The Bubba Gump Shrimping Company is a retailer that focuses on regional food. It does its business via restaurants and through other retail channels such as online purchasing. They gained a lot of attention thanks to a blockbuster film where the company’s name was used within the movie. This move was called Forest Gump and it featured a depiction of the creation of a company called Bubba Gump Shrimping Company taking the names of Forest and his friend Bubba. Its purpose in the film was to follow the dream of the passed Bubba. The real world business this though had a lot of sales increased due to the popularity of the film. However with the film now being over 2 decades old the extra attention they got as a result has long since faded. The company has recently integrated all of its data into a data warehouse. Prior the data was not all together and you could not link pieces of it together as easily. The major problem Bubba Gump is having currently is that for the past two years sales have been declining. This is a key problem in any business, once sales go down it becomes a slow fall towards eventual business failure. This is all unless something is done to increase sales or at minimum keep sales above costs in hopes of finding a solution down the line. To start I was provided with a sample of 500 customers that were selected from the analytics data warehouse (taken from various aspects of the Bubba Gump data so long as they made a purchase at least once from an outlet, restaurant, web store, etc.) and given a satisfaction survey. From the data provided by these surveys I hope to be able to find whether there are natural “clusters” within their customer population.

The purpose of doing this analytical work is to provide Bubba Gump Shrimping Company with the data and statistics they need in order to improve their company. Currently they are in a multi-year sales decline if that keeps up the business may be in danger. Prior to the data warehouse the company could not tell if one customer was making purchases on multiple of their retail platforms. Now customers can be assigned customer IDs that link their purchases and data throughout the Company. This means that their data can be used to better their customer experience and help gain more sales overall. Since the sales of the company have declined over the past few years it may be a good idea to do an analysis over time. They could examine the past few years to see if there was some cause to the decline, for instance maybe the company sold rancid shrimp from one of its providers and this led to a hit in sales. Having a view of the last few years could also allow the company to see if there was a trend, it is possible that the company’s sales had stalled or were already going down but were within the margin of error of the time. With the data stored within the data warehouse the company may be able to create a priorities list. Giving different priorities based on customers within certain areas of the company. For instance if a restaurant sales in a particular state are doing well but are doing worse in another they could give the customers in the area with better sales a higher priority. From there they could then examine what is causing the lowered sales at the lower priority area then try to improve on any issues they may find. The company can also use the 500 surveys to help guide these priorities when it comes to customer satisfaction. For instance if a particular product is not well liked it could be improved or removed to save on costs and allow the company to focus on more profitable products.

There are two data mining tools that I plan to use are Customer surveys and JMP. Customer surveys are going to be used to find out customer data and potential learn some of the potential issues using said data. JMP will be used to compile the data we obtain with the surveys and then create visual representations to better interpret the data. Using either tool can help with the research process but when used together they provide a better view of the overall data. With a better data quality it will allow for more precise and efficient changes to be made within the company.

In my report I will use bar graphs, frequency tables, and histograms. Bar graphs give us an actual visual of what the data is showing us. In this graph form it can be easier to interpret the data over looking at a bunch of numbers on a spreadsheet. Frequency tables provide the key values within the data for instance it provides the means, medium, minimum, and maximum. Histograms show the data in a way that allows us to see the general range of the data. This also makes it easier to notice any data that falls out of this range, the outliers.

There are a few research questions that can be created from the situation however I want to focus in on one. The question to be answered is “How can I use data analytics to identify clusters within the data and use those clusters to improve sales at Bubba Gump?” Using this research question I hope to go into the data and attempt to find those clusters. When found I will use data analytical tools to provide a visual representation of the clusters. Once that is done I will work with the data to try and find a way to help improve sales at Bubba Gump.

The research question becomes partially answered once the clusters have been found but the full extent of the question requires more detailed examination of the data. When I have devised a way to improve sales using these clusters I will consider the question answered. The hypothesis-generation was successful if the sales of Bubba Gump improve as a result. To measure my progress I will have to monitor the future data of Bubba Gump once my method is utilized. If the data shows an increase in sales using the method I can assume that the process was successful.

There are a few questions that can be created after the initial research. One of which would be “Are there a lot of outliers within the data?” If the data has a bunch of outliers within it they may want to attempt another survey in order to get data with a more clear representation of the range. With a bunch of outliers it becomes harder to make a clear judgement in regards to the data. A second question that could be asked is “Are there missing data types in the data?” If the data is missing categories or if there is not enough data available it is possible that the answer you come to may not be useful since you were missing key data.

There may be published sources of the hunt for the data clusters within Bubba Gump’s customer surveys. However I was unable to find any using basic search criteria. As well as I am under the impression this data and idea are completely fictitious meaning most of the similar things I would find would be other students’ work. If they fell short I would say it is because of a point I brought up earlier, the data may be missing key data categories. Since the data we were given were limited to what was on the survey we may be missing important data that could change the overall view of the data. Since I was unable to find any I do not believe they will be able to help guide my analysis.

With the analytical steps done during Exercise 5 multiple models were created in order to view the data. By going through a few of the variable sets and creating regression models you are able to determine or at least get an idea of what you want the final models to be like. Once you have the variables decided you can make the proper models that will make the report more meaningful. There are a few aspects out of my control, the main of which being the limitation of the data being provided. I am limited by only having the types of data the survey asked about. For instance if there were extra details that were added such as “Household Size” or something of the like you can use that to make any analytical processes more refined and provide a better guideline going forward.

During the analysis there were a few times that I ran into a minor problem. The problem was that the variables being compared would show very little in terms of the models. For instance when comparing some variables the lines of best fit had only slight angles on them. This makes it harder to see the actual trend unless you physically get closer to the model to see the slight declines. The other option would be to go through and look at the raw numbers but that is a less appealing option when visual representations are key. A few times to fix these problems I had to cycle which variables I was viewing for the sake of finding any change.

One pattern I saw that seems to warrant a bit more research is that being married led to more web visits. During the logistic regression the line of best fit should that web visits were more common when the customer in question was married. This may mean that a key factor to sales in general would be whether or not the customer had a family to provide for. An additional question on the survey asking about household size as I mentioned earlier may provide a more details on this topic to help with future research. If it is true that a customer would be more likely to purchase from Bubba Gump if they had a family it would lead to the question of if the company should focus on more family oriented sales, such as advertising saying you should bring the family in for a meal or something of the like.

Looking through the data you can see that a lot of customer values within the data are zero. A zero value makes sense in terms of raw data but when it comes to trying to use that data in the analytical process it can cause issues as it will skew data to the lower end. Though at the same time these zeros also make it clear in some cases of that data type’s meaning not being utilized. An example would be if the Web Visits had a majority of the data say zero visits then it should be clear that the web aspect of the company either needs work or needs to be removed. A way to resolve this would be to remove these zero values during calculations while keeping their totals in mind. This would allow you to use the non-zero values to get a good idea of the general data. Keeping in mind the fact that there were a certain number of zero values will also allow for you to make general judgements, for instance if a majority of the values were zero you can see a clear problem.

There are a few potential future methods to utilize with the data. The first would be to issue a few different types of surveys. These surveys would each focus on one aspect of the business allowing them to get a good amount of data on each part and do calculations later without missing out on important data by only picking bits and pieces. A second option would be to do analytical examinations of certain parts of the data. For instance maybe check all the data in a certain state or town then based on that data you can improve locations in that area to accommodate. These two options would provide meaningful data analysis while at the same time giving the company good points to focus on in each respective area. This could include service or locational changes that the data would help guide.

After doing analytical comparisons and viewing the data via models there are a few things to make note of. Firstly a fair amount of the data did not coincide with other factors. For instance the general locations of the restaurants did not seem to shift the data much overall. However there was a nice trend when it comes to marriage and purchases. It seems that a customer was more likely to purchase online or in general which means that if the company wants to increase profits they should focus in on these factor.

Since I am simply using the data to benefit a company I am not a part of the report will be factual and help to guide the company currently paying my fee. This means that the validity of my report as it is sound. In terms of reliability the report uses data provided by the company which means overall it can be considered reliable. However you do have to account for if the customer did not properly fill out the surveys or lied (whether intentional or not) which could lower the overall reliability if the data is fabricated in any way. My report is limited however by the amount of data provided by the company. With more than 500 sets of data I could create more refined analytical models. These models would then allow for further understanding of the data as it is.

When it comes to facilitating a potential decision in regards to results determined by this data I would like them to focus in on the key points. When looking at a set of data to determine a decision you want to focus in on the important items. If you look through every bit of data you not being able to decide properly. However if you focus in on key points that summarize large portions of the data you can make decisions with a fair amount of confidence. In this case I recommend focusing in the idea of being family oriented. Since more of their sales are seemingly coming from people who are married, meaning that they are buying food for at least one other person as well. If worked upon this idea should help to provide the company with an increase in sales.

The graphical representations of the data will be shown in a way that is easy to understand. If you can understand the data properly it will be easier to make decisions in regards to the data. I hope that the method in which the data is presented adds to the overall appeal of the report. Alongside the findings I believe the company may be able to utilize the data to shift their company in a way to increase overall sales. Once the figures and graphs are in place the results should be presented in a nice easily understandable way that may lead to further inquiry if needed.

Going forward the company may want to try and continue their analytical views of the data. If they use bigger sets of data say for instance one thousand customers instead of just five hundred they may be able to find trends or data clusters that were not visible with the amount they used. It may also be a good idea to focus some of the analytical steps one certain sets of data like for instance from a certain zip code or state to see if there are changes that need to be may in that area. A new hypothesis for this could be “Sales are effected by the states the restaurants are in”. For instance maybe the shrimp sales could be poor in states where the shrimp takes a while to reach the restaurant in question and are not as fresh overall. Another potential hypothesis “Focusing in on family oriented may lead to better profits”. With a high amount of people being married that make purchases it may be a good idea to look into becoming family oriented in order to produce better sales.

Below are some data models showing the comparisons created by whether or not the customer was married to their restaurant visits, web store spending, web store visits, and finally a comparison to third-party retailer spending. All data used in these models were provided by the Bubba Gump Company in the Bubba Gump data file.

**Logistic Fit of Married\_YN By RES\_VISITS**



**Whole Model Test**

| **Model** | **-LogLikelihood** | **DF** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Difference | 1.56637 | 1 | 3.132742 | 0.0767 |
| Full | 333.68550 |  |  |  |
| Reduced | 335.25187 |  |  |  |

|  |  |
| --- | --- |
| RSquare (U) | 0.0047 |
| AICc | 671.395 |
| BIC | 679.8 |
| Observations (or Sum Wgts) | 500 |

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Intercept | -0.7771573 | 0.2167823 | 12.85 | 0.0003\* |
| RES\_VISITS | 0.25435612 | 0.1433912 | 3.15 | 0.0761 |

For log odds of N/Y

**Logistic Fit of Married\_YN By Webstore\_Spend**



**Whole Model Test**

| **Model** | **-LogLikelihood** | **DF** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Difference | 0.05418 | 1 | 0.108364 | 0.7420 |
| Full | 335.19769 |  |  |  |
| Reduced | 335.25187 |  |  |  |

|  |  |
| --- | --- |
| RSquare (U) | 0.0002 |
| AICc | 674.42 |
| BIC | 682.825 |
| Observations (or Sum Wgts) | 500 |

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Intercept | -0.4444355 | 0.100876 | 19.41 | <.0001\* |
| Webstore\_Spend | 0.00034137 | 0.0010346 | 0.11 | 0.7414 |

For log odds of N/Y

**Logistic Fit of Married\_YN By WEB\_VISITS**



**Whole Model Test**

| **Model** | **-LogLikelihood** | **DF** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Difference | 0.00716 | 1 | 0.014326 | 0.9047 |
| Full | 335.24471 |  |  |  |
| Reduced | 335.25187 |  |  |  |

|  |  |
| --- | --- |
| RSquare (U) | 0.0000 |
| AICc | 674.514 |
| BIC | 682.919 |
| Observations (or Sum Wgts) | 500 |

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Intercept | -0.4408333 | 0.1257 | 12.30 | 0.0005\* |
| WEB\_VISITS | 0.0137957 | 0.115217 | 0.01 | 0.9047 |

For log odds of N/Y

**Logistic Fit of Married\_YN By THIRD\_SPEND**



**Whole Model Test**

| **Model** | **-LogLikelihood** | **DF** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Difference | 0.04169 | 1 | 0.083383 | 0.7728 |
| Full | 335.21018 |  |  |  |
| Reduced | 335.25187 |  |  |  |

|  |  |
| --- | --- |
| RSquare (U) | 0.0001 |
| AICc | 674.445 |
| BIC | 682.85 |
| Observations (or Sum Wgts) | 500 |

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **ChiSquare** | **Prob>ChiSq** |
| --- | --- | --- | --- | --- |
| Intercept | -0.4427539 | 0.1009108 | 19.25 | <.0001\* |
| THIRD\_SPEND | 0.00060094 | 0.0020772 | 0.08 | 0.7723 |

For log odds of N/Y

The data modeled above follows along with my previous hypothesis. My hypothesis was that it was more likely that a married customer would spend more and visit more on average. Starting with the comparison of whether or not the customer was married and their number of restaurant visits, on average the customer was more likely to visit more if they were married. This is proven by the fact the line of best fit is increasing as the displayed data goes further. When it comes to the comparison between whether or not the customer was married and their web store spending as well as their web store visits, it appears that on average a married customer is more likely to spend more money and visit the web store more often. The line of best fit increases in both of these cases meaning that there is a positive increase over time. Finally looking at the comparison between whether or not customers were married and their spending at third-party retailers it seems that married customers tend to spend more on average at their party retailers. The line of best fit in this data model is increasing over time meaning that on average married customers will spend more at third-party retailers. The fact that all of these comparison show positive increases helps to prove my hypothesis that married customers will be more profitable overall and that the company may want to focus on a more family oriented approach.

As a last set of comparison I will be providing a few comparison between other variables. The comparisons being done will be webstore spending compared to age and income of customers. All data used in these models were provided by the Bubba Gump Company in the Bubba Gump data file.

**Fit Group**

**Bivariate Fit of Webstore\_Spend By Age**





**Linear Fit**

Webstore\_Spend = 48.054521 - 0.1783736\*Age

**Summary of Fit**

|  |  |
| --- | --- |
| RSquare | 0.001351 |
| RSquare Adj | -0.00065 |
| Root Mean Square Error | 87.90191 |
| Mean of Response | 40.474 |
| Observations (or Sum Wgts) | 500 |

**Analysis of Variance**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Ratio** |
| --- | --- | --- | --- | --- |
| Model | 1 | 5205.1 | 5205.12 | 0.6737 |
| Error | 498 | 3847919.5 | 7726.75 | **Prob > F** |
| C. Total | 499 | 3853124.7 |  | 0.4122 |

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- |
| Intercept | 48.054521 | 10.03775 | 4.79 | <.0001\* |
| Age | -0.178374 | 0.217327 | -0.82 | 0.4122 |

**Bivariate Fit of Webstore\_Spend By Income**





**Linear Fit**

Webstore\_Spend = 48.046545 - 0.1265339\*Income

**Summary of Fit**

|  |  |
| --- | --- |
| RSquare | 0.000897 |
| RSquare Adj | -0.00111 |
| Root Mean Square Error | 87.92189 |
| Mean of Response | 40.474 |
| Observations (or Sum Wgts) | 500 |

**Analysis of Variance**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Ratio** |
| --- | --- | --- | --- | --- |
| Model | 1 | 3456.2 | 3456.21 | 0.4471 |
| Error | 498 | 3849668.5 | 7730.26 | **Prob > F** |
| C. Total | 499 | 3853124.7 |  | 0.5040 |

**Parameter Estimates**

| **Term** | **Estimate** | **Std Error** | **t Ratio** | **Prob>|t|** |
| --- | --- | --- | --- | --- |
| Intercept | 48.046545 | 11.98818 | 4.01 | <.0001\* |
| Income | -0.126534 | 0.189236 | -0.67 | 0.5040 |

Looking at these two new sets of comparisons. It seems that overall there is very little correlation between income and age when compared to web store spending. In the case of webstore spending compared to age the line of best fit is decreasing slightly over time, this means that there is a minor negative correlation between the two variables. In the case of webstore spending compared to income the line of best fit is decreasing slightly over time, this means that there is a minor negative correlation between the two variables. In both cases the lines of best fit are shifting very little over time meaning that the two values do not effect each other that much. This leads me to conclude that income and age will not overall change the profits of the company meaning they should focus on the previously mentioned family oriented approach as a possible way to increase profits.