Esmeralda Monteparo

DAT 220 Final Project

While taking a look at the business case, the problem was quite noticeable. Bubba Gump Shrimp Company started as a small, privately owned restaurant, and due to the enhanced exposure from being a main focal point in the 1994 movie, “Forrest Gump,” the business accrued an exponential amount of popularity and attraction, as well as expanded to where it operates several restaurant locations, sells branded merchandise through an online retail site, and wholesales their merchandise to other retail outlets. The name and recognition was the catalyst for the business’s rapid growth for the first few years. After those first few years, sales increased at slower rates, and then finally leveled off. Over the last couple years, though, sales have been declining.

Luckily, for Bubba Gump Shrimp Co, over these years they have collected a large amount of valuable data regarding the business, such as restaurant point-of-sales, web/e-commerce sales, customer information using loyalty programs as well as through online transactions with them or their third party retail partners. Using the data provided which consists of POS (cash register, credit card) data, customer database (collected from the restaurant’s loyalty program and online sales), web store sales transaction data, and customer and sales data from third-party retailers, it will help with constructing a data analysis/plan for revealing patterns and grouping like-data so that the cause of the problem can be clearly spelled out and then used to plan ahead for the future to avoid the risks and outcomes that occurred in the last couple years.

As per the prompt, now that the company has their data stored in a data warehouse, it will be easier to pull all the information in at once and view it all at the same time. This information can show us patterns of purchases from customers or groups of “like” customers (for example: patterns within certain communities), which can then be grouped between the restaurant, web channel, and third party retail partners to create data marts to view the trends. With these trends, it will be easy to address decision-making strategies to be able to improve the market strategy overall, as well as ensure that the business can avoid any more declines. There was also a survey for 500 individuals to fill out in exchange for purchase credits that can be used to gain insight on customer satisfaction, combined with extracting their details from the data warehouse, and to also aid in business improvements in targeting “clusters” or “subgroups” of specific communities.

In order to fully answer the question of why Bubba Gump Shrimp Co. is having a decrease in profits, we must understand how to utilize data mining to gain insight and help with resolving the issue. Data mining is defined as a “process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs. Data mining depends on effective data collection, warehousing, and computer processing.” (Twin, 2019). Since Bubba Gump Shrimp Co. is already set up to start analyzing the data due to having implemented a new data warehouse, the next step is addressing the tools and programs to be able to efficiently access, read, and analyze it.

One data mining/analytic tool I would recommend using would be RapidMiner. This program is a data science software platform that is used for business and commercial applications to provide analytic information through filtering, joining, merging, aggregating, (Guru99, 2019) and predictive models and graphs that are user friendly based off of raw data that is not online. (Wikipedia, 2019). A good tool to use for online or offline data would be DataWrapper. DataWrapper is a user-friendly, open-source web tool that can be used to create basic interactive charts by loading data into the program, and it translates it into pie charts, line charts, bar charts, and maps. (Visualising Information for Advocacy, 2019). These tools can take the data and output information based off of the results to allow the company to make their predictions based off of the patterns.

Once data is collected and analyzed, it becomes time to put it in a visual representation that will be easier for the user to understand, and more effectively see patterns, outliers or trends. Simplicity is key, especially when data mining many types of information is a new task at hand. I recommend using bar graphs and regression models for this reason, as well as the fact that they are easier to specifically visualize trends. Once these graphs are produced, they can easily be transformed or manipulated to the user’s preferences and focal points. In the case of Bubba Gump Shrimp Co., it will be beneficial to utilize this to see what has caused the decrease to occur, and then help with future decision-making to prevent this trend from continuing. (Heitzman, 2019).

Aside from the issue at hand being the decrease in profit, more specific questions need to be focused on. The one specific question that should be pursued is the reality of the customer loyalty program. Does the loyalty program offer good perks to interest the customers? Are the customers treated well or provided excellent support for their inquiries? Has the quality of the products remained the same since the increased spike of sales?

Following this question, as well as the research and graphical evidence, some follow-up questions should include asking whether or not the data and analysis tools are clean and accurate, as well as what adjustments can be made to cause a beneficial difference to profits and production. Bubba Gump Shrimp Co. has recently had a downwards trend in this respect, and the research, survey, and recent visualized patterns/trends will allow for awareness of the root causes of the situation, providing incentive to explore positive strategies for the future.

Utilizing the tools and recommendations above, there are numerous resources for a user to easily learn how to use them properly, as well as the ability to research different alternatives and similar situations. The tools offer support and assistance, and since they are widely used, questions and inquiries can easily be aided by other companies that use same or similar programs. These resources also ensure that what ends up being achieved are reliable, clean results.

**Analysis Organization:**

In order to piece the consumer population into different groups with comparable aspects, cluster analysis was what was used. Some aspects that were analyzed were age, income, spending amounts, and marital status. As an example of this technique, clustering can be used to segment consumers for the purpose of developing advertisement/marketing strategies towards certain consumers rather than an entire community in a broad specter, as to be more effective. Certain portions were found using every single characteristic. With these different groups, specifically those groups focusing on comparing amounts spent, the results concluded that a customer would spend more at the webstore the more they spent at the restaurant. With clustering analysis techniques, it is possible to utilize more than one variable to effectively group the consumers, rather than using only a single variable the way I did. Using only one variable would be one limitation to this study.

To form predictions based on what was spent at the webstore, the linear regression analysis was used. This model was incredibly useful for predicting the spending rates in comparison to the others, as well as evaluating how the different aspects directly/indirectly affected one another. Using this information, Bubba Gump Shrimp Co. can focus on certain groups of like-customers that are more likely to spend more, or to strategize new advertising and improve upon their marketing towards those who are less likely to spend more in order to appeal to those groups of clientele and increase their profits. Using many continuous variables, I was able to see that not every variable was efficient to getting the results I needed. I found that visits and sales for both restaurant and webstore were the most efficient when in comparison to each other, but what would have been more useful or less of a limitation would have been to use a few different variables outside of those provided in order to receive more accurate results.

To form predictions based on whether or not a consumer would make a webstore purchase, the logistic regression analysis was used. This could also be used to predict the potential for other groupings. For example, whether or not a consumer would purchase from a restaurant, or whether or not they would purchase a certain number of items. Using this type of information can increase profits and help advertising/marketing strategies by filtering what the most important factors are that connect with a certain group. I found that the logistic regression was harder to analyze, as the results were not incredibly accurate or easy to translate. One limitation for logistic regression is that it is restricted by the variables comprised, as well as the inability to analyze variables that are continuous.

**Sources of Error:**

There were some occurrences where the raw information was not reasonably reliable/sound. For an example of this, there were some variables that showed a customer who had a value for zero with web store spend, and a value of one that a webstore purchase was made. Due to this, it was difficult to determine which of these pieces of information were correct, so these were opted out of the data set. This type of error source will be checked prior to analysis. A second error source found incorrect in the data set was the variable of webstore purchases with y/n as the result being ordinal data. This hindered the usability of the data for certain assessments. This type of error can be corrected/used if the variable type is changed to nominal. The last error source found was the two variables for zip codes, one for the complete zip code and the other containing the first two digits. Some of the zip codes contain leading zeros that were removed. To correct this, leading zeros would need to be added to those with only four digits, variable type changed to character, and then recalculating the variable of zip code that calculates the first two digits. A possible error source that can’t be considered are errors that the customer makes. When the customer fills out the survey with accidental or incorrect information, this data gets transcribed into the data set. Human error exists in every form of analysis, so as a data analyst, it is wise to consider the potential reality of this being the case.

From now on, I plan to analyze the information contained in the data set in more of a step-by-step process. I plan to start with analyzing the data using the clustering technique to differentiate groups of consumers, and then follow it with various model techniques that best fit the situation to be solved for different types of predictions. Sometimes there can be information not found or information that doesn’t apply to different forms of analyses, so this would lead me to taking a step back and starting over again or bypassing certain steps. Possible parts of the data analysis process that I cannot control are time, and available data.

**Meaningful Patterns:**

The only patterns that I have come across from this data set are the more that consumers tend to spend at restaurants, the more they tend to spend at the webstore, as well as the more times they visit the restaurant or webstore, the more likely they are to make a purchase at the restaurant and webstore, respectively. Some questions that I have based on what these patterns indicate are as follows: Is there a correlation between the restaurant spending rate and the third-party spending rate? Is there a specific character trait between customers who spend more at the restaurant and web store versus customers who don’t? Given the answer to the most recent question, and if we can identify the main target of high-spending customers, is there room for improvement with our advertising? Is there a way to advertise and appeal to more groups of individuals in order to increase Bubba Gump Shrimp Co.’s profits?

**Inaccurate Depictions of Data:**

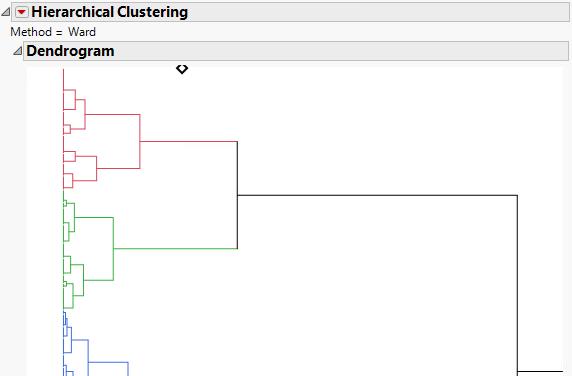
The only inaccurate depictions of the data set that came across were the graphs and models that were used by the regression models and matrixes. Since there was a large number of zero value results for the spending amounts, it made it challenging to analyze the charts. To resolve this, one would need to ignore this data from certain analyses.

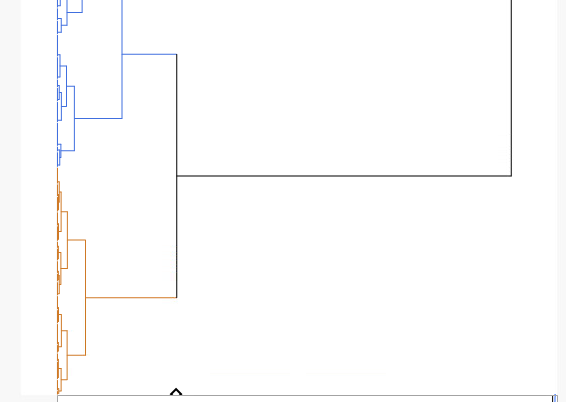
**Alternative Analytic Methods:**

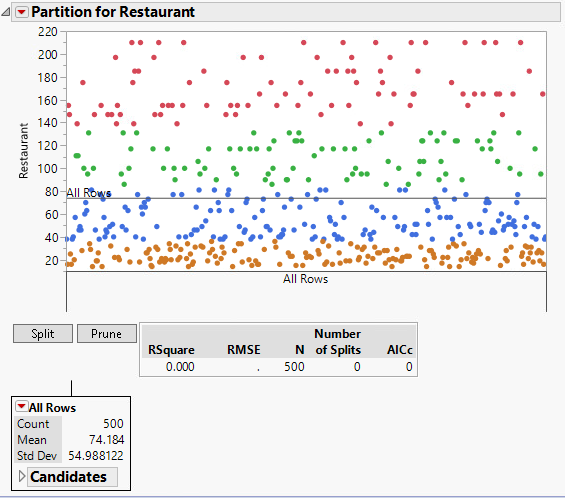
Based on the results, some alternative analytic methods could be different forms of hierarchal clustering, decision trees, k-means clustering, and stepwise/partial least square regression approaches. There are strengths and weaknesses to all of these that need to be considered dependent on the type of analysis presented. Some pay not require expectations of linear information which could be beneficial for those variables that aren’t linearly related with the response variables. Also, some methods are easier to interpret than others depending on who you are showing the analyses to, which is useful when it comes to displaying the proof behind the conclusions made and working towards improvements for the company. Lastly, gathering more or different types of customer data that can be compared against the customer location would be helpful due to the fact that having more valid information can produce more accurate results and draw a better picture. Some of these variables could be an area’s living expenses, and vicinity to the restaurants and third-party vendors.

**Display and Interpretation:**

The data that was analyzed to separate groups of consumers used to leverage advertising and increase profits is the6 Bubba Gump Shrimp Co. customer data. Since there are states that have fluctuating average restaurant spending amounts, a decision tree would be the best to use to sector states by the restaurant spending amounts of consumers who live in those states. For each state, the mean consumer restaurant spending amount was calculated, and then the data was analyzed by using this technique between the two variables of state (predictor) and restaurant spending amount (response). The sole reason behind this analysis was to differentiate which states had higher spending values and which states had lower and is shown in the first image below. In the second image, it displays the comparison between the restaurant spending amount variable and cluster. As shown in the image, there is a larger quantity of consumers who spend more in the very high cluster.

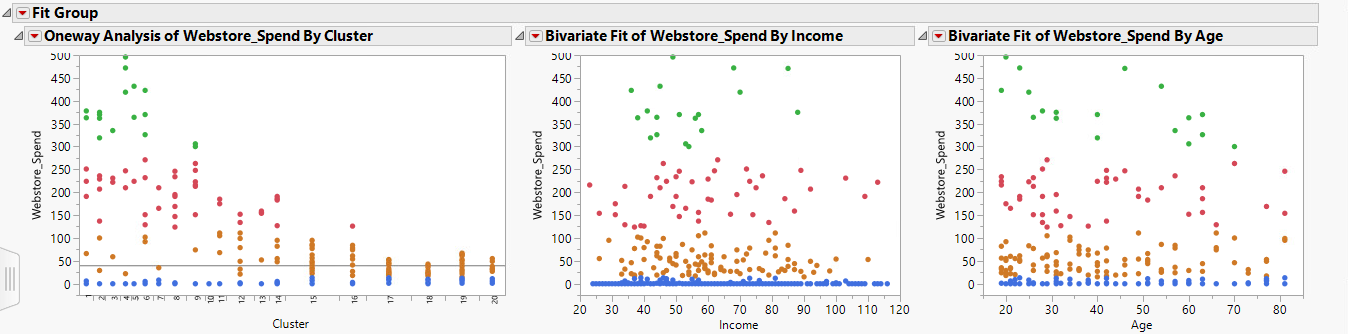


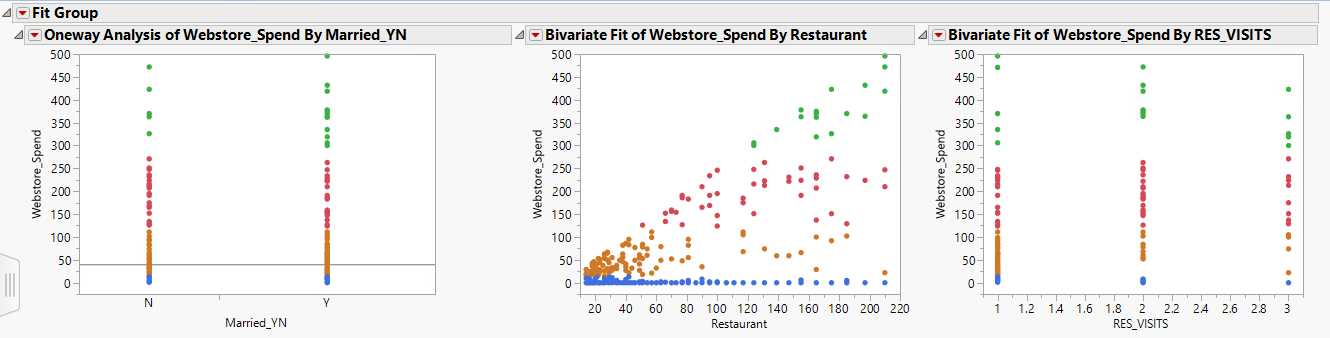




Based off of these results, Bubba Gump Shrimp Co. would most certainly benefit by focusing their advertisements on states that rank higher in order to increase return profits, since those in states that rank higher have a tendency to spend more at the restaurants.

One other analysis that was performed was a linear regression model. When it was discovered that webstore spending amounts and restaurant spending amounts were directly correlated, this technique was used by involving age, income, marital status, restaurant visits and spending amounts, and webstore visits and spending amounts as the key variables to compare against each cluster. In the below image, you will see the representation of the prediction of the amount a consumer will spend at the webstore when a purchase is made. However, this does not determine whether or not a consumer will/did make a purchase.





If the efforts of advertising focused on consumers that had a higher chance of purchasing from the webstore, it would be an efficient use of Bubba Gump’s budget because choosing a certain group of consumers that they want to market to would, theoretically, give them the highest profit return.

**Validity, Reliability, and Limitations:**

One limitation noted in the data set was that some locations only contained one customer. The specific variable that contained this issue was city. This problem affects the ability to generalize the analysis, especially when conducting the analysis while all locations with one individual residing happen to have a flawless correlation. Extracting a single customer’s information and survey results to represent an entire area population would be unwise and could skew the accuracy of the results. For states that contain an acceptable number of consumers within the data set, this form of generalizing should happen only after validation of the models being used.

In order to increase the internal validity for webstore spending amount models, each model can be built on two different sets: a starter set and a validation set. Once the sets are compared to each other, it will raise the level of validity and offer more of a guarantee that relations between the variables definitely exist in the data set.

Reliability of the data set analysis needs to be assessed based off the validation steps completed on the data, a level of control instilled when obtaining the data, and knowledge of whether or not the survey sample is actually illustrative of the consumer/group population. Bubba Gump Shrimp Co. gathered a survey set of 500 consumers, which is a small portion of their total consumer population. Due to this, there could be issues of whether or not the set actually embodies the reality of the situation and this idea should most certainly be considered when forming plans and strategies to improve the company based on the results of analyses performed. To increase reliability, there could be extra customer data taken into consideration that could potentially offer insight on the validation of the results.

**Resulting Decision Influence:**

The reveal from this data analysis can be used alongside an advertisement budget in order to determine which groups of consumers should be focused on. By using the cluster analysis technique, decisions can be made to distribute funds appropriately to areas that offer a higher potential for return. Information from current advertising strategies could be used in comparison with the results of data analysis in order to see whether or not the current strategies are efficient/effective. Bubba Gump Shrimp Co. will need to evaluate how evenly or not they are marketing across all areas, as this may be the cause to their decrease in profits. Fluctuating spending rates at the restaurant and webstore could be the result of varying marketing strategies.

**Visual Evaluation:**

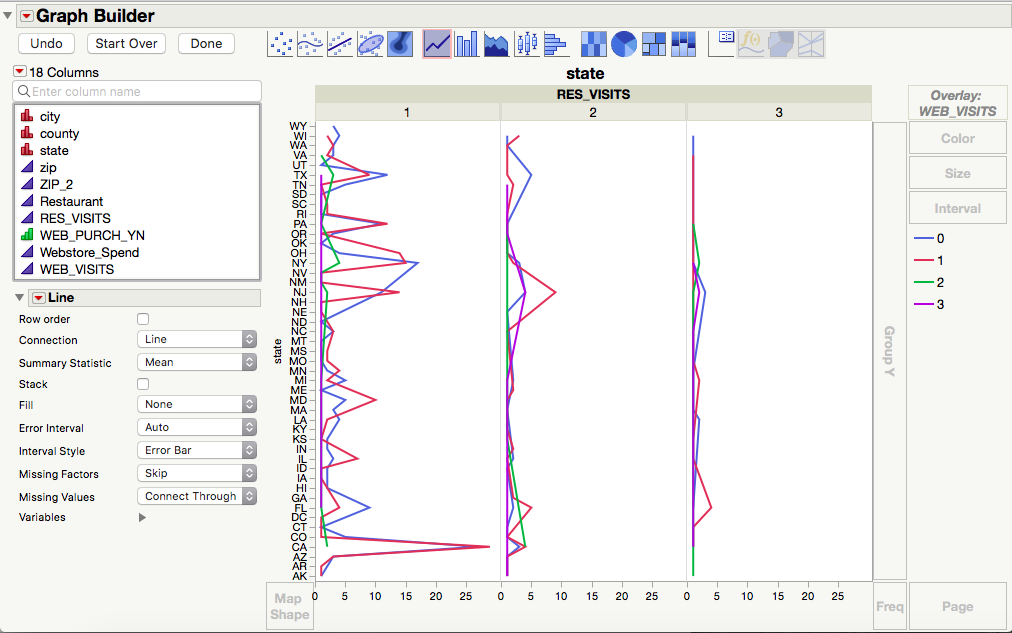
What is nice about the cluster analysis technique is that it can be shown to anyone, independent of whether they have knowledge of the data or data analysis process, and they will easily be able to understand what the visual graph is showing. What could always be improved is the visual perceptiveness of the graphs. Being more intuitive could help the audience understand it more quickly and wholly. Having different charts showing same or similar results could also help the audience understand the bigger picture as well. The more the audience understands the importance of what is shown in the analysis, the more likely they will be able to put a priority to the issue at hand.

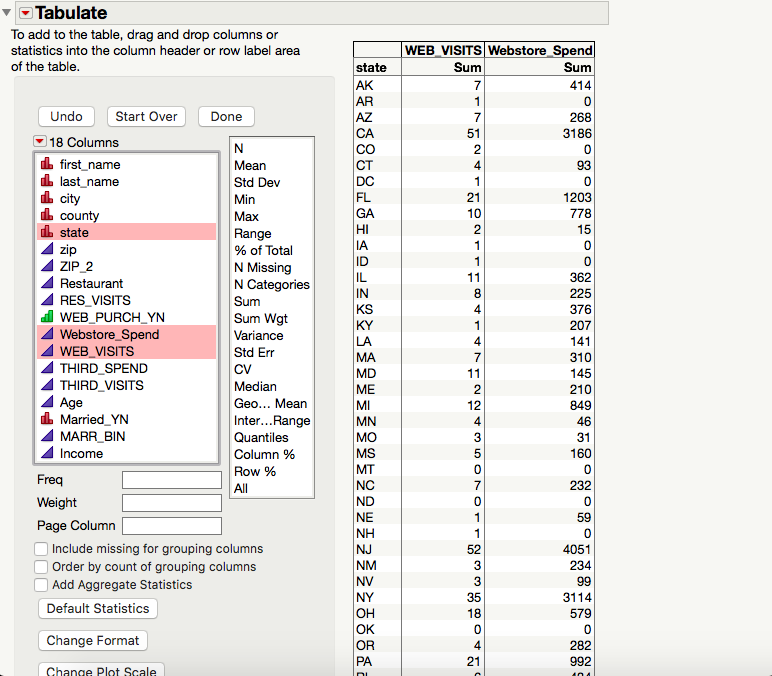
**Next Steps:**

The next steps that need to happen is comparisons of variables that were not included, as well as working on a method to obtain data that was not in the data set. An example of some pieces of data to work towards obtaining could be distance to/from the restaurant visited, number of items purchased, restaurant visits (y/n), loyalty program participants, or items purchased on sale vs. items purchased not on sale. An additional survey sample of the data warehouse could be obtained in order to compare it to the original data set, and then checked for consistency. Due to the fact that most of the models heavily rely on spending amount data, this data should be acquired on a regular basis using the new data warehouse or with the help of an additional program.

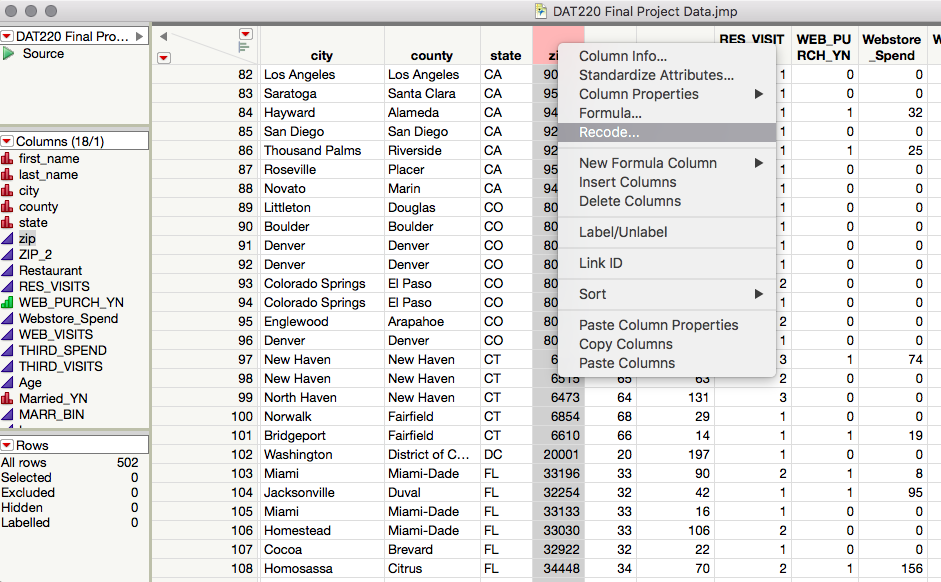
**Visualize and Summarize Data:**

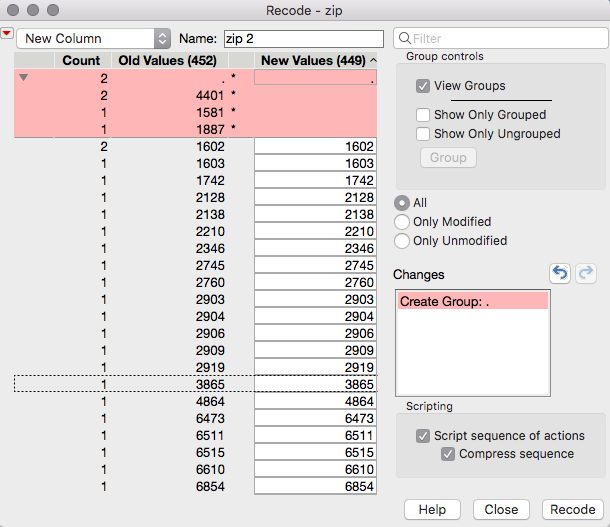
There are many advantages to the JMP program, but overall, the one advantage that it has that is most beneficial would be the fact that it allows the audience to be able to visualize the data analysis, no matter how well or not they understand or use the program. One of the features of JMP that becomes really useful is the Graph Builder platform. You can customize and change types of graphs with it by dragging and dropping variables in their respective axis, selecting which type of graph to display, elect to view summary statistics, add or remove additional variables, or edit the orientation. The tabulate function that JMP provides also creates tables to be able to be used in analysis. Using the same drag and drop features that the Graph Builder uses, tables can be created easily to display all sorts of information to compare against a multitude of variables. What is a great benefit to this would be when the data analysis is being presented, the graphs and tables can be made to be simple to understand for those who aren’t familiar with the program or objective, as well as be able to be shared with and viewed by individuals without the program utilizing JMP’s sharing functions.





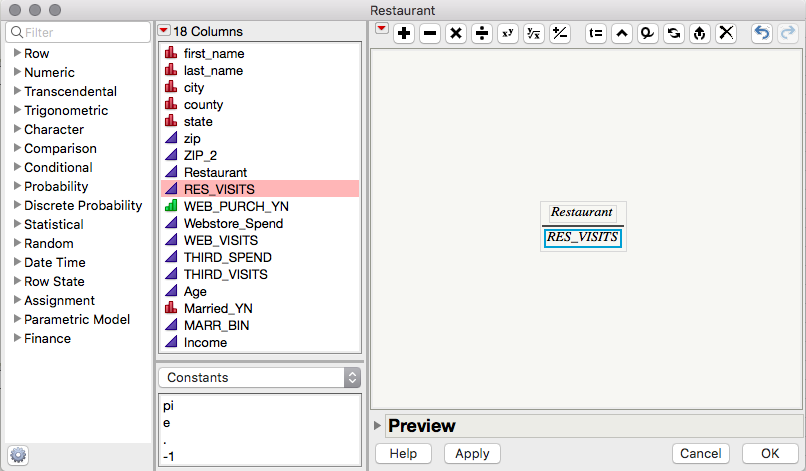
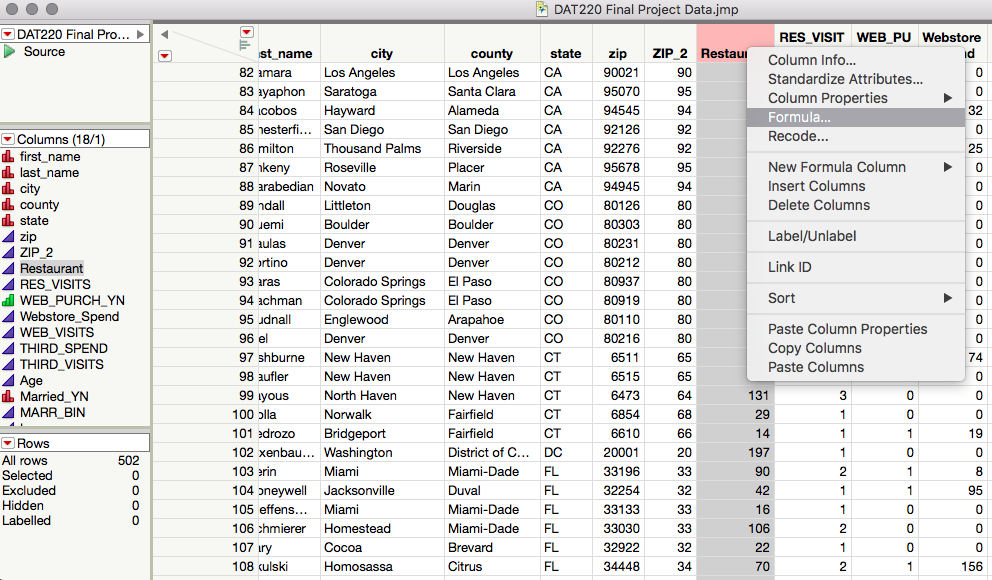
**Data Clean-up:**

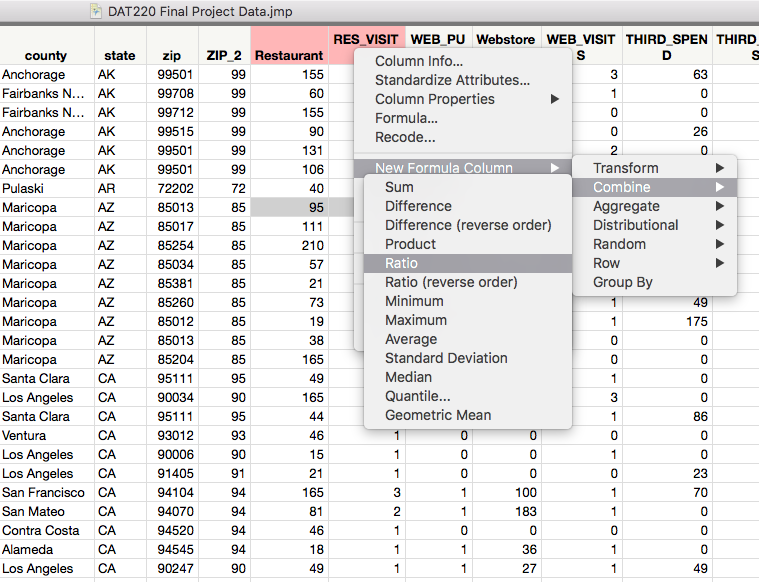
JMP has a functionality to be able to recode data. What this means is that when data is imported or entered and saved in data warehouses, at times there may be mistakes, misspellings, and inconsistencies. The recode function allows these errors to be grouped and corrected to eliminate as much human error as possible.

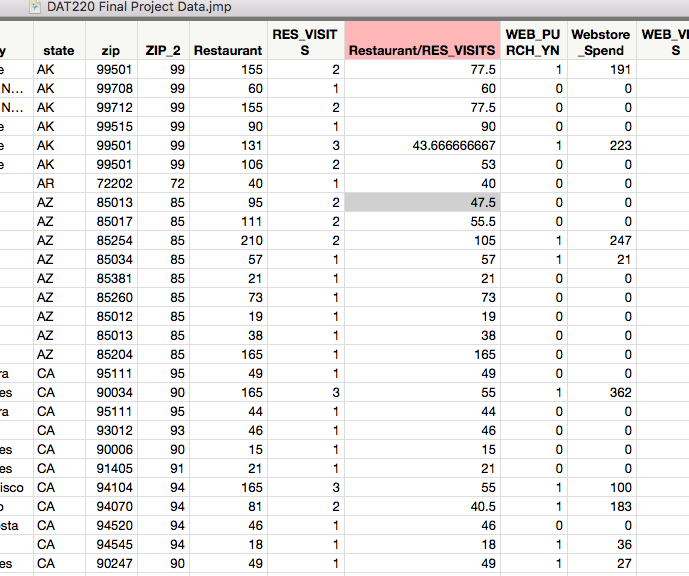


**Formulas:**

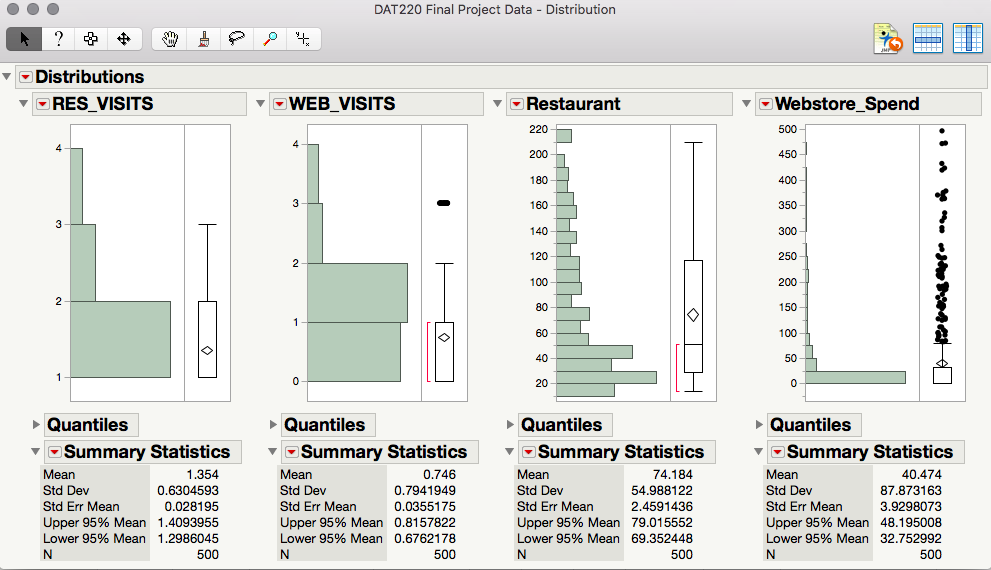
Formulas are easy to create in JMP, and even more advantageous if the results of the formula will then be used in analyses. This program contains a formula editor and an instant formulas function which take certain variables and compare them using a ratio function to produce percentage results. For example, this can be used to measure the percentages of visits at a restaurant versus at the webstore, or tip amounts to the bill amounts.

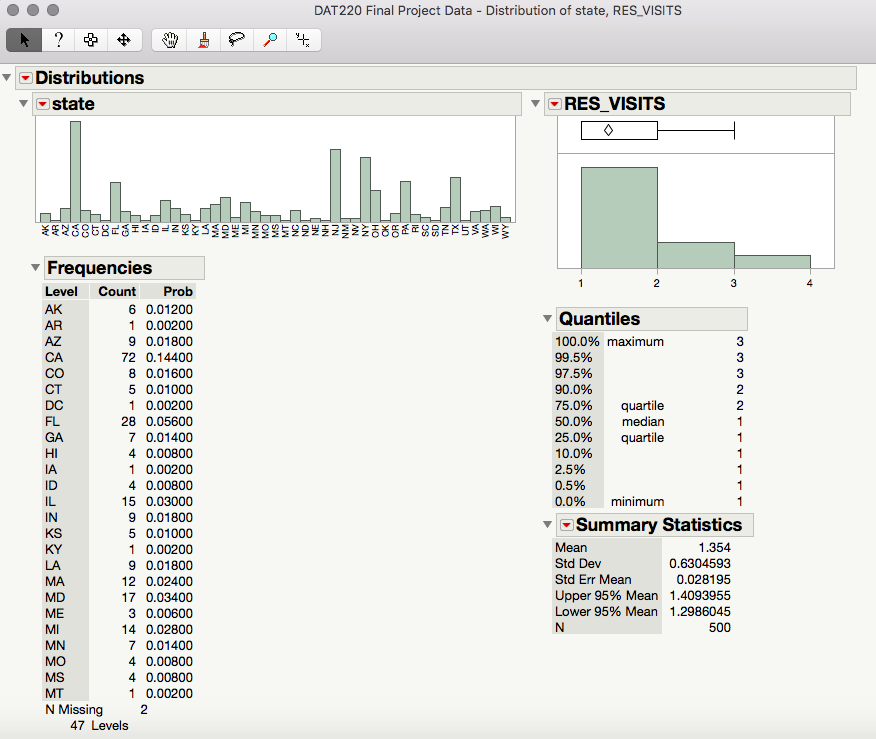






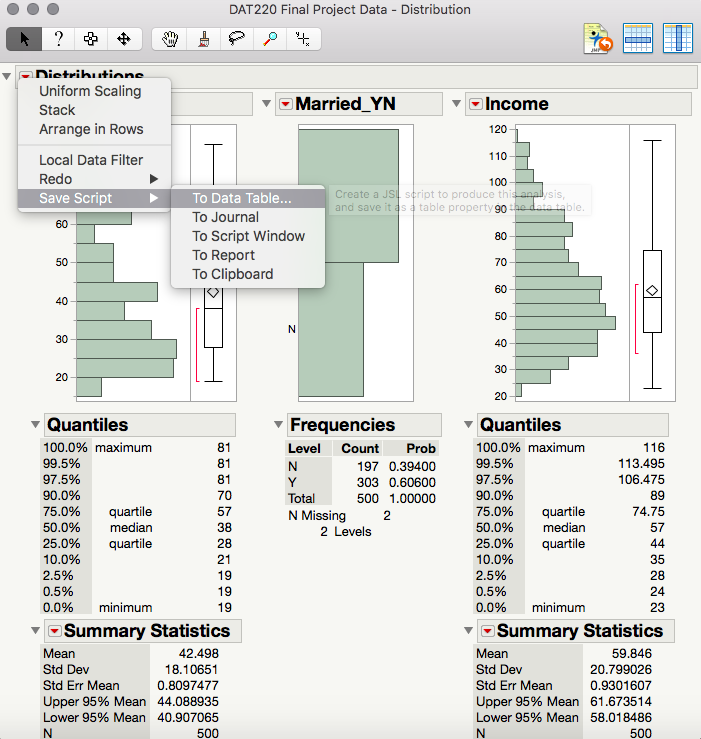
**Distributions:**

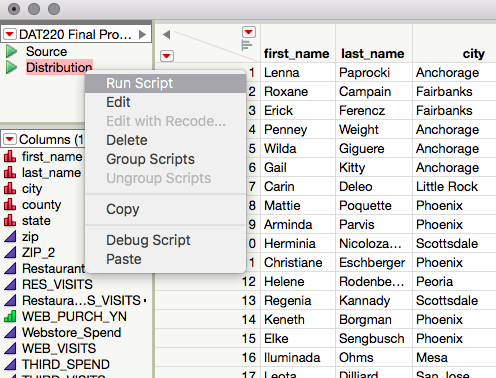
The distribution platform in JMP offers basic analysis function utilizing bar graphs to display multiple value distributions for each variable. This could be utilized, for example, to display differences in ages of consumers, or the number of people who visited the webstore. There is an option to display summary statistics, frequencies, quantities and more. This is a simple function to use to show simple graphs and functions for basic analysis. 

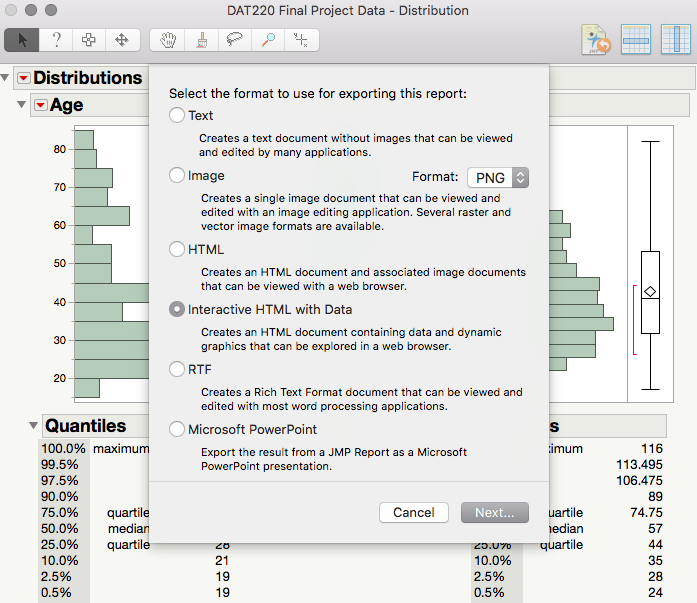
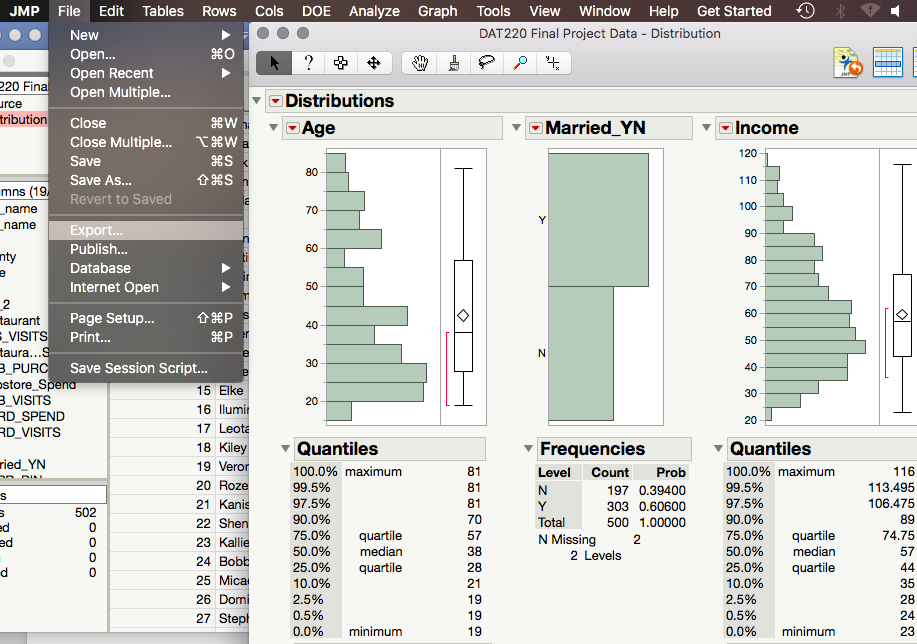


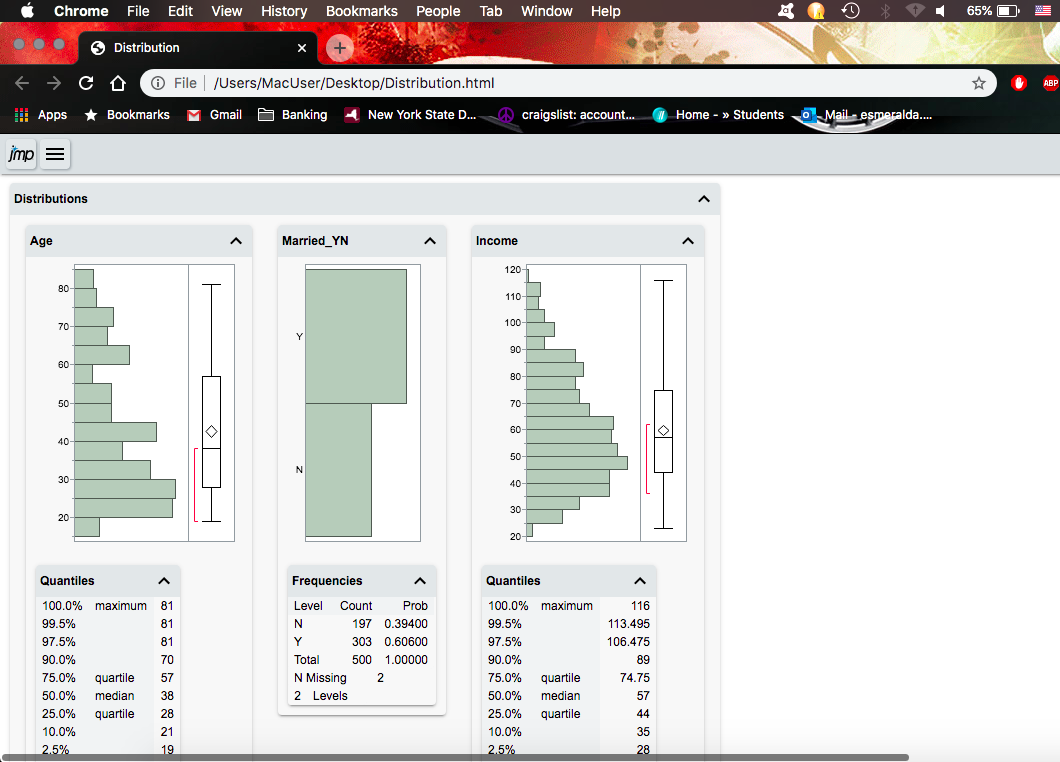
**Saving Scripts and Sharing Results:**

The last functionality to go over would be the saving and sharing scripts and results, respectively, and how they become handy when it comes to data analysis. JMP has the ability to save scripts that are run so that they can be accessed to produce over and over again. This becomes handy when a company has a recurring data analysis chart or table that they frequently use; it saves time from having to make/create it each time. As far as sharing results goes, JMP has the ability to output the results as an interactive HTML. This becomes useful when there are employees that need to access the data, but don’t have the program, or don’t fully understand how to work the program.









**References:**

Guru99. (2019). 25 BEST Data Mining Tools in 2019. Retrieved November 16, 2019, from https://www.guru99.com/best-data-mining-tools.html.

Heitzman, A. (2019, January 28). Data Visualization: What It Is, Why It's Important & How to Use It for SEO. Retrieved November 16, 2019, from https://www.searchenginejournal.com/what-is-data-visualization-why-important-seo/288127/#close.

Twin, A. (2019, October 8). Data Mining: How Companies Use Data to Find Useful Patterns and Trends. Retrieved November 16, 2019, from https://www.investopedia.com/terms/d/datamining.asp.

Visualising Information for Advocacy. (n.d.). Datawrapper. Retrieved November 16, 2019, from https://visualisingadvocacy.org/node/673.html.

Wikipedia. (2019, October 17). RapidMiner. Retrieved November 16, 2019, from https://en.wikipedia.org/wiki/RapidMiner.