Paper SAS722-2017

Investigating Big-Data Crime Scenes

Theresa Hoang Diem Ngo, SAS Institute Inc., Cary, NC

ABSTRACT

Statistical analysis is like detective work, and a data set is like the crime scene. The data set contains unorganized clues and patterns that can, with proper analysis, ultimately lead to meaningful conclusions. Using SAS® tools, a statistical analyst (like any good crime scene investigator) first performs a preliminary analysis of the data set through visualization and descriptive statistics. Based on the results of the preliminary analysis, both the crime scene investigator (CSI) and the statistical analyst (SA) can use subsequent, detailed analysis, along with scientific or analytical tools, to answer the key questions: What happened? What were the causes and effects? Why did this happen? Will it happen again? Applying the CSI analogy, this paper presents an example case study using a two-step process to investigate a big-data crime scene. Part I shows the general procedures that are used to identify clues and patterns and to obtain preliminary insights from those clues. Part II narrows the focus on the specific analytics that provide answers to different questions.

INTRODUCTION

The CSI TV show generally starts with a crime being committed (key details are omitted so that they can be revealed later in the show through scientific investigation). The clues and pieces of evidence are either hidden or in plain sight along with other noise and irrelevant information at the crime scene. A CSI team then comes to the crime scene and gathers evidence and clues for further analysis back at the lab. The lab has all sorts of super cool scientific tools that a CSI uses to process and identify pieces of evidence that are significant to the crime. The findings fill in the missing pieces to create an almost complete picture of what happened and/or who committed the crime. The show usually ends with a happy ending where the CSI solves the case. However, some episodes end unsolved with the cases jeopardized by lack of evidence, contaminated clues, improper procedures, and so on.

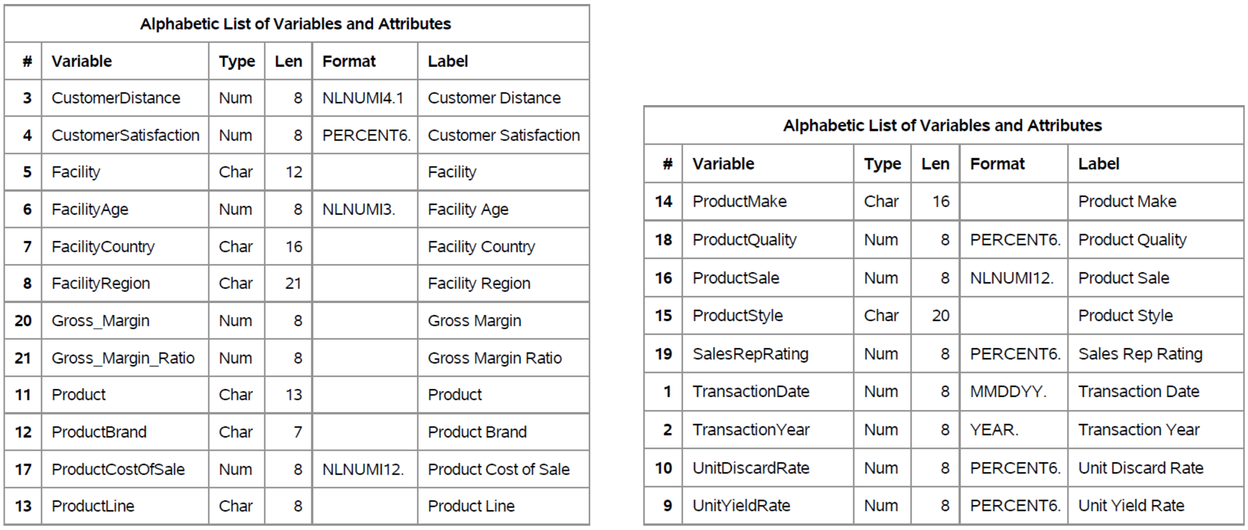
In the business world, data and analytics processes are often similar to the crime scene processes. Like crime scene investigators, business and statistical analysts use powerful SAS® tools to perform a variety of analytics on a raw data set to gain insights and solve problems that affect the company’s bottom line. The characteristics of a raw data set are analogous to the entirety of the initial crime scene; full of irrelevant noise and contaminated clues directly alongside telling evidence – biased and dirty. These general characteristics are common, but very critical in determining the results of the analysis. Even if the analysis performed correctly, the results can give false confidence and misleading conclusions due to the flawed data. The importance of filtering out irrelevant data cannot be understated. The two-step process of investigating big-data crime scenes begins with a preliminary data analysis (intended to filter out noise) prior to applying appropriate analytics. With that said, this paper is relevant to business and statistical analysts who are on the journey of analytic exploration. This paper discusses an end-to-end analytical process with a variety of techniques at a high level. For more in-depth understanding about these techniques, please refer to the References and Recommended Reading.

BIG-DATA CRIME SCENE

Big Data is not only big hype, but also has become the new normal across many different industries and government organizations. There are more data on more things in an ever-growing variety of formats, from structured data to unstructured text and video. The modern technologies can store large volumes of data across different databases, in distributed locations, connected via networks. The limitation of data storage capacities is no longer a concern. Big Data analytics has the potential to be extremely valuable to many companies, but is worth very little unless management is able to gain insights that help make value added decisions. Just as modern science is continuously providing new tools for the crime scene investigator to analyze crime scene evidence (3-D computer rendering, improved security video quality, cell phone video prevalence, and so on), analytical tools are becoming more powerful. This paper applies the latest SAS analytical tools to a real world business example.

The Acme Toy data set (see Table 1) is used in all the examples to demonstrate the visual analytics and statistical techniques. The data set contains 14 years of financial, manufacturing, sales, and marketing information. To give a brief background of the data set, Acme Toy is a manufacturing company that produces and sells toys. Its customers are vendors that resell their toys at different store locations. As a manufacturing company, Acme Toy's objective focuses on these main functions of the business:

- Finance: sales, gross margin, cost of sales, and so on
- Manufacturing operations: efficiency, unit discard rate, unit yield rate, and so on
- Customer service: sales representative rating, customer satisfaction, and so on



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
3	CustomerDistance	Num	8	NLNUMI4.1	Customer Distance
4	CustomerSatisfaction	Num	8	PERCENT6.	Customer Satisfaction
5	Facility	Char	12		Facility
6	FacilityAge	Num	8	NLNUMI3.	Facility Age
7	FacilityCountry	Char	16		Facility Country
8	FacilityRegion	Char	21		Facility Region
20	Gross_Margin	Num	8		Gross Margin
21	Gross_Margin_Ratio	Num	8		Gross Margin Ratio
11	Product	Char	13		Product
12	ProductBrand	Char	7		Product Brand
17	ProductCostOfSale	Num	8	NLNUMI12.	Product Cost of Sale
13	ProductLine	Char	8		Product Line

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
14	ProductMake	Char	16		Product Make
18	ProductQuality	Num	8	PERCENT6.	Product Quality
16	ProductSale	Num	8	NLNUMI12.	Product Sale
15	ProductStyle	Char	20		Product Style
19	SalesRepRating	Num	8	PERCENT6.	Sales Rep Rating
1	TransactionDate	Num	8	MMDDYY.	Transaction Date
2	TransactionYear	Num	8	YEAR.	Transaction Year
10	UnitDiscardRate	Num	8	PERCENT6.	Unit Discard Rate
9	UnitYieldRate	Num	8	PERCENT6.	Unit Yield Rate

Table 1. Overview of the Acme Toy Data Set

PART I: PRELIMINARY DATA ANALYSIS

MEETING THE MINIMUM DATA CRITERIA

In order to avoid drawing incorrect conclusions, it is important to first cleanse and prepare the data set. At a minimum, the collected data set should meet the following criteria.

- **A statistically large data** set contains at least 30 observations (records); a larger data set will help smooth out variations
- **An unbiased data set** is a good representation of the population; a biased sample data set is skewed to specific groups, characteristics, time frame, and so on
- **A clean data** set means good data quality; bad qualities such as errors, invalid and missing values, skewness, noise, and so on, will affect the results

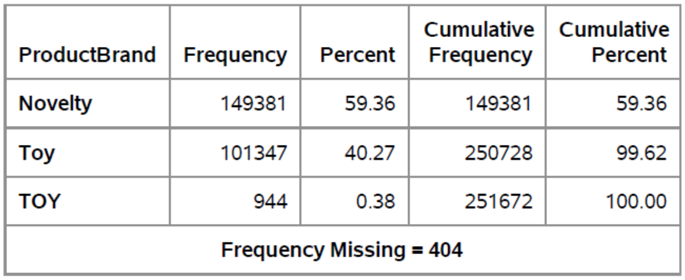
Just as a CSI must discern unrelated items from actual evidence associated with a crime, the statistical analyst must determine what, and how much, data is appropriate.

CHECKING FOR DATA PROBLEMS

A raw collected data set is naturally messy and dirty with a variety of different data problems and errors. As the CSI must ensure sufficient and valid evidence is collected (Is there sufficient DNA evidence to make a match?), the statistical analyst must also pre-screen the raw data in order to identify errors, duplicates, truncated/invalid values, missing values, and so on. Most of these problems are easy to identify and resolve. Four of the most common and effective data cleansing tasks are listed below. The individual data set might have additional problems that are not covered in this paper. *Cody's Data*

Cleaning Techniques Using SAS® Software by Ron Cody provides a comprehensive overview of cleaning techniques.

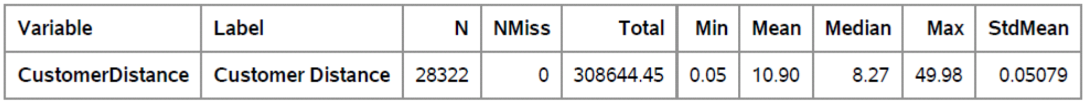
1. A very simple approach for identifying **invalid character values** is to produce a frequency table that lists out all the unique values (and their frequencies) of these variables. See the frequency listing for the variable PRODUCTBRAND in Figure 1. If valid values for PRODUCTBRAND are ‘Toy’, ‘Novelty’, and missing, any other values that appear on the frequency listing are considered errors. Depending on the situation, the uppercase value ‘TOY’ might or might not be an error. A judgment call must be made in such cases. For this example, a quick fix would be to make all values either lowercase or uppercase and then reproduce a new frequency table.



ProductBrand	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Novelty	149381	59.36	149381	59.36
Toy	101347	40.27	250728	99.62
TOY	944	0.38	251672	100.00
Frequency Missing = 404				

Figure 1. The Frequency Table Lists All the Unique Values for Character Variables

2. Another technique that checks for **invalid numeric data** is to produce descriptive statistics and plot a histogram. Descriptive statistics contain a variety of measurements such as mean, median, minimum, maximum, standard deviation, and so on. Figure 2 shows the descriptive statistics and histogram of the variable CUSTOMER DISTANCE (customer’s driving distance to a nearby facility). The minimum and maximum values of CUSTOMER DISTANCE are 0.05 and 50 miles. If negative values are observed, they can be flagged as errors since a driving distance cannot be negative. Using the descriptive statistics and histogram helps narrow the focus to only questionable values. Once identified, these data must be individually validated.



Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
CustomerDistance	Customer Distance	28322	0	308644.45	0.05	10.90	8.27	49.98	0.05079

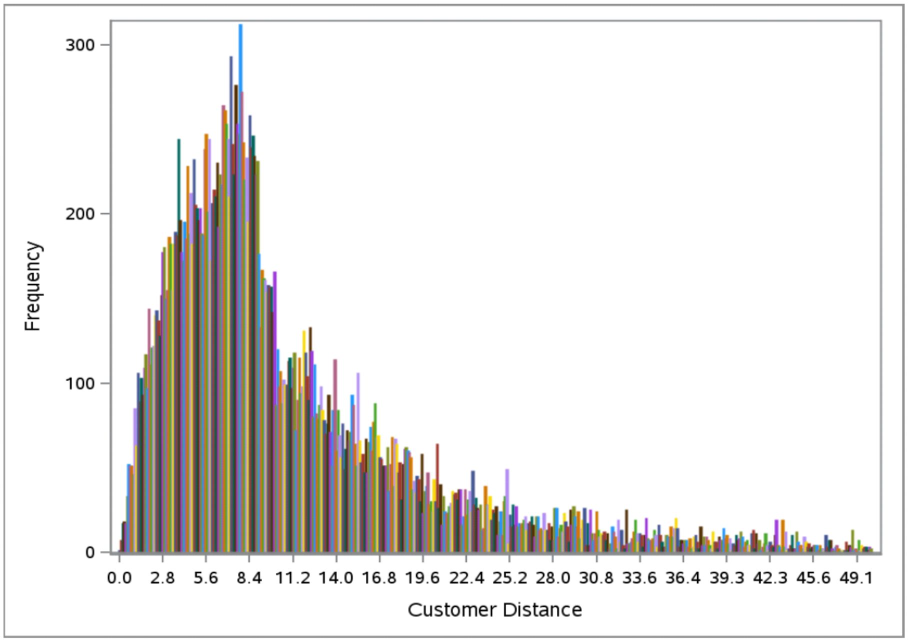
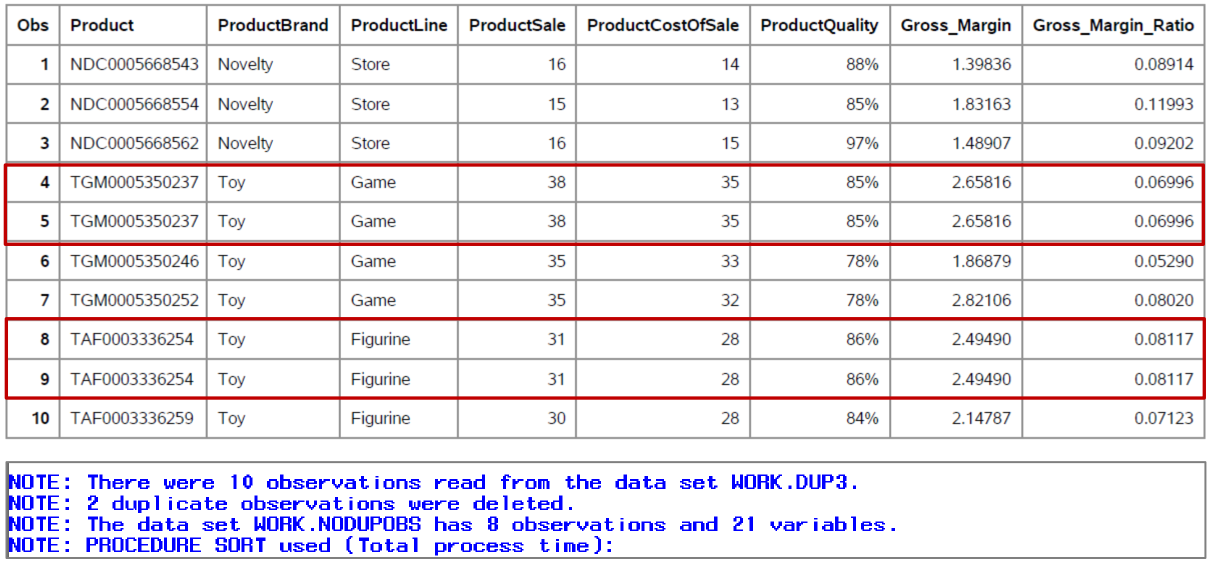


Figure 2. The Descriptive Statistics and Histogram Identify Possible Numeric Data Errors

3. Many data sets have **missing value**s for numeric and character variables. Values are generated missing intentionally or accidentally. In an accidental scenario, the cause could be reading invalid dates and/or character values with the wrong information. In the SAS Log, if there are messages related to invalid data values, it is an indication that something is wrong either with the data or the program. In an intentional scenario, missing numeric values can be imputed either using simple statistics or forecasting techniques. For missing character values, the best educated guess is to perform an internal research and/or reference other values that have similar attributes as the missing values. Both outputs in Figure 1 and 2 also produce the total count of missing values for numeric and character variables.

4. Besides checking for invalid values, it is important to check for **duplicate observations**. Eliminating duplicates is not difficult. Understanding how they were generated, though, is important to prevent it from reoccurring. One possible cause of duplicate observations is merging multiple data sets with missing, duplicate, or invalid ID variables. Figure 3 shows an example of duplicate observations in the red box. The SORT procedure is used to identify and eliminate duplicate observations. The SAS Log message notes that two duplicate observations were deleted.



Obs	Product	ProductBrand	ProductLine	ProductSale	ProductCostOfSale	ProductQuality	Gross_Margin	Gross_Margin_Ratio
1	NDC0005668543	Novelty	Store	16	14	88%	1.39836	0.08914
2	NDC0005668554	Novelty	Store	15	13	85%	1.83163	0.11993
3	NDC0005668562	Novelty	Store	16	15	97%	1.48907	0.09202
4	TGM0005350237	Toy	Game	38	35	85%	2.65816	0.06996
5	TGM0005350237	Toy	Game	38	35	85%	2.65816	0.06996
6	TGM0005350246	Toy	Game	35	33	78%	1.86879	0.05290
7	TGM0005350252	Toy	Game	35	32	78%	2.82106	0.08020
8	TAF0003336254	Toy	Figurine	31	28	86%	2.49490	0.08117
9	TAF0003336254	Toy	Figurine	31	28	86%	2.49490	0.08117
10	TAF0003336259	Toy	Figurine	30	28	84%	2.14787	0.07123

```
NOTE: There were 10 observations read from the data set WORK.DUP3.
NOTE: 2 duplicate observations were deleted.
NOTE: The data set WORK.NODUPOBS has 8 observations and 21 variables.
NOTE: PROCEDURE SORT used (Total process time):
```

Figure 3. Using the Sort Procedure to Identify and Eliminate Duplicates

EXPLORING DATA THROUGH VISUALIZATIONS

It is important to clearly define an objective for the analysis before exploring the data set. The objective could be a hypothesis or key questions that help guide the analyst through understanding the data. Once the data is properly cleansed and an objective is defined, exploring the data set can be performed very quickly with visualizations. Different visualizations help spot trends and patterns, find unknown relationships, identify opportunities, and more. As many CSI episodes feature highly-detailed, three-dimensional visualization recreations to help the investigator identify otherwise overlooked variables, visualization tools help the analyst identify trends that might not be apparent from simple, numerical representations.

For the Acme Toy Company example, the primary profit drivers are determined to be finance, manufacturing operations, and customer services. Having cleansed the data, the next objective is to gain insights into these three main function areas. Figure 4 and 5 show where the most sales and gross margin occur at the product and location levels. Figure 6 identifies which facility location has the lowest performance. Figure 7 shows the customer's driving distance to the facility in each city. The Acme Toy Company's marketing team can offer different promotions to get more foot traffic for facilities that cover longer distances. Figure 8 identifies the key driver to customer satisfaction. Some of these insights are expected, but such results serve to reinforce and quantify experience-based, empirical trends. Any new insights create potential opportunities to improve processes and increase profitability.

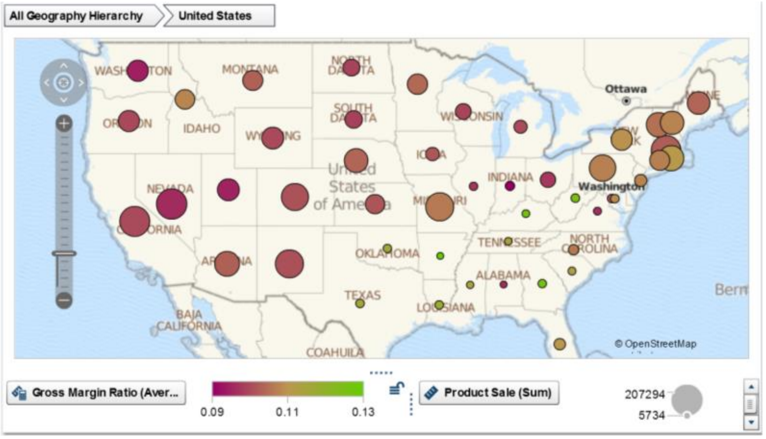


Figure 4. Most Sales Occur in Western and Northern Regions (Large Bubbles), Average Gross Margin is Better in Southeast (Green Bubbles)

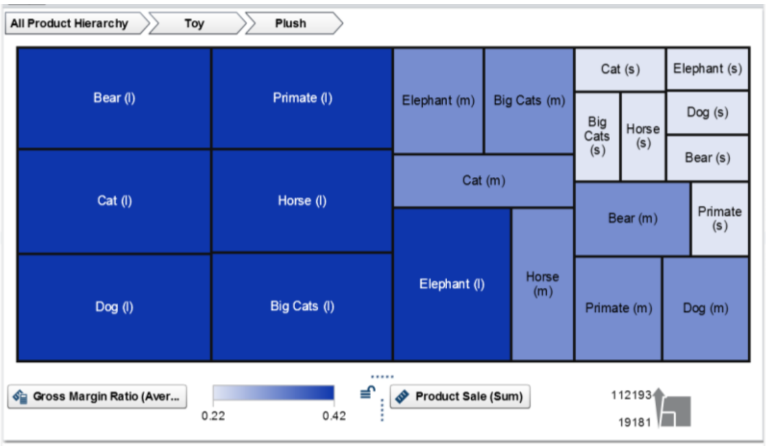


Figure 5. Highest Sales and Gross Margin Range from Large Plush Toys to Small Plush Toys

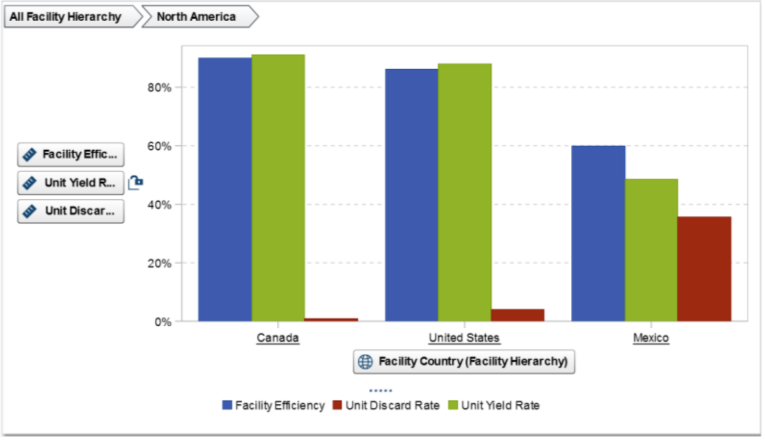


Figure 6. Mexico Has the Lowest Operational Performance

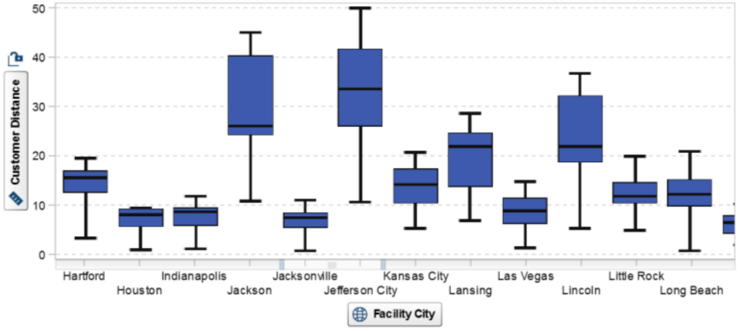


Figure 7. Offer Different Marketing Campaigns to Customers Who Live Farther Away from Facilities in Jefferson City, Jackson, and Lincoln

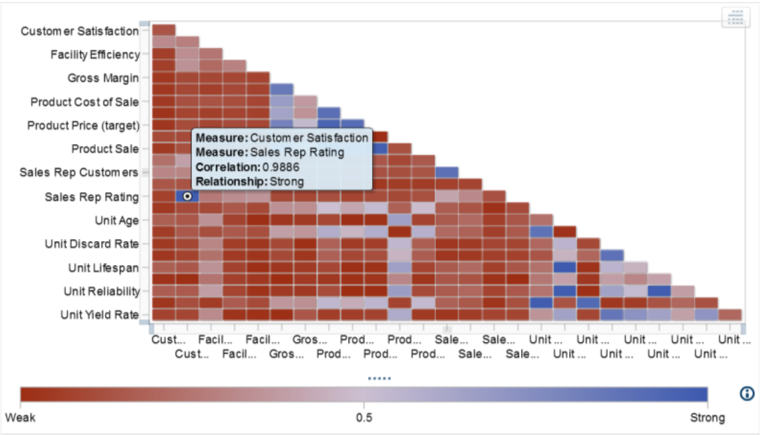


Figure 8. Customer Satisfaction and Sales Rep Rating have a Strong Positive Correlation

PART II: ANALYTICS

The preliminary findings from the Acme Toy data set provide some of the missing pieces of the puzzle. There are still some unknowns, but what has been learned up to this point provides guidance on the most promising areas to dig deeper and what specific statistical techniques to use. For the CSI, this is the point where they have identified the key pieces of evidence and must determine how they fit together (matching DNA evidence with a suspect, using the suspect description to analyze video evidence, and so on). The second objective for the statistical analyst is to further investigate the most promising areas for opportunities to improve sales, increase efficiencies in operations, and achieve better understanding of the customers. The following statistical techniques and analytics are a small subset of many options available to solve a variety of business problems. It is important to match the technique with the specific business objective/problem as closely as possible. For more in depth details about these techniques, please refer to the References and Recommended Reading.

FORECASTING SALES AT MULTIPLE HIERARCHICAL LEVELS

Acme Toy Company’s sale and gross margin vary across products and locations in United States (Figure 4 and 5). There are significant factors (such as price, location, product type, and so on.) that have positive or negative impacts on sales. The forecasting technique formulates an algorithm that take these key factors into account and produce accurate predictions with a certain level of confidence. With accurate demand forecasts, Acme Toy Company can manufacture the right quantity of products and place them at the right locations at the right times. This translates to reduced overstocked inventory, increased revenue, and improved customer satisfaction.

Figure 9 shows a data hierarchy of three forecasting levels (Sale, Facility, and Product). There are seven time series forecasts from the top level of total sale to the bottom level of product type (1 total sale + 2 facilities + 4 product types per facility). A SAS forecasting tool called Forecast Server has an incredible reconciliation process that makes the statistical forecasts (values in black) add up and down the hierarchy by applying the reconciliation effects (values in red). Having the forecasts at multiple levels synchronized is very important for every business. One of the seven time series forecasts (total sale) is shown in Figure 10. Table 2 is a recommended list of popular statistical models that fit different response data types.

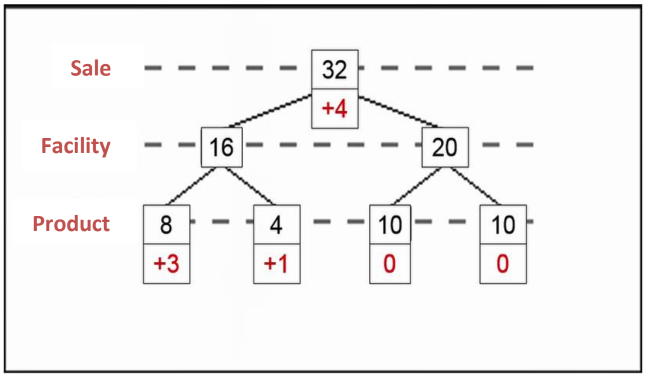


Figure 9. Generate a Forecast for Every Series at Every Level in the Data Hierarchy

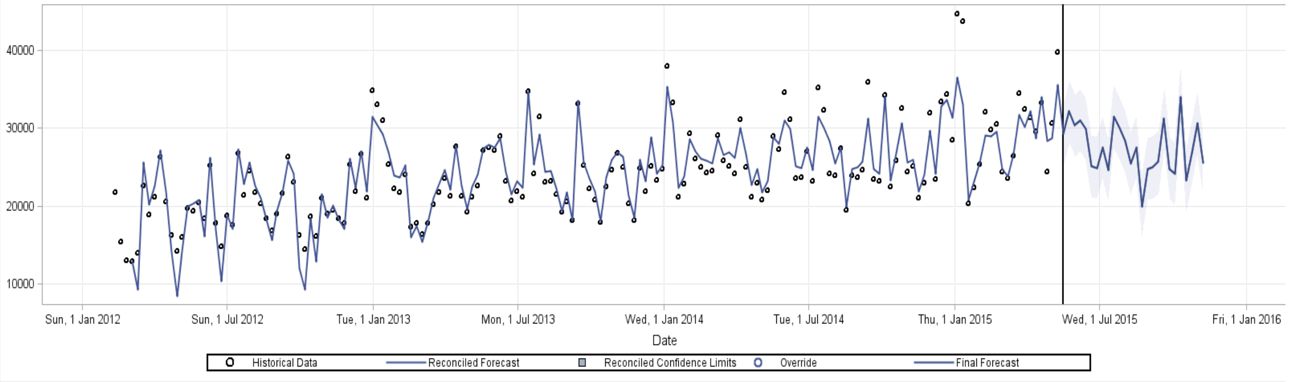


Figure 10. Generate the Total Sale Forecast on a Weekly Basis

Statistical Model	**Response Data Type**	**Comments**	Recommended Reading
Multiple Linear Regression	Continuous	Random errors $\varepsilon \sim N(0, \sigma^2)$; all pairs of random errors are independent	(1)
Time Series Model	Time Series	Random errors are correlated; time dependent	(2)
Generalized Linear Model	Continuous and Discrete	Response distribution is from the Exponential Family of Distribution	(3)
Logistic Regression	Binary, Ordinal, Nominal	Nonlinear relationship between a response and predictor variables	(4)
Survival Analysis	Censored (Survival) Time	Survival time is always positive and typically skewed to the right of the distribution	(5)

Table 2. Statistical Models Fit to Different Response Data Types

CLUSTERING FACILITIES WITH SIMILAR ATTRIBUTES

The preliminary insight of the manufacturing efficiency shows that Mexico has the lowest efficiency score compared to Canada and United States (Figure 6). It makes sense that the efficiency varies across different facility locations. The question is what or why certain facilities have higher efficiency than others. To answer this question, a clustering method uses algorithms to partition facilities into a number of segments based on similar measurements or characteristics of one or more attributes. Figure 11A and 11B are two different visualizations showing the same result of five cluster segments. The result indicates that facilities (Cluster 3 and 4, yellow and baby blue lines) that have been in business for more than ten years are likely to be at least 80% efficient. That implies that these facilities have more experience in operations and management. There are facilities less than ten years old that also operate at least 80% efficient. That would require a deeper look into what other attributes contribute to the efficiency of these younger facilities.

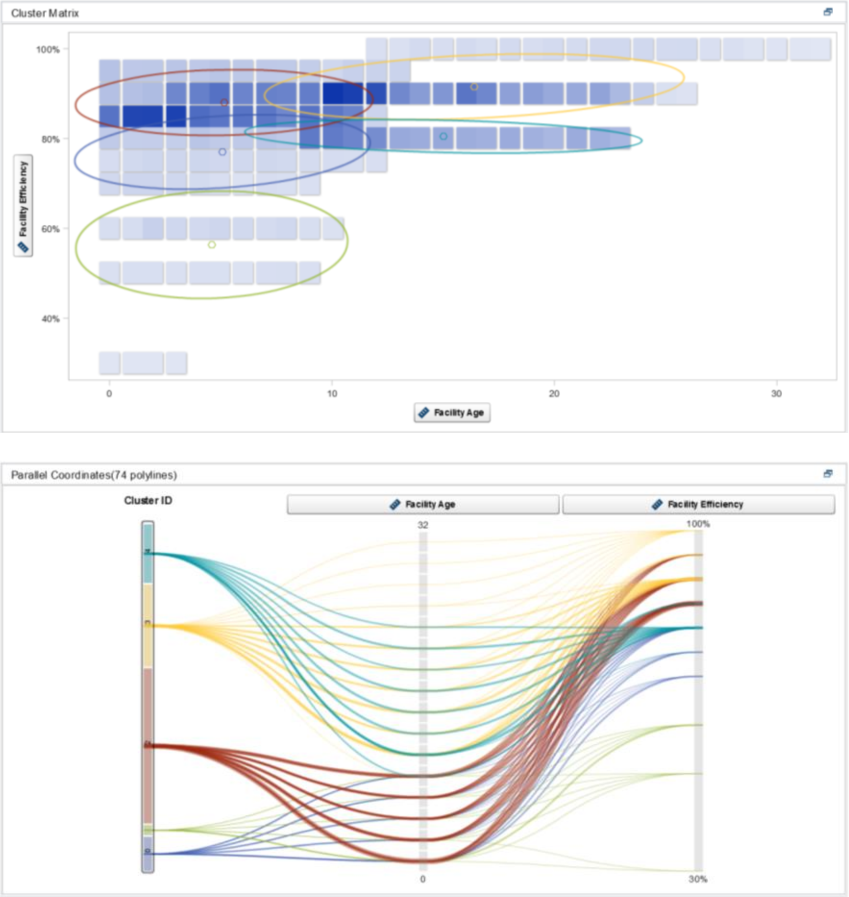


Figure 11A and 11B. Facilities with at Least Ten Years of Experience are Likely to be More Efficient

EXTRACTING MEANINGFUL INFORMATION FROM RAW TEXTUAL DATA

A major missing component in most business models is the ability to understand customer behavior. The ability to anticipate customers' needs and providing them exactly what they want, at the right time is a competitive advantage worth pursuing. Sales forecasting (structured data) is one method of predicting customers' needs. Other approaches such as text analytics uncover patterns and insights from customer feedback or survey responses (unstructured data). The challenges with unstructured data are to transform the free-form text into structured data and then translate the structured data into meaningful and actionable information. The process of transforming unstructured to structure data is not covered in this paper, refer to the Recommended Reading about text analytics.

Using the Acme Toy example, customers' textual comments have been transformed into structured data, which contains only the key elements of the products and services. Using a simple, word cloud visualization, Figure 12 shows key words presented such that the font size increases as the key word count increases in customer responses. Management might be happy to see that some of the most commonly used words are "happy" and "satisfy". The savvy analyst, however, can choose to focus on the root cause of the word "unhappy" appearing with similar frequency. Drilling deeper into the data set to find the root cause and implementing a solution to remedy and prevent future "unhappy" customers.

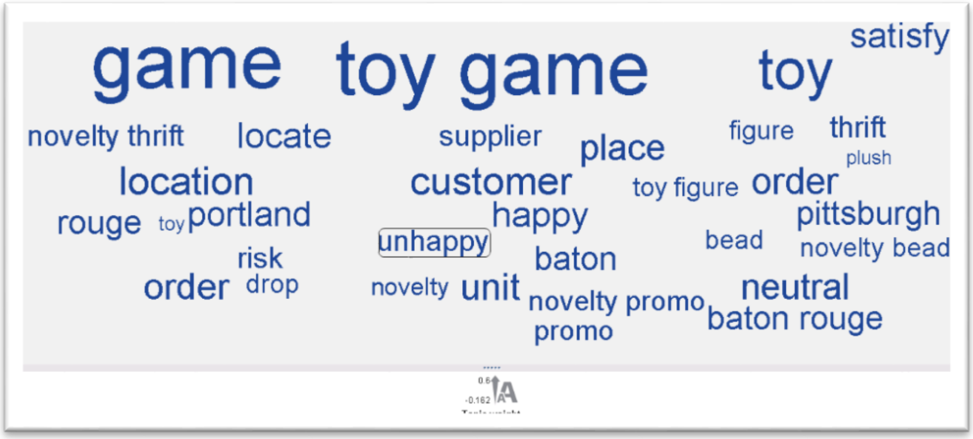


FIGURE 12. VISUALIZING AND IDENTIFYING AREAS TO IMPROVE CUSTOMER SATISFACTION

KNOW THE AUDIENCE AND GET TO THE POINT

While performing data analysis and analytics is critical, it is also important to effectively organize and present the results along with the recommendation. The key to a successful presentation is to know your audience and get to the point. Consider the audience background. It might not be appropriate to use technical terms to the business management. Always present the results and conclusions prior to showing the dashboards, reports, and outputs of the analysis. The message needs to be clear and simple; it should not get lost in the details of the analysis. At the end of the day, the business management (prosecutor) needs a solid case, without a doubt, with supporting evidence and proof to make a decision (get a conviction).

CONCLUSION

A CSI is just a glorified analyst. While data analysis and analytics might not seem as sexy or exciting as solving crimes, solving business problems requires the same techniques as crime scene investigation. Both start with defining an objective (increase sales/ catch a criminal) and forming a hypothesis and/or questions (margins are low in a new facility due to employee underperformance/ an unknown male individual robbed the liquor store). In both worlds, this includes gathering, processing, and eliminating certain clues and evidence. Next step is to use sophisticated tools for visualizing and connecting each piece of the findings and performing further analysis based on the initial findings with an appropriate technique. Lastly, organize and present the result along with recommendation simply and clearly to the decision-maker. Throughout this investigating process, some fundamental knowledge of the situation is needed (business experience/ relationship between the suspect and victim). These steps combined together forms a roadmap of analytical/investigation cycle. If used properly, they can allow reconstruction of a situation, its causes and effects, to which the analyst was not a first-hand witness.

REFERENCES

1. Cody, Ron. 1999. *Cody’s Data Cleaning Techniques Using SAS® Software*. Cary, NC: SAS Institute Inc.
2. SAS Institute Inc. 2016. *SAS® Visual Analytics: Fast Track*. Cary, NC: SAS Institute Inc.
3. SAS Institute Inc. 2015 SAS Course Notes. “Forecasting Using SAS® Forecast Server Software.”

ACKNOWLEDGMENTS

A greatest appreciation to Jonathan Klein for editing and contributing ideas on the CSI analogy. Special thanks to the readers for providing feedback and suggestions: Radu Miclaus, Erin McGowan, Michele Heath, and Marc Verville.

RECOMMENDED READING

1. Ngo, Theresa. 2012. "The Steps to Follow in a Multiple Regression Analysis." *Proceedings of the SAS Global Forum 2012 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings12/333-2012.pdf.
2. Ngo, Theresa. 2013. "The Box-Jenkins Methodology for Time Series Models." *Proceedings of the SAS Global Forum 2013 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings13/454-2013.pdf.
3. Ngo, Theresa. 2016. "Generalized Linear Models for Non-Normal Data." *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings16/8380-2016.pdf.
4. SAS Institute Inc. SAS Course Notes. “Categorical Data Analysis Using Logistic Regression.”
5. Klein, John P., and Melvin L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. New York, NY: Springer, LLC.
6. Zaratsian, Dan, Mary Osborne, and Justin Plumley. 2013. "Uncovering Patterns in Textual Data with SAS® Visual Analytics and SAS® Text Analytics." *Proceedings of the SAS Global Forum 2013 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings13/403-2013.pdf

CONTACT INFORMATION

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

Theresa Ngo
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
theresa.ngo@sas.com