Final Project

Lesley Potts-Langdon

DAT 220

Dr. Larson

April 29, 2018

**Plan for Analysis**

Bubba Gump’s situation is unusual in that the reason for is known for their sudden jump in their business but the reason for the decline is not known. From a business perspective it would have been nice to gain new customers from the name recognition and be able to keep them with great food and service but that may not have been realistic. The problem Bubba Gump’s would like to solve is the decline in business. They built new stores and invested in new retail opportunities based on their growth. When business leveled off that was to be expected but now with sales declining over the past two years they need to better understand their customers to develop a plan to increase revenue.

The objective is to better understand the Bubba Gump Shrimp Company customer. Identifying characteristics of the typical Bubba Gump customer is the first step to making a plan to increase loyalty in the existing customers and to draw new ones. The data from the loyalty programs, web sales and the satisfaction surveys will be used to compile the best overall picture of the 500 customers that have received the surveys. This data will be analyzed to look for clusters within the customer data. Once clusters are identified they will be analyzed more closely to determine the significance of the clusters. The next step will be to make some comparisons between the individual characteristics of the clusters. Finally presenting a comprehensive report to present to Bubba Gump Shrimp Company. Depending on the findings there may be similar clusters that could become, with some incentive, members of a more desirable cluster. For example, there could be two clusters which are identical with one exception, one cluster makes high dollar purchases from third party vendors and the other does not make any third-party purchases. It may be valuable to see if there is something to be done to expose that second group to third party vendors. Perhaps there are none in their area, it could give the decision makers at Bubba Gump some new locations to try to partner with some third-party vendors to increase revenue.

To complete this analysis there are several statistical tools to choose from, for this project JMP has been selected. Several other software packages were considered. JMP and SAS Business Intelligence were both developed by SAS. JMP has file management and statistical process control features while SAS Business Intelligence does not. JMP works on Windows or Mac and SAS Business Intelligence will only work on Windows. DOMO was considered but that is only available through a cloud server and cannot be installed. Finally, Tableau was evaluated but that is more geared towards internal data mining activities as opposed to an outside party analyzing other companies’ data for them. (statistical analysis software/compare, n.d.) JMP is the best software for this project because it will allow the data to be easily navigated and then produces clear and informative visuals for the final report.

Several data visualizations were used to assist in looking for patterns but the ones selected for the report will have to be both relevant, easy to read and to understand. A scatter plot showed a higher concentration of visits in the lower numbered restaurants even though the higher numbered restaurants had more visits per location. Principle Component Analysis shows correlations between data points. That information can lead to looking more closely for clusters. The final report will use either linear or logistical regression models to clearly show relationships between variables as well as any supporting graphs needed to show the client clusters they can target in order to increase revenue.

In general terms knowing what patterns exist in customer behavior can be used to target efforts and increase revenue at Bubba Gump Shrimp Company. They want to know if attention should be focused on obtaining new customers, bringing customers in for repeat visits to restaurants or web and third-party sales.

Specifically, the analysis will show what characteristics profitable Bubba Gump Shrimp Company customers have in common. The factors considered will be whether the customer is local or traveling, total number of purchases and from where, age, marital status and income.

The research question will have been answered if once natural clusters are identified within the sample group of Bubba Gump Shrimp Company customers. It is also critical that the executives at Bubba Gump Shrimp Company feel they are able to make sound business decisions based on the information provided. The goal is to determine who should be targeted with advertising to increase revenue. An increase in revenue cannot be guaranteed but if in the end specific groups that should be targeted to achieve a specific outcome can be identified the analysis will have been successful in achieving clients goal.

There are some follow up questions to help better serve the client. It would be helpful to know where the customers live in relation to the location they visited. One hypothesis is that the lower numbered restaurants see fewer repeat visits because customers dined there while traveling and it would be valuable to be able to confirm that. It would be helpful to have access to current sales and profitability information for each location. In looking at natural clusters one of the goals is to target customers that will fall into a desirable cluster. It would be a stronger analysis if it could be confirmed that customers determined to be in desirable clusters are visiting profitable restaurants.

Additional research will be conducted using Bubba Gump’s internal systems to obtain restaurant addresses and profit and loss data. A mapping program will determine the distance between the restaurant location and the customer’s home address. Census information may be used to compare the average Bubba Gump customer with national, state and local averages. This information should be everything needed to answer the follow up questions.

**Analysis**

An organized approach to analyzing this data begins with a cluster analysis as it identifies some areas that were worth further evaluation using other data mining tools. The first analysis was ran using all the data. Before attempting any analysis, it was helpful to bin the age and income data. Had that not been done there would have been way too many clusters to be able to analyze effectively. Even after making these adjustments using all the data in one graph was too cluttered to recognize any valuable patterns. Breaking the data down into two groupings both including webstore visits and purchases was the next step. The first was purchase information including the webstore data, restaurant visits and third-party visits. The second was demographics, income level, age and marital status along with the webstore information. In looking at the two graphs there was one group that stood out apart from the rest, webstore customers who visited a restaurant once and a third-party vendor multiple times. It was determined this could be better looked at in the linear and logistic regressions. In looking at some of the individual data points as compared to web purchases an interesting trend in the income data was uncovered. As the income bracket went up the number of purchases went down. There were 79 purchases from the lowest bracket, 46 in the second, 34 in the third, and 32 purchases in the highest bracket. That information may or may not be significant but it warrants a closer look. The value of the cluster analysis was in creating new questions to ask using other tools.

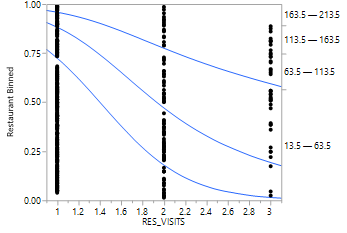
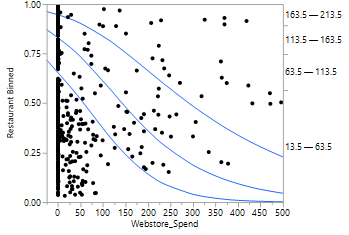
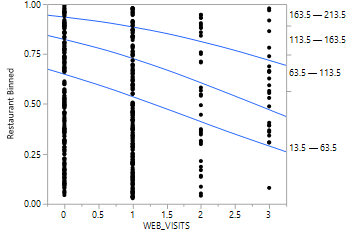
To get another look at the information found in the cluster analysis linear regression was the next step. A multivariate linear regression showed all the relationships. The first look was at the customers with multiple visits to third party vendors but what actually stood out was the comparison of webstore spending to restaurant visited. The overwhelming majority of webstore purchases come from customers who visited a low numbered restaurant and those who visited higher numbered restaurants only made large webstore purchases. Linear regression confirmed that the majority of webstore purchases were made by people in the lower two income brackets. As far as the cluster of customers that visited a restaurant once and third-party vendors multiple times, linear regression did not offer any additional information on this group except to show that a distinct minority of the customers who made webstore purchases visited third-party vendors multiple times. It is not yet clear whether this fact makes the cluster more or less significant. The logistical regression model will provide another perspective.

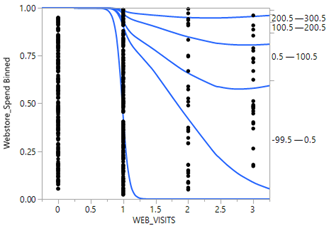
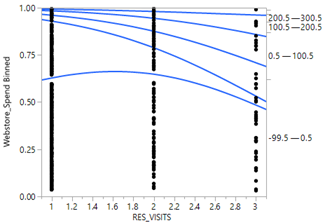
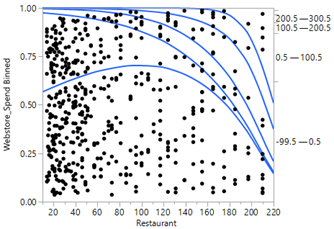
The logistical regression model was used to attempt to gain more information showing relationships between webstore purchases and visits to third-party vendors, restaurant visited and income bracket. Even though most of the customers who made webstore purchases visited a third-party vendor zero or one times a customer who visited a third-party vendor multiple times was more likely to make a webstore purchase. A similar trend is found when looking at restaurant numbers, there is a dense group of customers who made webstore purchases and visited lower number restaurants, however, customers who visited higher number restaurants were more likely to make webstore purchases. The relationship to income bracket was confirmed, a customer in a lower income bracket is more likely to make a webstore purchase than one in a higher bracket.

The progression from one model to the next was an organized, stepwise approach. While additional information was gained on each new question there are still some factors that can’t be fully understood. Even though not all answers are clear there are meaningful patterns. Some of the depictions were skewed because they included zero values others were hard to analyze because there were too many values. Binning values helped to improve the analysis. Webstore purchases were binned with the zero value in its own grouping so it is separated on the graphs. Age and income were binned to reduce the number of values making patterns clearer. Based on the results the next step would be to take a closer look to determine which factors combine into clusters that will be valuable target groups in order to increase revenue at Bubba Gump.

**Final Report**

After looking at the data from several different angles using different modeling techniques it became apparent that there is a significant difference in the behavior of customers that visited the stores numbered in the lowest quartile and the customers that visited the rest of the restaurants. In taking a closer look, this difference is most seen in how often they visit the restaurant and in their webstore habits. I will focus on how we can approach these two groups to increase revenue.

The best way to visualize the different behaviors is through logistical regression modeling. It shows us the likelihood that customers in certain categories with behave a certain way. Looking at these three graphs the restaurant numbers broken into quartiles are represented by the numbers on the left and right sides. The value being compared is across the bottom. In the first graph the comparison with the number of restaurant visits is shown. The guest who visited the oldest restaurants were most likely to only visit once while as the restaurant number gets higher so does the likelihood of a repeat visit. The second graph shows a comparison with webstore spending. It is a little skewed because it includes guests that do not spend on the web as all. The same thing is seen in the third graph which compare restaurant numbers to the number of webstore visits but includes customers that have never visited the webstore. In the next set of graphs, looking at webstore spending compared to other factors and will give a better picture in some cases because those customers who have not made webstore purchases are separated out through binning.

As we look at the graphs showing webstore spending the levels of spending are reflected in the numbers on the left and right and the value we are comparing to is on the bottom. The group below the lowest blue line represents the likelihood that a customer did not spend anything at the webstore. The space between the first and second blue lines shows a large group of customers who visited a low numbered restaurant, a large group who only visited a location once and another who visited the webstore once or maybe twice. Hierarchical Clustering is a great way to see if the groups are related. What was found is that while the largest cluster, number 23, is comprised of customers who visited a low numbered restaurant once but did not visit the webstore at all the next two largest clusters fit the groups of interest. Cluster number 17 is the second largest with 84 customers. These customers visited a low numbered restaurant once, visited the webstore once and spent a small amount of money while they were there. The next group, cluster 13 with 48 customers, represents those who visited a low numbered restaurant once and visited the webstore once but did not make a purchase. This data provides two solid categories of customers that Bubba Gump Shrimp Company can target with two different campaigns in their efforts to increase revenue.

Although some data is missing to say definitively there is enough to surmise that there are a significant number of customers who have only visited one of the original locations once because those restaurants are in touristy areas and frequented by people on vacation. A significant number of those guests enjoyed it enough to seek Bubba Gump Shrimp Company out on the web. For customers who visit the lower numbered restaurants a campaign promoting the webstore could work towards the goal of increasing revenue.

Customers who visited the higher numbered restaurants were much more likely to return for repeat visits. A campaign encouraging additional visits to their local restaurant could lead to an increase in revenue. These guests, however, should not be left out of the campaign encouraging webstore activity. When guests who visited higher number restaurants visit the webstore they make larger purchases.

The biggest factor effecting the validity and reliability of the report is the limited data supplied. Addresses for restaurants and customers could confirm the hypothesis that customers visited lower numbered restaurants while on vacation. That assumption is a large part of the analysis and not having the location information affects the reliability.

The report included recommendations to the client. Restaurant sale data would have been beneficial to present alongside this information. If there are specific underperforming locations it would be helpful to look at their specific performance as it relates to the survey data.

The logistical regression models are the primary visual for this report. They show the likelihood that a customer will behave a certain way and they are fairly easy to read and understand. When looking at the clusters there was not any benefit to showing the dendrogram, what was really needed was the chart showing the number of customers in each cluster and the characteristics of that cluster.

The next step would be to take a look at the sales and profitability data for each location. The report relies heavily customer attributes. An analysis of individual restaurant attributes could move Bubba Gump Shrimp Company further towards the goal of increasing revenue.