Introduction: Business Problem:

The main problem that presents itself in the prompt is the stagnation and decline of sales within the last two years. Fortunately, an integrated data warehouse has been created for current and future analysis. How can the newly acquired data be analyzed in a way that will help the future of Bubba Gump Shrimp?

Introduction: Analytic Method:

It seems like Bubba Gump Shrimp was not analyzing their customer base correctly before. Online services, third party retailers, and the restaurant’s customer base weren’t treated as the same group of customers. Combining these groups of customers but understanding their differences can create good insight to what the company is lacking.

The new strategy that will be used is combining the three major categories of customers within one data warehouse. While the customers should be divided up into three categories, there should be a customer ID assigned to each customer that is consistent. This would allow analysts to observe the buying habits of customers who crossover into the different categories.

Because a decline has taken place for the last two years, analysts can look at the historical data to determine the factors that could have caused this decline. The analyst can observe current and past trends. For example, Bubba Gump Shrimp offered a senior special for it’s first 5 years of business and discontinued it for the last two years. Analysis could go back to the data and if customer age was available, they could search to see if the amount of seniors who visited the restaurant sharply declined around this point.

The ideal results from this analysis would give insight as to what went wrong and why the company is headed in a decline. Hopefully, the decline was caused by an easily changeable factor. Some of these factors could include: cancellation of an item that could be brought back, cancellation of a promotion that could be brought back or at least updated, or a possible change in overhead management practices could also be a cause.

Analysis Tools:

To determine the data mining tools needed, what kind of analysis is being done and the data required must be examined. In order to do proper data analysis, the samples of data selected for this must be meaningful, relatively small compared to the whole of data available, and reliable. In order to get this kind of data, a few tools can be used: text analysis, the database, metadata, and the environment the data is contained in.

Data Visualizations:

As far as data visualization goes, I’m going to stick with what I am familiar with: graphs. Using JMP, I can collect many different kinds of data visualizations in the form of graphs. Primarily, I think I’ll use a histogram. Other tools are available in JMP that will be useful in addition to a histogram.

Research Question

The research question I want to work on is: “What is the cause of decreased business across physical and online stores?” Answering this question in particular will undoubtedly help the company understand what is going wrong, and probably give them an idea of how to fix the problem.

Research Measurement:

If my research querrey is solved with a straightforward answer, collected from stable, reliable, and meaningful data, then I will consider my attempts successful. In this scenario, I have no way of incorporating my answer into business practice, so I have no idea how successful implementing my solution would be.

Follow-Up Questions:

If a solution is found to my initial question, then how to implement a solution could be considered. For example, if the decline in active customers is due to a discontinuation of a product or service, can the service or product be brought back to bring in customers? There could be some issues with this hypothetical. When a product is no longer served at a business, such as a brand of breadsticks at a restaurant, it is often due to the fact that the brand has discontinued the product and it is unavailable. In the scenario where a service is cancelled, it may have cost the business more than it was worth.

Research and Support:

There does seem to be a large pool of information waiting to be tapped: the customers. As long as the customer’s information is kept within the database, then these customers can be reached out to. Surveys could be written up and sent out to customers in the form of emails, snail mail, or robo-calls. As long as a good selection of customers replies, this can give great insight into why they personally either decreased their activity with the company, or why they continue to use this business. Of course, this could have some issues. It is possible not enough customers will respond, or the answers aren’t entirely honest.

Analysis Organization:

As a visual person, I start my analysis by creating a graph. I choose the graph depending on what variables I want to compare. For example, if I’m not sure how data correlates, a cluster dendrogram is a good start to see how data is naturally correlated. I also have to acknowledge what variables should be compared on different graphs. Named variables and continuous variables, for example, usually shouldn’t be compared on a scatterplot matrix. Some of these aspects, however, are out of my control. Only, owning the data dump, I can’t easily create new variables with data I don’t have. I could create a new boolean variable for male/female depending on the names, but I couldn’t make a variable for the total amount spent at a restaurant by each customer because I don’t have access to that information.

Sources of Error:

Some of the variables seemed to have limits. The variable for number of reservation visits and number of web visits never hold a value greater than three. The only two reasons for this are either the data samples’ variables just happen to never reach a value greater than three or the variables are inaccurate. Either way, a maximum of three visits won’t tell me how or where the business is declining. It would be best to select a new batch of sample data to make sure the variables’ values aren’t limited in some way.

Meaningful Patterns:

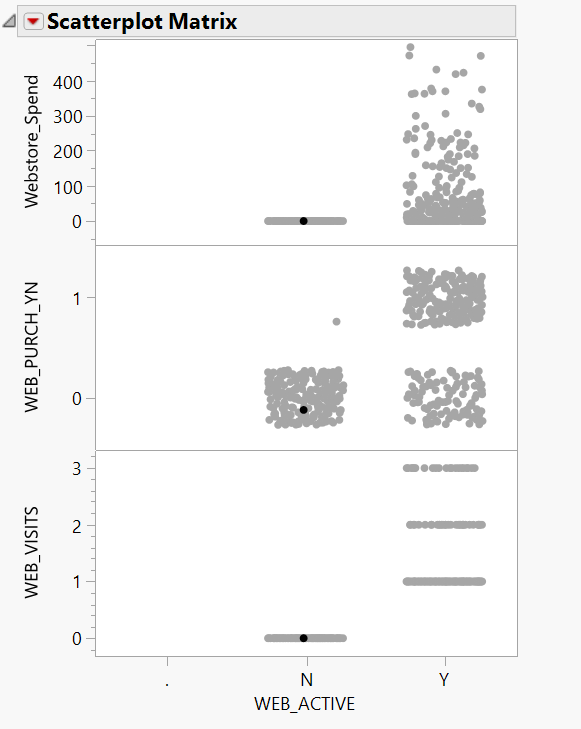
I’ve noticed that the income earned by the customers is very symmetrical and forms a bell shape. Those who earn the most are middle-aged, but there is also the most amount of variance in this age category. Older and younger people within this data set tend to make less on average. Utilizing this data could help the company appeal to those who have more money to spend at their restaurant, and therefore maximize profit. Another pattern I noticed is that the first numbered 60 restaurants have a declining amount of returning reservation visits. This could be indicative of a lot of things: of poor service, bad food quality, lacking management, etc. This is an issue that should be looked into deeper. A new variable should be introduced that consists of previous customer’s ratings. This rating variable could be compared to the reservation visit number variable to see if the customer experience affects whether or not they came back the next time.

Inaccurate Depictions of Data:

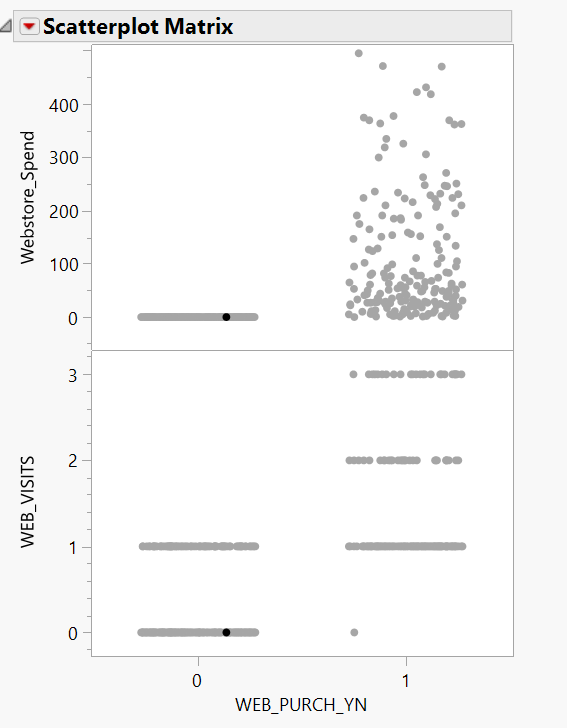
When analyzing the website data, I noticed that a lot of the customers who went to the in-person restaurants didn’t always use the website or purchase from it. This made a lot of customers show up as zero for web visits or web spending. When looking at the data through a graph, the data is very skewed towards zero. Instead of representing the data like this, I would recommend adding another variable that is boolean, true or false, for if the customer has been on the website before at all. Narrowing down the customers to only ones who have been on the website will allow for better website customer behavior analysis.

Further analysis:

After adding a boolean variable to determine whether or not a customer is web active, a new scatterplot matrix was made. This one uses 4 variables. The WEB\_ACTIVE variable is the x-axis, and the other three displayed are on the y-axis.



This is in comparison to the previous variables with this scatter plot which only used 1 x-axis variable and 2 y-axis variables.

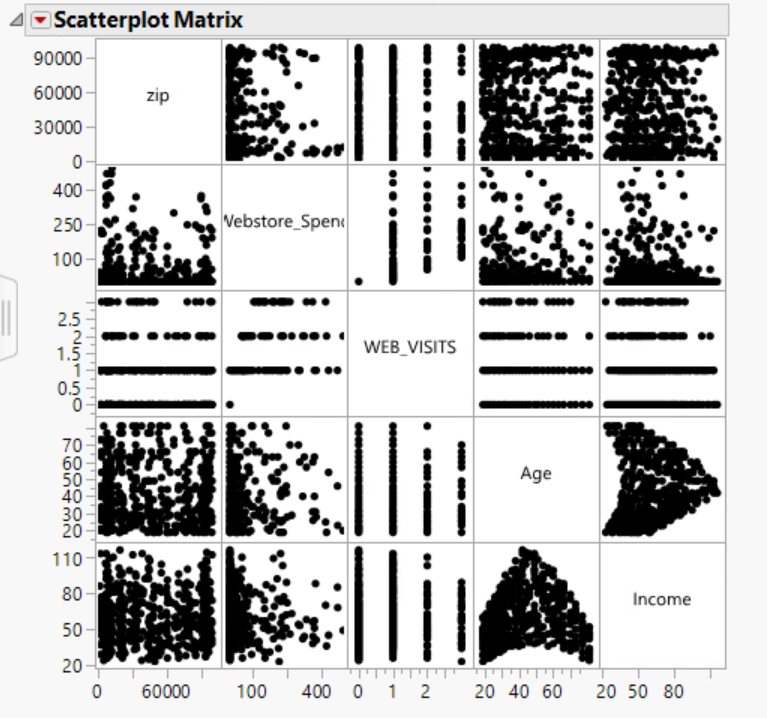


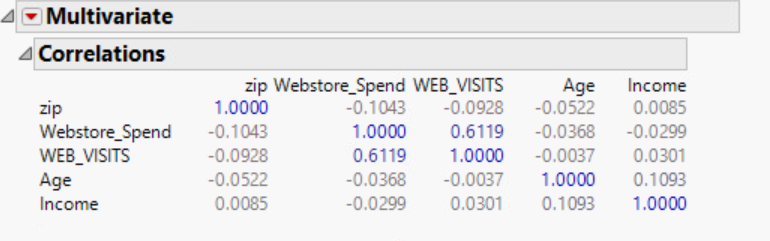
While these two results may look very similar, this additional variable helps create a distinction between web active users, if they make purchases, and how much they spend as opposed to non web active members. This eliminates the customers with zero web visits from inferring with the other data. This change in data display shows how adding this extra variable helps one greater understand the customer base.

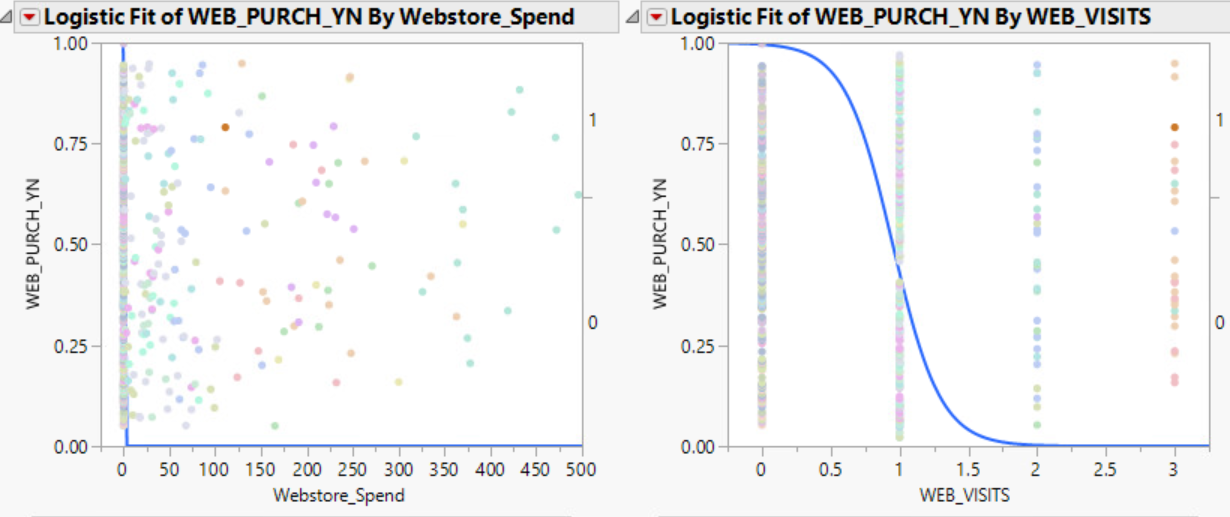
I did notice, however, the singular data point where there is web purchase with zero web visits. Scanning through the data, I found customer 198 had no web visits, no web spending, but was marked as yes for web purchase. As the only variable I created/ adjusted was the WEB\_ACTIVE variable, this was an error within the dataset.

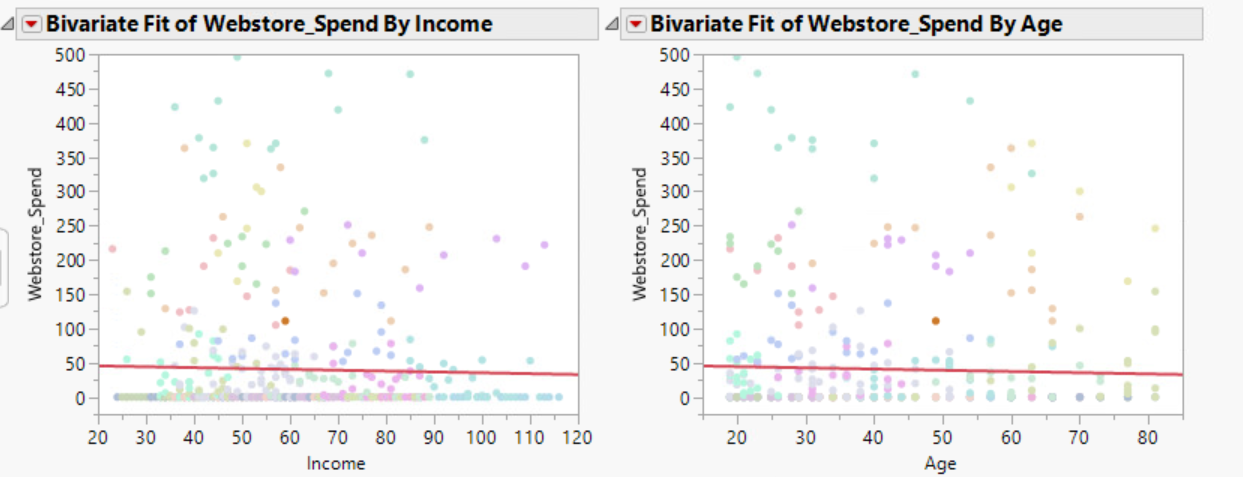
Alternative Analytic Methods:

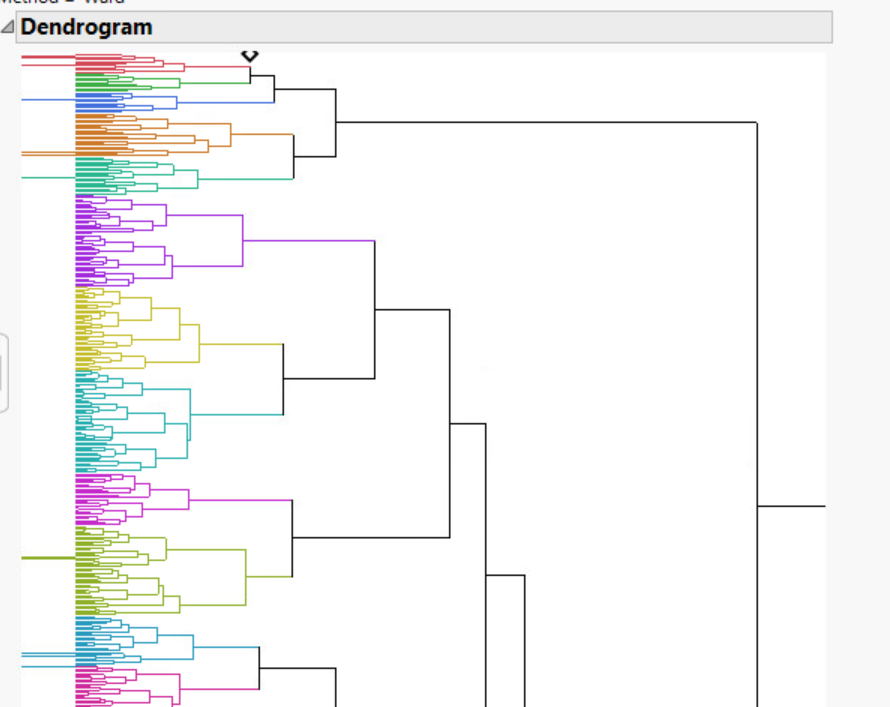
A lot of the information in this sample data is unneeded or jumbled together. Certain variables like names and marital status don't lend much to finding out the reason behind this business declining. A simple customer ID number assigned to each customer instead of a first and last name would reduce the number of variables involved. Also, if the customer ID number was given to customers in the order they became customers, then one could determine how long the customer’s have been contributing to the business. Other dated variables such as the last time the customer went to a restaurant would be helpful in determining the frequency of the customer visits and therefore value. The importance of the marital status is debatable, I think it could be removed without changing the implications of the other data.











Display and Interpretation:

The above graphs compare variables to visually display their relationships with the amount spent by each customer on the Bubba Gump website and the number of visits by each customer. Looking at these graphs, one can see that many of the customers never visit the website, and ,therefore, never spend any money on the website. The amount spent or number of visits doesn’t seem to be dependent on variables such as age or income. The customers from this sample data are also shown to never visit the site more than three times.

Validity, Reliability, and Limitations:

As far as validity is concerned, the sample data set selected is questionable. The data set itself is very small. It was never made known the size of Bubba Gump’s customer base, but for the 100+ physical restaurants and online store, this seems negligible. I think a larger data set would improve validity. The data with some variables doesn’t seem very reliable. The data is presented as survey information, which isn’t always reliable. It would be much more reliable to keep customer information from either visits to the restaurant or the website on file and use that data. Survey data can be incorrect. For example, if a customer underestimated or overestimated the amount of times they went to a particular restaurant, this could cause problems. This data set provided also has limitations. Certain variables like number of reservation visits and number of web visits are capped at a value of three. This makes it difficult to see where the business starts to decline. Looking at the data visually, it seems that the customers just suddenly stopped going to Bubba Gump’s after 3 visits. The same is true for the website. Having data related to dates could also help pinpoint where customer activity declines and what possible event could have triggered it.

Resulting Decision Influence:

In order for my senior to make the best decision, I would need my report to be easy to read, understand, and build off of. This would mean my visuals and interpretations are concise and insightful. A problem I would imagine some analysts having is overstimulating others with information. Outlining the potential problems to be solved with their help in a way they understand is best. It would also be wise to pair the information I’ve gathered with new data such as sales records from the website to back up my hypothesis.

Visual Evaluation:

I think the most helpful visual representation was the scatterplot matrix. It is very straightforward and allows for several variables to be evaluated at one time. The least helpful visual representation was the hierarchical cluster graph. Only a few variables can be included at once before it looks like a mess. Maybe with a small data set, this kind of graph would work a lot better, but not with a sample data size of 500. I could color the different sections, but this didn’t help too much. Labeling all the little sections would be too confusing and take up a lot of space. The bivariate and logistic fit graphs were more helpful mathematically rather than visually. However, the fit line equations and the information that section provided was useful.

Next Steps:

I think that once the possible errors of capped variables have been solved, a new hypothesis can be made. It appears that many of the customers for the website and in-person restaurants don’t have much crossover. I think that if more in-person customers were made aware of the website and potential deals there, then there would be more online sales. My advice to the company would be to send out advertisements about the Bubba Gump website to in-person only customers. This can be done via emails, snail mail, or advertisements in the restaurants.