8-2 Final Project

Steven Wade

DAT-220-Q6408 Fundamentals of Data Mining

August 23, 2020

Table of Contents

[Analysis Plan 2](#_Toc49082451)

[Business Problem 2](#_Toc49082452)

[Analytic Method 2](#_Toc49082453)

[Analysis Tools 3](#_Toc49082454)

[Data Visualizations 4](#_Toc49082455)

[Research Question 4](#_Toc49082456)

[Research Measurement 5](#_Toc49082457)

[Follow-Up Questions 5](#_Toc49082458)

[Research and Support 6](#_Toc49082459)

[Analysis 6](#_Toc49082460)

[Analysis Organization 6](#_Toc49082461)

[Sources of Error 7](#_Toc49082462)

[Meaningful Patterns 7](#_Toc49082463)

[Inaccurate Depictions of Data 8](#_Toc49082464)

[Alternative Analytic Methods 8](#_Toc49082465)

[Final Report and Presentation 10](#_Toc49082466)

[Display and Interpretation 10](#_Toc49082467)

[Validity, Reliability, Limitations 14](#_Toc49082468)

[Resulting Decision Influence 15](#_Toc49082469)

[Visual Evaluation 15](#_Toc49082470)

[Next Steps 16](#_Toc49082471)

[List of References 17](#_Toc49082472)

# **Analysis Plan**

## **Business Problem**

We want to increase revenue at Bubba Gump Shrimp Company and unify the customer experience across all avenues in order to increase throughput.

In recent years, sales have declined and name recognition as leveled off. If nothing is done to stop the declining sales and increase revenue, Bubba Gump Shrimp Company will no longer be able to operate at its current capacity.

We will be using unsupervised data mining techniques such as clustering and association in order to find patterns and identify possible optimizations or new revenue streams.

## **Analytic Method**

It isn’t apparently clear why once increasing revenue leveled off and is now declining. For this reason, unsupervised data mining techniques such as clustering and association will be used in order to look for patterns amongst the data.

Cluster analysis of the data can reveal groups and patterns as well as expose potential new groups of customers (“Cluster Analysis in Data Mining”, 2020). Cluster analysis can reveal purchasing patterns of customers as well – further providing insight into how a customer interacts with and behaves within the Bubba Gump Shrimp Company brand.

Association analysis will be used to provide insight into “co-occurrence events” (Cawley, 2014) such as what and how much merchandise was purchased when a specific location was visited. This information can be used to optimize physical store layouts as well as personalize individual online experiences based on customer habits.

The various analyses will be performed on data housed in the client data warehouse collected from point-of-sale systems, web sales, customer loyalty programs, and website data from retail partners. A subset of customers will be analyzed in order to gain a broader picture of overall performance and customer activity.

Analysis Tools

To perform the analysis, the software program JMP will be used. JMP provides a GUI interface to run statistical analysis of data, identify patterns, data visualization and scripting support with SAS, MATLAB, R, and Python (“Data Analysis Software”, n.d.). JMP supports importing CSV files and connecting directly to databases to build a dataset through queries and joins. The software is intuitive and relatively simple to use. Models and reports are simple to run and the defaults that JMP selects are often more than sufficient for any analysis being performed. JMP is capable of handling large datasets just as simply as small ones. For larger datasets, JMP makes it simple to generate subsets from a number of factors – random with a sampling rate, random with a sample size, and supports stratifying data for a more representative subset. An important strength of the JMP software is being able to easily and visually prepare data before analysis through various means (e.g. binning, missing data analysis, variable transformation). In addition to the built-in features, JMP’s greatest strength is its knowledge base and community. Finding support for questions is a simple web search away, or quick look-up in the built-in handbooks. A weakness of JMP is in the reporting. JMP assumes some statistical analysis from the end user (“Pros and Cons of JMP Statistical Discovery Software”, n.d.). Another weakness is its expense with its monthly subscription cost of $125 (“Data Analysis Software”, n.d.). Regardless of its weaknesses, JMP is a powerful tool and handles data analysis robustly and adaptively making it the ideal choice for this analysis. JMP will aid in this analysis by identifying missing data, providing data preparation tools for binning and variable transformation, and for running various models and generating graphs.

Data Visualizations

One graphic that will be utilized within this report will be an infographic with statistical totals related the sampling of the data. In this graphic will be bar chart of age demographics and the purchase amounts associated with each age group – this will help identify which age group to target in the future as it shows which one spends the most at Bubba Gump. A pie chart displaying the martial status statistics for the sampling will be utilized to provide insight into targeted marking of single or married customers. Lastly, a bar chart of sales totals for each of the revenue streams (point-of-sale, web sales, third-party retailers), which will identify which sales streams are the weakest allowing insight into where to focus investments in the future.

## Research Question

1. Do married customers use one particular sale stream (POS, online, 3rd-party) over others?
2. What customer profile (marital status, age group) engages the most with the Bubba Gump brand?
3. Does where the customer lives have any correlation with sales within the three sources of revenue (POS, online, 3rd-party)?

The research question from above that will be the focus of analysis will be identifying the most engaged customer profile. This will provide insight into what demographic to better target for future sales.

Research Measurement

The research questions are looking to define the most engaged demographic groups within the Bubba Gump customer base. Customers will be binned into groups for age and income. Cluster analysis will be performed in an attempt to identify commonalities between Bubba Gump customers. The success of the research will be measured whether a specific customer profile stands out from the rest as being more engaging than other demographic groups.

Follow-Up Questions

1. Do young married couples spend more money through online retailers than in physical locations?
2. Is there a sub-correlation between age and income and which sales medium customers prefer?
3. Is there a relationship between where the customer lives and their likelihood of making a purchase within a physical location?

## Research and Support

1. The first source to support this research and analysis will be the book *A Practical Guide to Data Mining for Business and Industry* (2014) which provides an overview of data mining including grouping and analytical techniques, data visualization techniques, and data preparation methods.
2. A second source that will aid in answering the research question is the article “Developing the profiles of supermarket customers through data mining” by Hokey Min (2007). This article directly relates to the research question and business needs in this analysis document as it relates to developing customer profiles.
3. Another source that will be beneficial in answering the research question is “Role of demographics, social connectedness and prior internet experience in adoption of online shopping: Applications for direct marketing” by Mohammad Bakher Naseri and Greg Elliott (2011). The article connects demographics’ role in online shopping and provides data models with an in-depth analysis of an example data set. This methodology can be applied to the Bubba Gump data set to find correlations, if any, with demographics and online sales.

# Analysis

## Analysis Organization

A manual stepwise regression model was applied during analysis. A series of regression analyses were run one at a time comparing two variables to one another looking for correlation and probability of a predictor. For instance, the age vs. web spend, income vs. web spend, and web visits vs. web spend variables – to name a few – were compared to one another in linear regression tests. In the age vs. web spend test, we were looking for a correlation between a customer’s age and how much they spent on the website, and when combined with a logistical regression test to determine probability, an estimated correlation between the two variables was established.

Aspects of this analysis that are out of the control of this tester are missing data such as dates of purchases, or individualized customer transactions rather than cumulative data. A further limitation of this analysis is the size of the data which is limited to 500 customers and may not be totally representative of the whole of Bubba Gump’s customers.

## Sources of Error

Prior to running the analysis, I ran a Missing Data Pattern on all of the rows in the given data set. JMP identified two erroneous rows with no data values present in the columns. I started by deleting these rows. Next I ran a formula on the “zip” column to pad leading zeros to the zip codes to ensure all values had 5 digits. Zip codes representing the Northeast were only four characters long. This is likely due to their representation as numerical values instead of strings, where leading zeros are omitted from numbers.

Meaningful Patterns

This analysis revealed a target demographic for the web store. The analysis showed that older customers spend less online as well as customers with higher income. This narrows the target for web sales to customers aging from 18-48 with a middle to upper-middle class income.

Another pattern revealed through the analysis was that there is a positive correlation between the number of web visits to web purchases, as well as restaurant spend and web spend. Customers that visit the Bubba Gump site more tend to spend more. Also, customers that spent more in restaurants were more likely to make an online purchase as well.

Based on the discovered patterns, an additional research question comes to mind: *what percentage of customers, aged 18-48, make purchases online?*

## Inaccurate Depictions of Data

An instance of inaccurate depiction of data were the amount of zero-dollar values in “spend” columns in the linear regression analyses. These were visually ignored when performing the analysis as they appeared to negatively weigh the graphs.

Trend lines and colors were added to the data graphs to separate groups and depict correlation. Prior to this, the graphical results were not immediately apparently.

Alternative Analytic Methods

Analytic methods not used in this analysis that could be beneficial to Bubba Gump would be to build a decision tree and a time series analysis. A decision tree could be used to segment customers into groups based on various levels of information. These groups can be specifically targeted separately for marketing and promotional purposes that are best tailored to each group. A time series analysis would be beneficial in identifying purchasing habits and patterns over time for customers and restaurants. The provided data set used in this analysis rolls up customer transactions into a single representation, whereas with a time series analysis would look at all individual records for each customer.

# Final Report and Presentation

Display and Interpretation

The first research question that this analysis focused on was “Do married customers use one particular sale stream (POS, online, 3rd-party) over others?”. Figure 1 below shows a comparison of marital status vs sales streams. As we can see in this figure, there is no clear distinction between married customers and a particular sales stream.

**Restaurant, Web, and Third-Party Spend by Marital Status**



Figure 1

**Total Sales by Source (Restaurant, Web, Third-Party)**



Figure 2

Figure 2 shows the sales totals for each sales stream. Restaurants made up the most source of revenue with web sales coming in second. Third-party retailer sales figures were last in this assessment. With restaurant sales doing well, and third-party sales relying on partners, focus is shifted to increasing web sales.

|  |  |
| --- | --- |
| **Age vs Web Spend**    Figure 3 | **Income vs Web Spend**    Figure 4 |

With no clear distinction for the first research question, attention moved to the second question: “What customer profile (marital status, age group) engages the most with the Bubba Gump brand?”

Figures 3 and 4 above examine demographics regarding age and income respectively to web spend. Income and age both show a negative association with web spend. The higher the income, the less spent via the web. Likewise, the older a customer is, the less they will spend online. Figure 5 below shows a positive association to web spend, showing that the more times a customer visited the site, the more money they spent on the web in total. This is backed up in Figure 6 which shows a 45% probability of a customer making a purchase on the web who visits the website once. This probability increases to 100% for customers who visited the site 2 or more times.

|  |  |
| --- | --- |
| **Web Visits vs Web Spend**    Figure 5 | **Probability Based on Web Visits**    Figure 6 |

Figure 7 shows a binned breakdown of customer ages. The majority of Bubba Gump customers fall between the ages of 18-48. Figure 8 displays a distribution of income based on the Bubba Gump survey data. Based on these distributions as well as correlations between age and income to webstore spend, the typical Bubba Gump customer, and ideal target group, is middle-class and is between the ages of 18 and 48.

|  |  |
| --- | --- |
| **Customer Age** | **Customers Annual Income** |
| Figure 7 | Figure 8 |

Analyses performed with regard to the third research question (as seen in Figure 9) did not reveal a clear correlation between customer locations and how much they spend with Bubba Gump.

**Average Spent in Restaurants by State**



Figure 9

## Validity, Reliability, Limitations

Model outputs included in this report disproved the first and third research questions – correlation between marital status and sales sources, and customer location and average amount purchased at restaurants. However, models revealed a valid target demographic in middle-income customers aged 18-48. The output of these models is accurate based on the provided sample data of Bubba Gump customers, and they are reliable in their results as confirmed in multiple linear and logistical regression tests. However, limitations in the results are possible due to the small sample size (500 customers) and due to the fact that customer data is cumulative instead of individualized in a time series of transactions. Despite any potential limitations, multiple models revealed the same results for a customer demographic that can be used by Bubba Gump’s marketing team to potentially reduce marketing costs and increase sales and revenue.

The tests conducted were designed specifically for the Bubba Gump survey data set and the Bubba Gump business problem. For this reason, the internal models were not tested on external data and likely do not lend themselves to generalization.

## Resulting Decision Influence

The data found in this analysis can aid in turning around declining sales and increase Bubba Gump revenue. The distribution of web store spend reveals that 64% of customers do not make a web purchase and 35% do not make a purchase via a third-party retailer. These figures represent underutilized or untapped revenue streams and room for growth. Increasing ad spend online to drive more visitors to the website will likely increase web sales. Target customers can be narrowed to customers aging from 18-48 with a middle to upper-middle class income. Sales can be further boosted with in store promotional campaigns for merchandise as well as in-store marketing that promotes the Bubba Gump online store.

Visual Evaluation

The visual style used was chosen to be simple and easy to understand. Most charts are color coded, such as the sales totals vs sales sources (restaurant, web, third-party), to easily distinguish between variables. Trend lines and probability lines have been included in the graphs as well to show projections and simplify graphs with many data points. Most importantly, the graphs are simplified, color coded, and labeled to present the models to business clients simply and help make the models easily understood and interpreted. The graphs for correlation between web spend and age and income could have been simplified and cleaned up more by removing the zero dollar spend data points from the graph while keeping them represented in the projection calculations. Other than that, the graphs themselves have been presented for the purpose of solving a business problem by a business community.

## Next Steps

After running this analysis, next steps going forward would be to re-evaluate the hypothesis and collect more data. As mentioned above, an additional research question came up during analysis: *what percentage of customers, aged 18-48, make purchases online?* With a new target demographic identified, further data should be collected about individual purchases and purchasing patterns, and additional research questions should likely be explored to narrow the focus of potential improvements to the Bubba Gump business.

# List of References

Ahlemeyer-Stubbe, Andrea, and Shirley Coleman. *A Practical Guide to Data Mining for Business and Industry*. Wiley, 2014.

Cawley, Keith. “When To Use Supervised And Unsupervised Data Mining - Machine Learning Times - Machine Learning & Data Science News.” *Machine Learning Times*, 17 Sept. 2014, www.predictiveanalyticsworld.com/machinelearningtimes/use-supervised-unsupervised-data-mining/4046/.

*Cluster Analysis in Data Mining: Applications, Methods & Requirements*. 28 Jan. 2020, [www.upgrad.com/blog/cluster-analysis-data-mining/](http://www.upgrad.com/blog/cluster-analysis-data-mining/).

*Data Analysis Software*, n.d.. www.jmp.com/en\_us/software/data-analysis-software.html.

Min, Hokey (2006/10/01). Developing the profiles of supermarket customers through data mining. *The Service Industries Journal*, 26, 747-763. doi: 10.1080/02642060600898252

Naseri, M., Elliott, G. Role of demographics, social connectedness and prior internet experience in adoption of online shopping: Applications for direct marketing. *J Target Meas Anal Mark* **19,**69–84 (2011). <https://doi.org/10.1057/jt.2011.9>

Pros and Cons of JMP Statistical Discovery Software. (n.d.). Retrieved August 20, 2020, from https://www.trustradius.com/products/jmp/reviews?qs=pros-and-cons