Data Analysis Artifact

Bubba Gump Shrimp Company Analysis

Zane Russell Brown

Southern New Hampshire University

Introduction: Business Problem

The Bubba Gump Shrimp company is a well-known retailer in restaurants and other retail channels. Bubba Gump quickly gained popularity thanks to a global blockbuster movie. The company used this popularity and opened several other restaurants, sells its own merchandise online and wholesales its branded merchandise to other outlets. Unfortunately, the rapid growth for the first few years started to decline. We have been tasked to analyze Bubba Gump's large collection of data and determine why the company's online sales leveled off and started to decline. By analyzing this data, we hope to find the reason for such decline, track patterns, determine associations and natural clusters and help the company solve this decline in online sales.

Introduction: Analytic Method

To help find the reason for the decline in online sales, the Bubba Gump company has recently integrated their data into a data warehouse. This data includes point-of-sale data, customer database, web store sales transaction data, and customer and sales data from third party retailers. Having all this data integrated into a data warehouse is extremely useful because data warehouses are collections of gathered data. Data warehouses also contain historical data and external data in order to help provide better forms of analysis. (Ahlemeyer-Stubbe, A., & Coleman, S. 2014). This consolidated data will help us analyze and find patterns and associations in sales transactions to customers in a range of channels. In addition to the information found in the data warehouse, a sample of 500 customers was selected and were given surveys. These surveys were then recorded.

Thanks to the amount of data that has already been collected and integrated into the data warehouse, we will be able to analyze the data and look at it in different ways. Analyzing this data can be done in a variety of ways. We are trying to determine why the online sales have started to decline. Therefore, it is important to track patterns, determine associations and outliers, understand if there are any clusters, and make predictions. These data mining techniques will help determine the problem at the Bubba Gump Shrimp company and provide insight and predictions to help solve the problem.

First, tracking patterns will be vital to solving the problem at Bubba Gump Shrimp. Tracking patterns and trends will help the company understand specific customer online sales transactions. Using these patterns, we can find out which products are being bought more often than others. (McDaniel, S. 2019). Using this data, the company can then determine whether to remove the items that are not being bought in order to save more money and then create products that customers seem to like more and thus create more sales in that area. Additionally, sequential patterns can also be analyzed from this data. Sequential patterns are useful in analyzing transactional data events. Stacey McDaniel (2019) states that sequential patterns can be used to determine a pattern or trend in customers when they are buying specific products. Meaning, some customers are likely to buy specific products after purchasing other products of the same genre. One example of these sequential patterns regarding the Bubba Gump Shrimp company is the possibility of a customer going to one of the Bubba Gump Shrimp restaurants, eating crab and/or shrimp and then purchasing a branded shrimp peeler or crab claw cracker from the online store or third-party retailer. We can see how powerful the sequential pattern data can be in finding patterns in certain events. We can then take these patterns of events and determine specific associations in these events. ‘Association’ is a specific data mining technique that

involves linking specific events found in the data and finding relationships between them using machine-learning. (Rouse, M. 2018). If these patterns and events are found in the data consistently, this can lead to significant associations and thus give us a better idea to what customers are thinking and what the next plan for the business should be. These patterns and associations can help determine what products are more popular than others or what products will be bought in combination with other products.

In addition to finding patterns and associations in the Bubba Gump Shrimp company's data, cluster analysis will be extremely beneficial to analyzing the data in the data warehouse as well as the sample taken from 500 customers. Cluster analysis is a data mining technique in which data is partitioned and grouped into categories in order to better analyze relationships, similarities and structures in the data. (Foley, B. 2018). The surveys from 500 sample customers provides a great form of clustering. In fact, the Bubba Gump Shrimp company has already clustered the data for us. The survey was conducted with customers who had made purchases from the Bubba Gump sales channels. These channels included restaurants, online shops, third party retailers, etc. These different sales channels are already one form of clustering that we can analyze using cluster analysis. Another form of clustering that can be made using the survey data is the customer satisfaction ratings. If Bubba Gump Shrimp asked customers to rate their satisfaction on a scale, we could cluster these satisfactions and compare them to product purchases or other patterns/trends that were previously discussed. These clusters could provide additional event patterns to help determine why online sales are declining.

Tracking patterns, determining associations, and clustering are all extremely useful techniques for analyzing the Bubba Gump Shrimp company's data. When analyzing all of this data, we may end up with some outliers. This is where outlier detection could be beneficial.

Outlier detection helps determine anomalies in the data. (McDaniel, S. 2019). Understanding these anomalies can help businesses create plans to prevent these events in the future if they are deemed to be a problem or capitalized on if they are deemed to be a benefit. Stacey McDaniel (2019) gives a good example of outlier detection. She states that if there is a spike in product purchases at a specific time of the day, companies could use this data to help create plans to boost their sales during these times. Using outlier detection techniques with along with external data, the Bubba Gump Shrimp company's data could also provide anomalies that the company could capitalize on. For example, we know that the blockbuster movie "Forest Gump" helped boost the Bubba Gump Shrimp company into what it is now. We also know that the block buster movie is aired on TV multiple times a year and is also streamed on specific platforms such as Netflix. Using external data of TV air times and specific release times on streaming services of this blockbuster movie and comparing this data to spikes in online product sales, we could find anomalies that point to specific times of the year that yield higher sales. These anomalies could help the company boost their sales if the outlier detection technique provides positive results.

Finally, after using all of these various analytical techniques along with the data warehouse, we can use this data to help create predictions. Predictions are created using patterns, associations, clusters, historical and current data, etc. These predictions can help companies create future plans about specific events or trends that are predicted to happen. (McDaniel, S. 2019). The consolidated data from the Bubba Gump Shrimp company provides all of the information we need in order to create clusters, find patterns, determine associations and possibly find outliers. This data can not only help find the reason for declining online sales, but it can also help predict future trends and events that could help the company stay ahead and create plans to keep their business flowing at a natural and economical pace.

Analysis Tools

The goal of data mining is to take a large amount of data and find meaningful patterns, then take these patterns and turn them into information for specific uses. In order to find these patterns, correlations and associations, data analysts use a variety of tools to create algorithms, statistical analyses, AI and database systems. (Uj, A. 2018). Data analysis involves specific tasks and questions and data analysts must use the correct tools to solve such problems. Data analysis tools can be thought of as a construction worker's toolbox. For specific tasks, a hammer might be used and for others a screwdriver might be used. It all depends on what needs to be done. This analogy can be used regarding the Bubba Gump Shrimp company. We are trying to determine why online sales have been decreasing. In order to do this, we must find patterns, clusters and associations in this data. *Rapid Miner,* and *Weka* would be great tools to perform the analysis on the Bubba Gump dataset.

Rapid Miner is an extremely useful tool when it comes to data analysis because it can perform a wide variety of functions. Rapid Miner's functions include data predictive analysis, visualization and pre-processing, filtering, and data cleansing. Additionally, this tool provides templates and repeatable workflows. (Shah, D. 2017). Having a tool that provides such a wide array of functions is a great asset when analyzing such a large data set like the Bubba Gump Shrimp company dataset. Rapid Miner is also extremely easy to use and implement. This is because this tool requires no coding software. However, if the need arises, languages such as Python can be integrated into the tool. (Shah, D. 2017). Each of these functions are and will be essential to the Bubba Gump dataset. Rapid Miner will allow us to pre-process the data set and prepare the data to be used in conjunction with Weka. This leads to the next useful function of

Rapid Miner. This data analysis tool also allows users to integrate with Weka software. Weka is an extremely useful tools to be used in conjunction with Rapid Miner. Weka is an open source machine learning software that helps with pre-processing, classification, regression, clustering, association, and visualization. Additionally, Weka is able to call algorithms from Java code. This is useful because Rapid Miner accepts and is written in Java code. (Weka, 2020).

Using both Rapid Miner and Weka, we will be able to pre-process the data and prepare it for proper analysis. These tools will allow us to cluster data and determine associations in order to find patterns and events in the online retailer data. From the Bubba Gump dataset, we can see that more customers visit and make purchases at in store locations and very little actually visit and buy products through the online store. We can use these tools to find significant patterns in these specific areas. In addition to finding significant patterns and events, these tools will also allow us to analyze the Bubba Gump dataset to get greater insight into perceptions and habits of customers, find the best product combinations, create effective product offers, and create predictions into how we can best improve the online stores sales and keep them up for the foreseeable future. (rapidminer, 2020).

Data Visualizations

Analyzation tools are vital to data mining and help us understand what patterns, events and associations exist in the dataset. However, sometimes understanding this data from the data itself can be quite difficult to grasp. This is where creating meaningful and understandable visualizations become the next vital step in the data mining process. Data visualizations help us see what the data looks like. This allows us to understand the data a lot better than if it was in its raw form. Rapid Miner and Weka tools provide functions for creating such data visualizations.

The Bubba Gump Shrimp company has provided us with a lot of data to work with. The company has recently integrated their data into a data warehouse. This data includes point-of-sale data, customer database, web store sales transaction data, and customer and sales data from third party retailers, historical data and external data. This data from the data warehouse will allow us to create comparisons, patterns, relationships, distributions, etc. All of these pieces of data can be represented in a variety of visuals. The data visualizations we will be using for the Bubba Gump Shrimp company are histograms, multi-set bar charts, choropleth maps, and pie charts.

Multi-set bar charts (Figure 1) are charts that contain two or more data sets that are compared side by side. These bars are measured by a numbered scale. A multi-set bar chart is used to compare a set of grouped variables or data that shares close relations. (Ribecca, S. 2019). For the Bubba Gump company, these multi-set bar charts could prove useful in comparing a variety of data. For instance, the Bubba Gump company can use these charts to compare sales data between in and online stores, number of visits between in and online stores, or number of products bought, or money spent between the two stores. Many other comparisons can be made with the Bubba Gump company data set using the multi-set bar charts. Additionally, we could add another set of bars to include comparisons with third party retailers as well.

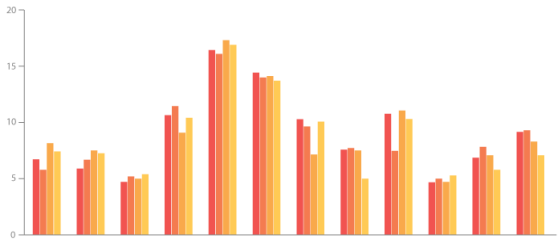


Figure 1. Multi-set Bar Chart., by Ribecca, S. 2019

Multi-set bar charts are great when we want to compare datasets side by side. When we want to visually represent data over a time period, data patterns, data gaps, or comparisons, we can use histograms (Ribecca, S. 2019). Histograms (Figure 2) work very similarly to multi-set bar charts, but instead of comparing data side by side, histograms provide a much simpler and easy to read visualization of the represented data. The Bubba Gump dataset contains a lot of information. Therefore, when we need to focus on one particular problem, such as time between purchases on the online store, how much money is spent per customer, different pages click amounts, etc., we can use histograms.

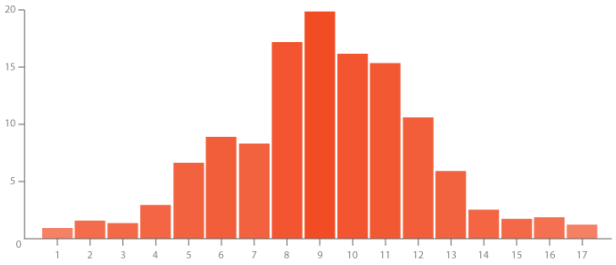


Figure 2. Histogram., by Ribecca, S. 2019

The Bubba Gump Shrimp company provides some unique data that can help provide helpful insights to their customers. They provide city, state, county and zip code data. This data can help see patterns about where more customers are buying products than others. These patterns can be compared to in and online stores as well as third party stores. In order to help visualize this data, we can use choropleth maps (Figure 3). These maps provide geographical areas that are colored or shaded in particular colors in order to represent the data. (Ribecca, S. 2019). Choropleth maps can help the Bubba Gump Shrimp company find areas that purchase

more products than others and advertise to those high priority areas. Additionally, they can see what areas make less purchases and use additional data to find out why or how to create better sales in those areas.

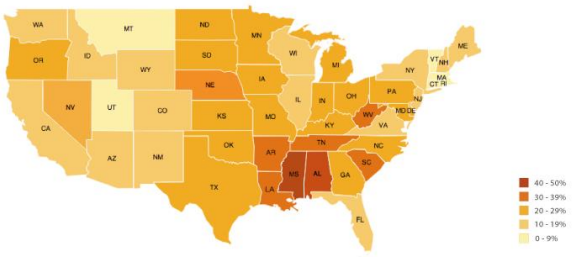


Figure 3. Choropleth Map., by Ribecca, S. 2019

Finally, to visually represent other kinds of percentages and proportion of the data set, we can use pie charts (Figure 4). These charts are classic forms of data visualization and can help provide an easy and simple visual aid to represent specific data. The Bubba Gump Shrimp company has a lot of data that could be used in conjunction with pie charts. For example, the percentage of people who buy products in a certain age range, percentage of purchases between third party, in and online stores, etc. Pie charts do have their advantages as well as disadvantages to representing data, but for specific percentage areas of the dataset, pie charts can be extremely useful.

Figure 4. Pie Charts., by Ribecca, S. 2019

Research Question

In order to create a research question to help analyze the data, we first need to understand the big question that has been tasked by the Bubba Gump company. The Bubba Gump Shrimp company is a well-known retailer in restaurants and other retail channels. Bubba Gump quickly gained popularity thanks to a global blockbuster movie. The company used this popularity and opened several other restaurants, sells its own merchandise online and wholesales its branded merchandise to other outlets. Unfortunately, the rapid growth for the first few years started to decline. We have been tasked to analyze why the company's online sales leveled off and started to decline. By analyzing this data, we hope to find the reason for such decline

Now that we understand the problem as a whole, we can now focus on a specific question that will help us research the problem and eventually lead us to a conclusion about the decline of online sales. If we look at the data, we can see that most of the sales being made are from the restaurant. This is understandable because the restaurant itself was boosted thanks to a blockbuster movie. We can also see that the in store and third-party visits are fairly consistent as well as the purchases from these locations. Unfortunately, the online store has very low traffic and even less purchases. Using this information, we can come up with several questions that could aid us in our research to find out about what is happening to cause such a drop in online traffic.

- What is the difference between the online stores and on sight stores?
- What are the purchase patterns of customers between both stores?

- Are there any significant external reasons for fewer traffic online than to on sight stores?
- Are most purchases made in areas with on sight locations? Do areas with no on sight store have less sales?
- What do customers like about in store and online stores? What do they not like?

These questions can help provide a base for our research. The data that the Bubba Gump Shrimp company provides can help us analyze each of these questions in detail. The last question "What do customers like and not like about online and in stores?" was added because the Bubba Gump company also provided a sample survey of 500 customers. This can help us find additional insight to the differences of the stores and the interests of the customers in these stores.

Research Measurement

The questions

- "What is the difference between online stores and on sight stores?"
- "What do customers like about in store and online stores? What do they not like?"

can be measured very easily by analyzing the data and the stores themselves. We can create additional data about what is similar and what is different about the on-sight store versus the online store. We can then take the sample survey data and compare what customers liked and didn't like to the data about the stores. This can help us find patterns in what customers liked visually, or functionally about each of the stores. To help reinforce the findings, we can look at historical sample survey data or run another random sample and run that data against the patterns

that were created from the first analysis. These patterns will provide useful insight to what makes these stores unique and what needs to be changed in order to get better sales.

The questions:

- What are the purchase patterns of customers between both stores?
- Are there any significant external reasons for fewer traffic online than to on sight stores?
- Are most purchases made in areas with on sight locations? Do areas with no on sight store have less sales?

Can be measured using the data provided by the Bubba Gump company and analyzing it multiple times and running it against the different analyses. By analyzing the data multiple times and creating different samples and running them against each other, we can find the answers to these questions by finding patterns, events and associations in the data. If the data continues to produce consistent results, we will know that the data patterns are reliable and correct.

Follow-Up Questions

Just like in any sciences, creating a hypothesis model and running tests in order to find an answer, pattern or event will lead to results and then eventually more questions. The analysis of data for the Bubba Gump Shrimp company is no different. We have our current research questions that we will be analyzing from the Bubba Gump dataset. Once we analyze the data, do some comparisons and create visualizations, we will have a better understanding of what is happening to the online store for the company. After we have a better understanding and have some patterns to work with, we can then create additional research questions that will start to

focus more on the online store and how to fix the issue as well as develop predictive analyses to help ensure the future of the online store. For example, we asked the question “What is the difference between the online store and the on-sight store?”. Once we understand the answer to this question, we can then create a new research question such as “What functionality, design or products need to be changed/updated in order to increase online traffic?”. These questions and analyses will continue until a solution has been made.

Research and Support

With the ever-growing market for Big Data as well as the rise of online marketing and retail, there are endless amounts of resources that cover the questions that were posed to the Bubba Gump company's problem. Two such resources include *The 15 Second Rule: 3 Reasons Why Users Leave a Website* by David Zheng (2019). And *Comparing retailer purchase patterns and brand metrics for in-store and online grocery purchasing* by John Dawes (2014). David Zheng's source addresses the question of what external reasons can cause fewer online retail traffic. He also goes into a lot of detail about "bounce rates" and his theory that most sites need 15 seconds to catch an audience's attention before they leave. Mr. Zheng provides a lot of very valid points and data about bounce rates and what makes an online store or website appealing to an audience. This is a great source for the design aspect as well as some product advice. While this is a great source to be used when thinking about the Bubba Gump company's problem, Mr. Zheng could benefit from going into more detail about functionality of a website that audiences prefer as well as any external reasons that an audience might not use a specific website. John Dawes source goes into details about the purchasing patterns of customers between online and on sight stores. Again, just like David Zheng, John Dawes provides a lot of very valid information and data that proves the point he is trying to make. One major point that John Dawes (2014) discusses is that research suggests that customers who shop online are behaviorally and demographically different to those who do not. This could provide major insights to how we analyze the Bubba Gump Shrimp company's data in order to solve the problem of decreased online sales. Although Mr. Dawes paper provides great insights in many areas of online vs. in store shopping, this paper specifically focuses on grocery shopping. I believe that this information could be used on a much wider scale than the one focused area.

Both of these resources are very good and what they are trying to prove and provide very compelling data. This data can be used to help analyze the Bubba Gump dataset as well as a lot of more online retail datasets. The world of data mining is growing rapidly every day, and this gives us endless amounts of resources to work with. Each of these resources provide a unique insight to analyzing data and should be used to help broaden analytical techniques.

Analysis Organization

We have been tasked to analyze Bubba Gump's large collection of data and determine why the company's online sales leveled off and started to decline. The Bubba Gump company has a wide range of data resources for us to analyze. However, this week we are analyzing the sample survey dataset that was given to 500 customers. In order to analyze the sample survey data, three different analysis models were used; simple linear regression, simple logistic regression and hierarchical cluster models. Because the analysis models that were used are regression analysis models, a stepwise approach was used. A stepwise approach according to Harrell, E. (2001), is an analysis method mainly used in regression models in which predictive variables are chosen and automatically carried out by a procedure. In the case of the Bubba Gump dataset, the JMP program helped with the stepwise approach.

At the beginning of the analysis of the sample survey dataset, in each of the analysis models used, every variable was used and analyzed. After each analysis, if a variable appeared to be unnecessary or provided no significance to the analysis question, (which focused on the online webstore) it was removed from the analysis model. This stepwise approach is known as backward elimination. (Wikipedia, 2020). For most variables, such as if customers were married or not, were obvious choices to remove from the current analysis. For other variables, they were

gradually removed after analyzing each visual chart. The goal of this backward elimination was to find any additional patterns in the data between specific variables in conjunction with the webstore variables. Unfortunately, due to the wide variety of variables in the sample dataset, this approach caused the stepwise regression to become cluttered and overfitted data. Due to this realization, a better approach to analyze this dataset would be a forward selection approach. Regardless of this issue, significant patterns and classifications were found during the analysis of the sample survey dataset which will be discussed in length below.

Sources of Error

At the beginning of this analysis, I was unsure where to start and what variables I wanted to use specifically. The first draft of analyses was done using more variables than were needed. Each of the models, simple linear regression, logistics regression and cluster models were all tested using the webstore variables as well as other variables in order to see what patterns could be found as well as to better understand the data. I assumed by using more variables than were needed, I would be able to find these patterns and understand the data better. Unfortunately, this led to very confusing results in the analysis models as well as unorganized and chaotic visuals that represented the data. These models were eventually scrapped and tested again with just the variables that were needed. For the most part, the sources of error in the analysis of the Bubba Gump Shrimp sample survey dataset was of human error. In order to test and understand the data, too many variables were used in these models and caused unwanted results. After learning from these errors, I used only the webstore variables that pertained to the specific question of analysis and this created a much more organized and clean representation of the dataset. This also helped to clearly visualize meaningful patterns in the dataset that will be discussed in the

next section. By having a clear and readable visual representation of the data, it is much easier to understand and analyze the data and then, if needed, create additional analyses questions to continue the analysis of the problem.

Meaningful Patterns

So far, we have used a variety of methods in order to analyze the Bubba Gump Shrimp company's survey data of 500 customers. For the previous analyses of the data, we used three different models to analyze this data. The models that were used was Simple Linear Regression, Simple Logistics, and clustering models. Each of these models provided observable patterns and many of these patterns' assumptions were backed up by the other models that presented the same patterns found in the previous models. Because we used these three different models in order to analyze the sample survey data, we could easily see and understand the patterns and what they data meant for the online webstore data variables.

For the simple linear regression model, we wanted to understand the web channel expenditure of various customers. To do this, we analyzed the WEB_VISITS and Webstore_Spend variables. These data variables were also given unique colors in order to better understand and see the data. When we analyzed these variables using the simple linear regression model, we were able to see the web channel expenditure patterns for the customers. A very large and dense pattern of the data represented that many customers only visted the webstore once. Because each customer was individually represented in the data, this large dense population also had a spend amount that ranged from 0 to 100 dollars. This dense population in the data provided us with a unique pattern that lets us know the average spending range for customers who only visit the store once. Another pattern that was represented in the simple linear regression model

was the second most dense part of the data which represented customers whom only visited the store once. And just like the first grouping of customers, the spending range was did not go above 100 dollars. In fact, the range was from 40 to 100 dollars. This gives us a larger spending pattern between two different groupings of customers. Finally, the simple linear regression model showed us that customers who came back a third time had more chance of visiting the store again and thus spending more money on the online store. This provides us with our third pattern which represents frequent and regular customers to the Bubba Gump Shrimp company's webstore.

For the simple logistics model, we were focusing on whether a customer will make a purchase from the web channel. To do this, we analyzed the WEB_PURCH_YN and Webstore_Spend variables. When analyzing these two variables using the Logistics regression model, the same pattern appeared that was found in the simple linear regression model. In this model, there was a large grouping of customers who have made purchases in the range of 0 to 100 dollars. This grouping makes up the majority of the purchases in the sample survey data. The customer groupings gradually become less dense as the chart moves outward, but this also backs up another pattern that we saw in the simple linear regression model. In the previous model, we can see that customers who made a second visit to the webstore, spent around the same amount as customers who only visited once. Because the groupings become less dense as it moves outward in the logistics model, we can assume that many customer who only visited once or twice will not likely visit again, therefore the data gradually represents frequent or regular shoppers to the Bubba Gump company rather than intervals of new shoppers.

For the hierarchical clustering model, the Webstore_Spend and WEB_VISITS variables were analyzed. The goal for analyzing these variables was to determine whether customers made

a purchase, visited the store or both. A large portion of the data represented a pattern that could also be seen in the Logistics regression model. The cluster data showed a cluster of customers who did not visit the store as well as did not make any purchases. The logistics model also represented this data pattern because both models let us know what grouping of customers had the pattern of not visiting the store and/or not making any purchases. The hierarchical clustering model also backed up a pattern that can be seen in both the linear regression model and the logistics model. This pattern is the grouping of customers who have visited once and/or twice and have only spent amounts in the range of 0 to 100 dollars. Throughout each of these models, this specific pattern is the most reoccurring. It is consistent throughout the models and therefore lets us know that This is a reliable and reoccurring pattern that should be analyzed more in detail in order to determine a solution to the decline in webstore sales.

Inaccurate Depictions of Data

Simple linear regression is a model that allows data analysts to understand the relationship between two variables. Simple linear regression models use dependent and independent variables and aim to solve for the dependent variable. (Devault, G. 2020). The linear regression models were a favorite in the analysis of the Bubba Gump Shrimp Company's survey dataset. This model provided a simple and easy to understand depiction of the data in regards to the chosen variables. This model created groupings of the customers web visits and their spending amounts and placed them neatly into the chart and were represented beautifully. After giving each customer grouping a unique color and symbol, the groupings were very easy to find and determine where they lie in the spectrum of visits and spend amounts. These groupings helped us find meaningful patterns in the data itself. These patterns were then compared to the

other analysis models that then confirmed that our pattern assumptions were true, valid and consistent in the dataset. This model depicts the data very well.

Simple logistic regression models are models that use binary variables in order to create prediction analysis of a particular dataset. (StatisticsSolutions, 2019). When creating the logistics regression model chart, WEB_PURCH_YN and Webstore_Spend variables were used. The WEB_PURCH_YN variables was significant for this model because logistics regression model charts use binary variables. After I inputted these variables into the JMP, I thought that I had done something incorrect and I didn't think that this chart represented the data correctly. I then switched the variable's axis to see if something positive would happen to the chart. The chart then became much worse than before. After some confusion, I decided to do some research on these models. After reviewing the textbook and reading about logistic regression models, I learned that this specific logistic regression model is just one type of this model out of a variety of models. Additionally, after creating a color system that allowed me to color each of the variables, I was able to understand that the depictions of the data were correct and quite easy to read and understand.

Hierarchical clustering models are very useful models that represent data through dendrograms. The data is organized into groups that make up leaves of a tree and thus make meaningful classifications. (Stephanie, n.d., 2016). When creating this model visualization, it was very easy to become overwhelmed due to the overwhelming and seemingly unending amounts of clades and leaves that make up the hierarchical clustering model. This cluster model's goal was to create meaningful classifications in the Bubba Gump sample survey data based on customer purchase habits. Because each customer was represented individually, the cluster model became quite cluttered and difficult to read. Before applying unique colors to the

natural clusters in the visual model, I would have claimed that this model inaccurately depicts the data due to very low readability. Fortunately, after adding specific colors to their respective clusters, reading the data became quite clear and easy to understand. This hierarchical clustering model also confirmed pattern assumptions that were made from the previous analysis models by representing these patterns in the classifications created in this model. Because this model is only focusing on creating meaningful classifications and not trying to predict anything from the dataset, I believe that this model accurately depicts the data and provides a useful and powerful tool in the analysis of the Bubba Gump Shrimp company's dataset. If the model was trying to create predictions in the dataset, I would say that the use of a decision tree would be a better fit for such a job.

Alternative Analytic Methods

The Bubba Gump Shrimp company's data has been analyzed using three models and these models have produced results that has given us specific patterns that are consistent in each. The results of these models have shown that there are patterns with customers who have only visited the webstore once or twice with a spending range from 0 to 100 dollars. Additionally, there is a pattern showing customers who return to the webstore for the third time will be more likely to come back to the store and thus spend more money that the first pattern of customers. These patterns also give us a clear classification of the data. We can classify these customers in groups and analyze them separately in order to gain better insight into the spending habits of these customers. Now that we have these results and understand the patterns found in the results, we can think about alternative analytic methods to analyze the data.

The results from the three analysis models has shown us that there are clear patterns and classifications in the customers habits. Therefore, two different alternative methods could be used in order to efficiently analyze these patterns and classifications. The two methods that could be used are Neural networks and Decision trees. Both of these models are designed to work like the human brain when analyzing data and both are very useful when analyzing classifications and patterns. However, because neural networks are extremely complicated to implement, this model should only be used when there is a large amount of data to be analyzed. The Bubba Gump survey sample data is not very big, therefore, decision trees should be used as an alternative analytic method for this dataset.

Decision tree analysis is a very reliable method for predictive and classification analysis. Decision trees are analytic models that take the shape of a tree with multiple branches and leaves. The branches and leaves are divided and segmented into their respective classifications in order to create specific predictions based on the analysis question. (Ahlemeyer-Stubbe, A., & Coleman, S. 2014). From the analysis of Bubba Gump sample survey dataset, we now know what classifications are and the patterns that are present in these classifications. In order to gain even more insight into these patterns, decision trees could be used to organize these classifications in the unique tree format used by decision trees and then segmented again into groups of customers, and the specific products that they bought in order to find patterns in customer spending habits and what products are more likely to be bought by first time buyers versus what products are being bought by return or frequent customers. This insight could help the Bubba Gump Shrimp company find out what products they should be focusing on what what products can be changed in cycles in order to produce more frequent buyers and predict how well these products will do in the future. Decision trees are extremely easy to build and implement

and this analysis method provide complex return on investment models in addition to the predictive models. (Ahlemeyer-Stubbe, A., & Coleman, S. 2014). Therefore, due to the ease of implementation, quality prediction analysis and additional models like ROI, decision trees could be a great alternative analytic method to analyze the Bubba Gump sample survey dataset.

Display and Interpretation

As we have discussed in previous sections of this paper, the Bubba Gump shrimp provided a large amount of data for us to analyze. These resources, as we have seen, have provided us with great insight to the different retailers of the company and lets us understand what is happening in areas like the online retailers versus the onsite retailers. These data resources, contained in the company's data warehouse, include POS data, customer databases, webstore sales transactions, sales data from third party retailers and a sample survey conducted by the Bubba Gump company. In order to get a better understanding of the Bubba Gump customer base, their habits, purchase history and traffic history, we decided to analyze the sample survey that was conducted from 500 of the company's customers.

To analyze this sample survey dataset, several different analysis models were used. These models included, multivariate correlations, scatterplot matrices, pairwise correlations, simple linear regression and simple logistics regression models and cluster analysis. Each of these analysis models provided significant results that led to the discovery of certain patterns and associations in customer behavior as well as certain predictions in relation to the online webstore. Each of these models and their results will be discussed and what these results mean regarding the Bubba Gump Shrimp company's online webstore.

Pairwise Correlations & Principal Components Analysis

In order to understand what variables correlated in the sample survey dataset, pairwise correlations and principal components analyses were used. During the analysis of the data tables and matrices, we were able to see several correlations in the data. We were able to see that in two different correlations, that customers were more likely to visit Bubba Gump stores, (both on site and webstore) after visiting the Bubba Gump restaurants. These two correlations in the data give us quite a significant insight to customers purchase habits and what conditions need to be met in order to get customers interested in buying Bubba Gump products. Additionally, another possible useful correlation was the relation between income and webstore visits. This could provide useful information for classifying customers based on income. One example of the data that represents these correlations can be seen in figure 5. We can use this data to help create deals for customers. For example, offering special coupons for online stores when visiting an onsite store, or create special stock for online stores that customers can't get from onsite stores.

Multivariate Correlations

	zip	ZIP_2	Restaurant	RES_VISITS	Webstore_Spend	WEB_VISITS	THIRD_SPEND	THIRD_VISITS	Age	MARR_BIN	Income
zip	1.0000	0.7467	-0.0602	-0.0906	-0.1043	-0.0928	-0.0053	0.0465	-0.0522	0.0849	0.0085
ZIP_2	0.7467	1.0000	-0.0259	0.0065	-0.0186	-0.0226	-0.0668	-0.0143	-0.0659	0.1090	0.0101
Restaurant	-0.0602	-0.0259	1.0000	0.5955	0.4534	0.2078	-0.0427	-0.0379	-0.0033	-0.0796	-0.0159
RES_VISITS	-0.0906	0.0065	0.5955	1.0000	0.2943	0.1839	-0.0801	-0.0846	0.0045	-0.0797	-0.0307
Webstore_Spend	-0.1043	-0.0186	0.4534	0.2943	1.0000	0.6119	-0.0059	-0.0034	-0.0368	-0.0148	-0.0299
WEB_VISITS	-0.0928	-0.0226	0.2078	0.1839	0.6119	1.0000	-0.0409	-0.0103	-0.0037	-0.0054	0.0301
THIRD_SPEND	-0.0053	-0.0668	-0.0427	-0.0801	-0.0059	-0.0409	1.0000	0.7422	-0.0827	-0.0129	-0.0601
THIRD_VISITS	0.0465	-0.0143	-0.0379	-0.0846	-0.0034	-0.0103	0.7422	1.0000	-0.0768	-0.0280	-0.0636
Age	-0.0522	-0.0659	-0.0033	0.0045	-0.0368	-0.0037	-0.0827	-0.0768	1.0000	-0.0570	0.1093
MARR_BIN	0.0849	0.1090	-0.0796	-0.0797	-0.0148	-0.0054	-0.0129	-0.0280	-0.0570	1.0000	-0.0278
Income	0.0085	0.0101	-0.0159	-0.0307	-0.0299	0.0301	-0.0601	-0.0636	0.1093	-0.0278	1.0000

Figure 5: Multivariate Correlations Data

Simple Linear/Logistics Regression Models & Cluster Analysis

In order to get a better understanding of customer's web channel expenditures, a simple linear regression model was used. Unlike the pairwise correlations and principal components analyses, every variable was not used in this analysis. Instead the webstore spend amounts and webstore visits variables were analyzed. This allowed us to see webstore expenditure patterns for each customer in the sample survey dataset. The visual representation of the data showed a large number of customers whom have only visited the online store once or twice. This classification of customers (who made purchases from the online store) made up most of the sample survey dataset. This classification of the customer population also was found to only have spent between 0 to 100 dollars. This dense population in the data provided us with a unique pattern that lets us know the average spending range for customers who only visit the store once. Additionally, the visual representation of the data showed what customers have made a third visit to the online webstores and made purchases. Customers who have visited for a third time were shown to be more frequent or regular customers and thus spend more money on the online webstore.

In order to focus on prediction models and whether a customer will make a purchase from the webstore, simple logistics model was used. Just like the simple linear regression model, only two variables were used. Both the webstore spends amount and binary variables allowed us to confirm our assumptions about customer habits from the simple linear regression model. The logistics regression models results show that a dense grouping of customers fit into the customer classification of one to two online webstore visits with spending amounts ranging from 0 to 100 dollars. The results also confirmed assumptions about third time visit customers and the possibility of them coming back as well as higher spending amounts.

Finally, in order to determine whether customers made a purchase, visited the store or both, a hierarchical clustering model was used. A large portion of the data represented a pattern

that could also be seen in the Logistics regression model. The cluster data showed a cluster of customers who did not visit the store as well as did not make any purchases. The logistics model also represented this data pattern because both models let us know what grouping of customers had the pattern of not visiting the store and/or not making any purchases. The hierarchical clustering model also backed up a pattern that can be seen in both the linear regression model and the logistics model. This pattern is the grouping of customers who have visited once and/or twice and have only spent amounts in the range of 0 to 100 dollars.

Combining the Results

Now that we have both results from the individual analysis models, we can create a much bigger picture that we had before. The pairwise correlations and principal components results represented correlations in the sample survey data such as, customers were more likely to visit Bubba Gump stores, (both on site and webstore) after visiting the Bubba Gump restaurants. The simple linear regression, simple logisitics regression and hierarchical clustering results represented the customers habits, and spending amounts. Three different clear classifications were made based on customer visits and spending amount ranges. These analysis results from each of the individual models provide a unique insight when analyzed by themselves. When these results are put together, we can now start to think about additional assumptions and ideas to help the Bubba Gump Shrimp company deal with declining online sales.

We know that from the correlations that customers are more inclined to make purchases from retail stores after visiting the Bubba Gump restaurants. We also know that most of the customers (from the sample) are customers who visit online stores only once or twice and have average spending amounts between 1 to 100 (omitting 0 dollars). Now that we know both of these analysis results, we can create additional plans that will help create more traffic for the online retail stores. Such plans include creating a reward system between the restaurant and webstores, specific coupons that can be retrieved only from the restaurants.

Validity, Reliability, Limitations

During the analysis of the Bubba Gump Shrimp company's sample survey data, several different analysis models were used in order to analyze different aspects of the data. The models were used to find correlations, patterns, associations and classifications. In order to maintain data

result validity and reliability, multiple analysis models were used during the analysis process. During the correlation analysis, pairwise correlations and principal component analyses were used. Once correlations were found in the pairwise correlation charts, these results were recorded and then compared to the results of the principal component charts. We found that in both analysis charts, the same correlations were found. This cemented our assumptions about the correlations in the data variables. During the analysis for customer web expenditures and customer purchase predictions, simple linear & logistic regression models were used as well as a hierarchical cluster analysis model. After running each of the models through the sample survey data, the results were then compared against each of the other analysis models. Again, the results (customer classifications) of each analysis model matched. Each of these models produced consistent patterns, correlations and classifications. This provided reliable and valid data analysis results for the sample survey data.

These consistent data analysis results allow us to have reliable analysis results. However, there will always be limitations during the analysis process. One of the limitations for validity is that additional validation methods could be performed on the data in order to have a better understanding of the relationships sample survey variables. Lift and gain charts are some of the validity methods that could be performed on the data to provide visual representations of the relationships between observed behavior (Ahlemeyer-Stubbe, A., & Coleman, S. 2014). Additionally, not all variables have been analyzed from the Bubba Gump Shrimp company's dataset. Thus far, we have analyzed the sample survey data that has covered customer purchase patterns. Additional analysis must be made with more data in order to completely understand the problems with the online retailer.

Resulting Decision Influence

As discussed above, we now understand certain correlations in the data as well as specific customer habits that correlate to these correlations. So, what should we be focusing on next? The answer to that is found in the results from our analysis models as well as the additional information that has been provided from the Bubba Gump Shrimp company. From our analysis models we know that a large number of the customer population shops at the online stores once or twice and likely will not return. However, a small number of this population has been shown to come back for a third time or even become a regular to the online stores. Within this population, analysis results showed that customers tend to visit the online stores and on-site stores after visiting the Bubba Gump restaurants. In addition to these correlations, customers also tended to spend money on the online stores after visiting the on-site stores. With this information, we now have unique insight into what customers purchase habits are. We have the information about how and why the customers visit both retailers, now the next step in our analysis to find out why online sales have been declining moves to *what* customers are buying, what products are sold at both retailers and what products are only sold at one of the locations. Additionally, other internal data such as pricing of products, sales that are found in each retailer, and even the designs of each retailer could provide significant information to the psychology of customer purchase habits. We know that customers are visiting the online store once or twice and then traffic dramatically declines after these initial visits. The data also lets us know that a significant number more of customers visit on-site stores rather than the online stores. The key to this decline in traffic is found in the question of products, sales, appealing designs, user-interface, etc. Thus, our next rational step in our analysis is to focus on products and sales between each of the retailers. Once we have gathered this data, we can analyze and determine if

additional analysis on the webstore itself (design and user-interface) is necessary. The Bubba Gump Shrimp company has provided us with plenty of raw data that can be found in the data warehouse. Here, we can use the POS data as well as other customer data to help us analyze purchase histories and patterns between both retailers.

Visual Evaluation

In the Display and Interpret section of the paper, the models that were used to analyze the Bubba Gump company's dataset were discussed. Some visual representations were also shown to help interpret what patterns, associations and clusters that were found in the dataset. Here, we will be discussing what visuals were specifically used and how successful they were during the analysis of the Bubba Gump sample survey dataset. As previously mentioned, the analysis models that were used to analyze the dataset were principal component analysis, pairwise correlations, simple linear regression and simple logistics regression models and cluster analysis.

Pairwise correlation & Principal component analysis

Below are visual representations of the principal component analyses that were used to identify correlations and associations in the sample survey dataset. Both the visual representations and data tables are represented below because each visual representation of the data provided a unique insight to the data.

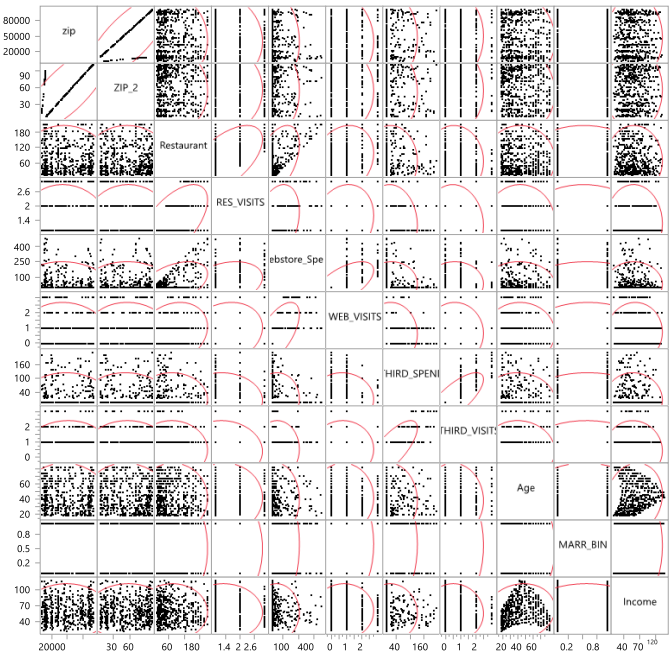
Multivariate Correlations

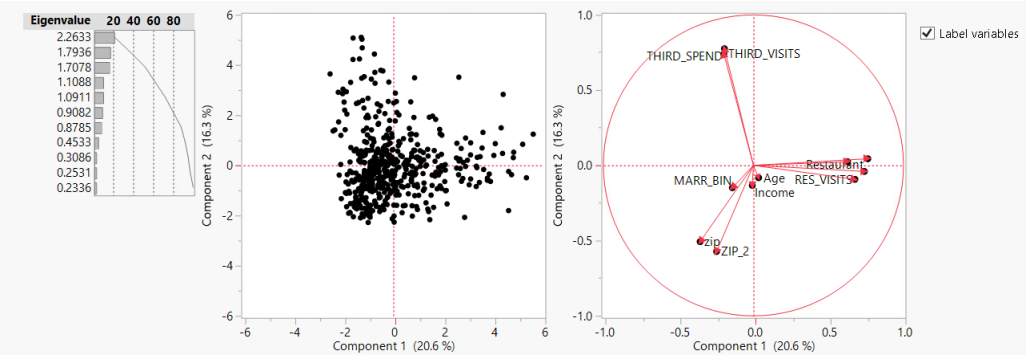
	zip	ZIP_2	Restaurant	RES_VISITS	Webstore_Spend	WEB_VISITS	THIRD_SPEND	THIRD_VISITS	Age	MARR_BIN	Income
zip	1.0000	0.7467	-0.0602	-0.0906	-0.1043	-0.0928	-0.0053	0.0465	-0.0522	0.0849	0.0085
ZIP_2	0.7467	1.0000	-0.0259	0.0065	-0.0186	-0.0226	-0.0668	-0.0143	-0.0659	0.1090	0.0101
Restaurant	-0.0602	-0.0259	1.0000	0.5955	0.4534	0.2078	-0.0427	-0.0379	-0.0033	-0.0796	-0.0159
RES_VISITS	-0.0906	0.0065	0.5955	1.0000	0.2943	0.1839	-0.0801	-0.0846	0.0045	-0.0797	-0.0307
Webstore_Spend	-0.1043	-0.0186	0.4534	0.2943	1.0000	0.6119	-0.0059	-0.0034	-0.0368	-0.0148	-0.0299
WEB_VISITS	-0.0928	-0.0226	0.2078	0.1839	0.6119	1.0000	-0.0409	-0.0103	-0.0037	-0.0054	0.0301
THIRD_SPEND	-0.0053	-0.0668	-0.0427	-0.0801	-0.0059	-0.0409	1.0000	0.7422	-0.0827	-0.0129	-0.0601
THIRD_VISITS	0.0465	-0.0143	-0.0379	-0.0846	-0.0034	-0.0103	0.7422	1.0000	-0.0768	-0.0280	-0.0636
Age	-0.0522	-0.0659	-0.0033	0.0045	-0.0368	-0.0037	-0.0827	-0.0768	1.0000	-0.0570	0.1093
MARR_BIN	0.0849	0.1090	-0.0796	-0.0797	-0.0148	-0.0054	-0.0129	-0.0280	-0.0570	1.0000	-0.0278
Income	0.0085	0.0101	-0.0159	-0.0307	-0.0299	0.0301	-0.0601	-0.0636	0.1093	-0.0278	1.0000

Scatterplot Matrix

Pairwise Correlations

Variable	by Variable	Correlation	Count	Lower 95%	Upper 95%	Signif Prob
ZIP_2	zip	0.7467	500	0.7052	0.7831	<.0001*
Restaurant	zip	-0.0602	500	-0.1471	0.0276	0.1788
Restaurant	ZIP_2	-0.0259	500	-0.1133	0.0620	0.5639
RES_VISITS	zip	-0.0906	500	-0.1769	-0.0030	0.0428*
RES_VISITS	ZIP_2	0.0065	500	-0.0812	0.0941	0.8846
RES_VISITS	Restaurant	0.5955	500	0.5358	0.6493	<.0001*
Webstore_Spend	zip	-0.1043	500	-0.1903	-0.0168	0.0196*
Webstore_Spend	ZIP_2	-0.0186	500	-0.1061	0.0692	0.6778
Webstore_Spend	Restaurant	0.4534	500	0.3808	0.5204	<.0001*
Webstore_Spend	RES_VISITS	0.2943	500	0.2121	0.3724	<.0001*
WEB_VISITS	zip	-0.0928	500	-0.1790	-0.0051	0.0381*
WEB_VISITS	ZIP_2	-0.0226	500	-0.1101	0.0652	0.6142
WEB_VISITS	Restaurant	0.2078	500	0.1223	0.2902	<.0001*
WEB_VISITS	RES_VISITS	0.1839	500	0.0978	0.2673	<.0001*
WEB_VISITS	Webstore_Spend	0.6119	500	0.5540	0.6640	<.0001*
THIRD_SPEND	zip	-0.0053	500	-0.0929	0.0824	0.9061
THIRD_SPEND	ZIP_2	-0.0668	500	-0.1536	0.0211	0.1360
THIRD_SPEND	Restaurant	-0.0427	500	-0.1299	0.0451	0.3405
THIRD_SPEND	RES_VISITS	-0.0801	500	-0.1666	0.0076	0.0735
THIRD_SPEND	Webstore_Spend	-0.0059	500	-0.0935	0.0819	0.8958
THIRD_SPEND	WEB_VISITS	-0.0409	500	-0.1281	0.0470	0.3616
THIRD_VISITS	zip	0.0465	500	-0.0414	0.1336	0.2999
THIRD_VISITS	ZIP_2	-0.0143	500	-0.1019	0.0734	0.7491
THIRD_VISITS	Restaurant	-0.0379	500	-0.1252	0.0500	0.3978
THIRD_VISITS	RES_VISITS	-0.0846	500	-0.1711	0.0031	0.0586
THIRD_VISITS	Webstore_Spend	-0.0034	500	-0.0911	0.0843	0.9394
THIRD_VISITS	WEB_VISITS	-0.0103	500	-0.0979	0.0774	0.8180
THIRD_VISITS	THIRD_SPEND	0.7422	500	0.7001	0.7792	<.0001*
Age	zip	-0.0522	500	-0.1393	0.0356	0.2436
Age	ZIP_2	-0.0659	500	-0.1527	0.0220	0.1414
Age	Restaurant	-0.0033	500	-0.0909	0.0845	0.9422
Age	RES_VISITS	0.0045	500	-0.0832	0.0922	0.9194
Age	Webstore_Spend	-0.0368	500	-0.1240	0.0511	0.4122
Age	WEB_VISITS	-0.0037	500	-0.0914	0.0840	0.9337
Age	THIRD_SPEND	-0.0827	500	-0.1692	0.0050	0.0647
Age	THIRD_VISITS	-0.0768	500	-0.1634	0.0109	0.0862
MARR_BIN	zip	0.0849	500	-0.0029	0.1713	0.0579
MARR_BIN	ZIP_2	0.1090	500	0.0215	0.1948	0.0147*
MARR_BIN	Restaurant	-0.0796	500	-0.1661	0.0082	0.0755
MARR_BIN	RES_VISITS	-0.0797	500	-0.1662	0.0081	0.0750





The pairwise correlations and principal components visual representations of the data provided a great deal of help when analyzing the sample survey data. The visual aids allowed us to easily understand the large number of individual customer variables. The visual were made up of scatterplot matrices, eigenvalue charts, multivariate correlations, loading plots and pairwise correlations. These variables helped us see what variables correlate and what do not. The pairwise correlations provide a great insight into what two variables correlate by providing the correlation data, count, lower and upper 95% and the significant probabilities. The loading plots also helped us see what variables were positively or negatively correlated for specific components. Although the data seems cluttered and chaotic at first glance, the data itself was represented very organized and readable. The data tables represented correlations with easy to read colors as well as bar charts and the scatterplot matrices gave us a good representation of how the data correlates between the variables and thus let us see how well each of the variables were correlated to each other or not. These visual aids were vital in the analysis of variable correlations.

Simple Linear Regression & Simple Logistics Regression Models

Below are visual representations of the simple linear and logistics regression analysis models that were used to identify web channel expenditures of customers and predict whether customers will make purchases from the web channel. in the sample survey dataset. Both the visual representations and data tables are represented below because each visual representation of the data provided a unique insight to the data.

Bivariate Fit of WEB_VISITS By Webstore_Spend

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.611933	0.553968	0.663993	<.0001*
Covariance	42.70581			
Count	500			

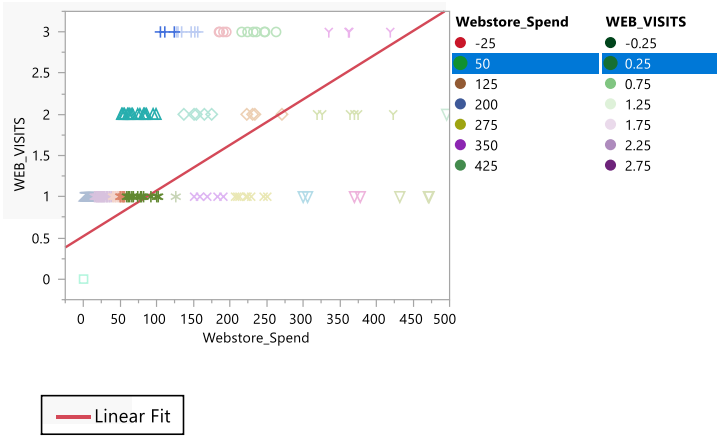
Variable	Mean	Std Dev
Webstore_Spend	40.474	87.87316
WEB_VISITS	0.746	0.794195

Linear Fit

WEB_VISITS = 0.5221534 + 0.0055306*Webstore_Spend

Summary of Fit

RSquare	0.374462
RSquare Adj	0.373205
Root Mean Square Error	0.628767
Mean of Response	0.746
Observations (or Sum Wgts)	500



Logistic Fit of WEB_PURCH_YN By Webstore_Spend

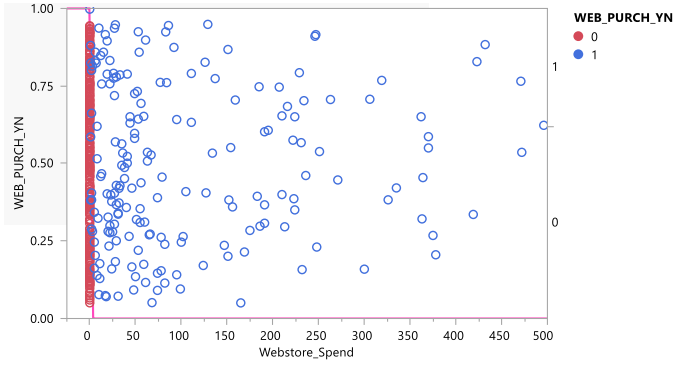
Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	325.78241	1	651.5648	<.0001*
Full	6.73496			
Reduced	332.51737			

RSquare (U)	0.9797
AICc	17.4941
BIC	25.8991
Observations (or Sum Wgts)	500

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[0]	5.73334128	1.0016168	32.77	<.0001*
Webstore_Spend	-24.434917	5753.9428	0.00	0.9966



The simple linear regression analysis model is a very useful analysis tool because it is a model that allows us to understand and predict the relationship between two variables by analyzing the dependent and independent variables and solving for the dependent variable. For this linear regression model, we aimed to understand customers web channel expenditures. The linear chart used two different variables, webstore spend amounts and webstore visits. Because the data was extremely dense in some areas, colors and symbols were added in order to distinguish specific classifications of customers. These colors and symbols allowed the linear regression chart to become an extremely vital tool in understanding the patterns in customer

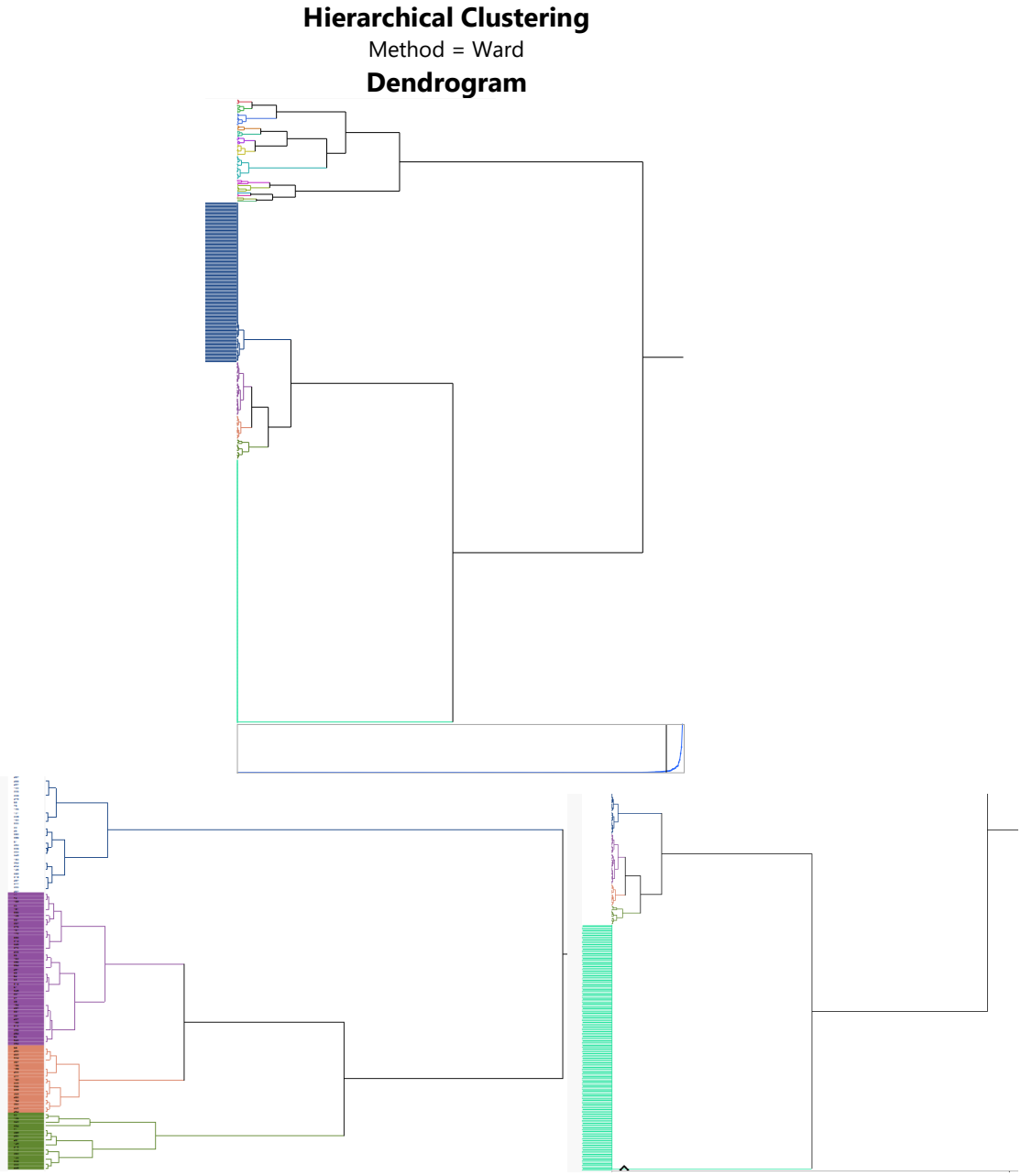
webstore visits and spending amounts. This model allowed us to easily understand the amount of times customers visit the online stores as well as the spending range of these customer populations.

The simple logistics regression analysis model is also a very useful analysis tool because it also is used for prediction analysis. Just like the simple linear regression model, two variables were used, web purchase (Y/N) and webstore spend amounts. This visual chart also had a lot of condensed data that was slightly difficult to read at the beginning of analysis. Therefore, colors were added in order to better see and understand what the data is showing. This chart also provided us with a good representation of what customers habits are when it comes to spending amounts and whether they made a purchase or not. Thanks to this visual chart, we were able to see a consistent correlation in customer behavior when it came to the amount of times they visited as well as consistent correlations in spending amounts.

Not only did both of these visual representations of the data provide unique and significant analysis results that will help in our future analyses, they also helped back up and support assumptions that were made when we first analyzed the simple linear regression model. By using both of these simultaneously, we were able to see the data represented in two different ways while also providing consistent results that allowed us to create patterns and classifications of customer populations in the sample survey dataset.

Cluster Analysis

Finally, below is the visual representation of the hierarchical clustering model that was used to find meaningful classifications in the sample survey dataset. The entire hierarchical tree as well as specific sections of the tree are shown.



For the hierarchical cluster tree, two variables were used, webstore spend amounts and webstore visits. The hierarchical clustering tree was a vital and easily readable tool that helped solidify our pattern and classification assumptions during the analysis of the two previous regression models. The tree is made up of multiple different clades and leaves which make up different classifications. Just like the previous regression models, colors were added in order to help with the identification of the different customer classifications. These colors used in conjunction with the hierarchical clustering model was a great tool in understanding how and where each customer variable fit in each classification. Looking closer and analyzing each of the colored classes provide unique and significant patterns that will help in future analyses.

Next Steps

The data analysis models that were used on the sample survey data from the Bubba Gump company has provided vital results that let us know the customers shopping habits and customer psychology when visiting each of the retailers. This information is the first step in understanding why online sales and traffic have been declining. Thus far, we understand that customers visit the online and instore sites when they visit the Bubba Gump restaurants. We also know that customers tend to only visit the online stores a couple of times while also only spending a certain amount of money. This behavior is vital to our analysis but doesn't give us the full picture to why the decline is happening. We know that customers tend to visit on site retailers and spend more money than on the online retailers. Therefore, our next question is *what* is the difference between the onsite store and the online stores? Possible problems could include difference in products, prices, and sales. Additionally, certain designs, colors and user interfaces could also provide problems for customers that could cause them to not return. For example, the

fast food chain McDonalds did some major research into color theory and what makes customers "hungry". The same thing could be happening with the online stores. So, the next step in our analysis to find out why there is a decline in the online retailers is, first, products, sales and prices. After this analysis, if necessary, design and user interface could be our next line of inquiry.

References

Ahlemeyer-Stubbe, A., & Coleman, S. (2014). *A practical guide to data mining for business and industry*. Chichester, West Sussex, United Kingdom: Wiley.

Dawes, J. (2014). Comparing retailer purchase patterns and brand metrics for in-store and online grocery purchasing. Retrieved from https://www.researchgate.net/publication/263605829_Comparing_retailer_purchase_patterns_and_brand_metrics_for_in-store_and_online_grocery_purchasing

DeVault, G. (2020, January 9). How Simple Linear Regression, Used to Analyze Quantitative Data. Retrieved from https://www.thebalancesmb.com/what-is-simple-linear-regression-2296697

Foley, B. (2019, May 4). An Introduction to Cluster Analysis: SurveyGizmo Blog. Retrieved from https://www.surveygizmo.com/resources/blog/cluster-analysis/

Harrell, F. E. (2001) "Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis," Springer-Verlag, New York.

Stepwise regression. (2020, April 7). Retrieved from https://en.wikipedia.org/wiki/Stepwise_regression

McDaniel, S. (2019). 16 Data Mining Techniques: The Complete List - Talend. Retrieved from https://www.talend.com/resources/data-mining-techniques/

RapidMiner. (2020, February 10). Retail Analytics Software. Retrieved from https://rapidminer.com/industry/retail/

Ribecca, S. (2019). The Data Visualisation Catalogue. Retrieved from https://datavizcatalogue.com/

Rouse, M. (2018, November 2). What is association rules (in data mining)? - Definition from WhatIs.com. Retrieved from https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining

Shah, D. (2017, November 19). Data Mining Tools. Retrieved from https://towardsdatascience.com/data-mining-tools-f701645e0f4c

StatisticsSolutions (2019). What is Logistic Regression? Retrieved from https://www.statisticssolutions.com/what-is-logistic-regression/

Stephanie. (2019, September 25). Hierarchical Clustering / Dendrogram: Simple Definition, Examples. Retrieved from https://www.statisticshowto.com/hierarchical-clustering/

Uj, A. (2019, March 12). The Top 10 Data Mining Tools of 2018. Retrieved from https://www.analyticsinsight.net/the-top-10-data-mining-tools-of-2018/

WEKA. (n.d.). Retrieved from https://www.cs.waikato.ac.nz/ml/weka/

Zheng, D. Z. D. (2019, February 26). The 15 Second Rule: 3 Reasons Why Users Leave a Website. Retrieved from https://www.crazyegg.com/blog/why-users-leave-a-website/

https://datavizcatalogue.com/