Dylan W. Sanderson

12/20/2019

DAT-220-J2600 Fundamentals of Data Mining 19EW2

Final Project

1. Plan For Analysis
2. Background and Introduction

In this paper, we will discuss the imaginary company, Bubba Gump Shrimp. This company faced a large shift in their business practice due to unforeseen exposure leading to unprecedented growth, but now their business is starting to hurt. They are turning to the modern method of data mining, and have already build a data warehouse, in order to fix their issue.

The overall business problem we are trying to solve is negative business growth. The past two years of business has seen a decline, and we are going to provide relevant data to the company that will allow them to reverse this trend. We will be pursuing this method through a data mart.

The analytic method we have chosen is one with historically positive results. Rather than attempting an analysis of the entire data warehouse as a whole, we will be using a data mart build from new external data paired with current internal data. The external data we will retrieve is based on 500 customer surveys offered to current customers in exchange for some store credit. The information in those surveys is paired with data already in the warehouse, namely their historical purchase information.

Using this data partition style holds many advantages. First and foremost, the actual calculative analysis and data processing will be significantly faster and less resource intensive. Second, using this model will allow us to test and validate the model, potentially through repetition, if necessary. Then we can be certain that the inferences we draw from the data set are globally valid across the population of their customer base. Lastly, using a surveyed approach will allow us to focus on the data points we have a specific interest in, catering our data mart to provide us with relevant, useful results. The data points we will receive at the end of this analysis will allow us to search for clusters within the customer base that indicate specific patterns or ‘subgroups’ that can be accessed with greater efficiency for increased profit.

1. Tools and Visualizations

As we continued into the data mining project for Bubba Gump Shrimp, we began to approach some vital decisions. The first decision we faced is what tools and techniques we will be using for the study. First and foremost, we will be using the software JMP Pro 12, as it is a powerful tool allowing us to breakdown massive data sets into a concise, effective analysis, through both statistically valuable information (like quartiles, RSquare, and other valuable quantities), and through visualizations. We will potentially look at some basic clustering methods, to see if anything stands out that isolates individual groups. However, the majority of our analysis will surround logistic and linear regressions, as they are very strong predictors of future trends. This is particularly valuable, as the information gained from the regressions will provide some of the strongest, most actionable data for the company.

To visualize this effectively, we will rely majorly on xy graphs with trend lines to demonstrate exactly what relationships are strong, and which ones are non-existant and, therefore, inconsequential. Despite the value of said graphs, we will begin with a dendogram representing a heirarchal cluster to explain why we performed the analysis as we did. As we are trying to make simple, straightforward visualizations, we will try to avoid 3D graphs and bubble plots, as they tend to be more visually busy without necessarily construing more data. The goal here is to show clear results, not massive amounts of data.

1. Specific Research Question

While we have already discussed the general reason for performing this analysis, and discussed some of the methods we will be using, we have yet to delve into specifics. The specific research question to be addressed is as follows: Are there any natural clusters or predictive trends within the customer base that can be exploited to encourage a positive profit margin? We will begin by looking at generalized spending trends based on the data we have, and subsequently perform multiple regressive analysis to see if there is any sort of predictive trend we can provide to Bubba Gump Shrimp.

Success and progress can be defined with this specific research question. Success will be measured as having successfully identified some actionable information to provide to the customer. Progress will have multiple definitions, however. We can consider it progress if we have identified a cluster that we can narrow in on, if we have identified a definitive trend, or if we have verified that some data point does not affect the profit margin at all. In this case, defining what isn’t useful is just as important as defining what is, and showing the company what socioeconomic demographics it cannot rely upon to predict a person’s behavior is just as important as finding what can be relied upon.

At this point there are a number of potential follow-up questions, though some are dependant on the mining operation. However, we can put forth some questions that should be answered no matter the specific results, namely, if this analysis is performed on this small sampling, will the same trends exist within other samples? The only way to answer this is to obtain the same data points about a different sampling of customers and perform it again. It is always a calculated risk using a sampling, but a proper cross-section can yield useful, universal results. We could also continue analysis of the sub groups identified, and search for additional trending within each group. Additionally, if we see a trend within a particular subgroup, we could use classification to take additional sample groups that fit that particular criteria (be it age, marital status, location, et cetera) and see if the same trends continue. Lastly, if our initial suggestions are taken, after some time, we should take another sample and verify the success of the operation and that the same trends are continuing.

There are two sources that provided the framework for the decisions made within this study. The first one that particularly aided in my understanding of sampling is the class text, A Practical Guide to Data Mining in Business and Industry, by Andrea Ahlemeyer-Stubbe and Shirley Coleman. The other source I leaned on for a bit more understanding of the different analysis types is guru99.com, with a number of brief overviews about data mining in general, providing a decent jumping point to start from. Both of these sources, while they don’t directly address specific situations like we are facing with Bubba Gump Shrimp, provide the framework of knowledge that allow these analysis to occur. The will guide the analysis by explicitly explaining the ‘what’ and ‘why’ of each type of analysis, so that we can decide where the next logical step takes us.

1. Analysis
2. Organization

The next question we seek to answer is if we followed an organized, stepwise approach to the analysis. The way we addressed this research question did follow an organized, sequential approach. As will be shown later, we began with basic distributions to gain a better understanding of the data, followed by using a dendogram to understand why our previous distributions look how they do. Lastly, we performed a series of regression based around the results of the dendogram.

1. Errors

Within our analysis, we did come across what had appeared to be an error in the data. As will be demonstrated and discussed later, our initial distribution made it appear as if the data set from Bubba Gump Shrimp had a huge number of outliers, indicating some error with the data. Initial review of the data points in question made it look as if faulty data had been collected, with many spots left blank, or at default values.

We addressed this with a rather simple concept. First, upon performing closer individual analysis, it seemed the data was actually sound, just unexpected. As such, we switched to a dendogram, and checked for heirarchal clustering to see if there were natural clusters formed along a split with the data points that were giving us a fuss, and there was. That particular heirarchal cluster was what led us to our two distinct sets of regressions later, and our eventual actionable information. As such, if we had not been adjusting our approach due to this unexpected parsing of data, we would not have walked down the distinct analytical path that brought us to the results we found.

1. Patterns

In the next section, we will be performing a more in-depth analysis of the results we will discuss here, with the actual visualizations we produced. We will spend a moment briefly discussing the patterns we discovered in our analysis and potential future directions this information takes us here.

After working through multiple layers of analysis, we found a few distinct patterns within the regressions. The majority of socioeconomic indicators we had data on, including age, marital status, and income all had no discernible effect on the spending habits of the customers. However, we did find that, when it came to online purchases, that the more times they visited the web store, the less likely they were to actually make a purchase. Additionally, we also found that newer branches tended to have more repeat customers than older branches.

Again, we will perform a more in-depth analysis on these ideas in the results section of this report, but we will start the discussion here. When it comes to both of these trends, the first question is, simply, why? Why does the website seem to decrease chances of a purchase the more a customer is exposed to it, and why do newer branches see more success than older ones? Answering both these questions will lead to specific, actionable information for the company. The second question that naturally follows is if these results are universal. If we perform this analysis on a second sampling, will the results be the same? The only way to answer that is to get a separate, distinct data set from the warehouse and a second survey, and repeat the analysis.

1. Inaccurate Depictions

While performing this analysis, we did encounter inaccurate data depictions that led us to believe we had faulty data early on. As discussed before, when we initially performed some basic distributions, we were met with a wildly high number of outliers, and quartiles that seemed nonsensical.This had actually led us to assume we had faulty data, and explore the potential source of the error. This was not the case, and once we understood that, we knew we had to readdress the data in a different way to find some valid patterns.

The primary resolution to our faulty depictions was to use a different analytical method, as the depictions actually led us to understand some of the main clusters that existed within the sampled customer base. As such, we will include said data within the report, as the logical flow is important. Since we knew that the distributions were being skewed by some of the customers not having ever made an online purchase, we instead performed a heirarcahl cluster, and showed the results in a dendogram, which ended up demonstrated exactly what we suspected: that our customer base had two distinct clusters based on whether or not they shopped online.

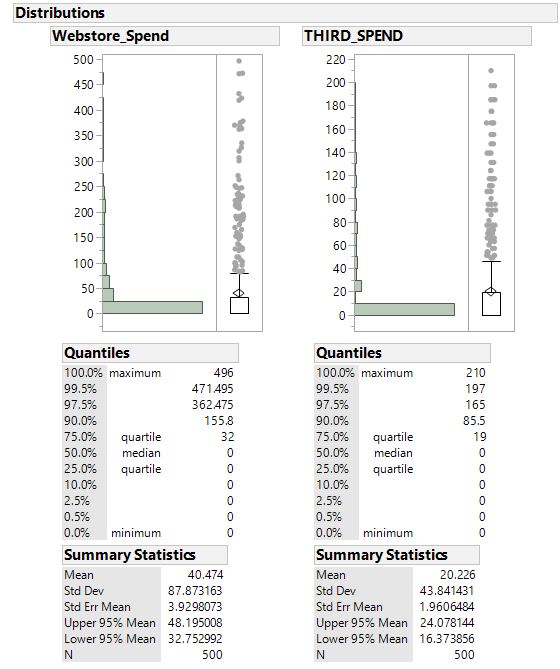
1. Alternative Analytical Method

Since we saw that distributions were ineffective with our particular clusters of customers within the data set, and our dendogram showed that we had a cluster of customers that shopped online and a cluster of customers that did not, we needed to pursue an alternative analytical method that was more targeted than basic distributions and quartile analysis. As such, we performed a series of targeted regressions, both linear and logistic, to find correlation, and therefore predictive trends, within the data sets. We performed two distinct sets of regressions, one targeting web store purchases, and one targeting repeated branch visits. With these regressions, we were able to identify multiple predictive trends of value to Bubba Gump Shrimp.

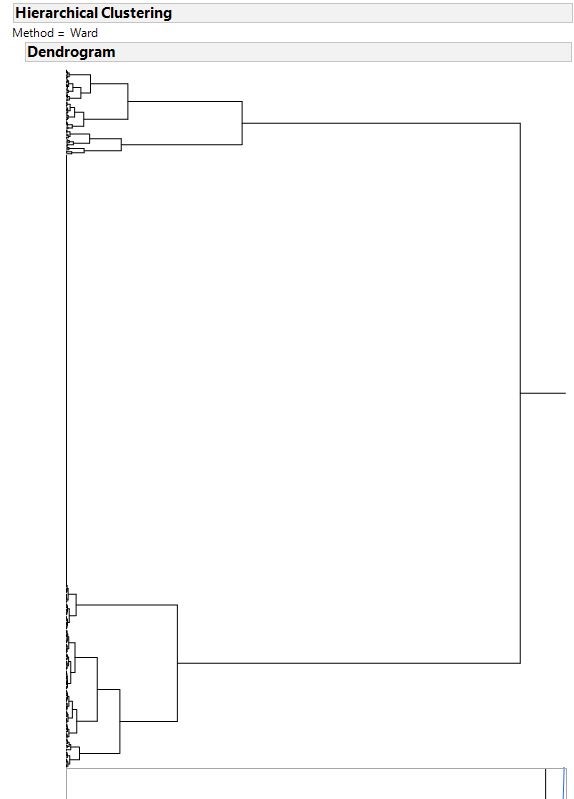
1. Final Report
2. Display and Interpretation

A series of trends were uncovered through multiple different data analysis techniques, but primarily through linear and logistic regressions. The majority of the regressions showed little to no correlation between factors, but some trends of import were noted. First, we will display and interpret the results of consequence. We will discuss various facets of the data, including it’s validity, and we will close this report by outlining what steps should be followed next.

The first analytical technique we used was basic distributions. The distributions themselves seemed to almost be faulty, as if we were operating off of incorrect data, but as discussed previously, this is because of how the subgroups within our data set are arranged. Because these distinct subgroups exists, we were able to move down a path towards more accurate results.

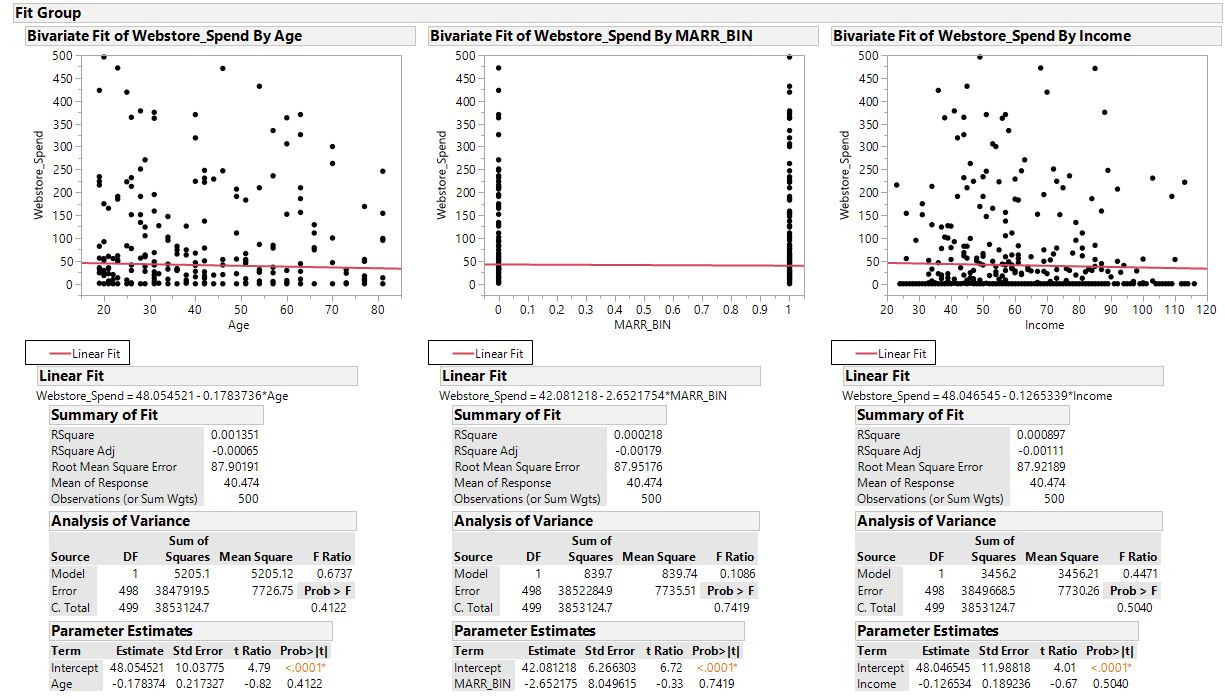


As can be plainly seen, there are an unusually high number of outliers, with the amount spent at the web store holding a median of 0. This occurred because we have in our data set a number of people who did not visit the web store at all, heavily skewing the results. Due to these results appearing as such, we were led to perform a heirarchal cluster, and the following dendogram confirmed our suspicions about the data set we were working with.



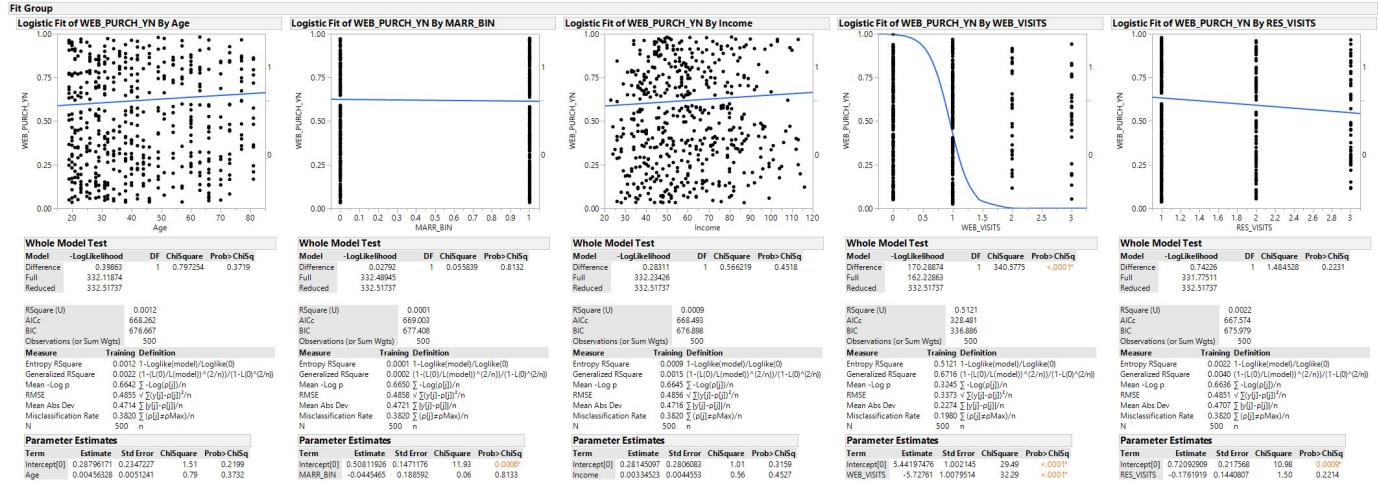
As can be quite plainly seen, there are two specific, distinct clusters within the customer base. The separating factor here is if the customer shopped at the web store or not. As such, this naturally leads to the two groups of regressions we performed. We had to consider the data set not as one source of profit, but two. The first profit source is web store purchases, which we performed the first sets of regressions on. The second is people making purchases at a restaurant, which we also performed a set of regressions on. These two distinct sets of regressions will be shown next, as they are the direct result of the analysis of the preceding dendograph.

When we began performing regressions is when we began to see usable trends. The first set of linear regressions is pictured below, when we attempted to correlate quantity spent in the web store with age, marital status, and income, to identify a socioeconomic factor that might aid the study.



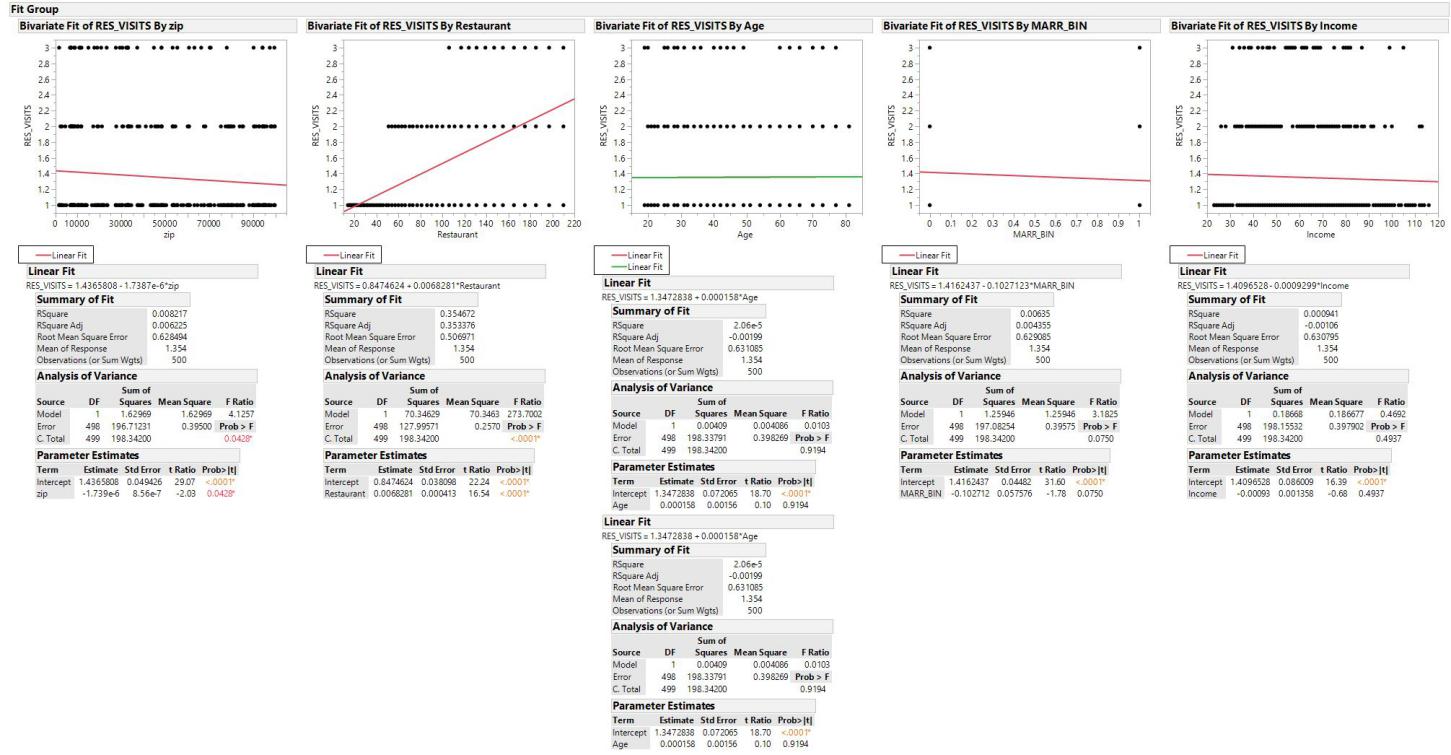
In all cases, we can see a nearly flat fit line, and an RSquare value of a fraction of a percent. We can interpret this as no correlation, which is actually valuable information to have. We know that these basic socioeconomic factors have had no effect on the amount spent in the web store, and can eliminate them as possibilities.

Next, we performed a set of five logistic regressions, to try to predict what factors increased or decreased the likelihood a customer would make a purchase of any size. Therefore, we compared the yes/no answer to having made a purchase to five factors: age, marital status, income, number of web visits, and number of restaurant visits. Ultimately, we don’t expect to see any trend with the first three comparisons, based on our last set of tests, so we hoped that the final two comparisons would yield some trend information.



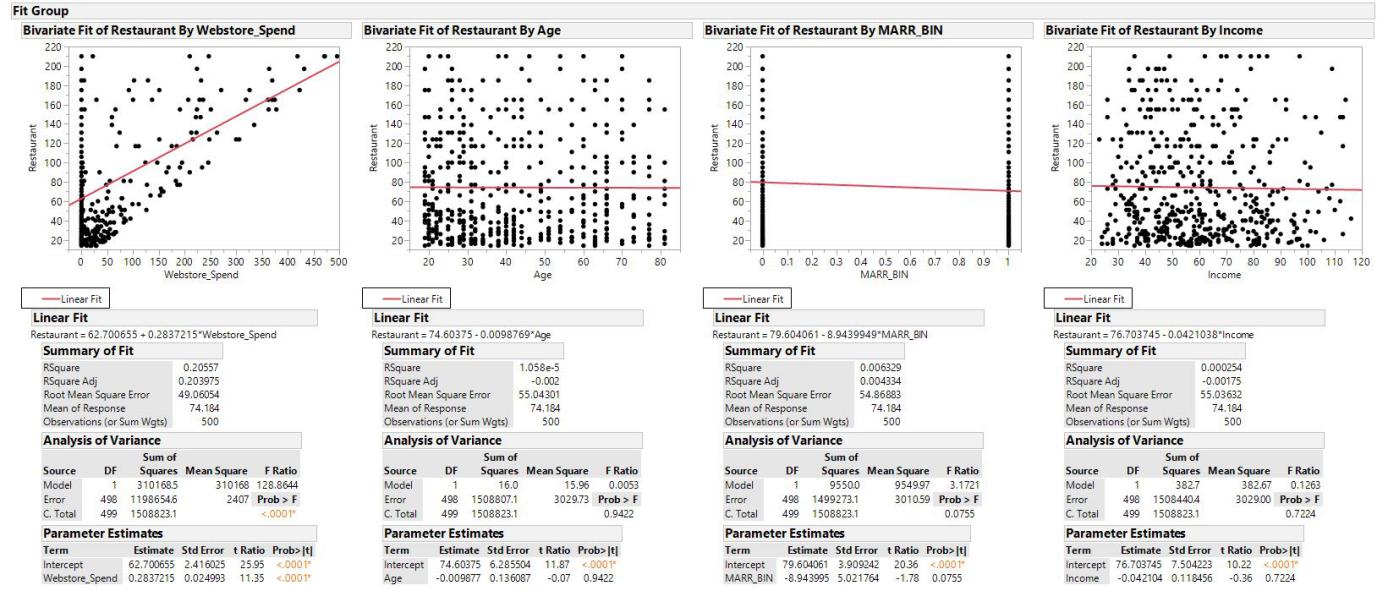
As expected, the first three analysis showed no trend, and the last one also showed no trend, but the fourth comparison showed a strong correlation. The web visits comparison had a p-value for the Chi-square of .0001, showing a very strong predictive trend. Surprisingly, this trend was negative, meaning the more often someone visits the site, the less likely they are to make a purchase.

Next, we noted that we have addressed one potential category, web store profits, but we have not addressed in store profits. Unfortunately, data was not included in the study for actual amount spent, but we do have a count for the number of restaurant visits. Therefore, we compared number of restaurant visits to zip code, restaurant number, age, marital status, and income. The results are pictured below.



From this data, we see only one correlation, and that is with restaurant number. While the RSquare value is only about 35 percent, it is the only value above a fraction of a percent. This is not a terribly strong trend, but it does imply that the higher number restaurants tend to see more repeat customers. As restaurants tend to be numbered sequentially based on when they are opened, we can assume that the newer restaurant are higher numbered, and tend to see more repeat customers.

Lastly, we did one final set of linear regressions, to test if any other factors correlated by restaurant, to see if certain restaurants attracted different demographics. We also included amount spend in the web store as our first value, to see if there was a unique trend present. We then also included age, marital status, and income.



Here we quickly noted a fascinating trend. Age, marital status, and income seemed to have no variation by restaurant, but amount spent in the web store had an RSquare of about 20.5%. Again, while it is not a powerful trend, it is of notable significance. The trend indicates a positive correlation with amount spent in the web store to the restaurant number, implying customers at newer restaurants are more likely to spend larger amounts online.

II) Validity, Reliability, Limitations

At this point, we have establishment the presence of a few interesting trends within our data set, but are these analysis valid and reliable, or are they too limited to be relevant? First, let us discuss validity, which we will need to break down into two separate ideas. We will address internal validity and external validity.

Internally, we can say the results are valid. We were able to show some trends across multiple different categories that were fairly strong predictors, and one comparison, namely, web visits to making a purchase, that was an incredibly powerful trend. We were also able to show many factors that did not correlate, which in many ways is just as valuable information, and just as valid.

Externally, the validity of the study is debatable. We are suffering from a few limitations that may damage our universality. First, our sample size is only 500 people, a minuscule sliver of the population of the United States. Second, we are only looking at data points for people who have the company’s loyalty program, which skews our results towards marketing to only that demographic. To increase the business effectively, we would need similar data on people who are not a part of the program or not customers at all, as those are untapped potential sources of profit.

We can say that the data and results are relativity reliable. Since a portion of the data comes not from surveys, but from point-of-sale information and the data warehouse, we can assume that data is majorly sound. The survey data itself relies exclusively on the honestly of the customer base, but they had no reason to lie, as they were rewarded for their time.

The limitations of the data were previously discussed in part along with external validity. Our sample size is small and a section of an exclusive group of people within the loyalty program. The other limitation that plaques this study is the data itself. If we had more complete data on amount spent in the restaurant, we could provide even more valuable analysis. At this point, we can provide analysis on the trending of a person to be a repeat customer, but information on actual profit margins is difficult to establish with our current data set.

III) Resulting Decision Influence

The direction we would urge our supervisors or the company, Bubba Gump Shrimp, to take this data is entirely dependant on what they wish to do with it. If they wish to see this data as stand-alone, there are a couple of notable trends that they could address to increase appeal to loyalty program customers. Since web visits negatively correlate with purchases, we can assume that the website needs to see some changes in order to keep customers coming back. Since the newer stores see more visits and customers spending more online, we can assume that the older stores need to be updated to be closer to the newer stores.

However, the true value of this data could be tapped into with additional analysis. First, when addressing the web store, data should be obtained as to what factors influenced a customer’s decision in making a purchase. Additionally, since quite a few customers did not visit the site at all, we could recommend they work with another company on Search Engine Optimization, or other such trends.

When it comes to the restaurants themselves, additional data on the differences between the newer and older stores would shed light on why customers are more likely to return to a new branch and spend more. Additional data from the company itself as to what design choices they enacted as time went on with newer restaurants would be useful in determining what branch features to emulate.

IV) Visual Evaluation

The visuals produced as the output for our specific data mining procedures are of debatable effectiveness. Part of this particular portion of the analysis is dependant on the customer and their particular familiarity with regressions. If they have no experience whatsoever in any sort of data mining or statistical analysis, we might prefer to redirect the visualizations into a simplified type, with less of the mathematical breakdown listed beneath each graph.

However, assuming they have basic statistical knowledge, or are open to a crash course, the visualizations are effective. Specifically, the addition of the fit line made all the difference, as it irons out exactly how the relationship exists. The RSquare or p-value for the Chi-square (depending on if it was linear or logistic) is a strong correlative indicator that aids in understanding how strong a trend we are examining, or disproving that a trend exists, in most cases.

V) Next Steps

Our initial hypothesis was to simply look for natural clusters or trends existing within this data set. We have the opportunity now to move into a more specific set of hypothesis based on this data, and to continue to pursue data mining operations to prove it. The hypothesis I would recommend would be to analyze the web store in depth to see why repeated visits are driving customers away, and to analyze the individual restaurants to see why the newer branches perform better than the older.

However, none of this will be successful without more data from wider-spread sources. We do not want to continue to perform analysis with limited external validity, as that will be of limited usefulness to the company. We should seek more data points from all customers, not just loyalty program members. We should also seek more information within each row, including information like categorical denotations of likes and dislikes of specific features, allowing us to perform regressive analysis on people’s complaints about the web store or branches.

Overall, we can state with certainty that, while our analysis did yield some useful information, it is not at it’s end. Rather, we have laid the framework for continued analysis that will yield beneficial information for Bubba Gump Shrimp. I am confident in the validity and reliability of our results, and would seek further information to address it’s limitations and be able to recommend long-term, profit-yielding solutions.

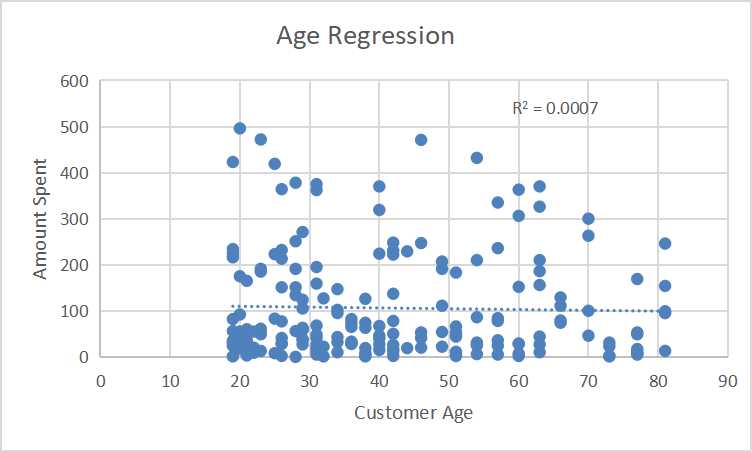
CAPSTONE ADDENDUM:

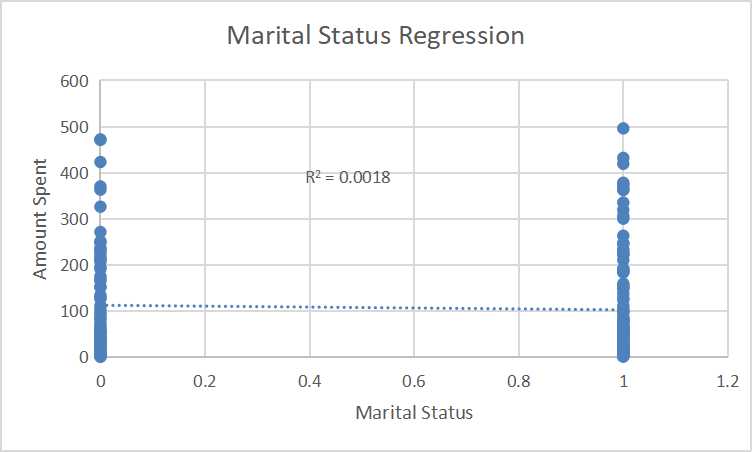
At this point, we are starting to scrape the surface of actionable data, and we have developed a suspicion that one of the biggest factors in the online business is actually how new the store is. We have also noted a trend in decreasing purchase sizes on repeated web visits. However, let us first begin by seeing if, overall, the newness of the store affects the number of web visits.

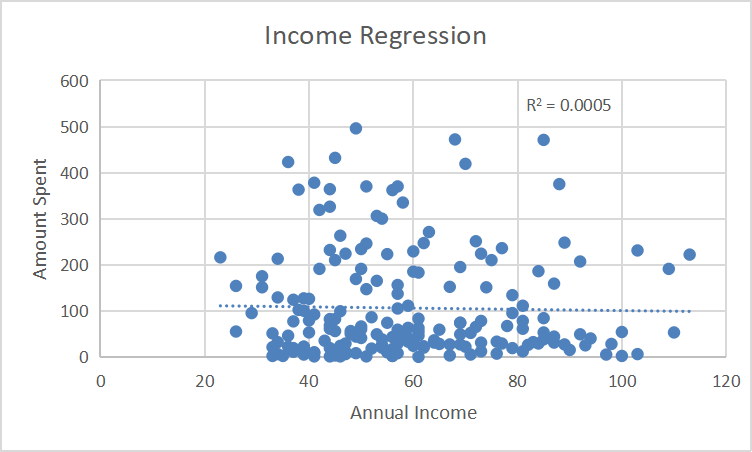


With such a low RSquare, we can likely assume that newness of the store does not affect number of web visits. But, there is another factor to consider. We have been looking at our data as a whole, and it may be time to examine certain trends more closely. As such, I subsequently sorted our data by whether or not the customer made a purchase, and pulled out those 191 rows. From there, we performed five linear regressions to attempt to uncover any trend that could point us to how to maximize sales to the willing customer base.

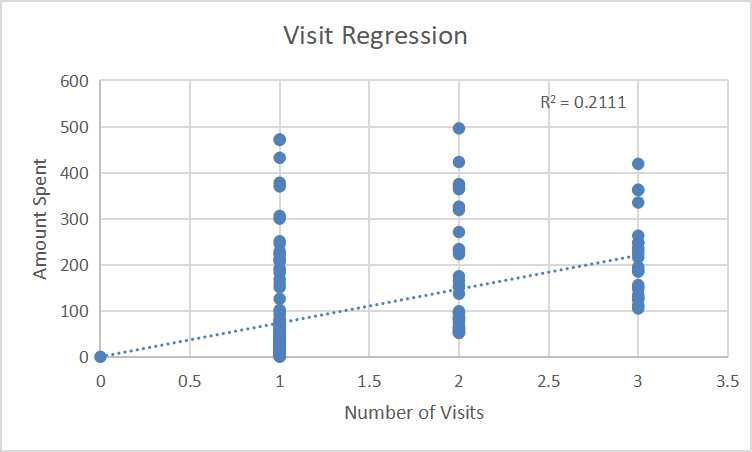
First, we will rapid-fire through the basic socioeconomic factors that we gathered information on, namely, age, marital status, and income. It is important to note that for the purposes of this analysis, the results of the marital status question were converted to numerical form, with a 1 representing yes, and a 0 representing no.



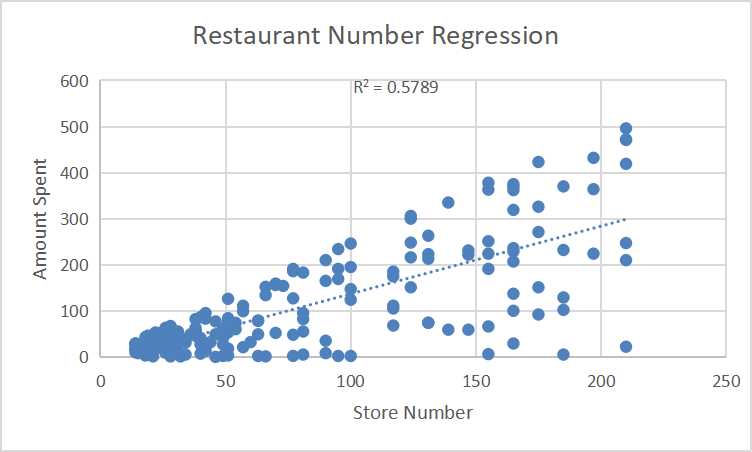




As can be plainly seen, all of these factos hold a fraction of a percent for the rsquare, definitively showing that socioeconomic factors are not altering the company profits from buying customers. However, the next two regressions began to show some more interesting trends. The first we compared was the number of visits to the purchase amount.



While 21% is not a terribly strong trend, it is the polar opposite of what we saw across the entire data set, where repeated visits decreased amount spent. This trend looks more like what we would expect from repeat customers, indicating that, while the website is driving away new customers, it is not necessarily barring current ones. As such, if a website redesign is implemented, we will need to be cautious that the company does not alienate it’s current customers while trying to increase it’s overall customer base. However, our last trend was the strongest, confirming a previous suspicion.



While we have already confirmed that the newness of the store (as new stores are made, they are numbered sequentially, meaning higher-numbered stores are inherently newer) does not alter number of web visits, we see an almost 58% correlation with store newness vs. amount spent. This leads us to once again confirm that customers in the newer stores are having a more positive experience, and are willing to continue their business online. Further on the ground analysis will be needed to confirm what the differences are in the newer stores, be it employee base or curb appeal from a newer building, but duplicating those factors is highly likely to increase the profit margin of the company.