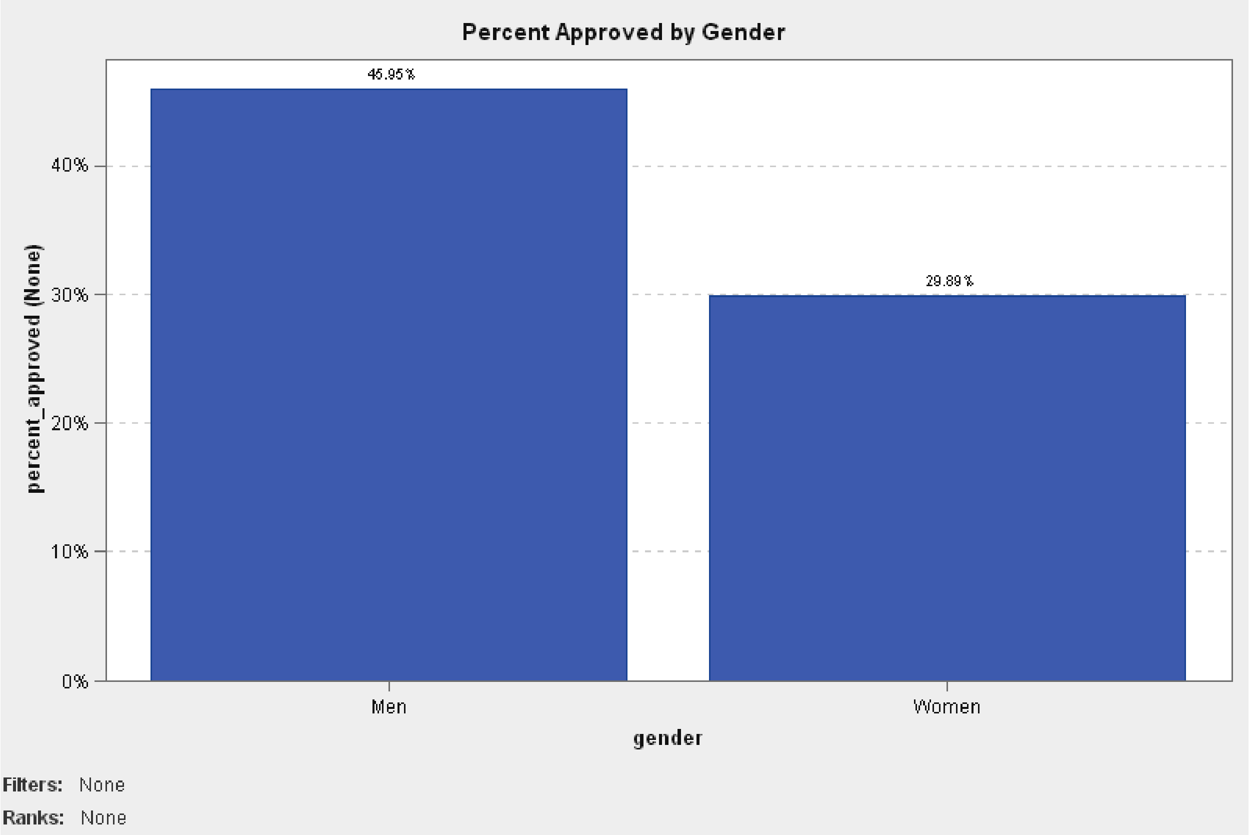


Figure 10 - Berkley Admission by Gender

Because the admission process was the responsibility of each individual department, the data was partitioned by department to find the biggest offenders. Once that was done, however, the problem seemed to flip; most departments seemed to have a significant bias towards admitting more women, not men. For the remaining departments, the difference was not significant (Figure 11). This new picture did not support the original conclusion drawn from the aggregated data.

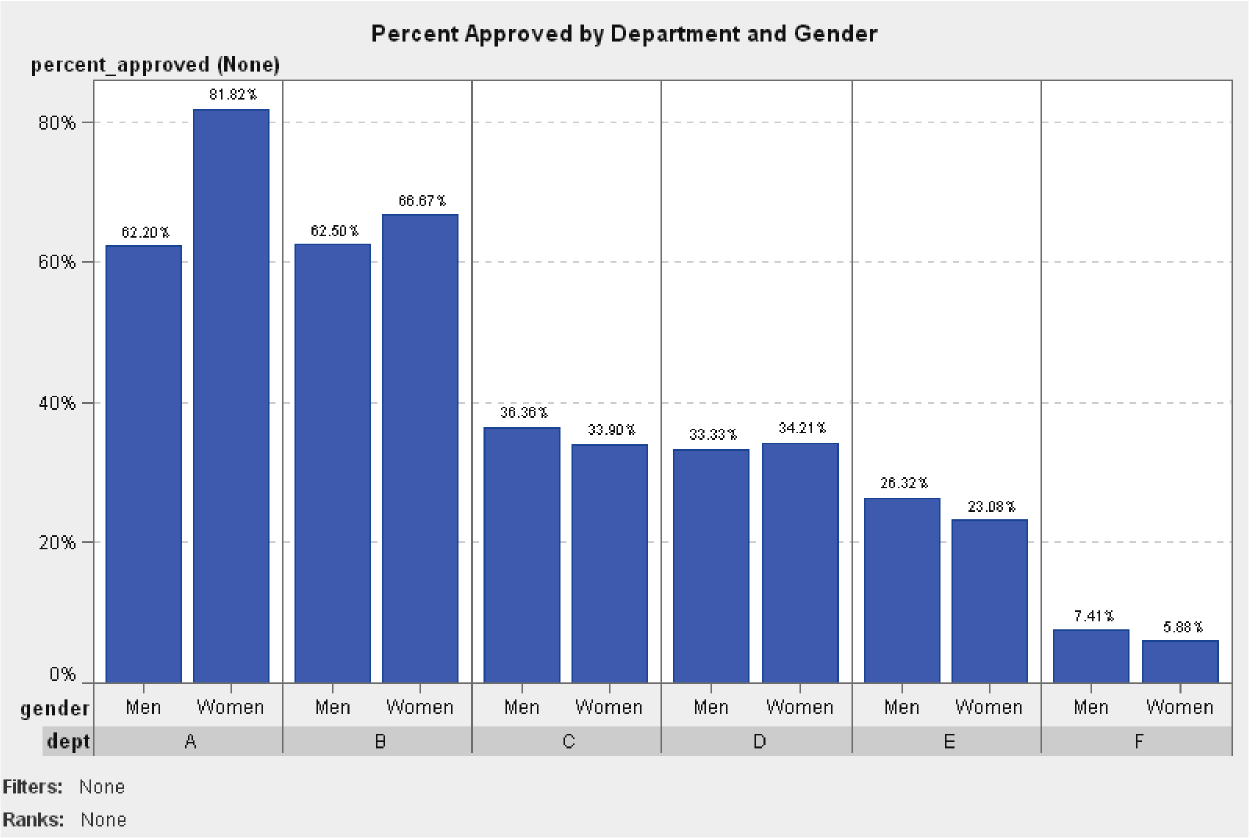


Figure 11 - Berkley Admission by Department and Gender

The explanation for correlation flip observed after the data was partitioned lies on the effect of the lurking variable that had not been considered in the original analysis. If the department variable is taken into account, the gender bias conclusion is reversed.

When looking at the acceptance rate per department, we see that departments A and B had high acceptance rates (Figure 12), they were easier to get in. These departments also happened to have a much higher number of male applicants, which explains why as a whole men had a higher chance to be accepted.

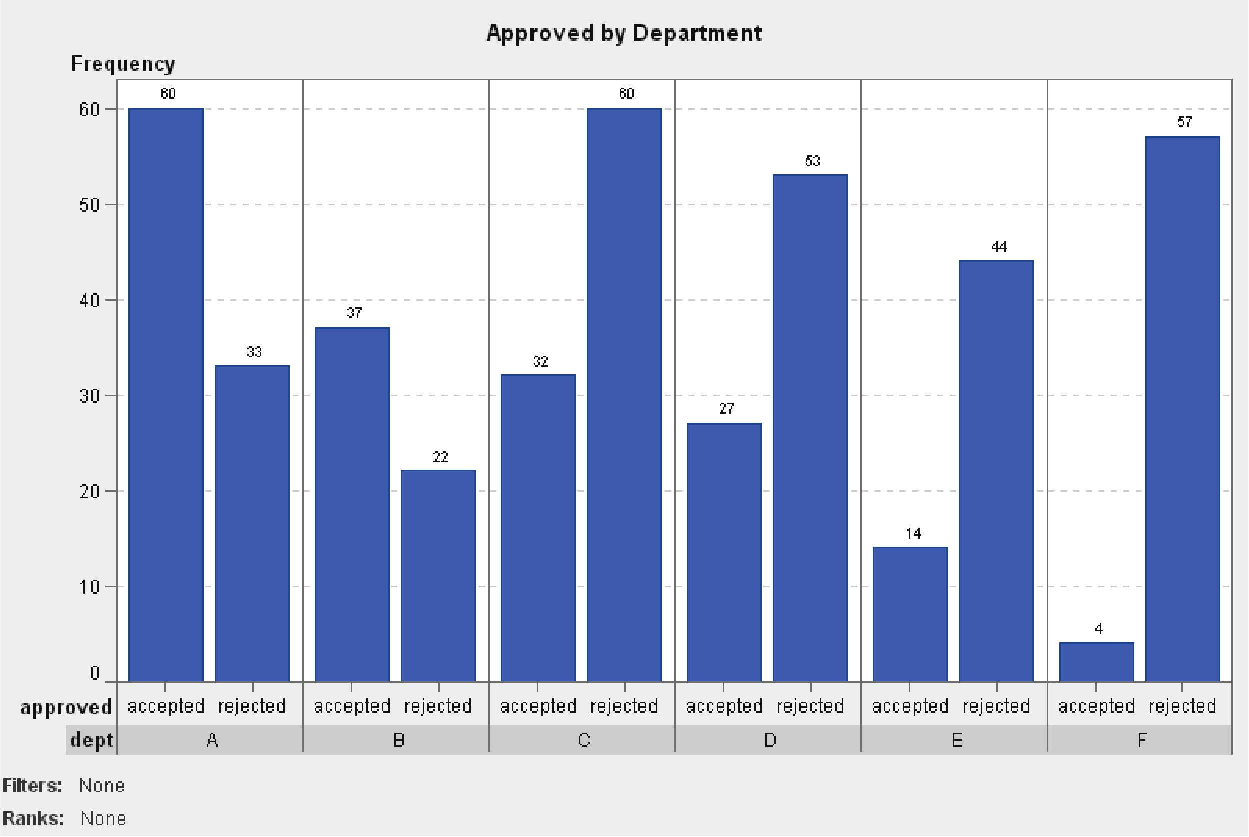


Figure 12 - Berkley Admission by Department

So women had indeed a lower chance of being accepted by departments A and B, but only because they were more likely to apply to other departments that had a lower acceptance rate.

The Simpson's paradox is also behind many other surprising data analysis mirages, from the erroneous selection of kidney stone treatments [Julious1994] to the misinterpretation of wage growth rates [Norris2013].

CONCLUSION

The increasing availability of power tools gave a boost to the DIY movement and led to an explosion of kitchen renovations and backyard tree houses. It also led to a large number of trips to the Emergency Room to deal with lacerated limbs and eye injuries. One needs to learn the assumptions – do not use inside the Jacuzzi – and limitations – don't try to cut through concrete – of a power tool to make the most of it without getting hurt.

Fortunately you are less likely to get hurt when doing data analysis. But your conclusions and insights can hurt your customers and your business if they are not accurate, no matter how precise is the math behind them.

We examined some scenarios and a few case studies illustrating how a data explorer can get lost pursuing mirages that disappear under close inspection. We hope that with the help of the SAS Visual Analytics Explorer you will make it through the data desert labyrinth and arrive safely at true oases of new insights.

REFERENCES

[Abousalh2013] Abousalh-Neto, Nascif A. "The Forest and the Trees: See it All with SAS® Visual Analytics Explorer" - http://support.sas.com/resources/papers/proceedings13/058-2013.pdf

[Bickel1975] Bickel, P. J. et al. "Sex Bias in Graduate Admissions: Data from Berkeley". Science, New Series, Vol. 187, No. 4175 (Feb. 7, 1975), pp. 398-404. http://www.unc.edu/~nielsen/soci708/cdocs/Berkeley_admissions_bias.pdf

[Fletcher2013] Fletcher, Michael. "Research ties economic inequality to gap in life expectancy". Washington Post, 3/10/2013. http://www.washingtonpost.com/business/economy/research-ties-economic-inequality-to-gap-in-life-expectancy/2013/03/10/c7a323c4-7094-11e2-8b8d-e0b59a1b8e2a_story.html

[Gould1975] Gould, Stephen J. "The Median Isn't the Message". http://www.stat.berkeley.edu/users/rice/Stat2/GouldCancer.html

[Julious1994] Julious, Steven A. "Confounding and Simpson's paradox". BMJ 1994;309:1480. http://www.bmj.com/content/309/6967/1480

[Lehe2014] Lehe, Lewis and Powell, Victor. "Simpson's Paradox". Visualizing Urban Data, UC Berkeley ideaLab - http://vudlab.com/simpsons/

[McQuown2005] McQuown, Gary. "PROC IMPORT with a Twist" SUGI 30, Proceedings - http://www2.sas.com/proceedings/sugi30/038-30.pdf

[McRaney2011] McRaney, David. "You Are Not So Smart: Why You Have Too Many Friends on Facebook, Why Your Memory Is Mostly Fiction, and 46 Other Ways You're Deluding Yourself". Gotham, 2011.

[Norris2013] Norris, Floyd. "Median Pay in U.S. Is Stagnant, but Low-Paid Workers Lose". New York Times, April 26, 2013 - http://www.nytimes.com/2013/04/27/business/economy/wage-disparity-continues-to-grow.html

[Sglavo2012] Sglavo, Udo. "Common Analytical Data Structures". SAS R&D presentation (internal).

[Thompson2014] Thompson, Derek. "The Rise (and Rise and Rise) of the 0.01 Percent in America". The Atlantic, 2/13/2014 - http://m.theatlantic.com/business/archive/2014/02/the-rise-and-rise-and-rise-of-the-001-percent-in-america/283793

[Wheelan2013] Wheelan, Charles. "Naked Statistics - Stripping the Dread from the Data". Norton and Company, 2013.

ACKNOWLEDGMENTS

I would like to thank Udo Sglavo for his insights on the critical role played by data structures in data analysis.

RECOMMENDED READING

- *Statistics Done Wrong* - http://www.refsmmat.com/statistics/index.html
- *Data Analysis: The Hard Parts* - http://blog.mikiobraun.de/2014/02/data-analysis-hard-parts.html
- *Wealth Inequality in America* - http://www.youtube.com/watch?v=QPKKQnijnsM

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Nascif Abousalh-Neto
SAS Institute Inc
SAS Campus Drive S4082 Cary, NC 27513-2414, USA
Phone: (919) 531-0123
nascif.abousalh-neto@sas.com
http://www.linkedin.com/in/nascif
https://communities.sas.com/people/Nascif%40SAS