**Samantha Pollard**

**7-2 Final Project Milestone Four**

**December 12, 2021**

**Southern New Hampshire University**

**Business Problem**

The business problem that we are trying to solve is how to increase the sales of Bubba

Gump products by improving the current marketing strategies being used. Specifically, to determine what’s the likelihood of products that are purchased, as well as what are the outlets used for purchase by customers. This strategy will allow the Bubba Gum to better customize the marketing strategies and techniques for each customer by

Targeting their needs with products they are more likely to buy or to use. The task for this project was to identify specific clusters of Bubba Gump customers through different demographic and available purchase information so that we could create a visualization to

describe theses said customer subgroups.

**Analytic Method**

The analytic method that is being used to identify subgroups within the

Bubba Gump customer pool will help to identify the natural clusters of customers that certain marketing strategies and sales may be more effective in increasing purchases.

Specifically, the customer surveys for this task. Customer surveys correspond with recent transactions that get analyzed to identify potential subgroups. The surveys yield a lot of useful information for the business purpose and to target marketing strategies to increase the sales through different venues. The surveys can also identify the browsing and purchasing habits, and the likelihood of all future purchases, and the likelihood of utilizing different outlets for purchases, depending on the survey. Other analytic methods were used during this analysis as well, with the most prominent was clustering. Clustering groups customers together based on similar characteristics, which is also helpful for this analysis. Methods also include plotting

variable frequencies to determine commonalities, scatterplots and to identify the trends, and regression for an analysis prediction.

**Analysis Tools**

An important mining tool that was used to perform the analysis includes the JMP.

This was chosen because it is a comprehensive and statistical analytics package and

Tool that we are currently using in this class. Given the scope of the project, the JMP is an

appropriate choice, especially for the data preparation and also performing the analyses of

customer clusters (Predictive Analytics Today Review Desk, n.d.). It would be handy for

joining of multiple purchase venues into a comprehensive purchasing analysis.

Given that the JMP allows for the visualizations and data modeling within the software.

.

**Data Visualizations**

Data visualizations used in the report will include all four (4) graphs. Geographic maps, bubble plots, bar graphs, and line graphs.

Since the identifying customer clusters and geographic maps are helpful when showing areas

where customers are located as well as which locations may that should be more

profitable for potential expansion. Bubble plots are also used in identifying customer clusters

based on gathered data. Bubble plots visualize the size of the clusters based on

stated attributes and help differentiate between purchases between different venues. Bar

graphs and line graphs benefit the analysis by showing the different attributes and

customer demographics that may be common.

**Research Question**

Specific questions that should be addressed are ‘what is the common customer

characteristics that a company should identify in order to increase profit?’ What will the

customer locations be and how to decide where additional venue offerings may be beneficial,

and all other possible solutions. Research questions will work to analyze the given

data for obvious patterns to ‘what do these customers have in common?’. What will also help to

target the marketing policies and procedures as well and to help identify how to increase

the potential customer base.

**Research Measurement**

Was the research question answered or if and how was the hypothesis-generation was

successful, the review of the visualizations that were created during the analysis of customer

clusters were they are an accurate representation of Bubba Gump customers? Progress shall be

measured by the number as well as the size of clusters, as well as the accuracy interval of the

analysis. If the research question is answered correctly, and marketing makes an appropriate and

effective response from the results of the analysis, Bubba Gump’s profits will improve.

**Follow-up Questions**

Follow-up questions that should follow from any initial research should include ‘what does

Bubba Gump have to offer to these customers in order to increase profits?’ and ‘Can Bubba

Gump increase its consumer base?’. Another question ‘what are the purchasing habits of the

customers that don’t visit the restaurant?’ Finally, another important question is ‘what

clusters are missing from the analysis?’

**Research and Support**

Several resources that could be utilized to support Bubba Gump and address

the abovementioned lines of query. One article, Avinashi et al. (2019), gives foresight in how to

use data analytics to determine a customer’s lifetime value for the company. Another article,

Rizzo (2015), details how to increase the efficiency of marketing strategies using date. A third

resource, Earley (2017), addresses the dangers of relying on automated analytics. A fourth, and

final, resource, Longbing (2008) focuses on the importance of customer behaviors to increase

profits.

.

**Analysis Organization**

This reflects an organized, stepwise approach. Once the data was gathered, it

underwent preparation through transformation and aggregation. This analysis, the subject is the

customer so the related variables needed to be identified accordingly. Once the data has been

prepared for analysis, different steps are taken for the analysis. First, hierarchical clustering was

performed with the data to try to identify customer groups with common demographic features.

Next, linear regression was performed to predict customer behavior with webstore

purchases and third-party purchases, specifically the amount of money spent.

A logistic regression was performed to predict likelihood of a customer making a webstore

or third-party purchase, but the model was unable to be created with the gathered data. Logistic

regression has similar limitations as a linear regression, with the additional limitation of being

unable to predict continuous variables. Additionally, while webstore and

third-party spending were tracked, restaurant spending was not. This affects the efficiency of the

analysis.

**Sources of Error**

A major source of error were several logical errors within the survey data itself. There

were inconsistencies within the data collected that required repair or removal. An example

was located within in the zip codes (both zip and ZIP\_2), in which the zip codes were missing

their leading zeros. Zeroes were added to the appropriate records to resolve this error. There

were other changes to make before having a successful run. It was more effective to change the

ages and income to ranges and create a Y/N variable for third-party purchases.

Another source of error was that all venue visits are maxed out at three. It was

unclear whether or not this is by design. It is impossible to resolve any errors in customer

responses because the data was previously gathered, and since the analyst cannot contact the

customers for verification or send out another survey with data mining. Inaccuracy could

severely mess up the data.

**Meaningful Patterns**

there were many meaningful patterns within the data. It is fairly clear upfront that the amount of money spent at each venue was higher for customers that spent more money at other venues. After some data manipulating, it was found that Bubba Gump tends to cater to lower-

income young people than any other demographic. Inaccurate Depictions of Data

There were a few inaccurate depictions that found within the data, all of which have been

mentioned in previous sections. The zip codes issue was noticed after a data run and needed to be

resolved. Inaccurate depictions within the regression analyses and the graphs . Because there

were so many zeros within the data for the webstore and third-party information, many of the

plots that were generated were harder to interpret.

**Alternative Analytic Methods**

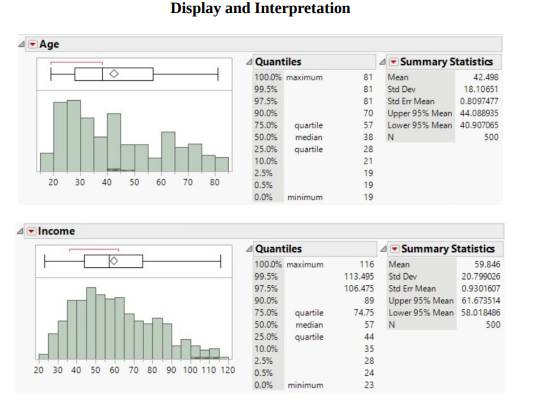
Alternative analytic methods were not used in this analysis, including kmeans clustering, decision trees, partial components analysis, multiple correspondence analysis,

and partial least squares regression. All analytic methods have their benefits and detractions, but

decision trees, in particular, would be useful in this analysis. A decision tree would be

helpful given the number of non-continuous variables used for customer descriptions and are

easier to read.



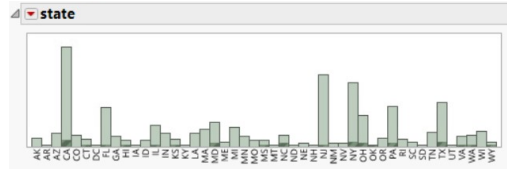
The two histograms plot the Bubba Gump customers’ age and income. These graphs

show that the customer ages range 19-81 years old, with the majority of customers aged 28-57

years old and the median age being 38. The income of customers range between $23,000 and

$116,000, with the majority of customers’ income being $44,000-$75,000 and the median

income is $57,000.



Next, the state of residence were plotted on a frequency chart. This shows

states where customers were more likely than others, and could help with location-targeted

marketing campaigns. The data that the marketing team could benefit by targeting customers that

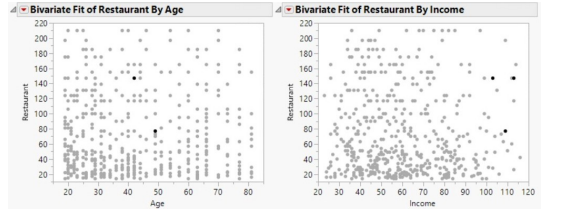
live in California, Florida, New Jersey, New York, Pennsylvania, and Texas. Analysis could also

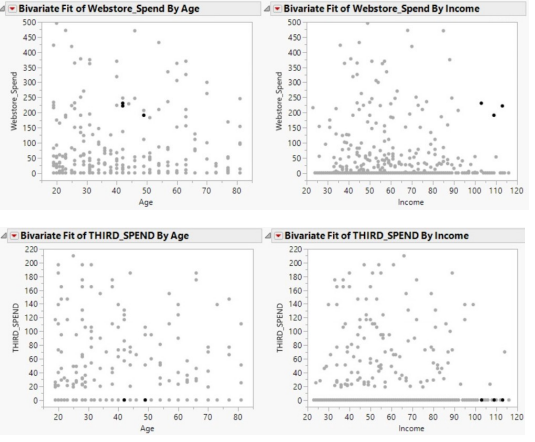
be run to show customer spending between each state (and further narrowed by city or zip) to

determine which restaurants make more money per customer, This age and income information

was plotted against the customer’s spending amounts in the restaurant, webstore, and third-party

venues, as shown below.



Reviewing the restaurant spending graphs, shows that most Bubba Gump customers are younger and lower income. This information is helpful to Bubba Gump in order to target their marketing audience. The income information is also helpful for Bubba Gump when considering their costs and production methods

From these graphs, show that third-party vendors are severely under-utilized by

restaurant customers, especially with customers over the age of 45.Looking at the webstore

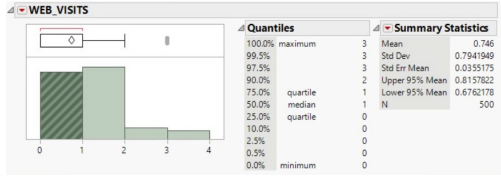
spending plots, show that more restaurant customers are also spending money at the webstores

than third-party venues, but plotting the frequency of webstore visits, as shown below, show that

almost half of the restaurant customers never even visit the webstore. This also shows an ample

opportunity to increase webstore visits by restaurant customers by making sure every restaurant

customer knows about the webstore





Next, was the correlations table that was created with all of the continuous variables, This table

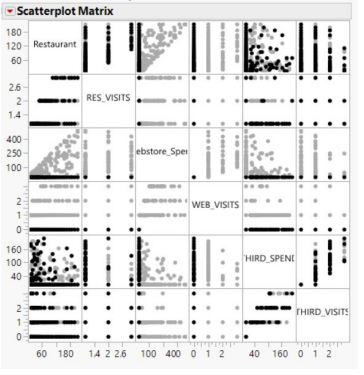
suggests that the strongest connection appears between restaurant visits with restaurant

spending, webstore visits with webstore spending, and third-party spending with thirdparty visits.

A scatterplot matrix was created with this information only the strongly correlated

Variables are shown. This shows that there is also a strong connection between

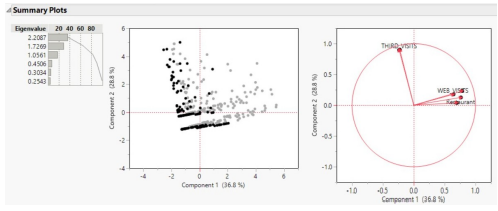
webstore spending and restaurant spending



Principal components analysis (PCA) with the same variables above, restaurant spending,

restaurant visits, webstore spending, webstore visits, third-party spending, and third-party visits,

as shown.



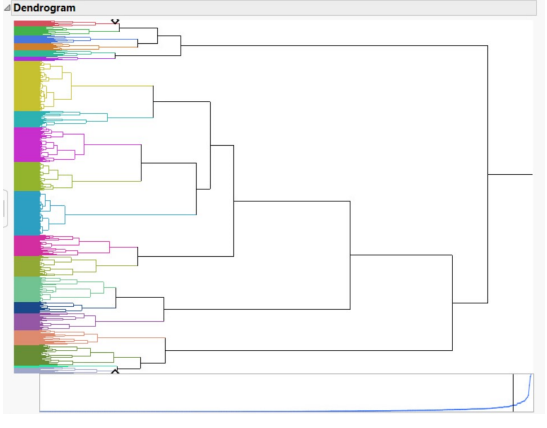
The PCA shows that the first two components account for 65.6% of the data variance.

It also shows that all of the variables correlate positively with component two, and that all of

them correlate positively with component one, except third-party visits, which correlates

negatively.

Cluster analysis was performed on the same group of variables.



The dendrogram, show that there are twenty natural clusters of customers using venue visits and

spending habits. The cluster means table, displayed below, gives a bit more information. It

shows that the most customers fall into cluster seven, followed by cluster eleven. 71 customers

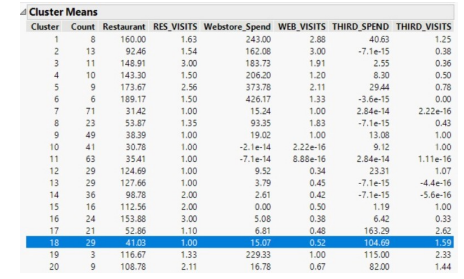
are grouped into cluster seven, and these customers spend an average of $31 in the restaurant, $15 in the webstore, and less than $1 at third-party vendors. 63 customers are grouped in cluster

eleven. These customers spend an average of $35 in the restaurant, but less than $1 at the

webstore and third-party vendors. The cluster means table also shows that the least number of

customers are grouped into cluster nineteen, with only three customers being a part of this

cluster. These customers spend an average of $117 in the restaurant, $229 in the webstore, and $115 at third-party vendors.



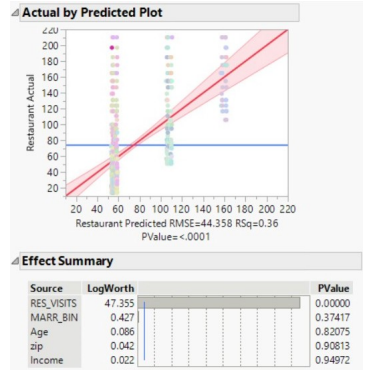
a linear regression model was created using restaurant

visits, zip code, age, and income to predict restaurant spending, but the R-squared value was too

low for it to be an effective model, as shown below. It also verifies our earlier conclusion that,

according to the p-values and t-values that age, zip, and income are not accurate predictors of

restaurant spending by customers in this case



Different models were created with the various continuous variables to try to predict

spending amounts through the various purchasing venues, but an accurate model could not be

found with the given time and data constraints.

**Validity, Reliability, Limitations**

This report is detailed but contains questionable results. The few conclusions drawn from the data are reliable and valid and will be beneficial to Bubba Gump in their effort to increase profits. There were several limitations in the time and data that prevented a more effective report. At first glance, it is obvious that the surveys were only given to customers that spent money at a restaurant, which heavily skews all gleaned results in that direction. Other data limitations have been discussed in previous reports, including the limit of venue visits being three. Time constraints prevented more detailed analysis in areas that could have used a further analysis, but these specifics are discussed in detail later in this report.

Inspecting results of the analysis and the shopping habits of these consumer groups suggests that Bubba Gump would benefit from the addition of a mobile app, increased social media presence, and loyalty programs (Wirecard FINAL PROJECT 16 Editorial Team, 2020). “Younger consumers are shopping conveniently from their mobiles” (Wirecard Editorial Team, 2020, para. 2). A mobile application would offer convenience for these customers and could streamline check-out at the restaurants. “Social media is also enabling consumers to share their latest purchases with friends, create a buzz, and potentially score the brand a new customer through their endorsement” (Wirecard Editorial Team, 2020, para.11). In addition creating a “social media-friendly” space within the restaurants can allow customers to work as advertisements. Lastly a loyalty program would give customers incentives for returning and for purchasing through the different venues. Marketing would be able to determine how to attract the identified customer groups.

**Visual Evaluation**

Contained within the report were the results in a understandable manner. The histograms were extremely helpful in identifying frequencies within single customer demographic variables, such as age, income, location, and marriage status. The scatterplots were influential in representing the relationships between actual sales per customer and other variables to help identify missed opportunities. The dendrogram and cluster means visualized specific customer groups and how prevalent they were within the customer sample.

**Next Steps**

There were several conclusions and possible solutions mentioned within this report based

on the available data. There were several limitations and issues mentioned throughout the report

that would be beneficial to addressed if the resources are available. There are specifically two

areas that would benefit from further analysis. The first area that should be further analyzed if

the resources are available, would be

narrowing the location scope of the customers. Identify specific restaurants and customers to

target with certain promotions and marketing campaigns by analyzing the customers by city and

zip code. Bubba Gump could also determine which, if any, restaurants should be closed or

downsized to decrease costs. Secondly additional analysis would be a closer look at the third-

party vendor effectiveness. The data is lacking from this specific venue, but a closer look at

profits from third-party vendors against vendor costs would help determine whether or not it is a

beneficial venue for Bubba Gump to sell through this venue.

**References**

Ahlemeye-Stubbe, A. & Coleman, S. (2014). A Practical Guide to Data Mining for Business and

Industry. John Wiley & Sons, Ltd.

Avinash, A., Sahu, P., & Pahari, A. (2019). Big data analytics for customer lifetime value

prediction. Telecom Business Review, 12(1), 46-49.

Earley, S. (2017). The problem of personalization: AI-driven analytics at scale. IT Professional,

19(6), 74-80.

Longbing, C. (2008). Behavior informatics and analytics: Let behavior talk. 2008 IEEE

International Conference on Data Mining Workshops, Data Mining Workshops, 2008.

ICDMW ’08. IEEE International Conference On, 87-96. https://doiorg.ezproxy.snhu.edu/10.1109/ICDMW.2008.95

Predictive Analysis Today Review Desk. (n.d.). JMP. Review: JMP in 2020. Retrieved from

https://www.predictiveanalyticstoday.com/jmp/

Rizzo, T. (2015). Maximizing results with data-driven financial marketing strategies. Journal of

Brand Strategy, 4(2), 114-126.

Wirecard Editorial Team. (2020, March 13). Shopping habits: Gen Z and Millennials. Wirecard.

https://www.wirecard.com/blog/shopping-with-gen-z-and-millennials