Portfolio Project – Module #8 - Option #1: Capstone Project – Final Report and Slide Presentation: U.S. Organization

MIS581 Capstone – Business Intelligence and Data Analytics

CSU Global

Dr. Jamia Mills

11/3/2019

Portfolio Project – Module #8 - Option #1: Capstone Project – Final Report and Slide Presentation: U.S. Organization

**The Organization**

Now, before I had even begun working on this assignment in earnest, I already knew exactly who I wanted to write about – Amazon. It’s difficult not to be impressed by Amazon, isn’t it? I think Figure 1 demonstrates what we all know to be true – what began as a humble online bookstore has evolved into a massive, global phenomenon that basically sells whatever it wants.

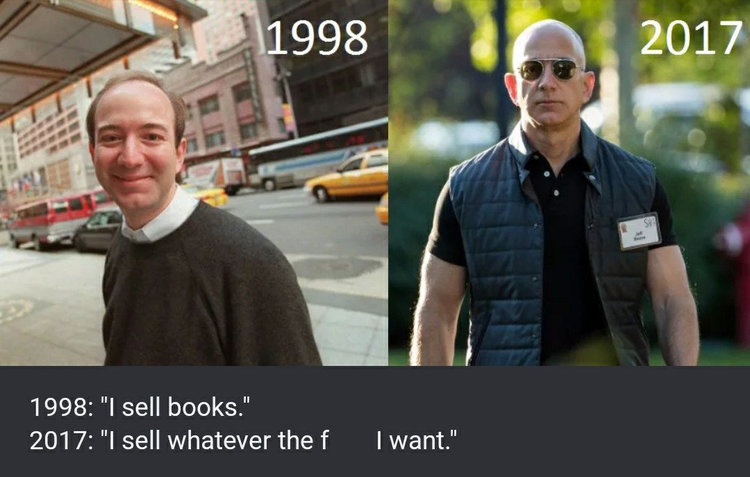


Figure 1. Amazon CEO Jeff Bezos then, vs now. Retrieved from https://amp.businessinsider.com/images/596cd260a47cb56b008b4a01-750-477.jpg. Adapted by C. Boyle, 2019.

The story is quite simple, really. Jeff Bezos, founder of Amazon, was unhappy at his Wall Street job – he regretted not getting involved in the Internet boom and decided to rectify the problem. In 1994, Bezos quit his job, moved to Seattle, and started on a business plan – soon thereafter, working out of the garage Bezos’ home, Amazon was born. As most of us know, Amazon began as an online bookstore, but in the early 2000’s, they decided to become much more than just a bookstore – in June of 2000, Amazon unveiled their “A-Z” logo, indicating their new commitment to providing a comprehensive selection of all manner of product. Ever since then, Amazon has branched off in several directions, and now has a variety of different revenue streams (Desjardins, 2017):

* Amazon Web Services (AWS)
* Retail Sales
* Third-Party Resellers
* Amazon Prime (and other retail subscription services)
* Whole Foods

We’ve got Amazon Prime subscribers and Prime Video (Amazon’s streaming service) subscribers. Amazon sells just about every physical product you can think of, including food, electronics, books, clothing, tools, and so much more. They offer cloud platforms, backup services, migration services, and other technology-related products. Plus, Amazon manufactures and sells some of their own products, and, while they certainly sell those products with their own platforms, other people sell those products, as well. Profitability between the different revenue streams varies wildly. For instance, in 2018, AWS and retail sales accounted for almost the same amount of operating income (roughly $7B USD); however, AWS only had to do $25B in sales to deliver as much operating income as the $140B in sales on the retail end of the business (Condon, 2019). The main takeaway here is that Amazon is making money hand over fist in myriad different markets.

Amazon’s headquarters is in Seattle, Washington – to the best of my knowledge (and the best of my research abilities), Amazon does not have any consumer retail locations of any type, unless we’re counting things like Whole Foods locations, fulfillment centers, etc. Amazon’s growth has been nothing short of staggering. In 2007, Amazon had roughly 10,000-20,000 employees, worldwide, and by the end of 2013, that number had just barely managed to blossom above 100,000; however, by the end of Q3 2018, Amazon had 613,000 employees spread across the globe (Richter, 2018). Revenue growth has been equally impressive – in 2008, Amazon did $19.1B in revenue, but by 2018, that number had grown to $232.9B (Macrotrends, 2019)!

So, why did I choose Amazon? Well, if I’m being honest, my approach to this assignment was to choose as broad a company as possible, with as broad a data set as possible, so that I have plenty of options with my final analysis. Let’s face it – Amazon sells just about everything. They’ve got their hands in so many different pies that it would be harder to find ways that data *isn’t* helpful than to find ways that it *is* – does that make sense? Plus, everyone is familiar with Amazon and, in my personal experience, all of Amazon’s customers love Amazon. I personally use Amazon Prime all the time, and I stream content with Prime Video on an almost-daily basis. I like Amazon, and it’s easier to work with things that you enjoy, right?

**Data, Tools, and Techniques**

The dataset that I chose for my analysis is called *Global Superstore 2018* and it’s an XLSX Excel file (Brennan, 2019) – no metadata was included in the XLSX file. I discovered it on a community forum for Tableau, a data visualization application, and from what I can tell, the dataset is included for practice purposes with some versions of Tableau. The dataset is 24x51290, so it features 24 variables (columns) with 51,290 observations (rows). For all intents and purposes, it’s a relatively simple, clean, and generic sales-related dataset. The dataset includes regional information, product information, sales information, and shipping information, primarily. The *Postal\_Code* variable doesn’t contain any values for countries outside of the United States, it would seem – every non-U.S. row contains a period, with no other values. Other than that, every variable appears to be correctly formatted, and with no missing values. Figure 2 is my attempt at a “data dictionary”, of sorts – a visual representation of the characteristics of the data in the dataset. In the dictionary, I included 5 different fields (NNLM, 2019):

* **Attribute Name** – The name of the variable.
* **Description** – A best-guess description of the variable, after looking at the data.
* **Field Length** – The maximum number of characters possible for a variable field.
* **Type** – The type of character permitted in the field.
* **Type 2** – My attempt to classify the type of variable based on what the data is intended to represent.



Figure 2. Data dictionary for Capstone Project dataset. Created by C. Boyle, 2019.

As you can see, the dataset includes a healthy mix of numeric variables and character-based variables. The *Postal\_Code* variable includes all “N/A” entries because, as I mentioned, there are only values included for entries that originate in the U.S. – since that equates to a huge chunk of missing dataset, I figure that I won’t get much use from the variable, considering all the missing values.

So, why did I choose this dataset? First, it’s a very generic data set that can be applied to a very broad business. Furthermore, the data is very granular – my initial surface-level analysis indicates that each row is a single line item from a total purchase, so I can really “get in there”, so to speak. Plus, it’s a huge data set, especially compared to most of the data sets that have been provided to me in various classes – I didn’t want to choose something with a limited number of observations on only a few variables. Obviously, I would have preferred to use a data set directly from Amazon, but I wasn’t able to find one; apparently, Amazon does not release their sales data to the public.

Next, let’s talk about some of the tools and techniques that will be utilized for this project. Without a doubt, I’m going to be using the following tools and/or techniques:

* SAS
* Tableau
* Descriptive Statistics
* Correlation Analysis
* Regression Analysis

Why these tools and techniques? Well, first, when it comes to tools, SAS and Tableau are a winning combination. In one hand, you have any extremely powerful statistical analysis platform in SAS, and in the other hand, a versatile and easy-to-use cloud-based data visualization software suite in Tableau. Furthermore, I’m most familiar (and most experienced) with predictive analytics – finding correlations, performing regression analysis, and creating predictive expressions has essentially been the foundation of my analytical education. I considered using either R or Python as my core statistical tools, but I ultimately settled on SAS because SAS is just so much easier to use. Over the course of my program, I’ve learned enough to know that Python and R are extremely powerful and customizable tools, and while I’ve certainly used both and could, no doubt, power my way through a project if forced, I simply enjoy SAS much more. I find it far less frustrating, and the learning curve much more forgiving. It’s worth noting that I could certainly create data visualizations in SAS, but I decided to incorporate Tableau into my project because it’s much more suited to the task. Make no mistake – just about anything you can create in Tableau, visually, can also be created in SAS; however, in most cases, I find that I can make the same visualization in Tableau as in SAS, but in a fraction of the time.

**Ethics, Privacy, and Security**

In the modern era of *Big Data* and data breaches, we commonly encounter terms like ethics, privacy, and security. While they’re all similar, they’re also distinctly different. *Ethics* is a branch of philosophy that involves the moral principles that govern human behavior – in other words, ethics is concerned with “right” and “wrong”, and how we determine what is “right” and “wrong” behavior. According to our textbook, four common elements comprise an ethical framework for *Big Data* (Davis & Patterson, 2012):

* Identity
* Privacy
* Ownership
* Reputation

This framework raises serious questions about ethics in the context of data management. Who controls access to the data? Who owns the data? How do we know if the data is trustworthy? What kind of uses should be permitted of the data? It’s important to remember that there are no iron-clad answers to these questions, which is what makes ethical dilemmas so ambiguous. After all, *Big Data* is ethically neutral; that is, technology and data have no preconceived notions of what is, or is not, ethical, because our ethics are derived from our values, and, obviously, *Big Data* doesn’t have values (Davis & Patterson, 2012). It’s only once you introduce the human element, and we impose our will upon the data, that ethics becomes a paramount consideration.

*Privacy*, on the other hand, is an entirely different animal. Privacy refers to our ability (or, perhaps more accurately, our desire) to maintain agency over our identity, and to control who has access to our personal information. According to the Generally Accepted Privacy Principles (GAPP), there are ten key components to *Big Data* privacy (Chapple, 2018):

* Management
* Notice
* Choice and Consent
* Collection
* Use, Retention, Disposal
* Access
* Disclosure to Third Parties
* Security
* Quality
* Monitoring and Enforcement

For example, the *Notice* principle suggests that subjects should be notified if their information is being collected and/or used, while the *Choice and Consent* principle suggests that data collectors should get consent from data subjects for the collection, storage, use, and sharing of personally identifiable information.

Last, but certainly not least, we have security. *Security*, in my opinion, is probably the most straight-forward and least ambiguous of the three; that is, security is simply the process of protecting data from unauthorized access and/or corruption. If privacy is about the right, or the desire, to decide how our personal information is used, security is about physically supporting that effort. There are many common data security measures, such as:

* Backup systems
* Encryption
* Password Protection/User Authentication
* Data Masking
* Data Erasure

Let’s address the elephant in the room. The assignment asks that I talk about the plans, tools, or techniques to address the ethical, privacy-related, and security-related challenges as they relate to analyzing data in the dataset that I have chosen for this final project. If we’re being literal, then I’m not really going to make any special considerations regarding ethics, privacy, or security. It’s like asking about my ethical considerations when dealing with the ever-popular Microsoft Northwind training dataset; in other words, I’m pretty sure that the dataset I chose is a fictional training dataset with no actual real-world relevance.

Still, in the spirit of the assignment, let’s pretend that I’m using a dataset that I got from a colleague at work – or, perhaps, from Amazon themselves. Let’s pretend that all the personally identifiable information contained in the dataset is about real people, and that the information is sensitive in nature. From an ethical and privacy standpoint, I would make sure to get all the necessary details from whoever provided me with the data, such as:

* In what context was the data originally collected?
* In what ways have users consented that the data be used, and/or shared?
* Is it ethical for me to have and/or use this data in the first place?

I might use SAS, Python, Tableau, or some other tool to create a custom dataset that omits all the names, or other identifiable information. Moreover, I might encrypt the data files. For instance, SAS offers advanced encryption standard (AES) on SAS data files. I would certainly make sure that no unauthorized users could gain access to the data, either through my devices (phone or computer), or through my hosted services (cloud-based analytics applications like SAS of Tableau). I would also make sure that I was only using cloud services that held data at rest in an encrypted form. I would never willingly share the data with anyone else. My mother taught me that it’s impolite to loan something that doesn’t belong to you. Again, at the end of the day, I don’t really have any of these concerns with my data, though, because I’m virtually positive that the data isn’t real – or, at least, the people the data is about are not real.

**Data Visualization and Descriptive Analytics**

Before I do any hypothesis testing, correlation analysis, or predictive analytics, I want to *look* at the data – and I’m not talking about poring over a bunch of numbers. Sometimes, it can be difficult to make sense of what we’re seeing when we’re trying to take into consideration hundreds (if not thousands, if not tens of thousands) of numeric variables. Consider my chosen dataset, for instance, which contains over 50,000 rows of data – it would take an eternity to draw meaningful conclusions about even the simplest of questions just by looking at the raw data. Imagine the following business questions:

* Which customer segment generates the greatest amount of sales?
* Which specific products are most popular?
* What do sales look like over the past few years for which data exists?

Using just about any data visualization software, we can have answers to these questions with no more than just a few drags and drops! Figures 3, 4, 5, and 6 demonstrate some examples of data visualizations that have been created in Tableau – Figure 3 shows *Sales* by *Segment,* Figure 4 shows *Sales* by *Product\_Name*, Figure 5 shows *Sales* by *Order\_Date* (specifically, by year), and Figure 6 shows *Sales* by *Order\_Date*, broken down by quarter. At the end of the day, humans are visual creatures, and it is much easier for our brains to make sense of large amounts of complex information when we condense and summarize it in a visual format.

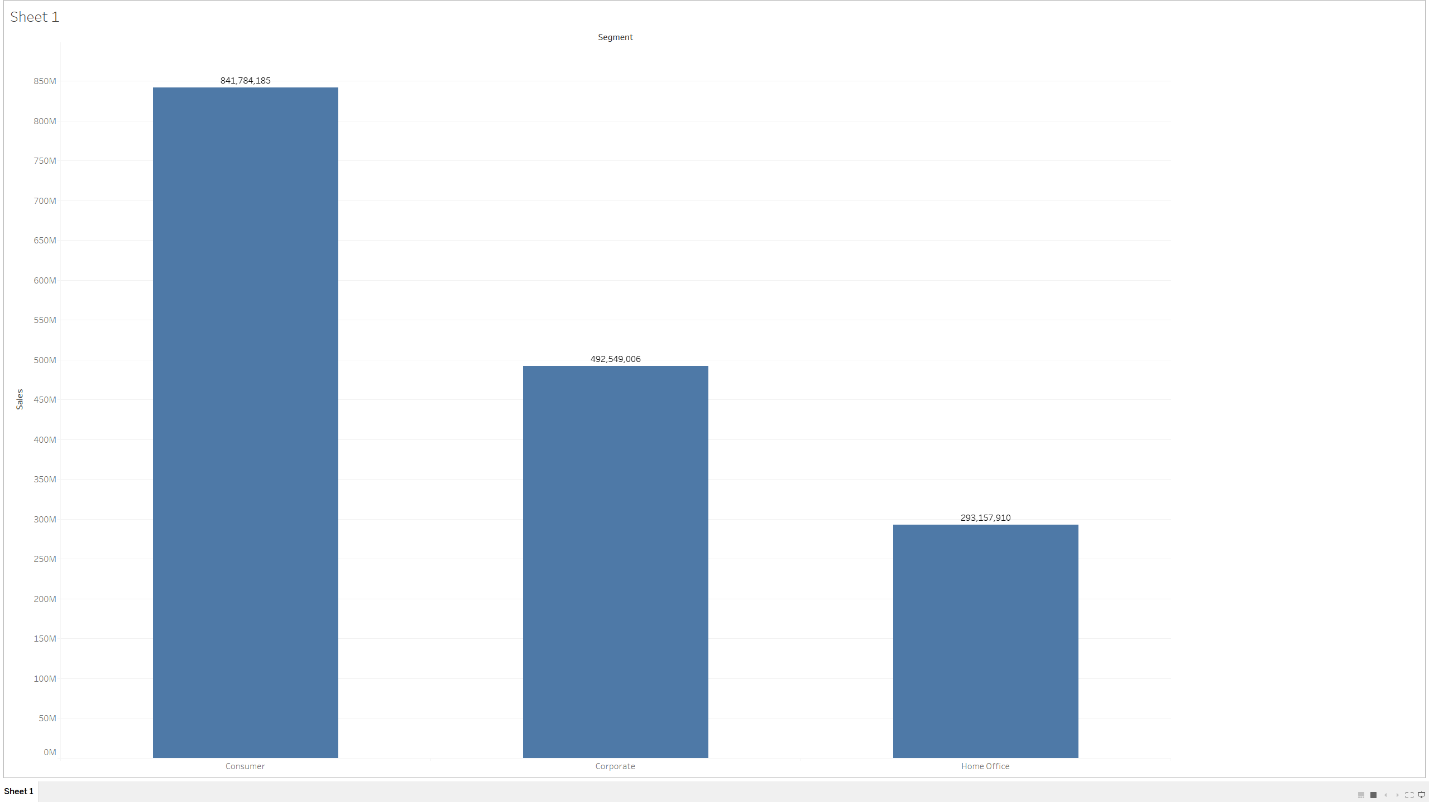


Figure 3. Visualization in Tableau. Screenshot by C. Boyle, 2019.

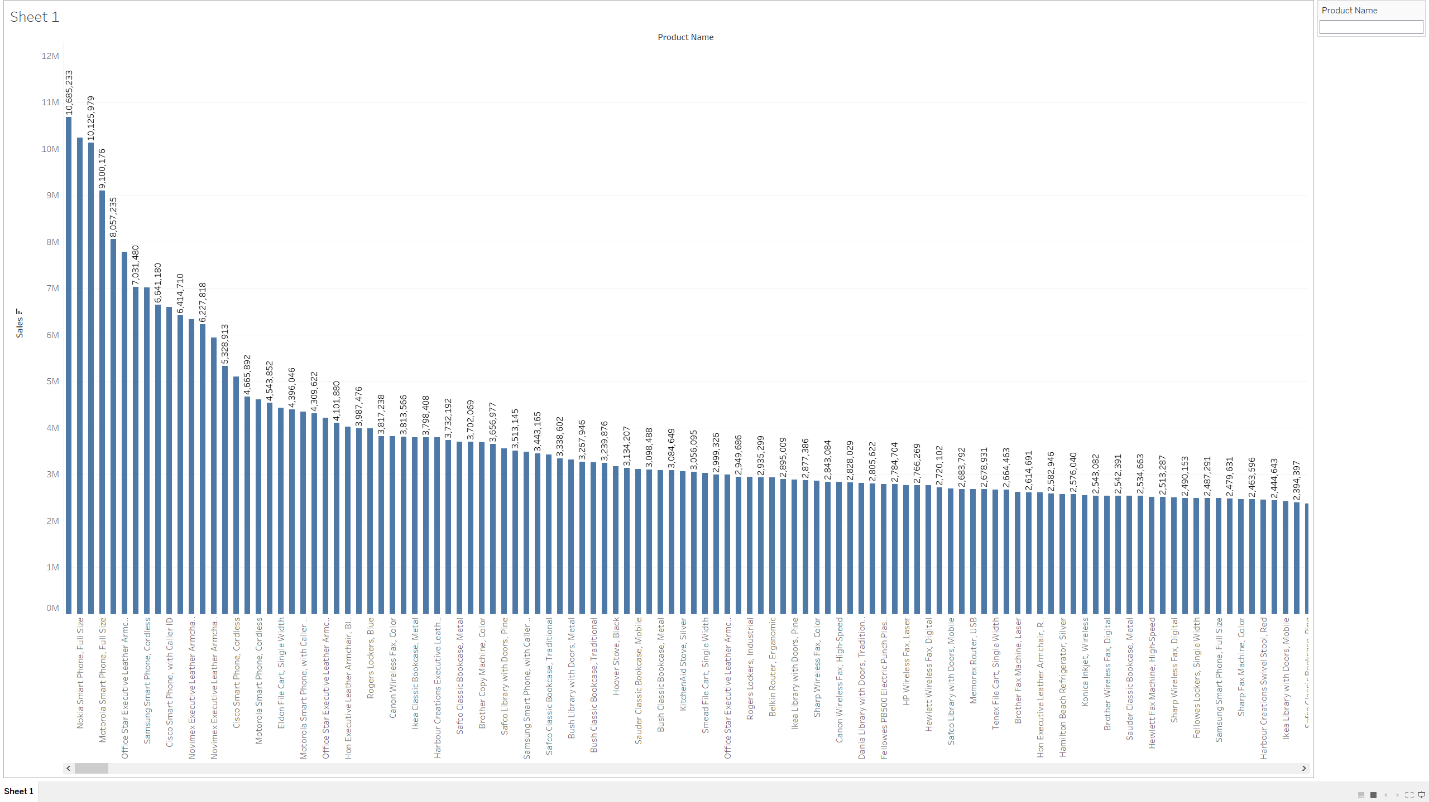


Figure 4. Visualization in Tableau. Screenshot by C. Boyle, 2019.

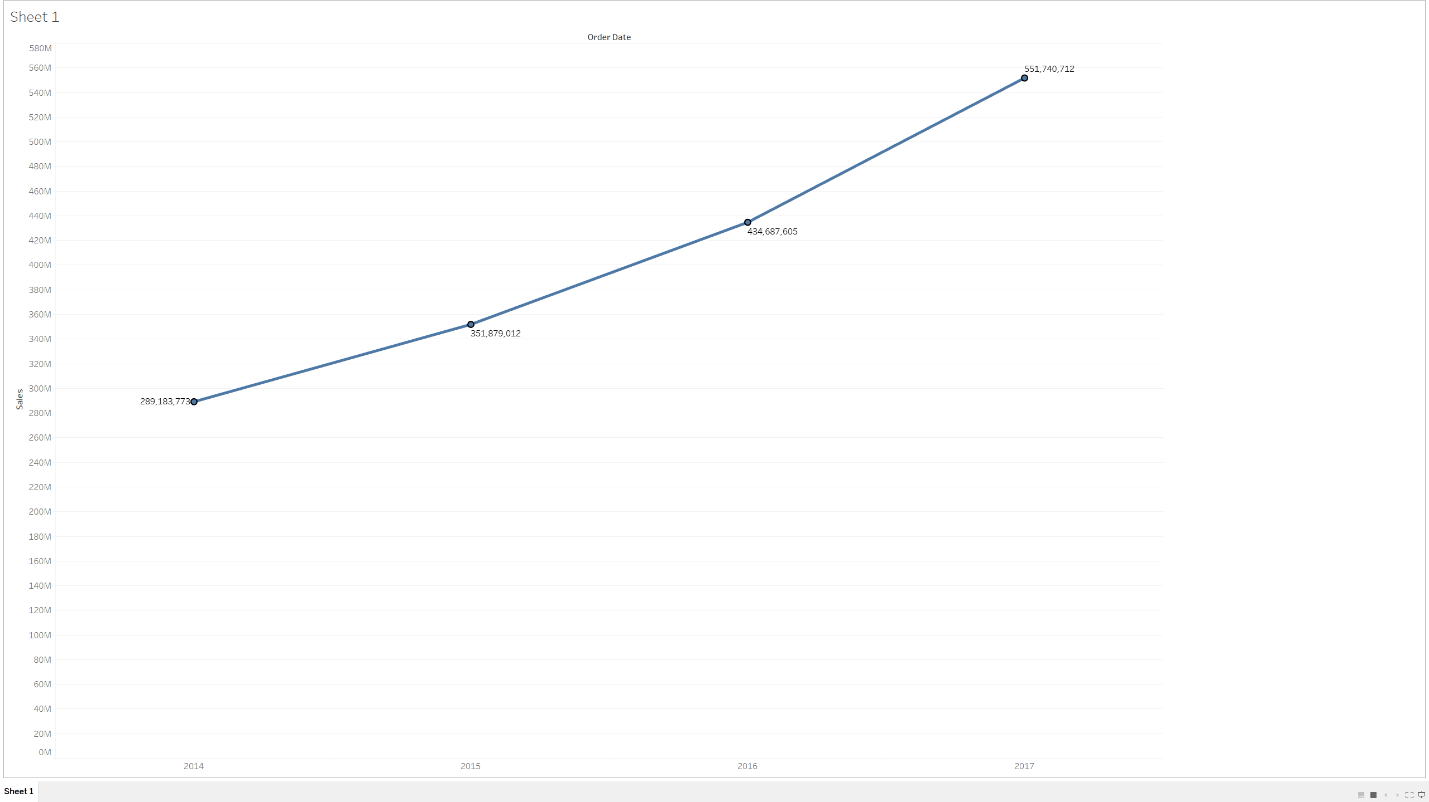


Figure 5. Visualization in Tableau. Screenshot by C. Boyle, 2019.

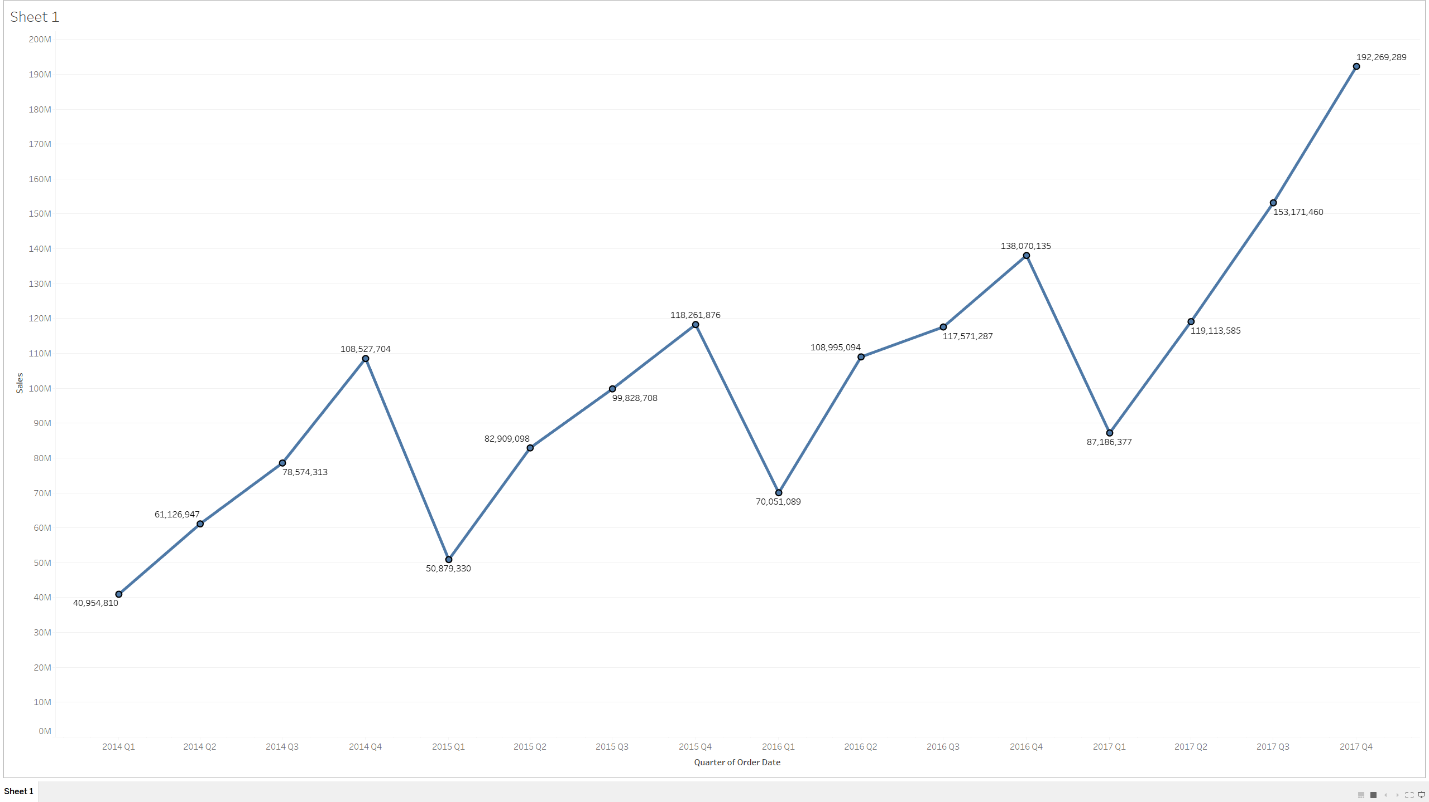


Figure 6. Visualization in Tableau. Screenshot by C. Boyle, 2019.

What kind of information is contained within these visualizations? Do they provide answers to our mock business questions? Well, according to Figure 3, the *Consumer* segment is responsible for the greatest amount of total sales; as a matter of fact, *Consumer* products are responsible for more sales than both *Corporate* and *Home Office* products, combined. Which products are most popular? According to Figure 4, the five products most responsible for driving revenue are all phones; moreover, we find that the top ten or twenty most popular products are almost exclusively phones and armchairs. Finally, what can our visualizations tell us about historical (and current) sales performance? According to Figure 5, revenue has been consistently growing, year after year; however, what is perhaps more interesting is the quarterly trend that we see in revenue, shown in Figure 6. Every year for which data exists, each quarter within the same year generates more revenue than the previous quarter. Furthermore, quarters in successive years always outperform that same quarter from the previous year.

**Hypothesis Testing with Correlation Analysis**

What is a hypothesis, exactly? Many of us have probably heard the colloquial definition – that a hypothesis is an educated guess. Depending on who you ask, you’ll receive a multitude of answers, but they all look and sound very much alike. According to Kerlinger (1957), a hypothesis is a statement of conjecture about the relationship between two or more variables – a relational proposition. According to O’Leary (2017), a hypothesis is a hunch or guess about the nature of relationships between variables, declared as a testable statement. In other words, a hypothesis is an interim explanation or proposed relationship, derived in part from existing knowledge, that has clear implications for testing and measurement. A good hypothesis should incorporate the following key characteristics (Mourougan & Sethuraman, 2017):

* Simple and specific
* Explanatory
* State the relationship
* Testable
* Consistent with existing literature

Hypotheses are important for several reasons. For example, hypotheses are testable, relational statements that provide direction to a research project (Mourougan & Sethuraman, 2017). Moreover, hypotheses provide a framework for the reporting of conclusions (Mourougan & Sethuraman, 2017).

One method of testing a hypothesis is to use correlation analysis, which is a type of statistical evaluation that examines the strength of a relationship between two continuous, numeric variables. Correlation analysis is particularly useful when we want to establish the existence of possible connections between variables in a dataset. Depending on the type of correlation analysis performed, some sort of correlation coefficient is the ultimate output – a coefficient of 0 represents no correlation, while coefficients of 1 and -1 represent perfect direct and inverse correlations, respectively. Furthermore, most correlation analyses provide statistical significance levels, which, in large part, enables us to determine whether we should reject our null hypothesis or not.

For the purposes of my analysis, I’m going to pose a very broad, simple hypothesis:

* H0 = There are no statistically significant correlations between any of the numeric, continuous variables in the selected dataset.
* H1 = There are statistically significant correlations between numeric, continuous variables in the selected dataset.

In order to test this hypothesis, I’m going to perform a correlation analysis using the CORR procedure in SAS. Figure 7 demonstrates the code used to perform the analysis, while Figure 8 demonstrates the correlation matrix that was created by the procedure. What is a correlation matrix, exactly? Well, a correlation matrix is a matrix of all the different possible combinations of one-to-one relationships among the variables in the dataset, and lists correlation coefficients and statistical significance levels of each relationship. By looking at the matrix, we’re able to get an idea of which variables have measurable relationships, if any – if the R value is between 0 and 1, there is (technically) a correlational relationship, and if the statistical significance is less than .0001, then the existence of that relationship is statistically significant, and not likely due to sampling error.

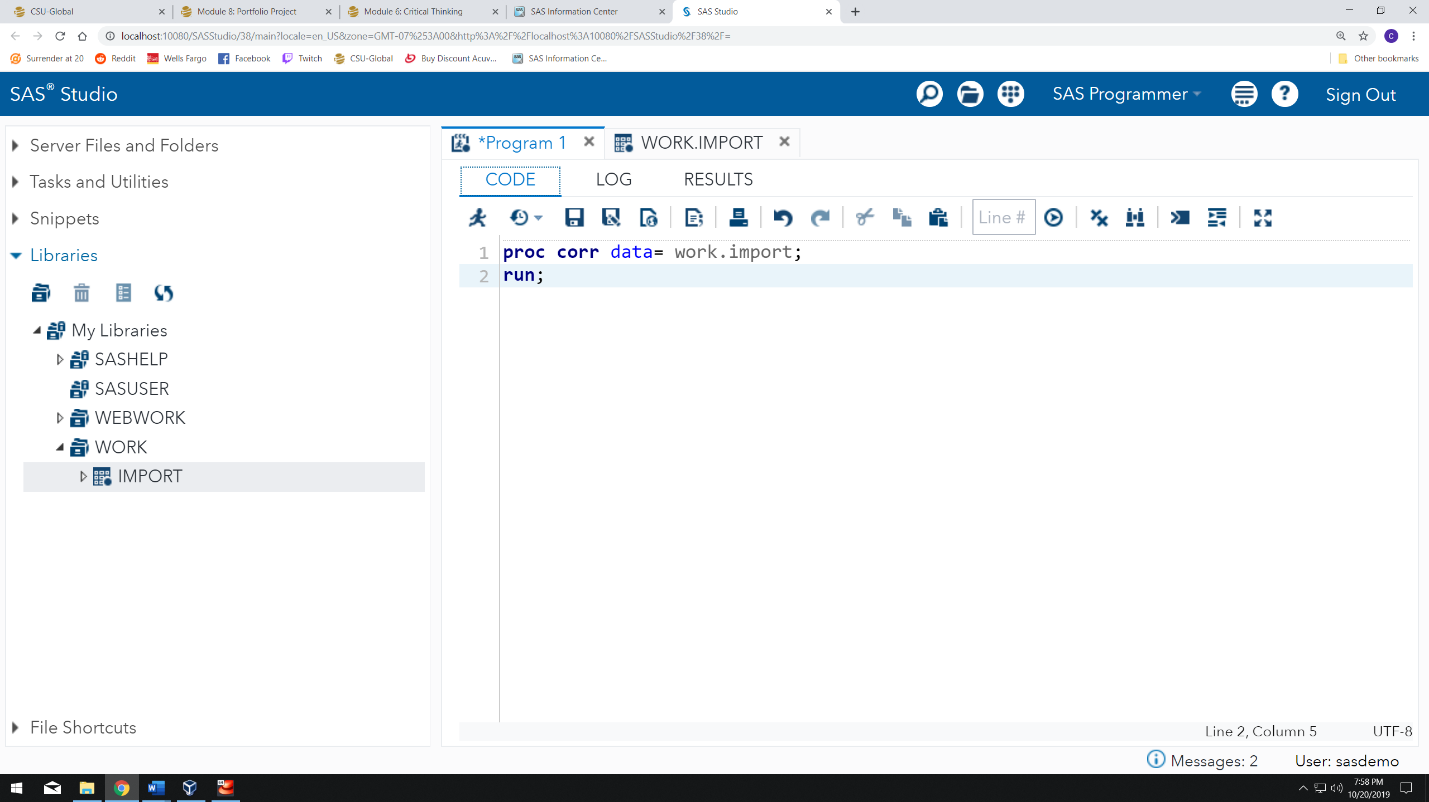


Figure 7. Correlation analysis code in SAS. Screenshot by C. Boyle, 2019.

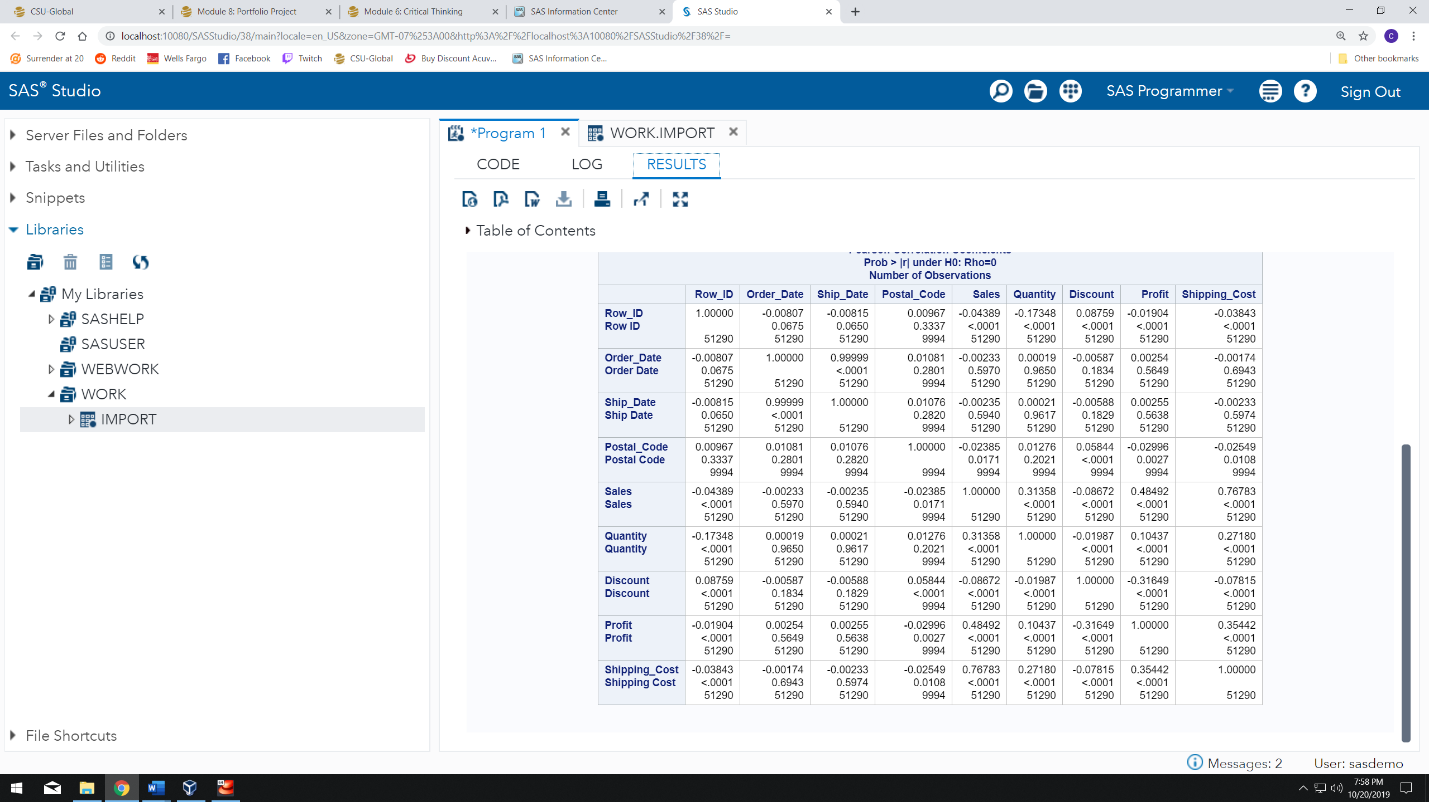


Figure 8. Correlation matrix in SAS. Screenshot by C. Boyle, 2019.

Now, according to the resulting correlation matrix, there are several statistically significant relationships between numeric variables in the dataset, such as:

* Ship\_Date:Order\_Date
* Discount:Quantity
* Shipping\_Cost:Discount
* Postal\_Code:Discount

Altogether, the correlation matrix tells us that there are 17 statistically significant relationships between the variables in the dataset; however, not all those relationships have a coefficient of correlation of notable strength. For example, there is a statistically significant relationship between *Discount* and *Quantity*; however, the correlation coefficient is extremely low, bordering on nonexistent (r = -0.019). As a matter of fact, according to the coefficient of determination (which is calculated by squaring the coefficient of correlation), as predictive variables, *Discount* and *Quantity* can only account for .036% of the variance in the other. Still, we have a clear answer to our hypothesis test – there are correlations between variables, and those correlations are statistically significant, so we can reject our null hypothesis and accept our research/alternate hypothesis.

Let’s narrow down the field a bit, shall we? The five relationships with the highest coefficients of correlation are:

* Sales:Shipping\_Cost, r = .768, p = <.0001
* Sales:Profit, r = .485, p = <.0001
* Shipping\_Cost:Profit, r = .355, p = <.0001
* Discount:Profit, r = -.316, p = <.0001
* Sales:Quantity, r = .314, p = <.0001

What do these numbers mean? Well, take *Sales* and *Shipping\_Cost*, for instance. According to the correlation matrix, there is a strong, direct correlational relationship between *Sales* and *Shipping\_Cost*; that is, as values of *Sales* increase, values of *Shipping\_Cost* tend to increase, as well, and to a substantial degree. In fact, for an organization such as Amazon, the ability to make predictions around shipping costs seems like it could be quite beneficial. We also learn some interesting things from the other correlations, such as:

* We know that as sales go up, profitability also tends to increase. Perhaps the margins on higher-priced items are better?
* We know that as shipping costs increase, profit also tends to increase. A lot of collinearity, perhaps, between *Sales*, *Profit*, and *Shipping\_Cost*?
* Unsurprisingly, we discover as people get more discounts, profitability tends to decrease.
* Increases or decreases in *Sales* and/or *Quantity* result in commensurate increases/decreases in the other.

**Predictive Analytics**

Now, let’s go back to our strongest correlation – *Sales* and *Shipping\_Cost*. How could we go about using these two variables to create a predictive model that allows us to estimate unknown values of one (*Shipping\_Cost*) with known values of the other (*Sales)*? For instance, how might we create an expression that takes the dollar value of a cart of items, and predicts the cost to ship that cart of items? First, I’m going to perform a regression analysis in SAS using the REG procedure. Regression analysis provides the only two outputs that we need in order to create our predictive equation – the slope, and the intercept. Figure 9 demonstrates the regression procedure code in SAS, while Figure 10 demonstrates the results of said regression procedure.

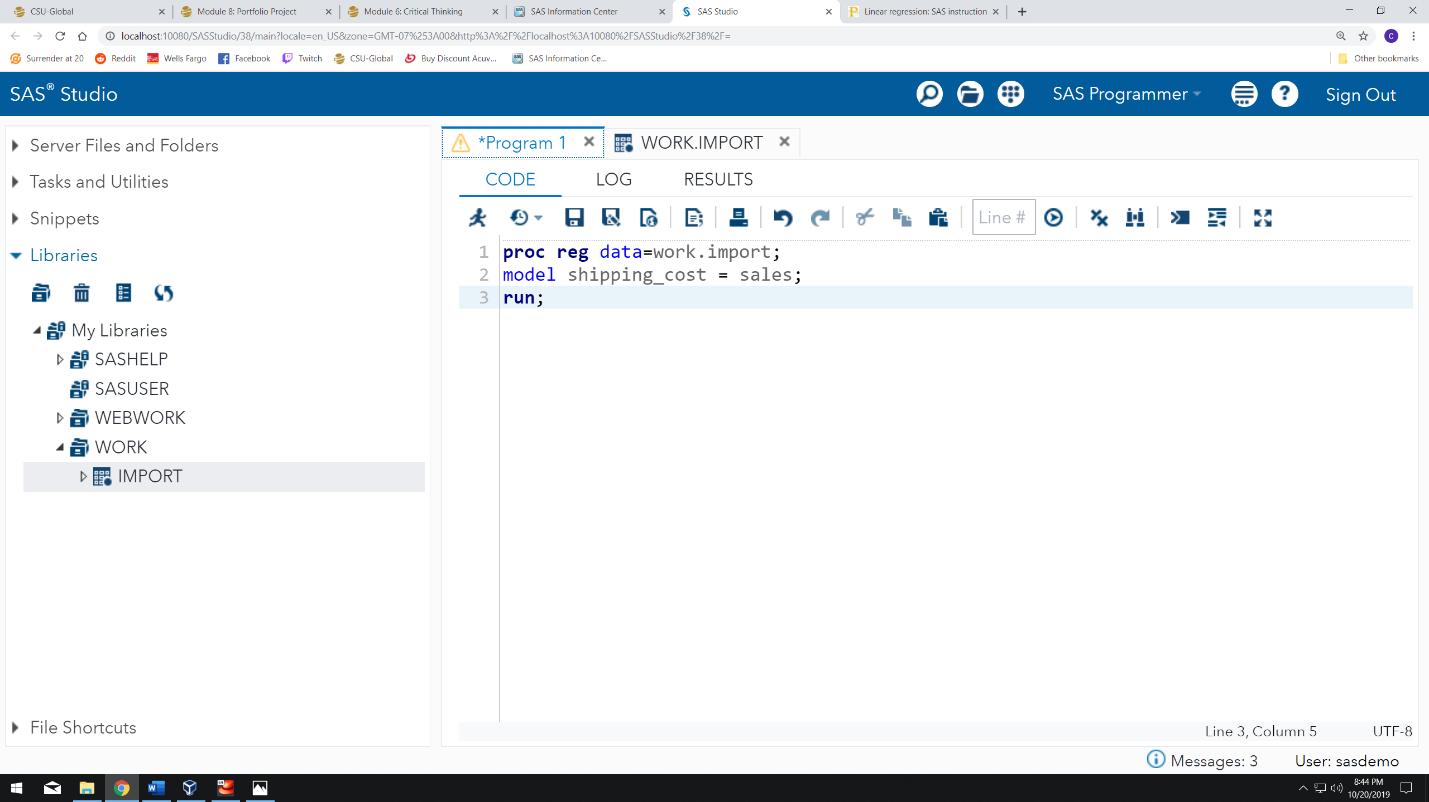


Figure 9. Regression procedure in SAS. Screenshot by C. Boyle, 2019.

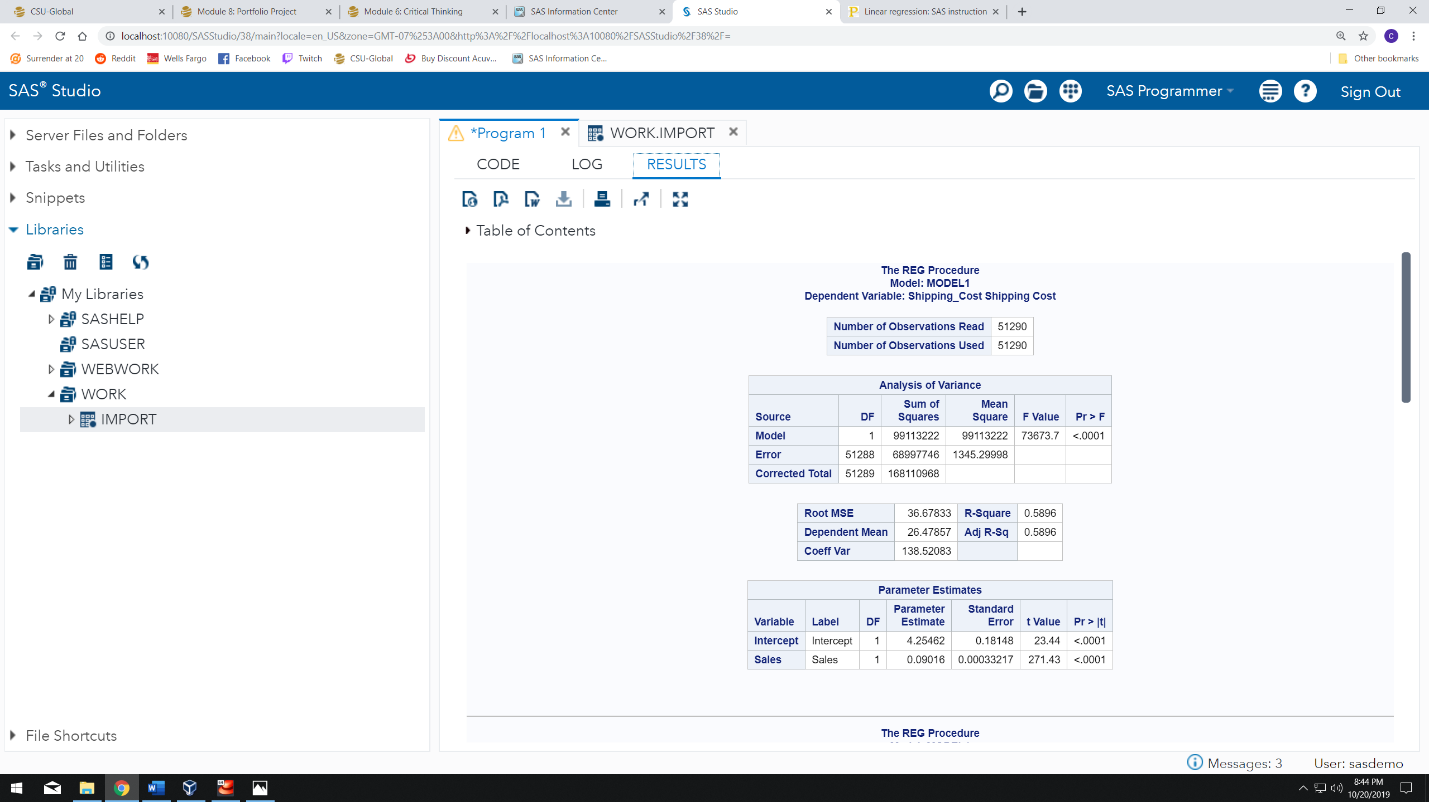


Figure 10. Regression procedure results. Screenshot by C. Boyle, 2019.

Once I have those two magical numbers, I can plug them into the following formula:

**Y = A + BX**

In this formula, Y is our dependent variable (*Shipping\_Cost)*, which is the variable we want to predict, while X is our independent variable, the value of which is known to us – A represents the y-intercept of the regression line, while B represents the slope of that same line (both the slope and the intercept are provided in the “Parameter Estimates” section of the procedure results. If we plug in all the corresponding values, we end up with the following formula:

**Shipping\_Cost = 4.255 + .09 (Sales)**

Let’s assume that a cart has $100 dollars-worth of items in it – how much is shipping going to cost? According to our formula:

**Shipping\_Cost = 4.255 + .09 (100)**

**Shipping\_Cost = 4.255 + 9**

**Shipping\_Cost = ~$13.25**

Apparently, according to the data, it would cost roughly $13.25, on average, to ship $100 worth of items.

References

Brennan, C. (2019, January 30). Tableau Community Forums. Retrieved from https://community.tableau.com/message/865707#865707

Chapple, M. (2018). Data Privacy. Retrieved October 4, 2019, from https://www.linkedin.com/learning/sscp-cert-prep-2-security-operations-and-administration/data-privacy?u=2245842&auth=true.

Condon, S. (2019, January 31). In 2018, AWS delivered most of Amazon's operating income. Retrieved from https://www.zdnet.com/article/in-2018-aws-delivered-most-of-amazons-operating-income/

Davis, K., & Patterson, D. (2012). *Ethics of big data*. Farnham: OReilly.

Desjardins, J. (2017, December 19). Breaking Down How Amazon Makes Money. Retrieved from https://www.visualcapitalist.com/breaking-amazon-makes-money/

Elliott, A. C., & Woodward, W. A. (2016). *Sas essentials: mastering Sas for data analytics*. Hoboken, NJ: John Wiley and Sons, Inc.

Field, A. (2013). *Discovering statistics using Ibm Spss statistics: (and sex and drugs and rock n roll) / Andy Field*. London: Sage.

Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science*, *85*, 78–85. doi: 10.1016/j.procs.2016.05.180

Kerlinger, F. N. (1957). *Research design and analysis of variance; a manual for students of education*. New York: School of Education, New York University.

Macrotrends. (2019). Amazon Revenue 2006-2019: AMZN. Retrieved from https://www.macrotrends.net/stocks/charts/AMZN/amazon/revenue

Meirelles, I. (2013). *Design for information: an introduction to the histories, theories, and best practices behind effective information visualizations*. Beverly: Rockport.

Mourougan, S., & Sethuraman, K. (2017, May). Hypothesis Development and Testing. Retrieved September 29, 2019, from https://pdfs.semanticscholar.org/9bd0/d555e809ac52142271fd04489e7d5e97e2ec.pdf.

NNLM. (2019). Data Dictionary. Retrieved from https://nnlm.gov/data/thesaurus/data-dictionary

O’Leary, Z. (2017). *The essential guide to doing your research project*. London: SAGE.

Richter, F. (2018, November 1). Infographic: Amazon's Workforce Is More Than Half a Million Strong. Retrieved from https://www.statista.com/chart/7581/amazons-global-workforce/