**DAT 220 Final Project**

**By Alexander Varljen**

**June 25, 2021**

**Plan for Analysis:**

1. Background and Introduction:
2. Business Problem:

The “Bubba Gump Shrimp Company” has experienced unexpected rapid growth in recent years, as customer demand was quite high do to a famous movie that featured the company. Over time, the business has amassed a great deal of customer data from its own internal systems, as well as from third party retailers. However, its growth has stagnated in recent years. Therefore, the business problem is to restart growth utilizing the data available to the business via the data analytic tools at our disposal.

1. Analytic Method:

The purpose of the analytic strategy that will be used is to integrate customer data from the Bubba Gump Shrimp Company’s enterprise data warehouse, which contains data from point-of-sale, its customer database, and its webstore sales transactions, as well as data from third party retailers, to analyze the complete customer experience. The end goal of the analysis is to find patterns in the customer experience via the observation of “clusters” via the data available. The discovery of these data clusters can then give the company insight into past customer patterns of behavior and allow the company to restructure its business to meet the needs of the consumer and, thus, hopefully restart the company’s growth. The information that the analysis will yield will be visualizations of any patterns and clusters that may be received from analyzing the data, which is produced by a sample size of 500 of the past customers, who have taken survey’s, along with the historical customer data mentioned earlier.

1. Tools and Visualizations:
2. Analysis Tools:

One analysis tool that will have a lot of utility is row selection and deletion. In this dataset, there is a large enough portion of the consumer data that has 0 visits to either the company web site or third-party vendors that it heavily skews the data, making useful measurements difficult. Another useful tool will be measuring quantiles. This allows for the break-up of data into useful chunks by percentages of value. These tools, as well as others, can be utilized with JMP, which is the software application that provides data mining and analytics tools and features, and will be the software application that I will be using to conduct my analysis.

1. Data Visualizations:

One useful visualization for the purpose of detecting patterns in consumer data is the histogram. This allows the viewing of the total number of occurrences of values that are near one another in value. For example, if one is measuring the number of purchases that are made that range from 1 to 100 dollars, a histogram might break up the number of purchases into groups of ten dollars from 1 to 100 to display a simpler visualization of an overall pattern, rather than the frequency of all amounts spent dollar by dollar. Another useful visualization could be a box plot. This allows for a visual representation of quantile data, including the mean, the middle fifty percent of an occurrence and any outliers. Another useful visualization that will be deployed in this analysis report are linear regression graphs. These take two continuous variables (e.g., “Amount Spent at the Restaurant” and “Amount Spent at the Webstore”) and see how they relate to one another. Another visualization that will be deployed in this analysis report are logistical regression graphs. These take a discrete variable (e.g., “Marital Status”, which is yes or no) and compare it with a continuous variable. Another visualization that will be deployed in this analysis is a scatterplot matrix. This can be used to create several linear regression graphs for all possible combinations of the continuous variables chosen. Another visualization that will be deployed is a hierarchical clustering dendrogram, which takes a discrete variable and groups customers by their relation to these variables with other “branches”, which start out as lines pointing to each amount of the discrete variable. For example, a hierarchical clustering dendrogram that uses “Web Visits” as its discrete variable will have a line coming from each variable. Then, the clustering process combines these lines into a “cluster”, which is the group of customers which have the closest relation to each other in terms of, in this example, the number of “Web Visits” they have made.

1. Specific Research Question:
2. Research Question:

My research question is: How can data clusters be analyzed in a manner where useful consumer spending patterns are discovered?

1. Research Measurement:

I will consider my research question answered if I am able to find past research that uses the analysis of clusters that is related to sales and the amounts spent on different products or services.

1. Follow-Up Questions:

One cogent follow-up question could be concerning whether an example used in a study has any close analogy to the kinds of sales and marketing problems that I am tasked with solving. Another question that could arise in my research would be whether I have the same tools at my disposal that were deployed by those conducting the study.

1. Research and Support:

I found one paper on car sales analysis that may be useful. It presents a general pattern of methodology for mining sales data, which is related to the task that I am pursuing for the Bubba Gump Shrimp company (Zhang et al., 2017). The drawback is that the information seems more general and does not talk about the tools that were used. Another interesting paper that I have discovered contains a study of determining consumer knowledge about a product or service based on sales patterns (Liao et al., 2012). The paper discusses techniques related to clustering patterns of consumer purchases and finding the relations between purchase decisions. The drawback is that the paper seems very advanced and uses some techniques and tools that I do not believe are available to me. However, there seems to be some simpler techniques used for clustering called “k-means”. Also, I have discovered that our eBook from MBS Direct has a chapter on cluster analysis in chapter six, which seems quite useful (Ahlemeyer-Stubbe, 2014).

**Analysis:**

1. Analysis Organization:

In my analysis, I used hierarchical clustering, linear regression, and logistical regression models to analyze customer behavior with respect to the Bubba Gump Shrimp web store. For the hierarchical clustering, I analyzed clusters of customer web store, as this seemed to be the simplest way to use hierarchical clustering to analyze customer web activity. For my linear regression models, I picked the continuous variables that seemed like they would have the greatest correlation with web store spending. For instance, I compared income vs. amount spent at web store. For my logistical regression models, I picked one binary that I thought was most related to the web store (i.e., “web purchase: 1 for yes and 0 for no”) and compared it to restaurant spending, as, after conducting linear regression, I realized that this was the variable that showed the most promise, as well as two other untried continuous variables.

One major limiting factor was that I had to be judicious with which variables I would present in the final analysis, as a more thorough analysis may have resulted in a large and, possibly, redundant number of hierarchical clustering, linear regression, and logistical regression models. Another limiting factor was that some potentially valuable data related to web store visits would have been hard to model with these tools, such as comparing the amount spent at the web store vs. customer zip codes, which could be useful in region specific marketing.

1. Sources of Error:

Much of the potential sources of error seem to be related to values of zero presented in the linear regression models. The most extreme example of this is in the model for “Amount Spent at the Web Store vs. Income”, where it appears that close to half of all the data points are at 0 on the x-axis, which likely has an outsized impact on the linear equations for the best fit line for the graph. In response to this, I decided to focus my attention on the linear regression model that showed the strongest and clearest positive correlation between two variables, which was that of “Amount Spent at the Web Store vs. Income”.

1. Meaningful Patterns:

I have discovered a strong positive correlation between the amount spent at a restaurant and the amount spent at the web store. I have also discovered a strong correlation between a customer being younger and web store purchases. I have also discovered that the most visits to the web store that a single customer has made is three visits. I believe a fruitful research question to ask would be one related to customer age.

It is not clear from the data set what age group should be targeted for marketing, as it relates to customer spending, but only that the younger a customer from the sample is, the more likely they are to make a web purchase. Perhaps there is a common practice that I am unaware of pertaining to how to generally group customer ages for business purposes (i.e., groups labeled “millennial”, “18 – 25 years of age”, etc.), or that it would be more advantageous to find groupings of age based off this survey sample, in and of itself.

1. Inaccurate Depictions of Data:

My main concern for a potentially inaccurate depiction of data in from both this analysis and the sample data in general is that 100% of the customers have made restaurant purchases. This means that the data set may be excluding customers who may, for example, make web purchases but who have never visited a restaurant. However, analysis of the data, in this case, does indicate that the more one spends at a restaurant, the more one spends at the web store. However, as the business problem relates to starting growth once again, I believe it is worth investigating which customers are likely to make web purchases without visiting a restaurant. It may be that, since there is a positive correlation between younger age and web purchases, some younger customers are willing to make web purchase without visiting the restaurant.

1. Alternative Analytic Methods:

The Group Purchase Method of Association Analysis may help give more accurate measurements for some of the positive correlations already found via the logistical and linear regression methods and the hierarchical clustering used for the analysis, so far. Group Purchase Methods can help in the analysis of what types of products are bought together by customers. Association Analysis can show us the probability that another product or service will be used by the customer based off the purchase of another product or service. This can be useful for putting a more concrete probability between, for example, in-person restaurant purchases and web store purchases and any other purchases in relation web store purchases. This can help the company not only find those who are more likely to make web purchases, but also what customers are likely to purchase together, thus making it possible to market multiple products to customers instead of focusing on one product or service at a time.

**Final Report:**

1. Display of Results and Interpretation:

1.a.: Correlations Between Restaurant Spending, Web Store Spending, Third-Party Spending, Age, and Income:

**Calendar

Description automatically generated**

1.b.: Linear Regression Graph of Amount Spent at Web Store vs. Amount Spent at Restaurant:

Chart, scatter chart

Description automatically generated

Restaurant = 62.700655 + 0.2837215\*Webstore\_Spend

1.c.: Linear Regression Graph of Amount Spent at Web Store vs. Income:

Chart, scatter chart

Description automatically generated

Income = 60.132917 - 0.0070889\*Webstore\_Spend

1.d.: Logistic Regression Graph of Web Purchases vs. Age:

Chart, scatter chart

Description automatically generated

1.e.: Hierarchical Clustering Dendrogram of Bubba Gump Shrimp Web Channel Activity:

Histogram

Description automatically generated with low confidence

Above are five visual displays labeled 1.a. through 1.e. The first display, 1.a., is a scatterplot matrix showing the correlations between Bubba Gump Shrimp Company customer restaurant spending, web store spending, third-party vendor spending, age, and income. The second and third displays, 1.b. and 1.c., show the relationship between web store spending and restaurant spending, as well as the relationship between web store spending and income, via simple linear regression graphs. The fourth display, 1.d., is a logistical regression graph that shows the relationship between whether a customer has made a web purchase and their age. The fifth display, 1.e., is a hierarchical clustering dendrogram that shows clusters between customers by the amount of web store visits each customer has made, which range from zero to three visits. From these graphs, we can see that the more someone spends at an in-person restaurant, the more a customer will spend at the web store, and that this is our strongest indicator of web store spending. We also find that an increase in income among customers does not cause and increase in web store spending. We find that there is an inverse relationship between the age of a customer and their willingness to visit and purchase products from the web store (i.e., the younger a customer is, the more likely they are to make web purchases). We also find that there is a second order cluster around those who have made two and three visits to the web store. In summation, these findings indicate that the target group for the sale of web store products are younger customers who have spent more at restaurants and made at least two web purchases thus far.

1. Validity, Reliability and Limitations of Results:

The validation and reliability of this report lies in the customer survey sample, the tools used, the breadth of the analysis, and the reliance on evidence as opposed to speculation. First, the data used was taken from a survey of five-hundred customers and contains several of the categories that the company itself thought where most relevant. Next, the data was analyzed with a variety of legitimate analytical tools, ranging from a scatterplot matrix to a hierarchical clustering dendrogram. Also, all the continuous variables were analyzed and compared to each other and then displayed, with special emphasis on the categories most relevant or interesting to the analysis. Also, the analysis preceded the conclusions, as opposed to any specific preconceptions preceding the conclusions.

One of the major limitations to this report was the sheer volume of displays that I could have included, but did not, due to convenience. To overcome this, I only included the visualizations that provided the strongest correlations and evidence of web store activity, or that where sufficiently compact enough that they would overload the reader with information. For instance, I did not include a graph displaying the relationship between marriage and web store visits, as no notable relationship was found. However, I did include a scatterplot matrix, as it was a compact way to display the relationship between all five continuous variables from the sample data, as well as a linear regression graph showing the relationship between web store spending and income, as I thought it presented an interesting implication (i.e., that those customers that can spend the most seem to be no more likely to purchase from the web store than other customers). Another limitation was that the survey data itself was composed of customers who all made restaurant purchases. This may present some unobserved phenomenon in customer spending from those who have made purchases of company products without making a visit to the restaurant. However, the analysis of this data set does indicate that the strongest indicator of web store activity is increased spending at an in-person restaurant.

1. Resulting Decision Influence:

To convince a client or superior of my recommendations, I would first present all the relevant evidence of the analysis, as it relates to the business problem, in a condensed and legible format which leaves out redundancies. For example, In the visual displays above, while the graphs and dendrogram is precise, I only included a level of information necessary to draw concrete conclusions from the data, such as the graphs themselves, titles and a few linear regression equations useful for future predictive power. This assumes that the client or superior, while competent, does not need an overload of technical information to evaluate the evidence. I would also discuss some examples of analyses that I left out and the reasons why they are not in the report. For instance, I left out a logical regression graph that showed the relationship between marriage and web store purchases, as no real correlation was found. This would show my client or superior that, while more analysis was conducted, only the most relevant findings are presented in the final report. Finally, I would relay my recommendations, after the client or superior has already reviewed the evidence of my analysis themselves, while going over the connections between the individual components of the analysis with the client or superior, as well as discuss any further hypotheses for anomalies or blind spots in the data set (i.e., such as income not being positively correlated with web store visits, as seen in visual display 1.c.).

1. Visual Evaluation:

In this report, I have only included the five visualizations which I thought where most relevant. The first visualization (i.e., the scatterplot matrix) provides evidence that my other four visualizations are the most relevant. I have done this so that, while the audience I am presenting to can put more trust in the conclusions, they do not have to view redundant information. All other visualizations, besides 1.c., are based on the strongest indicators of increased web purchases. The visualization in 1.c. of the report shows an anomaly that can be discussed or researched at a later period by myself. I have tried to make all the visual displays as simple as possible (i.e., without any of the tables of measurements that can go with the visuals), while keeping the essential and useful parts of the display for the audience (i.e., titles, best-fit lines, etc.). This was all done to be precise while keeping the information both concise and understandable.

1. Next Steps:

In my report, I have included a visualization (i.e., 1.c.) that shows the relationship between income and web store spending. In the graph, income seems to have a negative correlation between age and web store spending. I thought of this as peculiar, as income has a direct impact on how much one spends. In response to this, I have come up with two hypotheses, neither of which are mutually exclusive. The first is that there is a positive correlation between income and age (i.e., as customers get older, they accrue more wealth), and, as was already discovered, age and web store use is positively correlated. This, I believe, can be tested through the analysis of the current survey data, as data for both age and income is included in the data set. The second hypothesis is that an increase in wealth makes things like the communal nature and atmosphere of a restaurant more valuable then saving money or time via the use of a web store. This kind of question cannot be answered with the current survey data, but an additional survey that includes a report on a participant’s income and a question like “Please rank these categories by importance to you”, which would be followed by a few categories like “Restaurant atmosphere” and “Total cost”. The data collected and analyzed for the validation of these two hypotheses could provide useful information about customer income related preferences, which can further improve the Bubba Gump Shrimp company’s ability to market their products to groups based on income effectively.

**References:**

Qi Zhang, Hongfei Zhan, Junhe Yua. (2017). Car Sales Analysis Based on the Application of Big Data. *Procedia Computer Science.* Volume 107, 2017, Pages 436-441, ISSN 1877-0509. Retrieved from

<https://www.sciencedirect.com/science/article/pii/S187705091730412X>

Shu-hsien Liao, Pei-hui Chu, Yin-ju Chen, Chia-Chen Chang. (2012). Mining customer knowledge for exploring online group buying behavior. *Expert Systems with Applications.*

Volume 39, Issue 3. Pages 3708-3716, ISSN 0957-4174.Retrieved from

<https://www.sciencedirect.com/science/article/pii/S0957417411013753>

Andrea Ahlemeyer-Stubbe, Shirley Coleman. (2012). *A Practical Guide to Data Mining for Business and Industry.* [MBS Direct]. Retrieved from <https://mbsdirect.vitalsource.com/#/books/9781118981863/>