M

Milestone Four

Databases

Eric Clancy | 4/4/2021

Southern new hampshire university   eric.clancy@snhu.edu

Databases are very useful tools when someone knows how to use the data in them. The first item a analyst should know is what data to collect. With the proper explanations as to why and how that data will be used in their outcome. This can grant a better perspective for the individual that is participating in the study with the data. The databases that will be used are smaller data sets and can simulate larger data sets. First, I will go through the database named, “roller coaster.xlsx”. I will use the program JMP to execute the visuals in the following figures. I am also familiar with the open source program JASP which has been very resourceful for home projects.

**Bivariate Fit of Duration By Length**





**Bivariate Normal Ellipse P=0.950**

| **Variable** | **Mean** | **Std Dev** | **Correlation** | **Signif. Prob** | **Number** |
| --- | --- | --- | --- | --- | --- |
| Length | 3754.571 | 1589.455 | 0.800886 | <.0001\* | 78 |
| Duration | 139.5128 | 42.6985 |  |  |  |

**\*\*\*\*\*\*\*\*\*Figure 1**

**Multivariate**

**Correlations**

|  | **Duration** | **Speed** | **Height** | **Drop** | **Length** | **Inversions** |
| --- | --- | --- | --- | --- | --- | --- |
| Duration | 1.0000 | 0.3918 | 0.3413 | 0.4326 | 0.8009 | -0.1006 |
| Speed | 0.3918 | 1.0000 | 0.7640 | 0.9211 | 0.5025 | -0.2269 |
| Height | 0.3413 | 0.7640 | 1.0000 | 0.7492 | 0.4209 | -0.1781 |
| Drop | 0.4326 | 0.9211 | 0.7492 | 1.0000 | 0.5328 | -0.2277 |
| Length | 0.8009 | 0.5025 | 0.4209 | 0.5328 | 1.0000 | -0.3750 |
| Inversions | -0.1006 | -0.2269 | -0.1781 | -0.2277 | -0.3750 | 1.0000 |

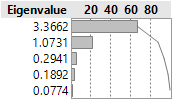
**Scatterplot Matrix**

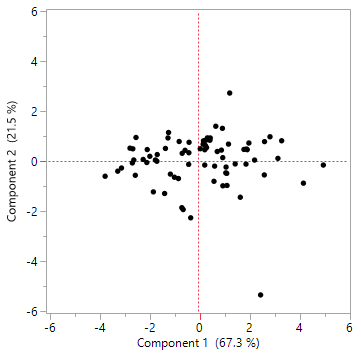


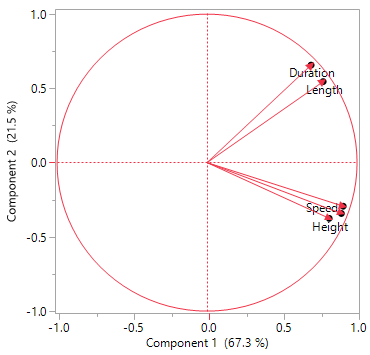
**\*\*\*\*\*\*\*\*\*Figure 2**

**Principal Components: on Correlations**

**Summary Plots**







**\*\*\*\*\*\*\*\*\*Figure 3**

With the gathered information from figures 1, 2 and 3 we can see the differences from a data set of roller coasters. Figure 1 is signifying the relationship between the duration in seconds you are on the roller coaster by the length of the roller coaster. As you can see the length of the roller coaster correlates to the duration you are riding the coaster. As the length increases it is seen that the duration of the ride is longer. The correlation shows close to +1 which shows an increase in the duration and length. The mean showing most coasters are around three thousand seven hundred feet and an average ride time of one hundred and forty seconds. This data set is coming from seventy-eight roller coasters.

Figure 2 is showing all the correlations between the variables from the data set. It is comparing duration, speed, height, drop, length and inversions. This matrix can show which points have a higher correlation than others. With the graph you can see that there is a significant correlation between the speed and drop of the seventy-eight roller coasters. When you look at the visual graphs of figure 2 you can see the oval is very tight when looking at the speed and drop. Almost all points are within the data set. Next in line would be that of figure 1, duration and length. Right behind that would be the height and speed of the roller coaster. Overall, the most significant values from the data set are the speed and drop. Which makes sense because the higher the drop of a coaster the more speed is gained from the gravity helping. This matrix also shows that the inversions of the roller coaster are not significant to all the other variables of the data set. As you can see in the matrix most of the ovals you can only see half of the oval trying to grab data from it.

In figure 3 there are three different graphs to see the data sets. These points are principal components of the data set. With this figure we are eliminating the inversions value because it did not bring much correlation to the data. The values closet to +1 are significant values for using the data on the roller coasters. Which is shown in the eigenvalues in the first graph as a percentage and the correlation. 3.3 range is at 67.3% which is having much data not that significant to use. At the 21.5% for being close to +1 this data can be used with significant relation to a roller coaster. Which shows in figure 2 matrix. The third graph in figure three shows the close correlations to the variables. Duration and length then speed, drop and height. When you look at the figure 2 matrix you can see those ovals are tighter than those distancing themselves. Like duration and height being the farthest apart on figure 3 and showing a larger oval in figure 2. Identification of principal components are beneficial to finding correlating sources in your data set.

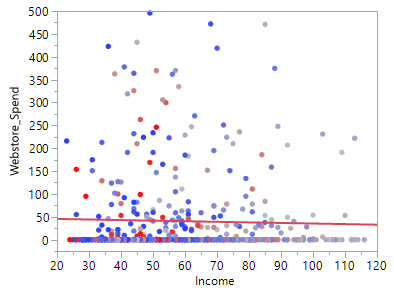
An online retailer would be able to use a vast amount of data collected. With the customer information the retailer can predict what you want to look at or at least interested in. The customer searches can be tracked to steer the guessing of the retailer to what you want to buy. Indirect contact with the customer with customer focused outlook when they utilize the site. With this creating a more personalized experience for the customer. Which can lead to the advertising that the company wants to invest in. Thee spending habits a customer displays can also increase the personalization for the retailer to home in on. An example would be from Netflix, the more movies you watch the more data collected. Then Netflix would be able to categorize movies that the customer would be more likely to watch. Two components of the data system would be searches and habits from the customer.

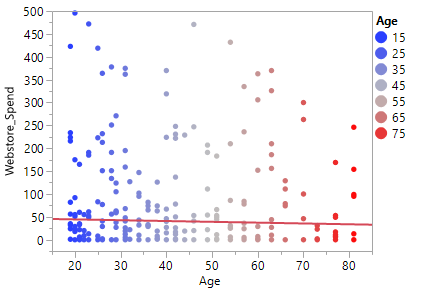
Data warehouse would be collecting all the customer and retailer information. The warehouse would act as a library for data analysis to choose what variables they need. The analyst would not need all the data stored just specific data to correlate to the study at hand. That would consist of also what was said above with the searches and habits. That would not be the only data collecting. From internal and external data such as stock quantity and price tracking. The retailer would have to keep up with competition by checking to make sure they can deliver the best prices for customers to return to their online retailer. The data warehouse would provide information to be able to make business decisions in the most effective manner.

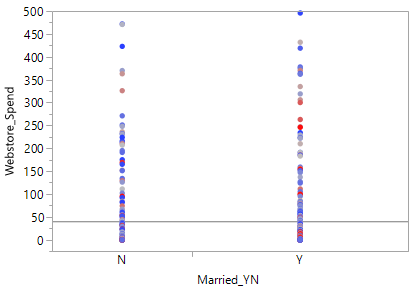
If the data collected within the retailer need more information the retailer would push for external data beyond their internal sources. This can be comparative to gauging the retailer to a larger scale. Perhaps a school is taking a poll to see what type of phone is most widely used at that moment in time. Information such as that will let a retailer know if they need to stock their shelves and focus an investment to a product. That was an example, I am sure there would be multiple schools, reviews, government agencies and social media statistics to justify the purchase and engage the product. With multiple sources to tap into that can prove difficult to generate data for a single organizational answer. Making a large business decision off external sources needs a good amount of proof. So, research on the secondary sources would be extensive compared to the reliable internal sources at hand. The more internal retailer information that can be collected the better for verifying the external data possibly needed for proof on a retailer’s business decision.

Next, we will interpret the Bubba Gump company that is a restaurant chain. 

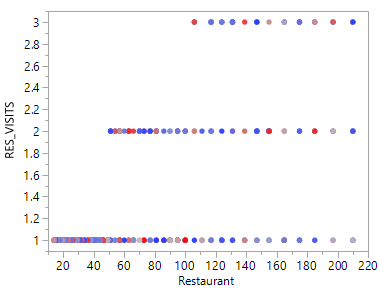
**Figure 4**

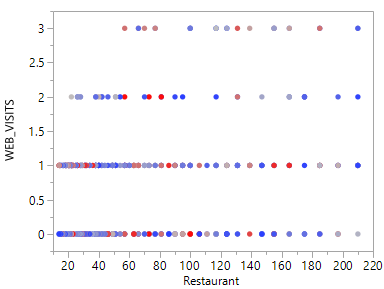




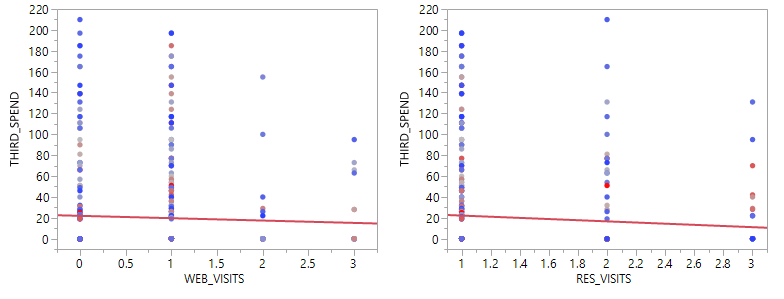


**Figure 5**

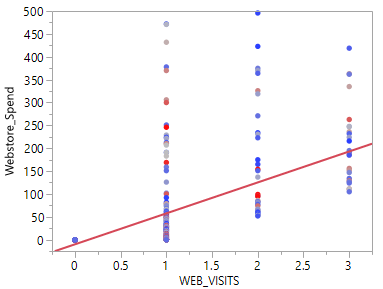




**Figure 6**



**Figure 7**

****

**Figure 8**

With the above information we can gain much oversight on how our webstore is performing. Let us begin with figure four showing the activity of the webstore. This is based out of the five hundred customers that participated in the survey. The number of visits is on the y axis and the money spent on the webstore is our x axis. All the blue in this figure is showing no profit being made by the webstore. It does show that up to three visits were made by the customers in the blue area. The green and red on the diagram is showing profit made by either one, two or three visits. This is also displaying that most of our customers are not using the webstore. We can focus on making the webstore more user friendly with a targeted survey on just that. The customers are fifty-fifty on whether they spend money on our webstore. We will go more in detail on this oversight of the webstore activity.

Figure five has multiple layouts to go over. This is showing characteristics with our customers and how much is spent on the webstore. The characteristics that were chosen for this comparison was income, age and if they were married. The income variable generated an outcome of decline with the higher incomes. Vast majority of webstore activity was from the forty to seventy thousand yearly income. This is showing the company that the range of income from the customers slightly reflects on if they are going to use the webstore. Not solid information but information none the less. Next, we can see the activity of the webstore by age of the customer. The population at forty and below are more inclined to use the webstore than that of ages fifty and above. There is a slight gap between forty and fifty but nothing major. The older generation seems to not use or spend as much when it comes to the webstore. The last variable used was the status of the customer. With the married status they were spending more on the webstore. This could be due to having a family and bringing the food or merchandise home. This shows that family (more than one) customers tend to gauge more on the webstore and can be beneficial to know for making family deals on the webstore.

With the next figure (Fig. 6) we can display the number of visits matched against the webstore to the restaurant. This is based on the restaurant identification to match the visits to them. With the results that we need to address. Some restaurants do not receive a third visit which can be disturbing to the company as to why. The webstore also shows a slight decline but not as major as the restaurant visits. The basis of this is regarding the webstore but wanted to show a discovery with the data to be later addressed. The focus on the webstore does show that the restaurants are more inclined to get webstore purchases rather than the restaurant visits. Steadily tapering off with the more visits which does need addressed. The first step with our webstore impacting better we should amplify or revamp our webstore to accommodate the customer easier.

The next two figures (Fig.7 and Fig. 8) I want to tie together. This is good information for some good news. Figure seven is showing the third spending with webstore and restaurant visits. Also wanted to show the restaurant visits due to the webstore generating a zero visit so the data can coincide. More third visit spending is higher at the restaurant but as before seen the visits are declining with the customers. Here it is showing the webstore visits need more of an impact on the customer to want to use the website again. Then when we see figure eight is when we see the good news. Figure eight is displaying that with the webstore visits increase so does the spending on the webstore. This may seem conflicting because I compared this to the third visit spending that did not show this. Overall, the webstore is receiving a higher spending rate with a broader outlook other than just one visit. The data set that we have can cover information that might give false outcome. This is good news for the webstore just as I said before we should probably focus on a targeted survey for the webstore. So, we can focus on customer wants and needs to better our customers experience with the website.

The company should focus on the survey results as a baseline to begin a trend analysis. As stated, "clustered" results would be informational to the business to find the current trend of the time. Be able to generate a farther outlook to what we need to do in the future. Delivering a survey for credits is a great start to understand the customer habits. The data warehouse that the company has now gives us some background to match to the survey given. The ability to grow in the future and adjust to what the next priority that is wanted. The data warehouse collecting the point of sale, web sales, customer information and sales transactions are enough to get started with the data analysis. We can focus on what the customer uses more being at the store or online services. We should begin with a descriptive analysis approach. Focus on the data at hand and see what the benefits can come of it. To focus on the customers and the point of sales generates use of that data alone. Future steps can be made to begin another focal point but let us start small then gradually increase the use of other data. With this method we can create a solid baseline for our data collected. We will conquer the steps to move in the direction Bubba Gump needs.

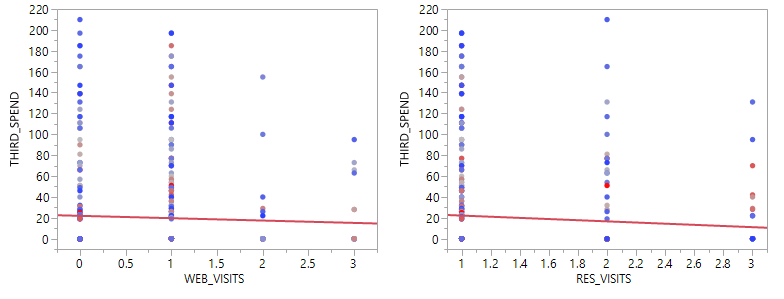
A question to be addressed is needing more data and less data. Let us focus on more of the purchasing methods and each purchase (web or restaurant) and gathering meaningful data that can better our results. With the give and take let us take away the marriage status columns (Married\_YN, MARR\_BIN) and add other column(s) for the purchases. Add RESTAURANT\_SPEND and HOUSEHOLD\_AMOUNT so the comparison can be made on which one is making more. For future analysis this will grant solid information when comparing the two, restaurant or webstore. Household amount can show trends with larger purchases and make more sense with them. With the data we have now I want to show an analysis to move forward with my decision and have management make the final decision.

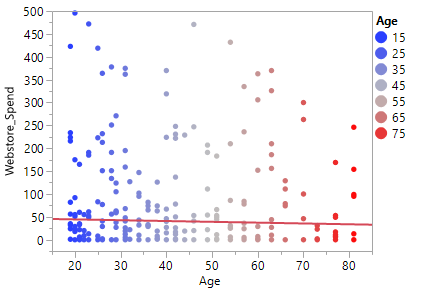
Measuring progress on the matter is not to lose the data we are using. To have this data on hand for future analysis will prove to be valuable. The data gathered from the web store is a focus now. Able to compare the data we gather now and project it towards our next step. Compare the web store to the restaurant purchases. Show some trends within our customers to make our data for the webstore more intriguing. This will show which direction we can take on making more profit for Bubba Gump Company. On top of making more profit making it easier for the customer to connect to Bubba Gump company. The webstore can grant that aspect for the customer to connect virtually and have a great experience while doing so.

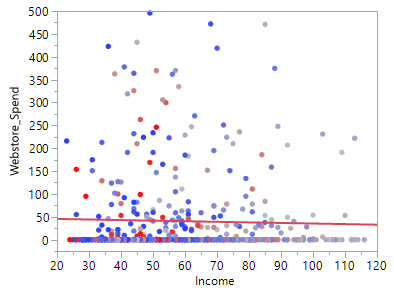
Following up on the comparison of web versus restaurant will benefit and amplify on the original data that is collected now. I am sure the best interest would be to make more profit and customer satisfaction. Let us find what the customer uses more and give them a more enhanced way to make their purchases. We can identify trends in the customer to show what age, income or status to target for sales. Will we be able to generate a new survey with different information that is needed? This will be able to use our data now and match it to a future dataset.

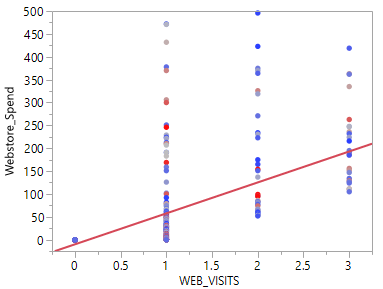
The resource that will come in handy is the data preparation for the next step for the company to visually decide to make forward progress. Keep a published document for monitoring the steps we are making to better the Bubba Gump Company. This can be traceable for your company to keep up to date information. Another resource we can incorporate is questions document brought up by management from the Bubba Gump Company to assess for discussion with the data analysis department. We can work together on this focus for the company’s behalf. We do not want to lead astray the company from their wants and needs. Let us create the background of the company’s interest and make the best effort to accommodate to Bubba Gump’s focus. We will also use some external data to compare to our information received from our customers. Additionally, incorporating the external data to support our data for analysis.

Bubba Gump Company’s stepwise approach is focusing on the value of the webstore. We have broken down some data from the data center to come up with some solid background to the webstore. We also have data that is out of our control with what we have, and we will go over both aspects. The effective material that we have collected reflects on the five hundred customers use of the webstore. This can display how many customers are using the webstore to purchase and shows how many customers use the webstore and do not purchase. This information can generate a new survey to expose the question to why they log onto the webstore but do not purchase. We can identify the population of the five hundred customers per income, age and marital status. The data collected on the population can reflect focusing on a generation that uses Bubba Gump’s products more. One element that we do not have control over is the restaurant visits and how much went to the restaurant versus the webstore. Our data is more focused on the webstore and should have more information pertaining to the webstore and tie the data with the webstore more. The difference of restaurant to webstore is out of our control when it comes to analyzing the data. We cannot tie a webstore visit to a restaurant visit expenditure with the third visit column. As shown above in data visualizations we can use the third spending for the restaurant spending in this case until more information is gathered.







****

There is a source of error I had to address when using the data that is present. That was focusing on the webstore and most data pulled reflected a downgrade to the information. The data was showing more customers using the webstore but not purchasing. It also was showing that the older generations were not using the webstore as much. The higher income customers were not using the webstore and was showing a decline in that data. All data sets generated was showing that solid decline in our information. Then with and overview of webstore visits to webstore spending the incline appeared. If the in depth digging of the information did not happen then one might see a failure in the webstore when there is light at the end of that tunnel. It is worth adjusting the budget to amplify on the webstore’s appearance and user-friendly experience.

There are meaningful patterns with the data pertaining to our customers. As stated above our customers gave information about their income, marital status, age, and location. The webstore activity and spending are the information that can be tied directly to what we need to focus on. This does raise question to be able to get the best out of the data by asking each spend and if restaurant or webstore. That information can pinpoint the data set to have heavier focus on a known outcome. Third visits need some more clarification to its meaning I believe. Modify the column to reflect more information to what we need. Manage a database with the proper information will always be a very powerful tool to gain outlooks on many situations.

Abiteboul, S., Hull, R., & Vianu, R. (1995). *Foundations of Databases*. Addison-Wesley.

Mood, A. M. (1950). Introduction to the theory of statistics. McGraw-Hill.

Blalock, H. M., Jr. (1960). Social statistics. McGraw-Hill.

Gerd Gigerenzer, Mindless statistics, The Journal of Socio-Economics, Volume 33, Issue 5, 2004, Pages 587-606, ISSN 1053-5357, https://doi.org/10.1016/j.socec.2004.09.033.

(<https://www.sciencedirect.com/science/article/pii/S1053535704000927>)