8-2 Final Project

Angela Ruhstorfer

Southern New Hampshire University

DAT-220: Fundamentals of Data Mining

Professor Singapura

April 25th, 2021

**Business Problem**

The overall business problem of this Bubba Gump Shrimp Company is: “Sales have declined in the last two years. How can we better understand Bubba Gump Shrimp Company’s customers and look for ways to create new revenue growth?”

**Analytic Method**

Various data mining techniques that will be used are Linear and Logistic Regression analysis as well as Cluster Analysis. “Linear regression is used to predict the continuous dependent variable using a given set of independent variables…while Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables” (Linear Regression vs. Logistic Regression – Javatpoint, n.d.). The ‘x’ variable is the predictor, while the ‘y’ variable is the response/outcome. (PennState Eberly College of Science, 2018). As mentioned, Cluster Analysis is another technique used. “Cluster algorithms try to choose the groups in such a way that the structural qualities of the groups are very different to each other whilst being very similar within the groups. Cluster analysis uses all the input variables to ‘learn’ the structure and determine the groups” (Ahlemeyer-Stubbe, 2014).

**Analysis Tools**

The data mining tool I will use to perform the analysis is JMP Pro 15. I am using this data mining tool because it was the one provided to use for this course and I am not familiar with any others. It is difficult to determine the relative strengths and weaknesses of JMP Pro 15 when we do not have another data mining tool being used in the class to compare it with. I did some research of user reviews of JMP Pro 15 and most noted that it is a fantastic tool for quick data visualization and detailed analysis as well as another plus of it being Microsoft Excel compatible for exporting data. I believe that these features will be great advantages when using JMP Pro 15 to perform the analysis on the Bubba Gump Shrimp Company data.

**Data Visualizations**

The data visualizations I will use in my report are pie charts, bar charts, box and whisker plots, and scatterplots. I believe pie charts would be helpful in figuring out the percentages of how many customers are in each state. Bar charts will be helpful in visualizing the number of visits per visit type (Restaurant, Web, or Third Party Vendor). A scatterplot will help visualize the amount of money spent against income. A scatterplot could be made for each visit type with these two variables in mind. A box and whisker plot will help in visualizing the ages of the customers.

**Research Question**

I want to note here that the rubric asks for a research “question” however, in the feedback provided on Milestone One, I was told to present this as a Null Hypothesis. I am not sure which I am supposed to provide, as I believe a Null Hypothesis is not typically written in the form of a question, so I am providing both:

Research Question: Do the variables of customer age and income impact the dollar amount spent on the webstore?

Null Hypothesis: If the income of a customer is greater than 60 and age is greater than 30, the amount spent online will be more than those of income less than 60.

**Research Measurement**

I will determine if my research question was answered or if my hypothesis-generation was successful by comparing the income and age of the customers against their money spent for all three visit types. I will measure progress by creating data visualizations of my findings. As mentioned, a scatterplot will help visualize the amount of money spent against income. A scatterplot could be made for each visit type with these two variables in mind.

**Follow-Up Questions**

Does location have an impact on the dollar amount spent? I believe this question should be asked alongside the original research question, as housing in different locations has an impact on how a customer would spend their income. For example, someone who makes 60 in Georgia (GA), where housing is relatively cheaper, would spend their money much more differently than someone who makes that same amount in California (CA).

**Research and Support**

I was not able to find much in terms of outside resources on Bubba Gump Shrimp Company’s data. Most of the database sites I visited required a purchased subscription in order to view their information on the company. One I did find was Dun & Bradstreet , which mentioned that Bubba Gump Shrimp Company’s annual revenue so far in 2021 has been $10.14 million USD. The other information provided, like number of employees, is unfortunately not valuable data in the analysis we are trying to complete. The Dun & Bradstreet site does not provide really any information other than their current revenue of the fiscal year. I will be honest and say I could not find any published resources that address my line of inquiry for this project.

**Analysis Organization**

The formulated question is: “Which customers are likely to shop at the web channel and how much they are likely to spend?”. Resource evaluation is that we were provided a sample of 500 customers from Bubba Gump Shrimp Company’s analytics data warehouse. I do not believe the initial question needs to be reformed, as it is still appropriate to apply the question based on the data evaluation. Deeper understanding of the data: examples data categories I would consider ordinal and can be sorted / grouped are age and income. Data categories I would consider nominal would be WEB\_PURCH\_YN and MARRIED\_YN. From the analysis results completed on the data set provided, the average amount spent online is $40.47. The Age range that visited the site the most were ages 31-40 (Age Bin 4). The Income range that visited the website the most was 71-80 (Income Bin 8).

**Sources of Error**

I wanted to mention an error, or should I say, missing piece of data, I noticed in the Example Analysis, for "Visits per channel" graph. I noticed that the Web visits and third-party visits had a "0 visits" row, but the Restaurant visits did not have this row. Even if the Restaurant visits had 0 respondents for this row, I believe it should still be recorded here, so we know for sure that there were zero respondents that never visited the restaurant. I look at this graph now, and it looks like it is missing that data. I believe there should be some clarification on the Income column; is this Net Income or Gross Income? I believe this data can affect the income of the average customer for the Bubba Gump data. In the Bubba Gump data, I see a column named "Restaurant". Is this spending done in the restaurant? If it is following the same nomenclature as the "Webstore\_spend" and "THIRD\_SPEND" columns, then this column should be edited to "Restaurant\_spend" for clarification.

**Meaningful Patterns**

When completing last week’s analysis, looking at the Dendrogram and cluster means, it appears the largest cluster is #5, while the smallest cluster is #2. From the Parallel Plots created, it appears that data points are intersecting in cluster #2, while all the other cluster’s bars seem to go in the same direction. From the analysis results completed on the data set provided, the average amount spent online is $40.47. The Age range that visited the site the most were ages 31-40 (Age Bin 4). The Income range that visited the website the most was 71-80 (Income Bin 8). From these results, I would say the best way to increase Web Sales is to focus advertisements on these groups

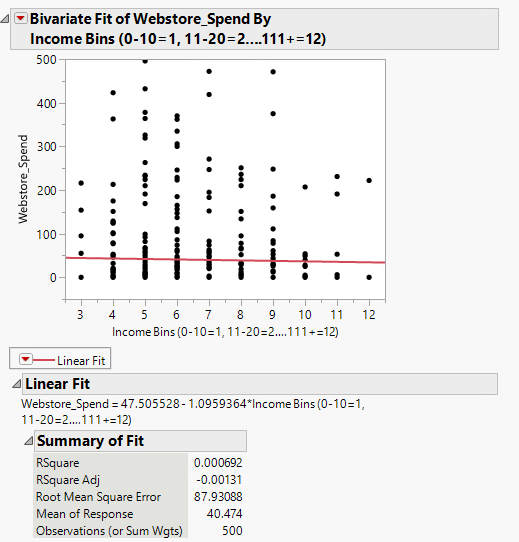
**Inaccurate Depictions of Data**

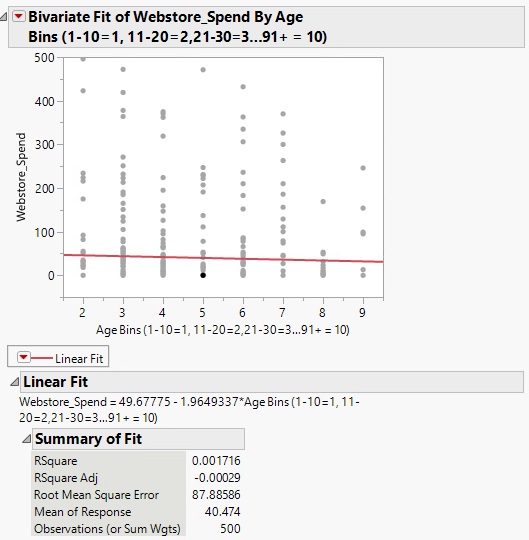
I did not find any inaccurate depictions of data, besides the sources of error mentioned prior. From professor feedback, my cluster analysis did not seem to be done properly. I was able to correct this by editing the variables chosen to create the cluster analysis, yielding more accurate results.

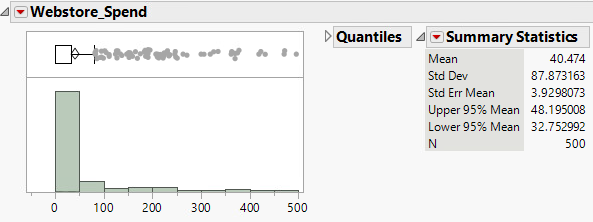
**Alternative Analytic Methods**

Another alternative method that could be used here is group purchase method. From Ahlemeyer-Stubbe, “Group purchase methods analyze what products, services or activities customers do together. It is also referred to as market basket analysis” (2014). For example, we could use this method to determine how high of a chance a customer might purchase from the webstore versus the restaurant and third-party? From Ahlemeyer-Stubbe, “A common feature of all rules [in group purchase methods] is that they relate one element of the receipt to one or more other elements” (2014). I believe this alternative method would be helpful because it relates to the type of data we are analyzing, customer behavior.

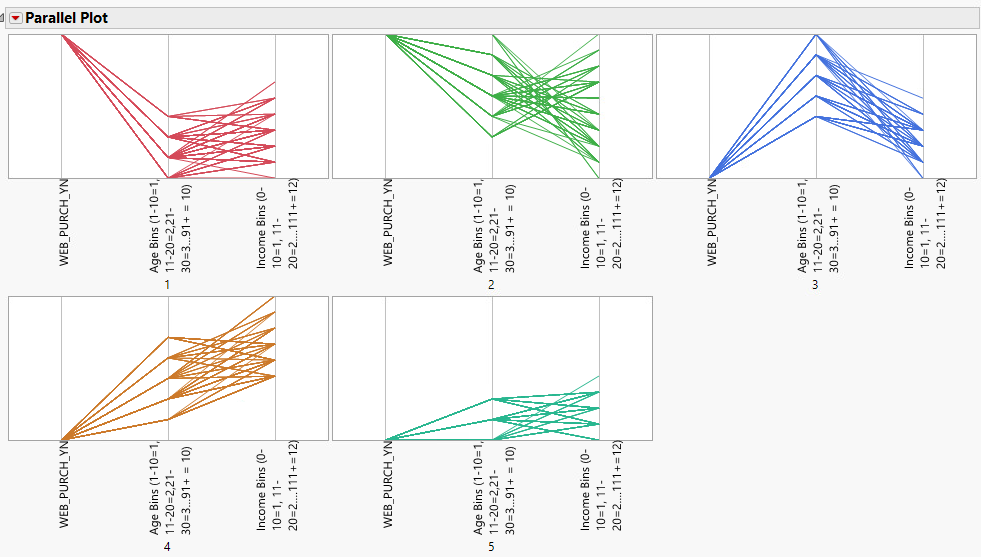
**Display and Interpretation**

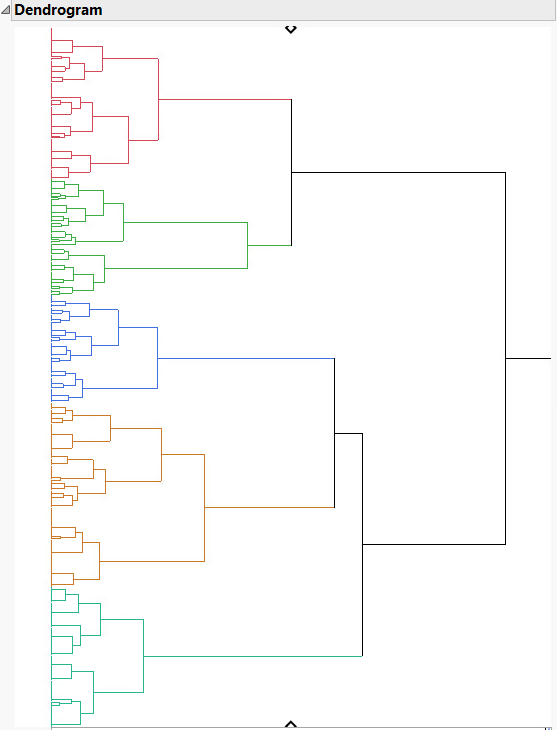


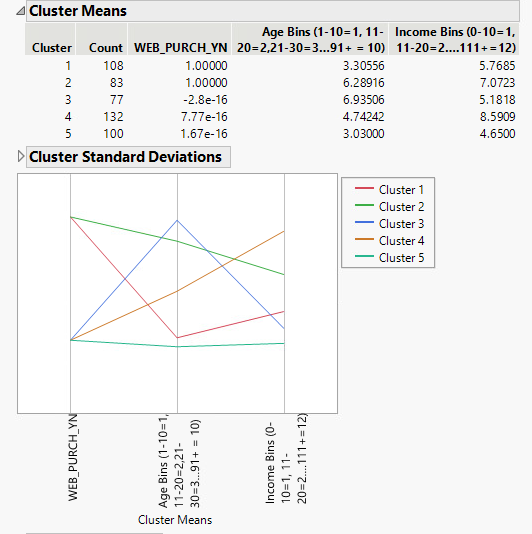




WEB\_PURCH\_Y/N vs Age Bins vs Income Bins Clustering:







From the analysis results completed on the data set provided, the average amount spent online is $40.47. The Age range that visited the site the most were ages 31-40 (Age Bin 4). The Income range that visited the website the most was 71-80 (Income Bin 8). From these results, I would say the best way to increase Web Sales is to focus advertisements on these groups.

**Validity, Reliability, Limitations**

I chose the displays because these were the ones taught to us in this class. There may be other displays that could prove useful within JMP, but we did not learn them. I chose the cluster analysis because it shows a relationship or pattern between multiple variables, in this case, Web purchases, web spending, age, and income. I chose a histogram to show the average money spent on the webstore. The Fit Y by X graphs show us a Line of Fit, which matches with the results of the clustering done.

**Resulting Decision Influence**

As previously stated, the Age range that visited the site the most were ages 31-40 (Age Bin 4). The Income range that visited the website the most was 71-80 (Income Bin 8). From these results, I would say the best way to increase Web Sales is to focus advertisements on these groups since these appear to show the most activity online.

**Visual Evaluation**

From the regression graphs (Fit Y by X graphs), we see the linear fits for Webstore Spending vs Age as well as Webstore Spending vs Income. Both show that the Webstore Spending average does not appear to go over $50.00. I did a Histogram to show the average spent in the Webstore, and it is under $50.00 at $40.47.

**Next Steps**

Since we now have the groups that visit and spend money at the webstore the most, I would say next steps for the company is how do they reach these groups of customers? The best way to figure out this data (and honestly would have been nice to have in this assignment) is to ask these customers “how did you hear about us?” with specific choices they can choose from (i.e. Social Media, TV Ad, Physical Mail, etc.). That way, they can gather their responses and push out the best form of advertisement to these groups aimed at increasing Web Sales.

**References**

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

Bubba Gump Shrimp Co. Restaurants, Inc. (n.d.). Dun & Bradstreet. Retrieved March 21, 2021, from https://www.dnb.com/business-directory/company- profiles.bubba\_gump\_shrimp\_co\_restaurants\_inc.3e79205ac65c7df97b8594bda4ec6b9e. html

*Linear Regression vs. Logistic Regression – Javatpoint.* (n.d.). Www.Javatpoint.com. Retrieved March 28, 2021, from https://www.javatpoint.com/linear-regression-vs-logistic- regression-in-machine- learning#:%7E:text=Linear%20regression%20is%20used%20to,used%20for%20solving %20Regression%20problem.

PennState Eberly College of Science. (2018). *2.1 - What is Simple Linear Regression? | STAT 462.* The Pennsylvania State University. Retrieved April 3, 2021, from https://online.stat.psu.edu/stat462/node/91/