8-2 Final Project Submission

Olivia Bielawski

Professor Wanda Bradley

DAT-220 Fundamentals of Data Mining

Southern New Hampshire University

August 25th, 2019

1. **Business Problem**

Bubba Gump Shrimp Co. has grown to a retailer for regional food with thanks to unexpected exposure from the movie, “Forrest Gump” But, even with initial popularity among the movie’s release, data has shown that business has declined within the past two years. Recently, the company’s data has integrated into a data warehouse, and will have the ability to reflect a customer’s shopping experience. This can be at a restaurant, on the online web-store, or through a third party vendor. The first step in this data mining exercise is to collect initial data. Survey data was collected from 500 customers who have made purchases through Bubba Gump. Along with these survey answers, the method of purchase was also recorded as well as history of purchases. To solve the decline in business, Bubba Gump Corporate has requested a cluster analysis of the data, and can better understand its customer feedback to create more revenue in the company.

1. **Analytic Method**

For the strategy of this analytic method, the company is taking customer feedback and utilizing it to their advantage. This is a common method of Data mining among companies, and visualizing where the problems are in the company and how they can be resolved. Because this information is coming from consumers of all sub-groups, Bubba Gump can gain very helpful and trustworthy feedback from its average consumers on customer satisfaction. One method that can be used is a Cluster analysis. This is where the user can discover natural, “clusters” within the data from the survey takers, and can assist with visualizations of data results. Also, In order to produce accurate results, Analytic Methods such as Regression methods can be utilized to find correlation data between two variables.

1. **Analysis Tools**

One tool that can be utilized for performing this analysis is JMP, which is a computer software specifically for the use of analyzing data. This can be very useful because of the programs ease of access for importing and exporting data. JMP can be executed into an excel spreadsheet, and perform the calculations to provide a mean, maximum, and overall average of the data. This can be very useful for the large amount of Data that Bubba Gump Shrimp co has collected in the means of organization.

1. **Data Visualization**

Data visualization is important because it gives a visual representation of how the data has changed over time. For example, if Bubba Gump Shrimp Co. has increased sales over a specific period of time, then that will show in a line graph. Also, histograms can be very useful when conducting data. A histogram is a numerical summary of data collected in an organized fashion. This can also be utilized because it can be a key asset into understanding the growth and decline from the data provided by customers.

1. **Research Question**

From the survey data taken from customer surveys, what relationship can be driven by the total spending at Bubba Gump, and where how customers buy product from Bubba Gump. This can be from either an on-site restaurant, the online web-store, or from a third party vendor. From this data, Bubba Gump can produce a solution intended to increase customer spending, and an increased Company Revenue.

1. **Research Measurement**

One way that this research can be displayed is through a form called regression analysis. This type of analysis displays In order to accurately analyze the data, the end user needs to determine which customer variables can be clustered. Clustering Data can be useful because it provides a visual depiction of how these variables relate to one another. Ina Regression analysis, one can see where the data is increasing or decreasing, and a general summary of what is called a Fir Line.

1. **Follow Up Questions**

The following are follow-up questions for the Data analysis

1. Does Third-party Vendor spending have a positive impact in relation to web-store spending?
2. How can customer information affect spending? This can be in relation to age, income, and online capabilities.
3. **Research and Support**

Along with using the Data provided by Bubba Gump Shrimp Co, there will be other resources used for this inquiry. One method is the article, “Use of ridits to analyze categorical data in preference studies” by N.Pouplard, E.M.Qannari, and S.Simon. This article is very useful because it explains how to analyze data, and how to organize it in verbal, organized categories. (<https://www-sciencedirect-com.ezproxy.snhu.edu/science/article/pii/S0950329397000207?via%3Dihub>)

1. **Analysis Organization**

One analysis tool that was very useful was clustering data. This is a practice where rows that have similar characteristics are grouped together to produce more accurate results. In the cluster diagram, the best method of approach was to cluster by age, income, and third-party spending. This is because the two variables share similar qualities that can be further analyzed. However, some conclusions were difficult to come to due to missing data in variables. Because of this, various cluster practices needed to be performed. Also, some aspects of the data was uncontrollable, For example, the data for the customers future spending are uncontrollable and impossible to predict.

Another analysis organization tool is logistic and linear regression. These are very helpful tools because they display a visual representation of a X and Y axis, and how these variables correlate with one another. For example, in a Linear regression of web store spending vs age, a negative fit line was concluded. This fit line tells the user that these variables have little to no effect on the amount a customer spends in the web-store. An example of logistic regression can be summarized with an analysis of webs-store purchases by web-store visits. For this analysis, it was concluded that customers who visit the web-store have a very high chance of making an online purchase.

1. **Sources of Error**

One source of error that can be concluded is the lack of data taken from the 500 customers that were surveyed. Even though the data was derived from customers who had made an online purchase, it was concluded in the analysis that a good amount of customers had not spent anything in the online web-store. Also, another source of error can be the lack of data. Along with the data taken from online purchasing, a date and time stamp can also be helpful. This is because the user can identify if the purchase date was different from the web-store visit date. This data could be missing because a customer did not input this data while taking the survey, and therefore it would be a blank field.

1. **Meaningful Patterns**

A One-way analysis is a great example of Meaningful patterns because it shows how two variables correlate to one another. It is shown on a x and y axis, and includes data points as long as the variables contains them. An example of this is from the analysis of web-store purchasing by web-store visits. These are two very important data variables taken from 500 survey-takers, and they are from individuals who have both visited and made an online purchase.

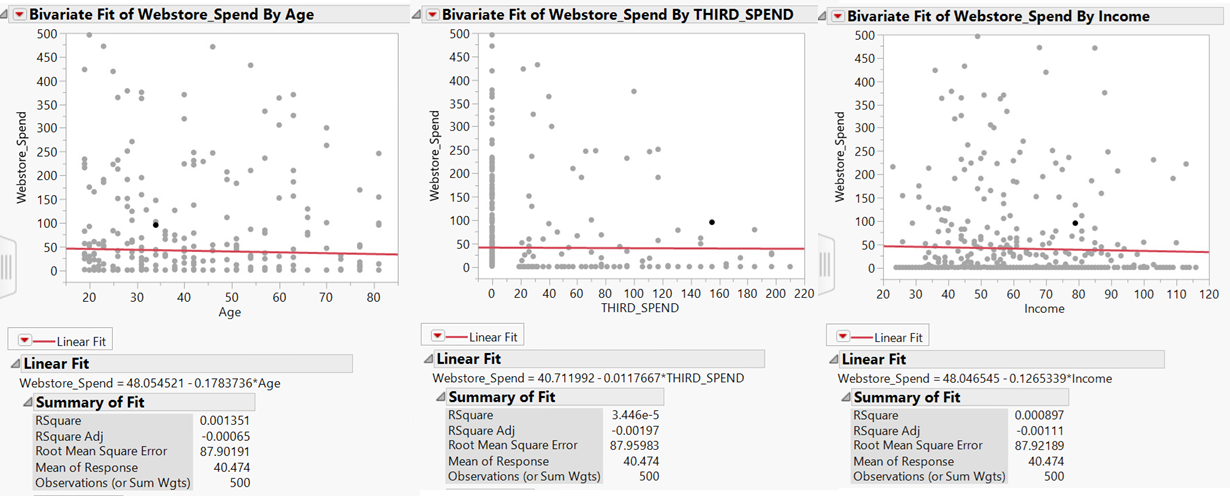
1. **Inaccurate Depictions of Data**

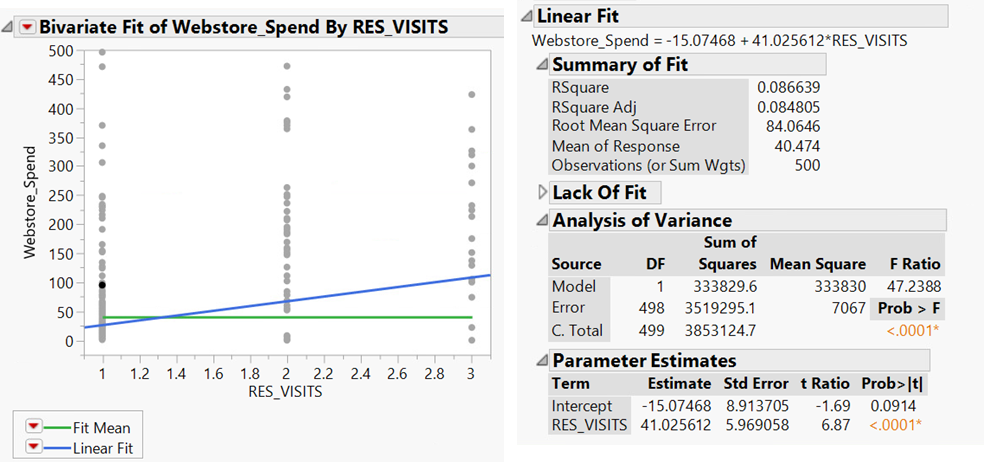
An inaccurate depiction of data can be seen from the logistic representation analysis, specifically from customer spending. From the data represented, there were higher misclassification rates. Thus concluding that not enough information could be derived from this analysis. This could be due to the lack of variables in this logistic analysis, and another analysis should be performed so the correct and more accurate variables can be analyzed. The solution to resolve this is to exclude customers who have not spent anything so that a more accurate result can be generated.

1. **Alternate Analytic Methods**

Linear regression can be an alternate analytic method because it can derive more individual data. Also, a linear regression can display results such as a fit line, mean status, and a regression line. As an example, the linear regression analysis of web store spending (Y-Axis) by the number of restaurant visits (X-Axis) that a customer makes is very helpful. The analysis displays that there is an increase of web store visits.

Clustering data can be a useful Analytic method because it analyzes more variables than two. Also, clustering data uses hierarchical clustering to group different variables, producing more accurate results. Clustering can provide a visualization of how this data is portrayed, and what the end user can derive from these visual results.

1. **Display and Interpretation**



1. **Validity, Reliability, Limitations**

One limitation of this report is the amount of spending. The user that have not spent anything on the web-store are not included in this analysis, making the data very difficult to interpret. In order to prevent errors in data research, all three aspects are very important because one can make sure the company has the most accurate information possible. For example, in the data analysis set of web-store spend by Third party Spend, it can be interpreted that someone who spends through a third party is most likely going to spend on the web-store.

1. **Resulting Decision Influence**

One way to facilitate decision making is to prepare all the data points, and hold a meeting to discuss the results. This can be at first with a superior to discuss the results from the survey-takers, and analyze what changes need to be made. Or, another resolution is to analyze how the web-store profits are competing against the Third Party Spending, and see what changes need to be made. Based off these model results, it is clear that online spending can play a larger role in company revenue. Bubba Gump can host more advertisement for their online web-store to engage their customers to also begin shopping online.

1. **Visual Evaluation**

These results were presented well because they display a fit line. With a fit line, the user can indicate the trend of business when comparing these two variables. For example, in the figure that displays the web store spending (Y-Axis) by the number of restaurant visits (X-Axis) that a customer makes. As seen in the analysis, the number of web store visits increases.

1. **Next Steps**

One next step could be to further analyze the data from the 500 survey takers, and de-bug any possible errors in the analysis. This is to make sure that all the data is completely accurate when presented is any meetings. A possible new hypothesis would be: using the research data from the 500 survey takers to re-make and re-advertise and web store to new clients.