Byron Laferriere

DAT-220

28 FEB 2021

Final Project

**Final Project Milestone One**

**Introduction:** ***Business Problem***  
The business problem that will be solved for Bubba Gump Shrimp Co. is where the decline in sales over the past two years has been happening. We will solve the problem by using the data warehouse information that the company has collected, to help provide analysis the company can turn into information that will help the company turn its declining sales around this year.

**Introduction:** ***Analytic Method***  
The purpose of using cluster analysis to solve business issues is to study customer spending habits across the three platforms that Bubba Gump Shrimp Co. operates through. Instead of having three separate entities, which were not communicating customer information, there will now be a data warehouse that stores all of this information and provides keen insight into where customers are spending money and possibly highlight ways to get them to spend more. Building profiles on customers will allow the company to begin sending coupons during the customers usual spending hours or holiday seasons, in hopes of increasing sales. Cluster analysis will help reveal which variable relativity is highest, so that action by the company may be taken accordingly.

**Final Project Milestone Two**

**Analysis Tools**

To perform this analysis, the JMP statistical software by SAS Institute will be used. The software has many capabilities like data visualization, statistical modeling, and data analysis tools that use histograms, regressions, and distribution fittings that will be needed to perform the cluster analysis for Bubba Gump Shrimp Co. By using these tools, large amounts of information can be easily displayed and interpreted, so that assumptions may be made. In particular this analysis will use histograms, dendrograms, and some regressions in combination with hierarchical clustering, to highlight exactly where the business is declining in its sales generation from year-to-year and how it can reverse this process. There are some weaknesses present in these tools, like regression modeling, which can be skewed with variable selection to enhance results. Therefore, we can only use these models as a basis to continue research further through more accurate tools that use clustering.

**Data Visualizations**

I will be using histograms along with quartiles, standard deviation, box plots, hierarchical clustering, regression models, and stem and leaf plots. The purpose of using this many statistical techniques to evaluate the data is to find things like normal distribution and patterns that can be recognized in customer spending habits. Standard deviation can tell us how often customers tend to behave in similar fashions or when they will go against the grain, when it comes to dining out at the restaurant or even online shopping. These types of behaviors can be affected by receiving coupons in the mail and these types of statistics can help highlight how affective this is. Using the clustering analysis will shed light to variables that have higher relevancy and can be grouped accordingly within the Bubba Gump customer population. Data visualizations are a useful way of portraying large banks of information, that need to be digested quickly by the viewer, which is why this report will include so many different types.

**Research Question**

The research question that is being addressed through this report will be focused on why sales have been decreasing over the past two years at Bubba Gump Shrimp Co. Using the data set provided, this analysis hopes to answer this question by extrapolating the necessary data from the new data warehouse created by the company. The question I will specifically focus on will be the transference of sales from restaurant customers that are dining in, to those who purchase merchandise after visiting a location. “Is there a traceable connection that can be made in sales generation, from giving a dine-in customer a coupon for online merchandise?” This will be the focus that I will approach this issue with.

**Research Measurement**

To determine if this research question is valid, I will need access to the data from all three outlets that Bubba Gump Shrimp Co. releases products through. From here, I will be able to combine data and see just how often customers who eat at the restaurant, also purchase merchandise from one of the outlets. If there is no traceable connection to be made, I will be forced to admit that this will not be a viable solution to the issue at hand. If there is a recognizable pattern that is present, it will verify that there is a chance to increase sales through coupon dispersion to customers visiting restaurant locations for the first time. Statistical validity and increased sales will measure the success of the research.

**Follow-Up Questions**

Some possible questions that could follow up the exploration into this data set could be:

“Is it going to be more beneficial for Bubba Gump Shrimp Co to switch to online retail only and follow suit with Amazon and other large online retailers?”

“How can customer loyalty be increased throughout each region?”

“Would more Hollywood exposure or commercials benefit the company again, like it did from the movie Forrest Gump?”

**Research and Support**

According to an article written in the Wall Street Journal, ever since the beginning of the coronavirus outbreak in America, online retail sales have been skyrocketing. Credit card usage for online retail shopping has increased 88% a month since April of 2020 (Pacheco, 2020). This research done highlights a need for Bubba Gump Shrimp Co. to continue its efforts into the success of their webstore and online retailers. The data analysis will conduct research into the current percentages of customers that purchase merchandise online, after visiting a restaurant location. If there is a connection that could be made, this could help Bubba Gump reverse its downward trending sales over the past two years. E-commerce is becoming an extremely large market to get involved with and this could be a good route for the company to explore further.

**Final Project Milestone Three**

**Analysis Organization**

When preparing to begin the analysis, there are two main types to consider that we have covered. They are the descriptive analysis and the predictive analysis. First, the descriptive analysis defines characteristics of the data set and answers questions like, “where does the problem lie”. This is the best way to begin an analysis, so that assumptions and predictions may be made in the predictive analysis. The predictive analysis uses tools like regression to create predictive models, forecasting, and simulations from the data set (Unknown, 2019). In this analysis, we have completed the descriptive analysis and have begun moving into the clustering and supervised learning methods where information is observed from both sides of the problem describing selected variables leading to a known target. This information will provide Bubba Gump with answers regarding the decline in sales over the two previous years. Once in the predictive analysis, unnecessary variables will be eliminated from the report to enhance the credibility of its results. The stepwise approach calculates variables that make a significant contribution to the model (Ahlemayer & Stubbe, 2015), which therefore validates that this analysis reflects an organized, stepwise approach and has been following the guidelines provided.

**Sources of Error**

There were some incomplete values that were included in the original data set, which could not be included in the results to obtain truly correct conclusions. The clustering seemed to struggle to find groups from the sample survey data that were closely entwined. This error could be due to the large number of variables present in the data set and trying to cluster meaningful information from columns of data with little relation present. To fix this, columns needed to be eliminated from the results, which helped clarify the models. The columns ‘ZIPCODE’ and ‘MARR\_BIN’ were eliminated for clarity. There was an error present in zip code, where some of the clients from the northeast only had four-digit area codes input into the system. This could be a typographical error or an error resulting from combining data from different sources. The other column married bin was eliminated due to a lack of specificity. We do not know if 1 means married or 0 means not married, indubitably. Therefore, it is not wise to include this information in the presentation, as it could negatively impact the results of the analysis. To achieve the best results, the only columns that will be used throughout this analysis are: “age, income, webstore spend, restaurant, and third-party spend.” By doing so, we can ensure that we are eliminating closely related variables by association and generating data that will be useful to Bubba Gump Shrimp Co.

**Meaningful Patterns**

When working with the different types of regression models, it was noticed that there was possibly a meaningful correlation between online shopping and age. This seems like it could be an obvious conclusion, except for the range that exists. From around 20-45 years old, there was an existing correlation of webstore spending. The closest clustering of variables present in the regression model was from 25-35. Another noticeable trend in the data was present when comparing salary by webstore spending. One would think that a customer who made more money annually would obviously spend more, but that was not true in this survey data. Webstore spending actually decreased from customers who make 50k to those who make 60k yearly. Combining these two noticeable patterns can prove to be useful for the company to conduct further research and find a target audience, in hopes of battling declining sales over recent years. After both tests had been conducted, another pattern stood out in the results. It was noticed in the linear regression models that webstore spending decreased as salary went up. However, in the logistic regression models performed to inquire whether a customer was likely to make a webstore purchase, it showed that there was greater chance that higher salaried individuals may make a webstore purchase. The pattern present between these two models shows that individuals making more money, are less likely to become returning customers. They may make one purchase, but the customers which should be sought out should be making purchases more than once. When conducting a K-means Clustering Analysis using the same variables where there was a relativity noticed in the regression models, it was confirmed that the results discovered were correct. The highest earned income cluster in the data set, was not the largest spending among any of the three outlets that Bubba Gump offers merchandise through.

**Inaccurate Depictions of Data**

To overcome inaccuracies in the data set, there are multiple steps that need to be taken to resolve the issues. The formatting used for the column ‘income’ should be changed to a decimal point for accuracy purposes. Many columns could also be renamed for better clarity of what the column intended to be used for. For example, the column ‘zip2’ is the first two numbers present in the customers zip code, which is intended to be used for geographical purposes and regions. This could be renamed to something that would allow data set viewers to know this as well. Instilling ranges for certain columns would be helpful also for clustering purposes. This could be applied to the ‘age’ column and the ‘income’ column. There should also be a column added giving each state in the data set its own integer. Regression models react better to these types of input over descriptive columns containing the state name.

**Alternative Analytic Methods**

An acceptable alternative to the methods conducted so far would be using a decision tree. Decision trees are tools that use tests to determine which path through the tree is taken. In this scenario, a decision tree could be implemented with conditional probabilities comparing clustered groups against one another, as it progressed through the tree. Using the selected remaining variables from the data set, any number of clusters could be chosen against each other and plugged into the tree to determine which cluster tested best as the tree progressed through restaurant spending, webstore spending, and third-party spending. This would determine clusters that should be sought out to help increase Bubba Gump’s sales over the coming years.

**Final Project Milestone Four**

**Display and Interpretation**

**Clustering Analyses:**

**Bi-plot of K-means Cluster Analysis : Webstore spending by age**





**Eigenvalues**

|  |  |
| --- | --- |
| 1.0367544 | 0.9632456 |

**K Means NCluster=3**

Columns Scaled Individually

**Cluster Summary**

| **Cluster** | **Count** |
| --- | --- |
| 1 | 49 |
| 2 | 224 |
| 3 | 227 |

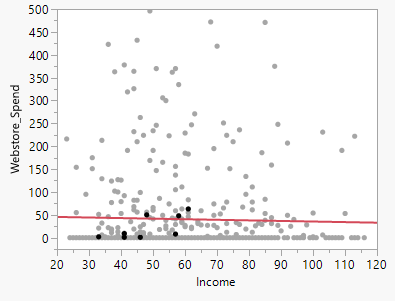
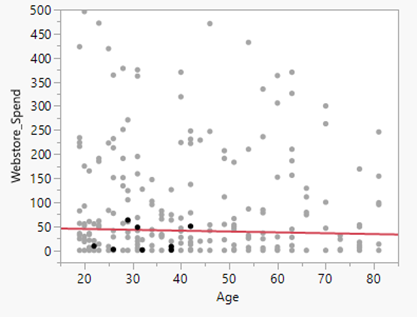
| **Step** | **Criterion** |
| --- | --- |
| 11 | 0 |

**Cluster Means**

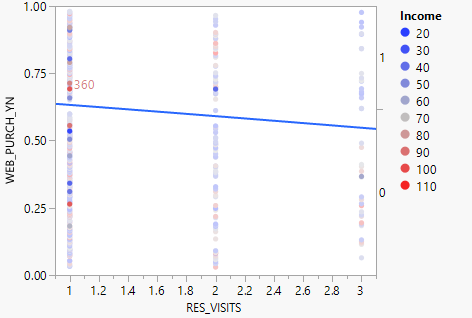
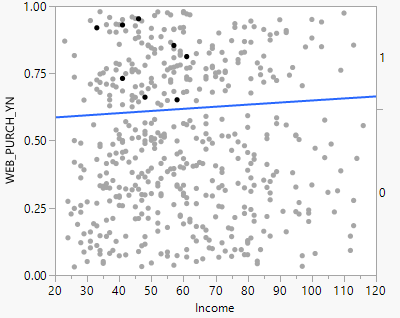
| **Cluster** | **Webstore\_Spend** | **Income** | **Restaurant** | **THIRD\_SPEND** | **Age** |
| --- | --- | --- | --- | --- | --- |
| 1 | 271.591837 | 57.1428571 | 155.653061 | 21.3469388 | 39.7755102 |
| 2 | 15.84375 | 69.7544643 | 68.4464286 | 14.5535714 | 57.2053571 |
| 3 | 14.8898678 | 50.6519824 | 62.2599119 | 25.5814978 | 28.5726872 |

**Regression Models:**

**Bivariate Fit of Webstore\_Spend by Income Bivariate Fit of Webstore\_Spend by Age**

****

**Web Purchase by Income Web Purchase by # of Restaurant Visit**

****

With the established intention of increasing web channel sales, these linear regression models have pinpointed potential clientele that would be able to fulfill this desire. There is a slow decline in both regression models that shows how an increase in salary and also in age, decreases in webstore activity among customers. The models have shown that the most active age group that uses the web channel is from 30-40, with a few remaining outliers outside of this age grouping. The same can be said about the web channel spending by income. We can see a clear grouping exists between 50-60k yearly, with some outliers existing in the data beyond that. With this in mind, Bubba Gump Shrimp Co. can now take this information and create an aggressive marketing scheme that is focused on this clientele bracket. Another noticeable trend in the data was present when comparing income by webstore spending. One would think that a customer who made more money annually would obviously spend more, but that was not true in this survey data. Webstore spending actually decreased from customers who make 50k to those who make 60k yearly. In the final regression model attached number of visits to a restaurant was compared against the likelihood of a webstore purchase being made. It is clear from the model, that customers are most likely to make a purchase after their first visit to the restaurant.

Taking this information into the clustering analysis that was performed considering this newly discovered information, we see that the conclusions made from the regression models do actually test to be true. Although cluster one only had 49 of the 500 customers selected, cluster one spent more in restaurants and webstore than the other two clusters combined. The age range and salary noticed in the regression models was also cluster one. The largest cluster of customers chosen was a younger age group, that made the least annually but spent the most at third party vendors of any cluster. The second largest cluster was the oldest generation of the data set and also the highest income earners. This group was more likely to spend money at the restaurants and the webstore than the youngest cluster. The cluster analysis has now confirmed some assumptions made when viewing the regression models and can provide Bubba Gump with viable information that can be used to begin increasing sales through demographic-specific marketing schemes.

**Validity, Reliability, Limitations**

First considering validity of the analysis, the internal validity needs to be considered. The regression models chosen contained very low R-values and would not be credible due to this. Therefore, assumptions could not be made off of these results alone. Internal validity requires a relationship between variables to be trustworthy and consistent. There should not be any external validity until this can be verified as true. Once the five variables had been chosen, however, there was a confidence level that could be trusted and internal validity was established through statistical verification. This was the result of running many tests to see which variables had the highest relativities. External validity in this analysis could be assumed after successful clustering had been performed and trends in the data were confirmed. Reliability of the data comes through this process, where two different models confirmed the same assumption. There were some limitations present in the survey sample taken, which could have potentially resulted in an inaccurate depiction of the entire customer base for Bubba Gump Shrimp Co. For one, the sample taken was only 500 random customers. A much larger sample would be helpful in testing the results found from this analysis against the larger sample. This is most noticeable by the cluster sizing in the sample of 500. Cluster one contained 49 customers in the age range relative to 39, while cluster two had 224 customers around the age 57, and cluster three contained 227 customers who were around the age 28. Increasing the population size for the data set would help to balance these clusters out more evenly and possibly reveal more mineable information that could be used.

**Resulting Decision Influence**

Upon completing the analysis, the best way to communicate the findings to someone responsible for making decisions would be to show them the statistics on age groups that are more likely to spend money at the online retail outlets. Since increasing sales of the webstore is the focus of this analysis, it is important that the company not waste any money advertising to the groups that are less likely to purchase any merchandise through the webstore. It is noticed that webstore purchase likelihood decreased as restaurant visits increased, so a new customer sign-up campaign could also possibly be exploited. Using the results of the clustering analysis have shown a possible combination of factors that can be combined to combat the flatline in sales Bubba Gump is experiencing recently.

**Visual Evaluation**

Now that the analysis is finished, I have chosen to use regression models and clustering analysis to best communicate the findings. The K-means Cluster Analysis was conducted using webstore spend versus age as the principle components. It can be seen in this model that as age increased, webstore spend trended towards the negative. The next cluster analysis provided shows the five variables that have been chosen to complete the business analysis in three clusters. This model provided a clear representation of the data set and highlighted trends in spending among the three different outlets that merchandise is sold. The regression models were attached also to show the trends in the variables selected. Upon finding these trends in the regression models, I was able to conduct proper clustering analyses to begin confirming assumptions that had been made off the regression models. These models present the clearest interpretation of the data set provided from the surveys. These models clearly show the company where it can make improvements on future sales, however.

**Next Steps**

The next steps that should be taken by Bubba Gump Shrimp Co. involve taking the same variables used in the cluster analysis and applying them to a larger population set. Although this sample of 500 did indeed provide valuable insight, a larger population mean to analyze would provide more meaningful data for the company to act upon. Once this is completed, Bubba Gump could begin target marketing off of the resulting information. From this survey sample data, the company could conclude that 39 year old’s who make 57,000 yearly, visiting for the first time, could be the best customer base to pay attention to. This could be confirmed through further research into the matter.

**References**

Ahlemeyer-Stubbe, A., & Coleman, S. (2014). VitalSource Bookshelf Online. Retrieved January 06, 2021, from <https://mbsdirect.vitalsource.com/>

Pacheco, I. (2020, October 06). How coronavirus changed the retail landscape. Retrieved February 28, 2021, from <https://www.wsj.com/articles/how-coronavirus-changed-the-retail-landscape-11601976600>

Unknown. (2019, March 25). Difference between descriptive and predictive data Mining (with COMPARISON CHART). Retrieved February 28, 2021, from <https://techdifferences.com/difference-between-descriptive-and-predictive-data-mining.html#:~:text=The%20descriptive%20and%20predictive%20data%20mining%20techniques%20are,historical%20data%20as%20the%20chief%20principle%20for%20decisions>.