CS499 ePortfolio: Databases

Ryleigh Failla

**Business Problem**

Bubba Gump is a successful regional retailer of food that became a household name due to exposure from the blockbuster film, Forest Gump. After the initial surge of growth following the film’s release, the sales plateaued and eventually began to decrease. For the last 2 years, the sales have been decreasing. Failure to secure means of revenue growth may leave Bubba Gump with no choice but to downsize or decrease operations. The business question would be, “How can we use internal and external data gathered about our business to attract new and returning customers, increase sales, and promote customer retention?”

**Research Question**

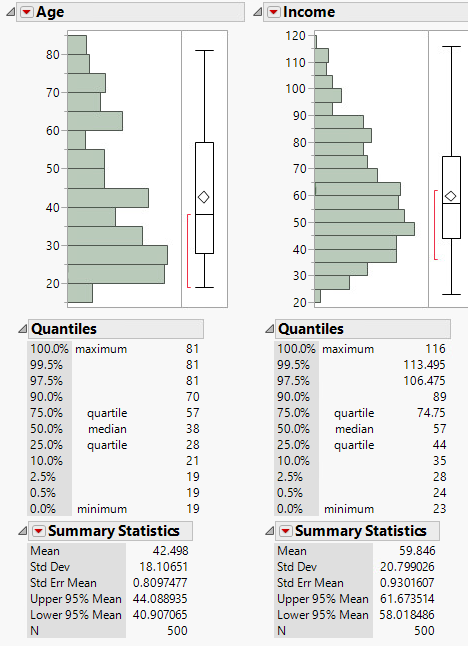
The research question that needs to be answered is, can Bubba Gump customers be grouped into different clusters and how do we appeal to said clusters in order to create sustainable revenue and business growth?

**Research Measurement**

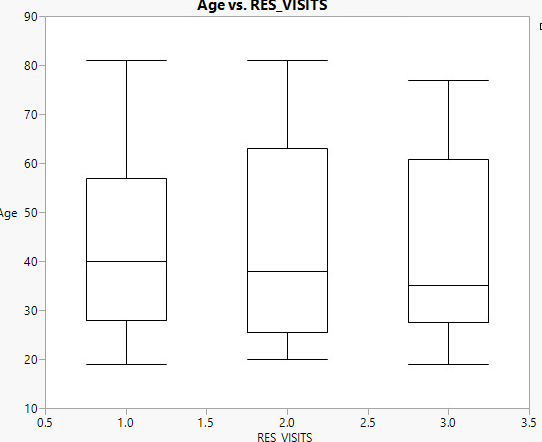
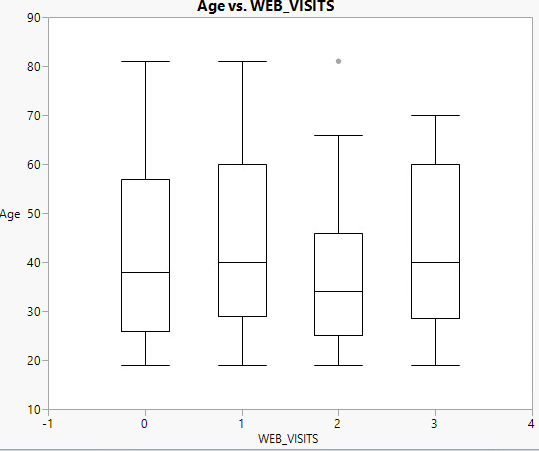
We can generally tell if we have successfully answered the research question by looking at trends in the data and assessing potential relationships or subgroups that emerge from within the data which we can then apply in a business setting. For example, if the data is scaled and represented in a histogram is the histogram symmetrical and bell shaped? Also known as a normal distribution. In addition, are there outliers that skew the data?

**Display and Interpretation**

To initially assess the data, I created basic distributions of customer ages and income to get a sense of what the sample data looks like. This will influence how we interpret the results because for example, if 50% of the customers are age 18-28 but they only account for 10% of sales, we know that we need to do a better job of targeting that age group. The distributions are depicted below.



Next, I made a box and whisker plot based on age groups to analyze any relationships between them as pictured below.



Further, I binned the ages into 7 different groups of 10 years apart to see what trends there were between webstore spending and third spend. That way we can group the customers into different age groups rather than their individual ages.

**Validity, Reliability and Limitations**

From the results we can see definitive subgroups within the customer base age groups which gives us a means of grouping the customers in order to better target who needs to be re-engaged as customers as well as who we need to draw in as customers. The limitations would be that the data set could be larger. Also there is no additional process of validation other than the check points provided by JMP.

**Analytic Method**

There are 3 main channels of data that are being analyzed; customer restaurant purchases, online purchases, and third-party retailer purchases. These channels of data are integrated within a single data warehouse for analysis. Furthermore, there is a sample of the customer population consisting of 500 customers that have been selected from the data warehouse and issued surveys to complete. Based on the way that the customers complete the survey cards, it’s inevitable that “clusters” of customers will emerge. By utilizing internal and external data, not only do you gain an overall view of the customer experience from their perspective, but you are able to assess the subgroups (hereinafter “clusters”) that emerge from the data. You can use external data to determine if there are any common characteristics or qualities of these customers, and further us that information to change business practices to enhance their customer experience and market to them. To summarize, by identifying similar groups of customers, we can see what types of customers that Bubba Gump attracts and how the groups generally behave based on their survey results.

**Analysis Tools**

To analyze the data, I will use JMP to cluster groups of customers and reveal ways that the customers can be grouped. The survey data is stored in an excel spreadsheet which is compatible with the JMP software. To accomplish the aforementioned goals, the most appropriate form of data mining tools would be supervised data analysis. Supervised data analysis is defined as a means to “estimate an unknown dependency from known input-output data” (Ahlemeyer-Stubbe & Coleman, 2014). Using input variable such as quantities purchases by customers, amount they spent, and location they purchased at, we can observe output variables that are targets and can assist in visualizing any trends or associations that will make customers more likely to respond. In addition, Knowledge-discovery techniques can also be used. More specifically, cluster analysis, association rules, statistical methods, decision trees, association rules. Clustering will help to group the customers and how they can be appealed to, since not all customers respond the same to sales tactics.

**Data Visualizations**

When it comes to data visualizations, it is important that the data is modelled in a way that gives “credible rank ordering of customers in terms of associated variables” (Ahlemeyer-Stubbe & Coleman, 2014). With the data that is available and targets for the business model, some appropriate means of data visualization would be histograms, whisker plots and cluster analysis. These are effective ways to compare variables. Since we are trying to find out what customers respond to it will be easier to visualize these trends and associations through models that utilize comparison methods.

Sources

Ahlemeyer-Stubbe, A., & Coleman, S. (2014). A Practical Guide to Data Mining for Business and Industry. Chichester, West Sussex, United Kingdom: Wiley