Data Mining: Bubba Gump Shrimp Company

Leah Launiuvao

Southern New Hampshire University

DAT 220

Table of Contents

Introduction **3-4**
Business Problem 3
Analytic Method 3-4

Plan for Analysis **4-7**
Analysis Tools 4
Data Visualizations 4-5
Research Questions 5
Research Measurement 5-6
Follow-up Questions 6
Research and Support 7

Analysis **7-10**
Analysis Organization 7-8
Sources of Error 8-9
Meaningful Patterns 9
Inaccurate Depictions of Data 9-10
Alternative Analytic Methods 10

Final Report **11-23**
Display and Interpretation 11-20
Validity, Reliability, Limitations 20-21
Resulting Decision Influence 21-22
Visual Evaluation 22
Next Steps 22-23

References **24**

INTRODUCTION

Business Problem:

Popularized by a successful film, your company Bubba Gump Shrimp found increasing success in the restaurant and retail market. After a period of growth, you have found yourselves struggling to maintain profitability, reporting a decline in sales over the last two years. Your company has preserved historical data across the various aspects of their organization, in the hopes that it can be consolidated with new data collected through customer surveys. Through a data warehouse, we can establish connections between facets of your organization and customer experiences. What comparisons can be drawn by observing this integrated data, and how can these patterns increase long-term profitability?

Analytic Method:

To address this question, we will examine the internal sources you have provided, including data related to point-of-sale, customer information, web and retail transactions, and newly acquired customer satisfaction surveys. The data will be analyzed through graphs, charts and plots using the JMP database software. Connections between item prices and customer satisfaction may be established. We will use these findings to guide questions that promote further research such as…Do certain price points warrant a more positive customer response over others? In terms of profit derived from retail and web store sales, customer's means of purchasing may suggest expanding or reducing your use of alternative sales platforms. By grouping data into clusters based on customer demographics and your company offerings, we can establish who the target audience is, increase marketing, address peak seasons, and revive

the Bubba Gump Shrimp Company.

PLAN FOR ANALYSIS

Analysis Tools:

The internal data provided by Bubba Gump Shrimp Company will provide the basis for analysis. By comparing customer records with previous purchases, we can make inferences to increase profit. JMP data mining software will be used to integrate the data. The program is user-friendly and provides a platform that allows users to manipulate and arrange data for the purpose of statistical analysis. Within the program, we can create visual translations of data through charts, plots and histograms to present data findings in a focused, graspable way.

Data Visualizations:

Data visualizations provide a digestible look at complex data. Through a series of graphs and charts, we can establish connections between customer demographics, sales platforms and purchases. This enables your company to tailor your marketing, sales and inventory to more focused criteria.

Bar charts allow us to illustrate a range of measurements in a comprehendible way. It makes it easy to distinguish outliers and draw general conclusions about the data set. Stacked bar charts can be used to compare multiple variables against the same constraints. For example, examining the occurrence female customers verses male customers in relation to purchases across all sales platforms. Line graphs would be useful if the purchase dates were included in the data set. Then, trends would be visible across a span of time. They can still be used to determine

the relationship between selected variables with a measurement of which to base them off. If we are analyzing total numbers across the sales platforms, a pie chart would give a clear visual representation of how they compare. A variety of visual representations of the data is necessary to make comparisons in a meaningful and understandable way.

Research Question:

A focused research question provides a basis for determining the value of associations between otherwise abstract variables. Correlations that provide insight into customer engagement, profits or avenues of revenue can provide your Company with the insight to make informed decisions for the future. Valuable information can be derived from clusters by identifying patterns or significant figures. Associations between survey variables can inform future marketing decisions by helping your company to narrow the target audience. The research question should allow us to decipher what data is relevant to our investigation:

- *Which clusters demonstrate meaningful associations between customer criteria and increased customer spending?*
- *Is there any connection between customer age and a tendency to revisit Bubba Gump restaurants? How does this possible correlation speak to the restaurant's atmosphere and customer base?*

Research Measurement:

Identifying correlations between data clusters creates a foundation for further investigation. Using the data visualization tools suggested above, the distribution of data can be

interpreted by looking for significant clusters attached to certain variables. Remarkable statistics, percentages or connections of data can be beneficial, whether positive or negative. By identifying your company's largest demographic, you can better tailor marketing campaigns and brand identity, while making efforts to expand your reach to other audiences. Confirming the associations established through data mining can be done by acquiring and analyzing more data. Progress can be monitored by tracking sales and by issuing new customer surveys regularly. As the data set grows, more accurate predictions can be made, and more in-depth relationships can be established.

Follow-Up Questions:

- What aspects of the restaurant does the occurrence of repeat customers speak to?
- How can marketers tailor sales and ad campaigns to fit their customer base based on the results generated by the analysis?
- Are customers predominantly married or unmarried? Is this data representative of the restaurants' environment?
- Do some restaurant regions acquire substantially more customers than others?
- Is there any correlation between restaurant visits and subsequent online and third-party purchases?
- Are third-party sales a relevant source of income for the company?
- What age range represents the company's target audience?
- If the survey included purchase dates, would we see a fluctuation in accordance with particular seasons or holidays?

Research and Support:

To utilize the JMP software most effectively, I came across videos on the JMP website, providing tutorials that teach the user how to navigate through the program basics (JMP Learning Library). The assigned course reading titled *A Practical Guide to Data Mining for Business and Industry* will also guide my investigation by outlining the various forms of data preparation, data mining concepts, visuals and interpretation techniques (Ahlemeyer-Stubbe & Coleman). In order to make the best use of the data collected, an article titled, *From customer data to value: What is lacking in the information chain?* examines the best ways to approach customer data segmentation (Crié & Micheaux, 2006). This source will help me to distinguish between relevant and nonrelevant associations. However, it is implied that the data collected be representative of their recommendations, where in this case the data has been collected in advance.

ANALYSIS

Analysis Organization:

Data was collected from customer surveys and transformed to establish consistent variables for analysis. Customers can be defined by age, location, marital status and income. Characteristics of Bubba Gump Shrimp provide insight into restaurant location, occurrences of revisits, and spending methods and patterns. To establish connections between various customer attributes and aspects of company data, I can employ cluster analysis methods. Each means of analysis utilizes unique methods which draw comparisons among selected characteristics. By testing each variable individually for statistical significance, I will follow an organized, stepwise approach.

Hierarchical clusters are grouped based on highest points of similarity. Under this lens, we should be able to pinpoint relevant points of customer data. Hierarchical cluster analysis can also be used to map out customers based on characteristics provided in the survey. These segments may provide insight into which customers are likely to spend across the various platforms, based on characteristics such as age, income, marital-status, and location. This data can be used to target marketing campaigns and emphasize popular locations and spending avenues.

To try and determine the strength of the relationship between webstore spending and variable characteristics provided by the survey, I will use various linear regression models and their summaries. This data can speak to target audiences to tailor prices and market campaigns accordingly. Zip codes can provide insight into both customer location and indicate where promotions for online sales should be directed. To make predictions about the likelihood of customers making webstore purchases on the basis of various criteria, logistical regression provides a model for analysis and probability. We can make predictions about spending in relation to location, customer age, marital-status and income.

Sources of Error:

The data in its current form may cause inaccurate distributions of spending, as values of zero are placed for customers who did not spend money on the website or through a third-party retailer. In one scenario, the survey shows a customer made a purchase from the webstore, but the amount was recorded as zero. Multiple versions of the same variable provide greater room for entry error. Some categories were redundant and not necessary for analysis. With only 500 entries, it was unnecessary to include variables such as city or full zip code. To perform an analysis not swayed by these appendages, I only used variables such as two-digit zip, spending

patterns, age, marital status as binary number and income. I disregarded non-numeric variables such as customer name, married Y/N, etc. By analyzing variables in pairs, it is easier to decipher possible relationships. Overall, the analysis would be stronger with a larger dataset.

Meaningful Patterns:

In my preliminary analysis, I have recognized a few instances of connection among the variables. Understandably, customers with more visits to Bubba Gump restaurant were more likely to have spent more overall. Restaurant visits were positively correlated with webstore visits and webstore spending. This could raise questions about marketing efforts outside the restaurant atmosphere. What can be learned from the current marketing model provided by restaurants, directing customers to the website? For customers who did not attend the restaurant but made purchases on the webstore, what informed their choice? In terms of customer demographics, younger generations, and those who fell within an income bracket of 30-60k (mid-income range), were more likely to spend on the website. Is this an indication of the website's design or user-friendliness? Married customers appear to be more likely to spend a greater amount at the webstore. Is this an indication of combined income? There appears to be the potential for other relationships amongst the data, but an analysis would benefit from a greater dataset to observe these connections further.

Inaccurate Depictions of Data:

Since the customers who did not spend at the webstore or third-party retailers were included in the analysis of spending patterns, it created an inaccurate depiction of the trends. When many points fall along the minimum line, it creates the illusion of a high concentration of spots. For example, the bivariate fit of webstore spending in relation to zip code. The points which represent spending are scattered across the bulk of the window. Along the bottom, zip

code is represented by points along the x-axis, where there appears to be a line of dots along the bottom which represent zero dollars spent. These distorted results are also evident in the bivariate analyses assessing webstore spending in relation to other variables such as marital status and income. To eliminate the confusion these points present, I can create an analysis that disregards values of zero. Other inaccurate depictions of data include outliers that are unusually far from the main population of data points. There doesn't appear to be many points outside the normal concentration of data points aside from some webstore purchase amounts. These points can be analyzed by assessing the standard deviation, and any points that are more than three times the standard, can be considered anomalies. These anomalies can be disregarded in comparative analysis but should be analyzed for error or significance.

Alternative Analytic Methods:

In addition to the hierarchical clustering that was performed, other forms clustering allows us to of discover patterns within the data and group the data according to those similar attributes. Centroid-based clustering incorporates K-means clustering techniques to identify clusters based on a central point. This can also be beneficial for identifying outliers. Decision-trees branch out potential outcomes, enabling us to evaluate different scenarios from different attributes. We can use this algorithm to determine the significant features of the Bubba Gump data set, presented in a way that is easy to understand.

FINAL REPORT

Display and Interpretation

A correlation analysis amongst continuous variables from the data sample provides an overview of potential relationships for further analysis (**Figure 1**). Correlations of significance

can be seen between restaurant visits and restaurant spending, webstore spending and restaurant spending, webstore visits and webstore spending, third-party visits and third-party spending. Less prominent correlations can be found between age and income, and marital-status and zip code. While the stronger correlations may indicate more valid connections, issues surrounding sample size and selection may undermine the results of correlation between variables. This evaluation provides a basis for further attribute comparison, especially between restaurant and webstore platforms.

Multivariate

Correlations

	ZIP_2	Restaurant	RES_VISITS	Webstore_Spend	WEB_VISITS	THIRD_SPEND	THIRD_VISITS	Age	MARR_BIN	Income
ZIP_2	1.0000	-0.0259	0.0065	-0.0186	-0.0226	-0.0668	-0.0143	-0.0659	0.1090	0.0101
Restaurant	-0.0259	1.0000	0.5955	0.4534	0.2078	-0.0427	-0.0379	-0.0033	-0.0796	-0.0159
RES_VISITS	0.0065	0.5955	1.0000	0.2943	0.1839	-0.0801	-0.0846	0.0045	-0.0797	-0.0307
Webstore_Spend	-0.0186	0.4534	0.2943	1.0000	0.6119	-0.0059	-0.0034	-0.0368	-0.0148	-0.0299
WEB_VISITS	-0.0226	0.2078	0.1839	0.6119	1.0000	-0.0409	-0.0103	-0.0037	-0.0054	0.0301
THIRD_SPEND	-0.0668	-0.0427	-0.0801	-0.0059	-0.0409	1.0000	0.7422	-0.0827	-0.0129	-0.0601
THIRD_VISITS	-0.0143	-0.0379	-0.0846	-0.0034	-0.0103	0.7422	1.0000	-0.0768	-0.0280	-0.0636
Age	-0.0659	-0.0033	0.0045	-0.0368	-0.0037	-0.0827	-0.0768	1.0000	-0.0570	0.1093
MARR_BIN	0.1090	-0.0796	-0.0797	-0.0148	-0.0054	-0.0129	-0.0280	-0.0570	1.0000	-0.0278
Income	0.0101	-0.0159	-0.0307	-0.0299	0.0301	-0.0601	-0.0636	0.1093	-0.0278	1.0000

FIGURE 1.

A bivariate analysis of webstore spending in relation to restaurant spending illustrates this relationship further (**Figure 2**). The dots representing money spent on the website, and money spent at the restaurant follow a positive linear line (red). Many of the dots fall within a reasonable parameter of density, with a high concentration of dots in the bottom left corner and across the bottom of the chart indicative of the customers who did not purchase from the online store. This is confirmed by a chart which specifies a p-value of significant correlation (**Figure 3**).

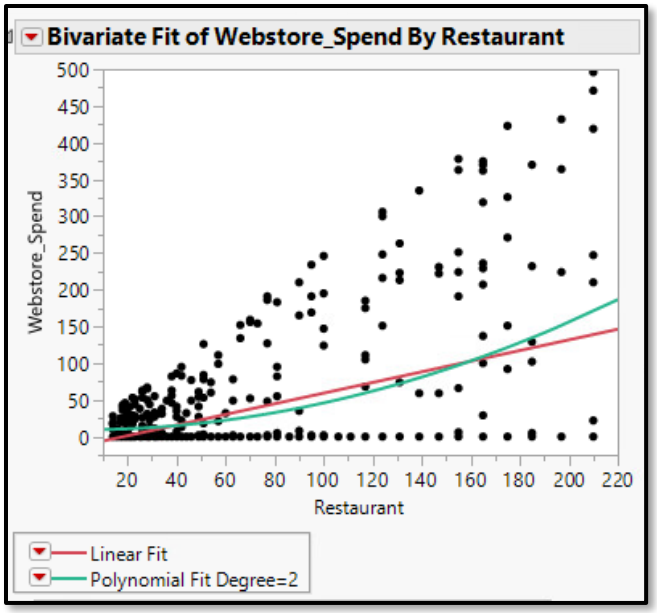

FIGURE 2.

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.453398	0.38085	0.520398	<.0001*
Covariance	2190.81			
Count	500			

Variable	Mean	Std Dev
Restaurant	74.184	54.98812
Webstore_Spend	40.474	87.87316

Linear Fit

Webstore_Spend = -13.27585 + 0.7245478*Restaurant

Summary of Fit

RSquare	0.20557
RSquare Adj	0.203975
Root Mean Square Error	78.40066
Mean of Response	40.474
Observations (or Sum Wgts)	500

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	792086.0	792086	128.8644
Error	498	3061038.6	6147	Prob > F
C. Total	499	3853124.7		<.0001*

FIGURE 3.

Creating a hierarchical cluster of the website spending creates a dendrogram with 4 natural clusters (**Figure 4**). When I performed the analysis with categorical clusters of spending in terms of low, moderate and high, I found a clear divide between those who did not make purchases and those who have made significant ones, confirmed by the length of the lines connecting the cluster branches. This may suggest a lack of awareness surrounding their online platform, or other factors hindering potential customers. A hierarchical cluster of restaurant spending (**Figure 5**) allows us to see the distribution of spending of restaurant patrons. These results are more dispersed, as all customers from the sample visited a restaurant location.

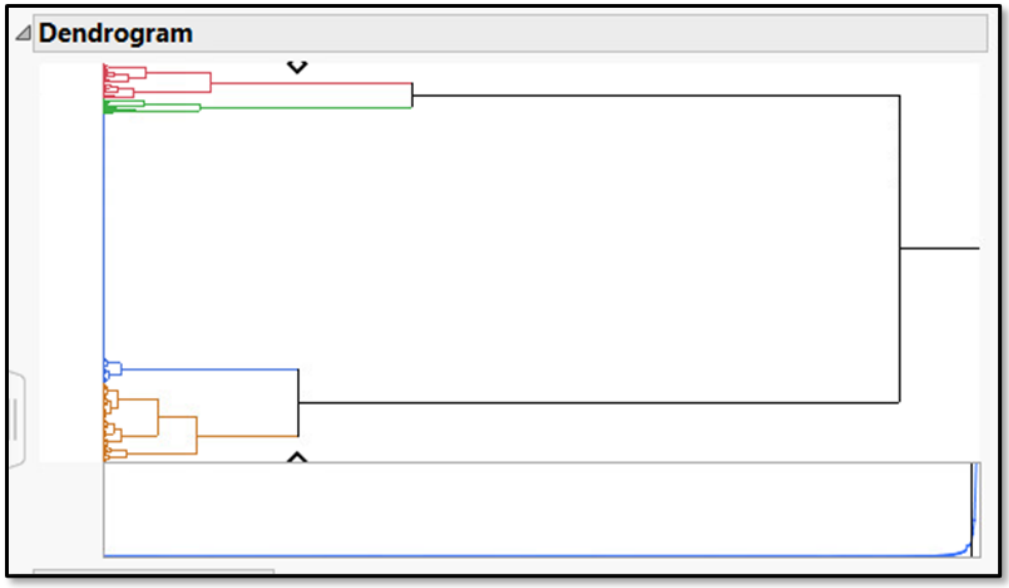

FIGURE 4.

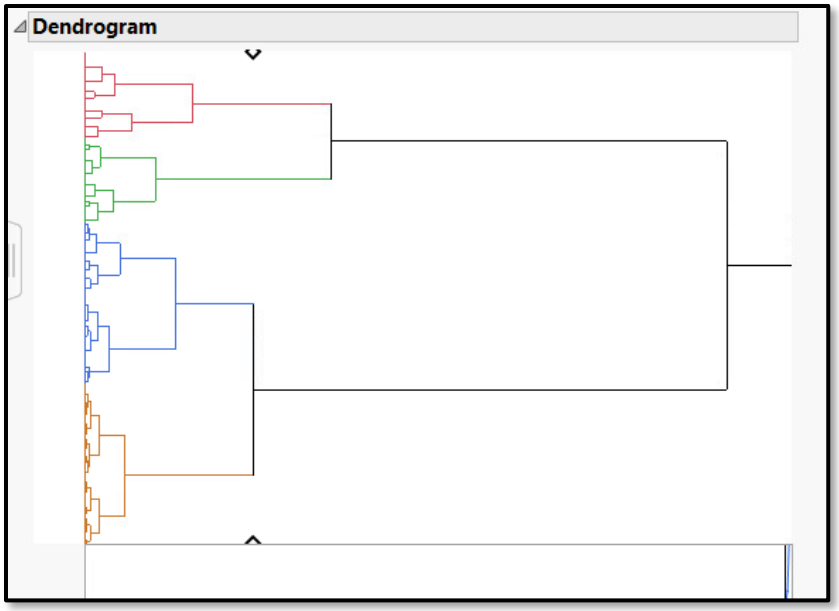

FIGURE 5.

I then sought to test the strength of comparison between restaurant spending and customer age. The graph below (**Figure 6**) shows the density of customers based on the overall amount spent. It can be seen that the highest concentration (indicated in the darkest shade) is located around the 20-45 age range, at a $0-50 purchase range. I then explored the potential connection between restaurant profit and customer income. **Figure 7** shows the greatest customer purchase density around the 35-70k income range, with purchase amounts between $0-50.

FIGURE 6.

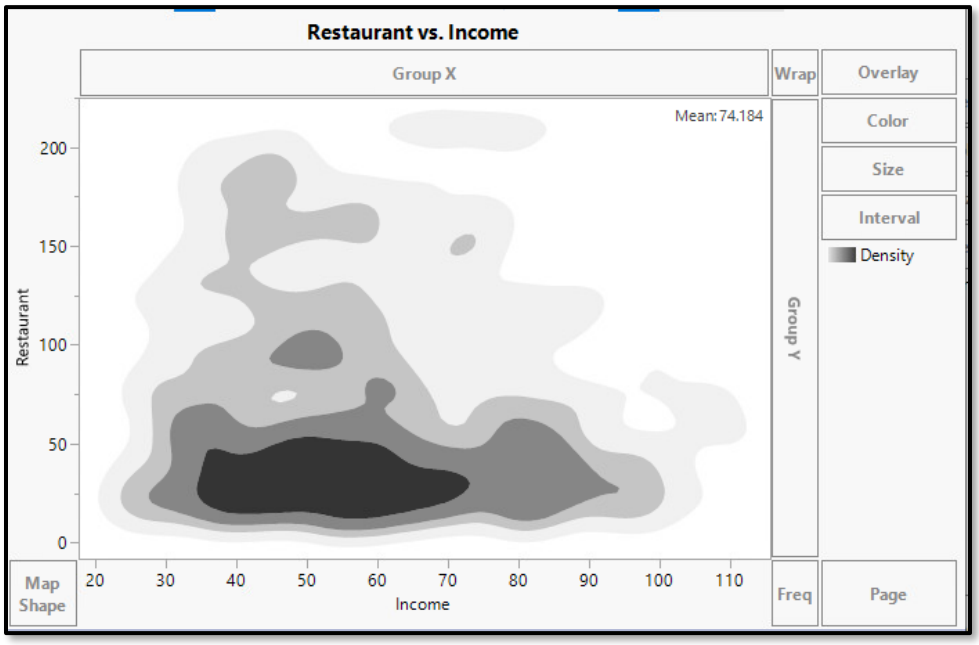

FIGURE 7.

A comparison of married customers vs. non-married customers across all platforms can be observed by the bar graph featured below in **Figure 8**. There is a much higher percent of married customers indicated by the bar on the right, making up 61% of the customer base. This

may speak to the restaurant atmosphere and Bubba Gump Shrimp as a brand, identifying as more of a family atmosphere or couple's destination rather than a location for singles.

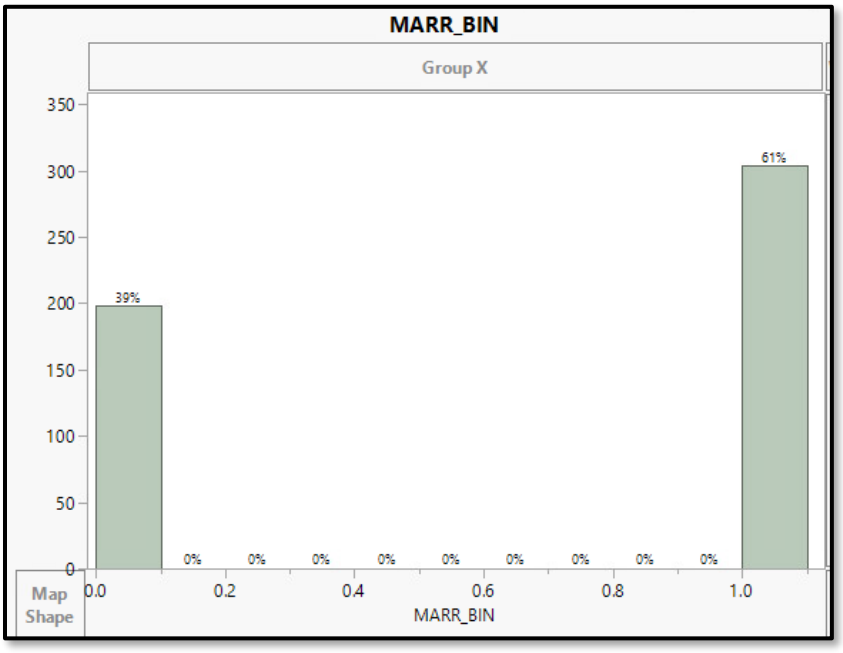

FIGURE 8.

If we consider location as a factor of spending, a map of the United States in **Figure 9** below illustrates the states with the least customer interaction. These color-coded states each represent one or two customer interactions from the data sample. This may indicate a more limited access to restaurants, or deficient marketing in these areas.

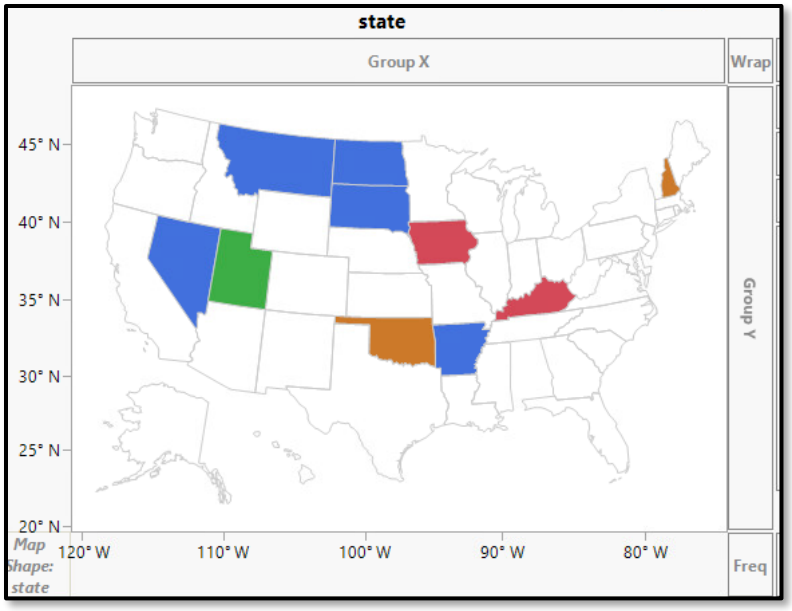

FIGURE 9.

A map which represents the maximum amount spent at restaurants by state can be seen in **Figure 10** below. According to the key, states in red have generated the highest profits, light blue/purple in the middle, and blue represents the states with the lowest profit. Research and marketing campaigns can be conducted with a focus on these states, to gain insight on factors contributing to their success.

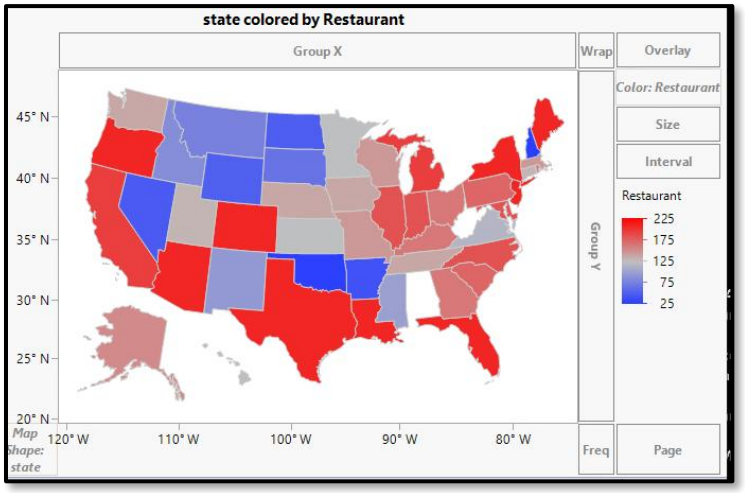

FIGURE 10.

The same graphics can be applied to webstore and third-party purchases seen in **Figures 11 and 12**, indicating areas of profit or states with little-to-no profit return.

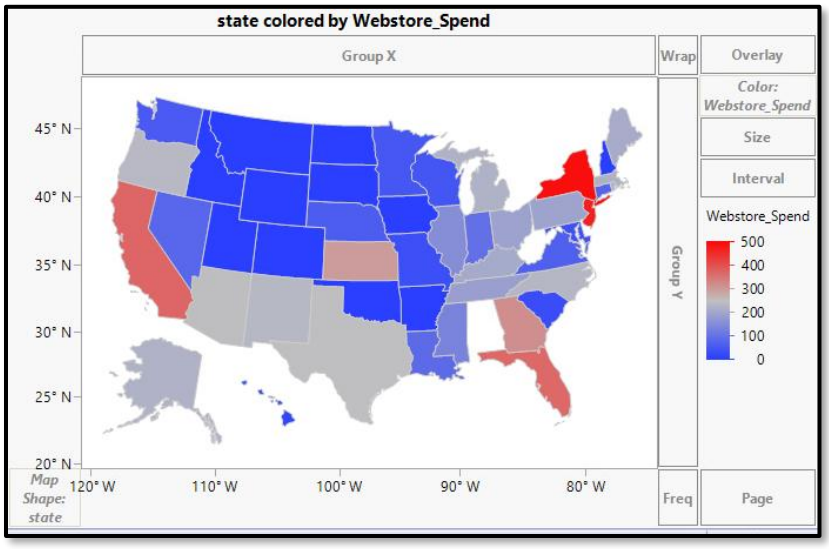

FIGURE 11.

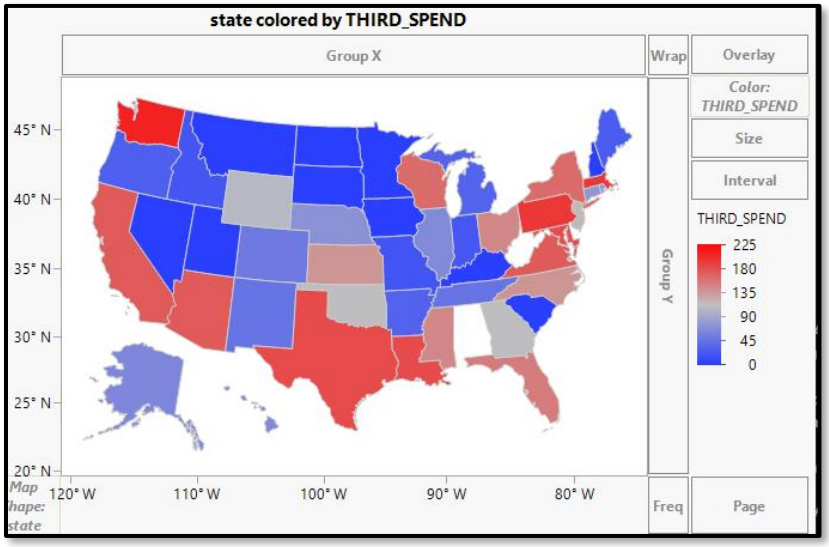

FIGURE 12.

To make predictions about the likelihood of customers making webstore purchases on the basis of various criteria, logistical regression provides a model for analysis and probability.

Comparing spending patterns across customer ages suggests that a younger demographic is more likely to purchase from their webstore (**Figure 13**). This might pertain to younger generations comfort with technology, or other factors which would require further research.

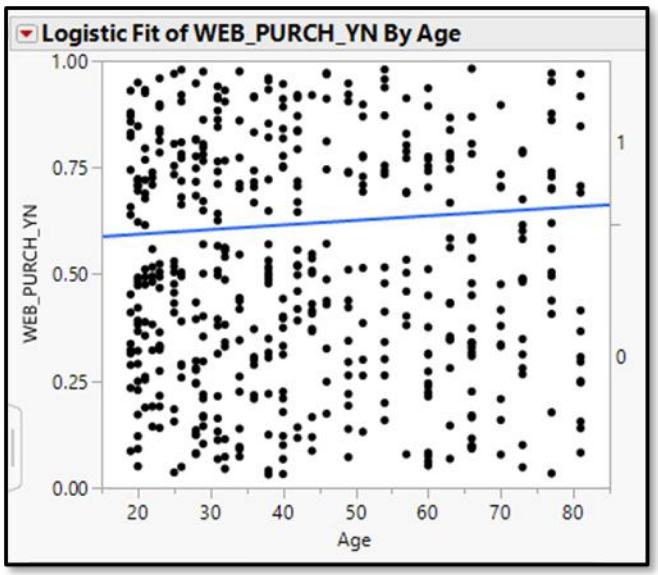

FIGURE 13.

Consistent with buying history, married customers are more likely to make purchases from the website (**Figure 14**).

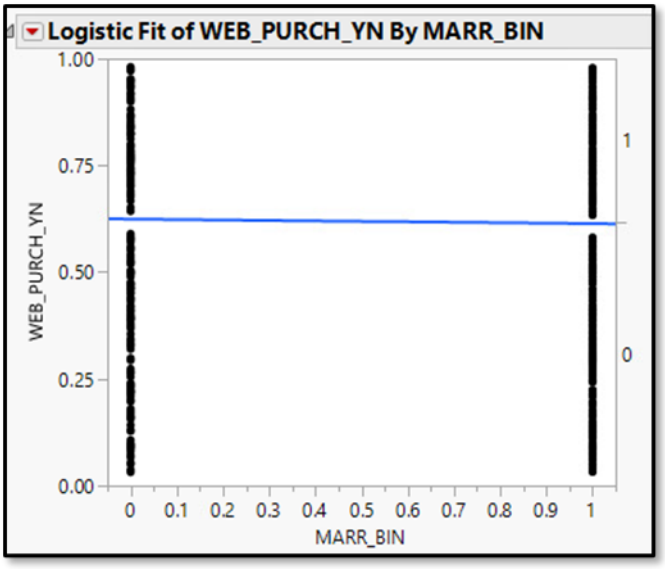

FIGURE 14.

Predictions regarding likelihood to spend at the webstore in relation to customer income suggest that Bubba Gump Shrimp has a higher probability of connecting with lower income customers (**Figure 15**). This would require further analysis, but could point to price points, brand identity, marketing targets or customer locations given the connection between restaurant visits and web visits.

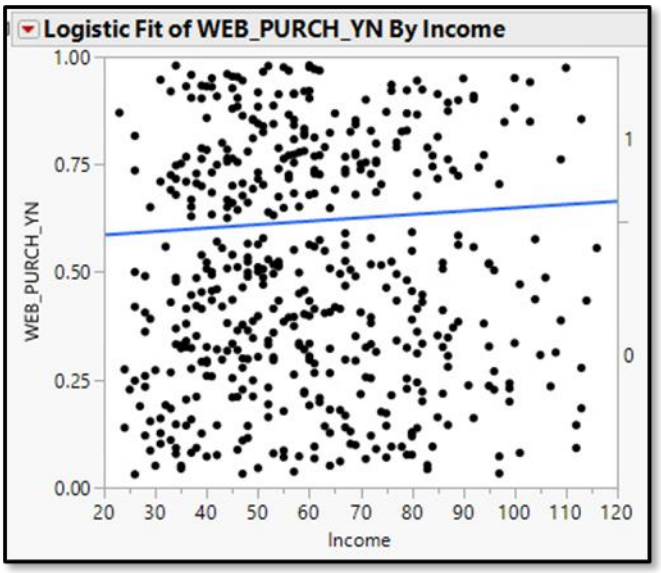

FIGURE 15.

Validity, Reliability, Limitations

Although we have established some meaningful connections, it is difficult to gauge the accuracy in applying the results to the entire Bubba Gump Shrimp customer base. In terms of data mining, 500 entries is a rather small dataset and the distribution of data may present the information disproportionately. Without extensive amounts of data, the predictions and assumptions made are meant as a foundation of which to explore a theory more in depth. Reliability of the dataset is based on user's entry accuracy, as well as any transferring of customer information. Without access to the customer database from which the sample was

derived, it is difficult to assess the validity of the results of the analysis. A comparison of the sample against other samples from the dataset may reveal other connections or trends. The objective of analysis is to provide insight into decreased sales. However, it is not possible to discern spending trends across time without knowing when purchases were made. More data related to purchase specifics and timelines could reveal more value insights. Upon performing the hierarchical cluster analyses amongst various sets of continuous variables, the results were expansive which made clusters difficult to decipher. The volume of clusters and indistinguishable results were limiting to this model of study. Since data correlations don't necessarily suggest causation, they are not enough to dictate business decisions, but rather expose trends that are worth exploring further.

Resulting Decision Influence

Although there are some limitations to constructive analysis with the data sample provided, some inferences can be made by further examining significant clusters. Through the hierarchical clusters distinguished above, the company can categorize customers by spending habits across each platform. Further observations about locations and demographics can be made regarding the customers who make up each cluster. Research and marketing campaigns can be employed to target these specific groups, to expand the current customer base, pinpoint areas of deficiency and tailor the brand to a target audience. Some correlations worth exploring further such as restaurant and webstore purchases, locations of greater/less customer spending and prime customer demographics specified above pertaining to age, marital status and income. I would explain to the client or supervisor the benefits of supplying a larger sample size or various samples for comparison. With variables that better address concerns about decreasing sales, they

could make more accurate predictions about spending patterns and factors that may be contributing to their losses.

Visual Evaluation

Overall, the visualizations chosen above will provide a valuable representation of the findings. I have provided descriptions which define the graph results, and explanations to express the significance to their business. The visuals that track spending patterns by state are self-explanatory. Other depictions on their own might require some basic understanding of graph and data mining fundamentals. The graphs which illustrate logistical fit may be difficult for the client to interpret without explanation. A more expansive understanding of the JMP software and data mining principles would allow me to create visuals that represent that data in a more digestible way. Although there might be more dynamic visual options, I think the graphs and depictions chosen accurately represent the findings.

Next Steps

Based on the connections gathered thus far, further research should be employed with an emphasis on groups connected to higher customer spending. States with higher rates of spending can speak to successful marketing campaigns, local food cultures, customer demographics based on those locations, and other factors which may be contributing to their success. These findings can facilitate measures for increasing spending in states with less customers. Research that explores the connection between these assumptions requires more data. Bubba Gump Shrimp can seek to obtain other means of acquiring data through customer surveys or post-visit reviews. Customer reflection on their experience with Bubba Gump can provide valuable insight into

aspects of the business which contribute to declining sales, or rather, factors which facilitate customer return. Surveys can address influential elements such as price points, menu items, restaurant atmosphere, customer service, expediency, and location. Incentives for leaving reviews and for survey participation can encourage customer engagement. These new data sources will strengthen the existing database. Another essential data incorporation includes records maintained by the company. Visit summaries which include items, dates and prices, allow data researchers to track spending habits and monitor sales more thoroughly, and is of greater significance to the business problem: declining sales.

Obtaining more data will allow us to not only confirm the results of our preliminary analysis but expand on these connections and discover new relationships. More data may validate the positive correlation between restaurant and webstore spending or introduce other factors to support this hypothesis. The new points of comparison acquired through surveys and reviews may attribute areas of customer dissatisfaction to declining sales.

References

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

Crié, D., Micheaux, A. From customer data to value: What is lacking in the information chain?. *J Database Mark Cust Strategy Manag* **13,** 282–299 (2006). https://doi.org/10.1057/palgrave.dbm.3240306

JMP Learning Library. Using jmp. Retrieved March 19, 2021, from https://www.jmp.com/en_us/learning-library/using-jmp.html