Bubba Gump Data Analysis

Katie Stapleton

Fundamentals of Data Mining

June 28, 2020

Table of Contents

[Analysis Plan 2](#_Toc509487733)

[Milestone One: Introduction 2](#_Toc509487734)

[Business Problem 2](#_Toc509487735)

[Analytic Method 2](#_Toc509487736)

[Milestone Two: Analysis Process, Tools and Visualizations 3](#_Toc509487737)

[Analysis Tools 3](#_Toc509487738)

[Data Visualizations 3](#_Toc509487739)

[Research Question 4](#_Toc509487740)

[Research Measurement 5](#_Toc509487741)

[Follow-Up Questions 5](#_Toc509487742)

[Research and Support 5](#_Toc509487743)

[Analysis 8](#_Toc509487744)

[Milestone Three: Preparing Data and Analysis 8](#_Toc509487745)

[Analysis Organization 8](#_Toc509487746)

[Sources of Error 9](#_Toc509487747)

[Meaningful Patterns 11](#_Toc509487748)

[Inaccurate Depictions of Data 14](#_Toc509487749)

[Alternative Analytic Methods 15](#_Toc509487750)

[Final Report and Presentation 18](#_Toc509487751)

[Milestone Four: Data Presentation and Report 18](#_Toc509487752)

[Display and Interpretation 18](#_Toc509487753)

[Validity, Reliability, Limitations 25](#_Toc509487754)

[Resulting Decision Influence 28](#_Toc509487755)

[Visual Evaluation 30](#_Toc509487756)

[Next Steps 33](#_Toc509487757)

[References 35](#_Toc509487758)

# **Analysis Plan**

### Milestone One: Introduction

## **Business Problem**

Bubba Gump desires to discover applicable methods to increase the company’s revenue growth and to gain a higher comprehension of its customers. Previously, Bubba Gump experienced a rapid growth in business due to unexpected media exposure. However, their sales plateaued over a few years and have now declined for the last two years. Bubba Gump has recently complied their various data channels into a data warehouse designed for data mining. The company only performed data analyses of channels independently rather than combined as a whole. For the data mining process, the data analysis method used is unsupervised learning.

## **Analytic Method**

The data warehouse provided by Bubba Gump is analyzed using the unsupervised learning method. Unsupervised learning identifies underlying patterns without stating a specific focus (Ahlemeyer-Stubbe & Coleman, 2014). Unsupervised learning is suggested for situations where the data trends and/or specific concentrations are impractical to propose (Soni, 2018). Bubba Gump is unaware of the reasons behind their revenue loss. They are aware of their financial loss, but have no definitive reason that caused the loss to happen. Bubba Gump has established a clear vision to gain higher customer awareness and find methods to increase revenue. Bubba Gump has not provide specific methods they wish to implement or a particular focuses regarding the customer analysis. By using the analytical method of unsupervised learning, the data analysis and models are a basis for developing methods to increase revenue. As the same time, unsupervised learning presents information about customer including unpredictable findings or trends (Soni, 2018). Bar graphs, histograms, scatter plots, and line charts are included to visualize the information and possible patterns.

### Milestone Two: Analysis Process, Tools and Visualizations

Analysis Tools

JMP Pro is the software tool that is used to analysis the data provided by Bubba Gump. JMP dynamically links the data, statistics, and graphics (JMP, n.d.). JMP supports the modeling techniques of different regressions, which are used in this analysis (JMP, n.d). JMP does not require a coding background for reasonable usage, but coding can still be used (TrustRadius, n.d.). Datasets in different formats can be upload into JMP without rewriting or designing the dataset. Some users said that it takes a large amount of time to find answers regarding issues and general usage (TrustRadius, n.d.). Other users have said that the regression analysis produces accurate results, but use different parameters than most software (TrustRadius, n.d.). The interface does have mixed reviews about its design (TEC, n.d.). Users have claimed that it can be hard to navigate and excluding high-level advanced statistics. In this analysis, the results are not compared across other software, so the differing regression parameters are not an issue. In this analysis, standard to medium level statistics/analysis are used, not advanced. As for issues, the data and visuals are not overly complex. Issues requiring hard research are unlikely to occur.

Data Visualizations

Overall, the analysis focuses on comparison and understanding data relationships. Common visualizations for comparison and relations are bar graphs, histograms, lines charts, and/or scatter plots (Boost Labs, 2019). The bar graphs are used to compare values within the same category (Klipfolio, 2019). Subsets of the same values can be included within the same chart (Klipfolio, 2019). However, with too many bars, it may be confusing and have minimal usefulness (Klipfolio, 2019). Histograms are used to compare two values that are each divided into intervals (Lund Research, 2018). Histograms can display frequencies, patterns, gaps, or unusual values (Klipfolio, n.d). Intervals must be evenly divided (Lund Research, 2018). With uneven intervals, the histogram misrepresents the data (Lund Research, 2018). Line graphs compare values to display patterns, relative changes or progressions (Ribecca, n.d.). Visually, data points are charted in intervals, with a continuous line connecting the point together (Ribecca, n.d.). Line graphs are difficult to analyze if too many lines are present (Harmath, 2014). Scatter plots attempt to find if a correlation exists between two variables (ImportIo, 2019). If the plot points occur more frequency in certain area(s) of the chart, then a trend or relationship may exist (ImportIo, 2019). Trend lines can be added to scatter plots provide a better visualization of the trend (ImportIo, 2019). Scatter plots best applied for continuous data only (Valcheva, 2017). Over-plotting can result in false interpretation or provide no results (Valcheva, 2017).

## Research Question

The research questions are designed to address issues and ascertain insightful information (Biddix, 2018). The purpose of the information is to help Bubba Bump achieve their goals of understanding their customers and increasing revenue. The follow research questions are addressed:

1. What is the relationship between age groups and visits amongst customers?
2. What is the relationship between income and spending of customers?
3. How much money do customers spend at Bubba Bump?

Research Measurement

A research questions should be turned into statement/hypothesis to measure its process and success. The statement must be logical, reiterate the research question, and answerable in yes/no manner (Ahlemeyer-Stubbe & Coleman, 2014; Enago Academy, 2017). The statement needs to be measured in its credibility before any conclusion is decided (Enago Academy, 2017). To be conclusive, evidence must clearly support all parts of the statements, not just sections of it (Enago Academy, 2017). Evidence is the measurability of how much the finding support the statement (Enago Academy, 2017). As the supportive and verifiable evidence increases, so does the statement’s credibility. However, if any evidence contradicts or shows minimal support of the objective statement, the statement should be considered inconclusive/null (Ahlemeyer-Stubbe & Coleman, 2014). In short, a successfully-answered research question results in a single objective statement with conclusive findings that directly and fully answer the research question (Enago Academy, 2017).

Follow-Up Questions

* Does the number of restaurant visits increase as the age of customers increases?
* Does the number of website visits decrease as the age of customers increases?
* Does the likelihood of website purchasing increase as income of customers increases?
* Does the number of customers decrease as the website purchases amount increases?
* Does the number of customers decrease as the restaurant purchases amount increases?

## Research and Support

First outside sources about consumer behaviors and demographics are researched. The understanding the consumer behavior is critical to long-term success in the (Kumar, 2018). Finding patterns and trends can be compared to find irregularities between the Bubba Gump and other restaurants/business (Thibodeaux, 2019). Demographics classifications can be misleading (Walker, 2017). The consumer behaviors and demographics may not apply to restaurants. The findings may be too generalized or too complex to assist the analysis (Walker, 2017). The classification does not necessarily reflect the customers habits (Walkter, 2017).

Another source is business-to-business (B2B) publications. As the title suggestions, B2B publications are focused on business affairs (Uzialko, 2019). The publications are more likely to be professional and more objective than other forms of media (Uzialko, 2019). However, B2B publications are also used for marketing and sales. Instead of objective information, the publications may contain information to promote concepts or products to generate sales for other companies (Uzialko, 2019). One general publication website that may help in the analysis of Bubba Gump is the [Business News Daily](https://www.businessnewsdaily.com/). It is more generalized, but still may provide information that can assist the analysis Website publication are available that focus specifically on restaurant materials/articles. Three website to consider for the analysis are [Nation’s Restaurant News](https://www.nrn.com/), [Restaurant Business](https://www.restaurantbusinessonline.com/), and [Modern Restaurant Management](https://modernrestaurantmanagement.com/) (Crane, 2018).

The final outside source is external censuses and surveying. Censuses are based on gathered information instead an individual’s personal opinion or research (Australian Bureau of Statistics, 2013). One reliable source for facts about economics, spending, and income is the census from the U.S. [Bureau of Economic Analysis](https://www.bea.gov/). A generalized census source is the [U.S. Census Bureau](https://data.census.gov/cedsci/), which provides census regarding business expenses, sales, income, population and more. The data and findings can analyzed with a higher objectively than published articles or books (Australian Bureau of Statistics, 2013). Bubba Gump’s information may be more comparable to external resources. Surveys and census may not provide appropriate scales in comparison to the data provided by Bubba Bump. Censuses and surveys also may have outdated information, contain gaps, or not represent the total values (Australian Bureau of Statistics, 2013).

Analysis

### Milestone Three: Preparing Data and Analysis

## Analysis Organization

The data was prepared for the analysis by using stepwise regression. Stepwise regression is a method for constructing the data model from the dataset through the selection and/or elimination of independent variables (Investopedia, 2019). From the three available methods in stepwise regression, the method of backwards elimination is applied to the Bubba Gump dataset (Investopedia, 2019). Backwards elimination starts with all possible variables (Smith, 2018). Each step involves removing the least significant variable from the dataset (Smith, 2018). Variables are eliminated using formulas, computer software, and logical human rational (Ahlemeyer-Stubbe & Coleman, 2014). The process stops when the remaining variables are statistically significant to the analysis (Smith, 2018).

Using the backward elimination does come with its concerns. Variables are eliminated using formulas, computer software, and logical human rational Outliers in the variables may skew the computations and software visual. The human eye may misjudge the results or patterns, leading to the wrong decisions (Cocchi, 2019). A variable suitable for the research may be removed due to its current stage of relations to other variable at the given step (Cocchi, 2019). Whether the variables from the dataset were originally useful or not, hidden patterns may be missed by removing certain variables (Cocchi, 2019).

With the Bubba Gump data set, all variables were included at the start of the backwards elimination. At each step, the entire dataset was analyzed to select the least significant variable. Variables were eliminated based on the values/data within the variable, relationship to other variables, significant to the analysis and research questions, practical application, and plausibility of resolving errors. Ten steps occurred during the process, meaning that ten variables were eliminated from the equation. By using backwards elimination, the remaining variables within the dataset should relate to the research questions, possess practical application, and contain the representations. Possible issues include missing possible hidden data pattern, misinterpretation of the variables, and misjudgment when selecting the variable to eliminate.

## Sources of Error

Certain variables turned out as dependent variables instead of independent variables. The city variable and county variable are dependent variables. More than one location can have the same city name and county name. However, the location cannot be properly determined without the independent variable of state initials variable. The analysis is a national wide study, not regarding a specific location or region. Naturally, customer demographics stored by a company include the entire address of customers, such as for delivery or sending promotions. The entire address is important to the company, but not necessarily of the same importance in the data mining analysis of Bubba Gump. Instead of revision, the city variable and county variables were removed from the dataset. The dataset contains variables with better representation of the customer geographic location.

Certain variable contain a large amount of incorrect values. The “zip” variable is the five-digit zip code representing a physical location. The zip variables transformed into another variable titled “zip 2”, mostly likely with the intention to provide better grouping. Zip 2 variables reduced the zip code down to the first two starting digits. The zip variable should have five digits in every value. However, in a few state locations, the first digit is zero. Since the data is digital, the source of error is most likely a computer program(s). Computer programs are often designed to strip entries of “unnecessary” values. Without user specifications, a data program usually views a zero place before a set a numbers as an unnecessary value and removes the zero from the remaining values. The values from zip 2 variable most likely originate from zip variable. The computer is programmed to select the first two digits as a uniform value to generalize/group locations of the customer. Because the zero is missing in certain zip codes, some codes contain the second and third digit instead of the first and second digit of the code. The errors are repairable, but would be time consuming. The errors were not repaired. The variable was removed from the dataset during backwards eliminations based on grouping issues and insignificance to the research questions.

Plausible duplicate variables are in the data set. The variable representing the marital status of the customer is listed twice in the data set. The married variable is using the letters for yes/no. The other married variable is using binary coding for yes/no (1/0). A few scenarios may have led to duplicate variables. The marriage status may have come from two different sources. The both data sets are include to display the binary representation of the marriage status. The data was converted binary, but the yes/no variable was simply not excluded from the dataset. To resolving the issue, both variable were compared to the analysis methods to narrow it down to only one variable. Due to the computer software and design of the analysis, the married binary was selected. It was noted what the binary numbers represented in the marriage binary variable.

Impractical relationships between the variables exist in the dataset. Third visit implies a yes/no or binary variable, but use the values of 0 to 3. No explanation is available to reference the meaning behind the values. The application of the third spend variable becomes impractical without variables regarding spending from other visits (besides the first visit), such as the second visit. It is hard to pinpoint a specific source of error with such little information and the oddity of the error. Overall, some aspect related to the variable was lost during data transformation, such incorrect binding of data sets or misnaming. Unfortunately, though the data may have been useful for the analysis, the third visit variable was eliminated because its representation cannot be even moderately devised.

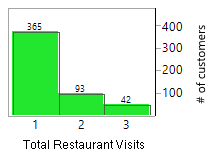
A reoccurring issue is properly titling of variables. The “restaurant” variable does not state exactly what the values represent. It may refer to restaurant number, money spend on the first visit, or total money spent at the restaurant. The values in the restaurant variable were compared to values in other variables and the logical association between variables. Based on the examination, the restaurant variable most likely refers to the amount of money spent at the restaurant. Under the assumption that the restaurant variable represents money spent, the restaurant variable and webstore spend variable are potential misguiding in their titles. It is not clear if the amount spent refers to a summary or the original visit. Due to the range of value in each variable, the restaurant and webstore spend variable are both treated as the total/summary of money spent. The third spend variable needs a more specific title. It does indicate the amount of money spent on the third visit. However, the title does not indicate were the money was spent, such as in-store versus online. The source of error most likely originated from the merging of data from Bubba Gump’s various sources. The location (online versus in-store) was evident at the original source of the data. However, when merged, it was not noted where the data specifically originated from. The application of the third spend variable becomes impractical without variables containing spending from other visits (besides the first visit), such as the second visit or reasonable intervals. The third spend variable was eliminated from the dataset.

Meaningful Patterns

One research question addressed the relationships between age groups and visits amongst Bubba Gump customers. Before comparing customer age to visits, a histogram counted the number of customers in each age group. The purpose was to finding any underlying pattern in the age group that may influence the relationship between age and visits. In the histogram, the age of the customers was grouped in 10-year intervals, starting at age 18. The histogram is left-skewed. For Bubba Gump, the number of customer is highest at the starting age group of 18 to 28. The number of customers decreases as the age of the customer increases. Two additional histograms were executed relating to the research question. The first histogram counted the total number of restaurant visits per customer. The histogram is left-skewed with a range of one to three total restaurant visits.. The number of customers decreases as the total number of restaurant visits decreases. From the sample of 500 Bubba Gump customers, most customers have only visited the restaurant once (see Figure 1).

**Figure 1**

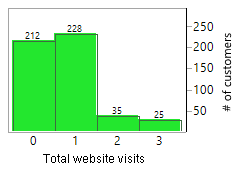
**Total Restaurant Visits per Customer** (Histogram)



The second histogram counted the total number of website visits per customer. The histogram is left-skewed with the range of zero to three total website visits (see Figure 2). The number of customers decreases as the total number of website visits decreases. Notably, approximately 42% of Bubba Gump’s customers have never visited Bubba Gump’s website. Approximately 45% of customers have visited the restaurant one time.

**Figure 2**

**Total Website Visits per Customer** (Histogram)



Another research question addressed the relationship between the income levels and spending of Bubba Gump’s customers. Along with searching for patterns, the histogram help identify outliers when reviewed against the income-to-spending relationship charts. A histogram counted the number of customer in each income level. The income of Bubba Gump customers ranges from $10,000 to $120,000. The income level is grouped into $10,000 ($10k) intervals. The histogram is left-skewed. Overall, the number of customers per income level decreases as the income level increases.

Based on the patterns discovered, Additional research questions to consider are:

how much do customers in each age group spend at Bubba Gump of if the age of the customer increases/decreases spending. It may help to expand the research by pairing other variables together or focusing the research on potential target markets. Some examples are:

* What is the relationship between the marriage status and restaurant visits amongst customers?
* What is the relationship between marriage status and spending amongst customers?
* Do customers from the age of 18 to 48 spend more money on the website than restaurant visit?

## Inaccurate Depictions of Data

Histogram visual the frequency of values through column-charts, often grouped into bins (ThoughtCo, 2019). Binning is a process that groups continuous data into data ranges, usually separated by an interval (ThoughtCo, 2019). Even though mathematical formulas exist, selecting the data range of a bin can be challenging (AnswerMiner, 2017). The length of full span of values, the starting/ ending value of the span, width/range of the bins, and the logical application of the range to what the value represents (AnswerMiner, 2017). In this analysis, the tools would auto-generate bins. The bins were accurate mathematically. However, multiple times the bins were illogical considering the data it represented. When creating the histogram for the age of customers, the software auto-generated 5-years intervals. The intervals started at the age of 15. It was illogical because the age of an adult in the United State is 18 years old and the youngest value in the data set is 19 years old. A new bin formula was created that started the age at 18 and used 10-year intervals to more accurately represent age groups. The website spending had binning issues. Many customers have spent no money on the website. The binning started $0 to $25, continuing at $25 intervals. However, the range was a misrepresentation of the data. Many customer have spent no money at Bubba Gump. The value of $0 is better represented independently, but the variable need to maintain a reasonable interval. The formula was adjusted to start at -$24 with the range ending at $0. The intervals provided a better depiction of the values. However, when the bin -$24 to $0 was written on charts, it had to be manually altered to show $0 as the value it was depicting.

When developing scatterplots (linear regression) for income levels and spending, the wrong variable was selected. The variable containing the binned version of the customers income was used for the factor (x-axis) with the spending as the response (y-axis). The scatterplot was visually flawed. The purchases were scattered correctly, but income was visually incorrect. The plot showed straight, dotted vertical lines per income group. The original scatterplot was disregarded. The scatterplot was attempted again. However, the income variable (non-binned) was selected for the factor (instead of binned income). When reviewed, the scatterplot then had the correct variables for factor and the response.

When developing a scatterplot (logistic regression) for the likelihood of website purchasing and customer income, the wrong variables were selected and variables were also placed on the wrong axis. On the first attempt, the website spending ($/binned) was selected as the factor (x-axis). The income levels (binned) were selected as the response (y-axis). The graph was a contingency analysis instead of logistic regression. On the second attempt, the website spending (Y/N, binary) variable was the factor. The income level (non-binned) variable was the response. The graph showed straight, dotted vertical lines. Before continuing, the information about building a logistic model was reviewed. With logistic regression, the factor should be a continuous variable and the response should be a categorical variable. On the third attempt, the income (non-binned) variable was selected as the factor. The website site visit (Y/N, binary) variable was the response. The logistic graph was created correctly with the proper visual and variables.

Alternative Analytic Methods

Companies often use data analytics make better decisions based on relevant information discovered during the analysis (Vessel, 2018). Companies are not limited to only one method to analyze data. Alternative analytics methods are available for data analysis. Some analytic methods serve as the basis of another method (Vessel, 2018). The variance is usually the specific “question” that the method focuses on (Vessel, 2018).

One set of analytics methods to discuss is the methods of descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics. The methods can be used individually or in-secession. With a common goal, the set of methods forms a strong basis for answering common business questions (Vessel, 2018): What happened? Why did it happen? What will happen? What can be done? Instead of being described as separate methods, the methods are may be referred to as stages (Mehta, 2017)

The Bubba Gump analysis is using a descriptive analysis, but it is helpful to understand the analysis method. Descriptive analysis summarizes prior data to find meaningful information (Gibson, 2018). It discovers what happened in the past and uncovers the relationships (Vessel, 2018). Descriptive analysis is often describe as the simplest form of analytics (Mehta, 2017). It is commonly the basis of more advanced analysis methods. The functions from descriptive analysis include stating the business metrics, identity the data, extract and prepare the data, analysis, and presentation (Vessel, 2018). Its application to business is track Key Performance Indictors (KPI), revenue reports, and sales reviews (Gibson, 2018). Common tools are surveys, transforming data into groups, clustering, histograms, and bar charts (Mehta, 2017). The results can identify issues and unforeseen patterns (Gibson, 2018). For example, it may identify a production issue and allow the company to repair the system instead of sending out a product recall.

The next method is diagnostic analysis. Diagnostic analytics builds upon descriptive analytics to discover why it happened (Vessel, 2018). It identifies the patterns and attempts to dig deeper to find the cause or additional relationships (Vessel, 2018). The analysis may finish faster than expected with prepared data from the descriptive analysis (Gibson, 2018). Tools used are regression analysis, filtering, time-series analytics, and probability theory (Vessel, 2018). The analysis can find the reason why revenue is decreasing or increasing, a sudden increase in customers, or why customer complaints have increased.

The following method is predictive analysis. It switches the focus from past into future outcomes (Gibson, 2018). Predictive analytics focuses on predicting what will happen in the future (Vessel, 2018). The types of analysis techniques include neural networks, machine learning, and statistics (Vessel, 2018). It is important to the source of the data, validating results, and test-out predictions (Vessel, 2018). The analysis can predict sales trends, revenue increases or decreases, or supply/demand reports for logistics.

The final method is prescriptive analysis. Prescriptive analysis focuses on future decisions and solutions (Vessel, 2018). Prescriptive future usually prescribes multiple futures, not just one (Mujawar & Joshi, 2015). It takes predictive analysis to the next level. Prescriptive analysis recommends certain courses of action and shows the likely outcome of each decision (Mujawar & Joshi, 2015). With the help of modern technology such as machine learning, prescriptive analytics uses tools such as optimization and simulation algorithms (Mehta, 2017). Prescriptive analytics requires the application of business rules and government regulations (Vessel, 2018). However, it is important to test the results to help prevent large mistakes (Vessel, 2018).

# Final Report and Presentation

### Milestone Four: Data Presentation and Report

Display and Interpretation

Cluster bar charts examined the relationship between the total number of Bubba Gump visits and the age groups of customers. In the charts, the total visits per customer are clustered within each age group to find trends between and within age groups. Figure 3 analyzes the relationship between Bubba Gump restaurant visits and customer age group. Restaurant visits are most popular with adults from age 18 to 47. The number of visits tends to decline as the age of the customer increases with a slight peak at the age group of 58 to 67. Within each age group, the number of customers decreases as the number of restaurant visits increases. Many Bubba Gump customers in each age group have visited the restaurant only once.

**Figure 3**

Figure 4 and Figure 5 analyzes the relationship between Bubba Gump website visits and customer age groups using a different type of cluster bar chart. The histogram from Figure 2 is a recommended alternative to review for website visits exclusively. Figure 4 represents the number of customers per total visits. Figure 5 breaks down the information into more detail. The chart uses the total customer count (per age group) and transforms the number of total visits into percentages within the age group. This helps to explore how popular Bubba Gump website visits are among each age group. Figure 4 showed inconsistent results on patterns between website visits and age groups. It is more common for Bubba Gump customers to visit the website zero to one time than two to three website visits. Keeping the count in mind, Figure 5 provides a better view of the likelihood and/or popularity of customer visits to Bubba Gump’s website. Returning website customers are most likely to be customers from the age groups of 28 to 47 and 58 to 67, with the highest percentage of third website visits. Customer from the age group of 18 to 27, 48 to 57, and 78 to 87 are least likely to visit the website. Customers from 68 to 77 are a unique age group. The customer are the most likely to visit the website out of all age groups, but usually only visit Bubba Gump’s website once. Figure 4 and Figure 5 do not show any distinctive overall patterns between age and customer visits.

**Figure 4**

**Figure 5**

Scatterplots examined the relationships between income levels and spending of Bubba Gump customers. The income and purpose amounts were not grouped, but a color-based legend was created afterwards to visual the purposing amounts better. Figure 6 examines the relationship between the income and the money spent at Bubba Gump’s restaurant. The income of Bubba Gump’s customer ranged from $20,000 to $120,000. The restaurant spending ranged from approximately $25 to $ 250. Outliers do exists, but are debatable considering the decrease in customers with higher incomes. Figure 6 indicate that no relationship exist between income and restaurant spending. The alternative is Figure 6 indicate that, excluding outliers, higher income customers spending slightly less money than customers with lower incomes. Figure 7 examines the relationship between the income of customers and the amount of money spent on Bubba Gump website purchases. This includes customers who have not visited the website. As the income of Bubba Gump customers increases, the range of the money spent online decreases. The range starts at $0 to approximately $450 for lower income customers. The range steadily decreases to only $0 at the highest income of customers, with one outlier. Figure 8 used a logistic scatterplot to explore the probability of customer making an online purchase in comparison to income levels. The scatterplot compares the income level and if the customer has made any website purchases previously (“Yes” is blue dots / “No” is red dots). The purpose is discovering any relationship between the two variables. Figure 8 shows that customers at lower income levels are least likely to make website purchase. Customers are higher income levels are more likely to make website purchases. As the income of customers increases, the likelihood of website purchasing increases.

**Figure 6**

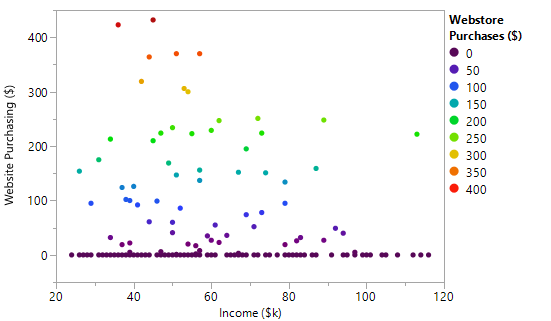
**Restaurant Purchases by Income** (Linear Scatterplot)





**Figure 7**

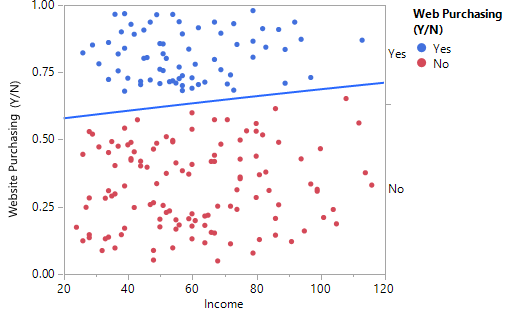
**Website Purchases by Income** (Linear Scatterplot)





**Figure 8**

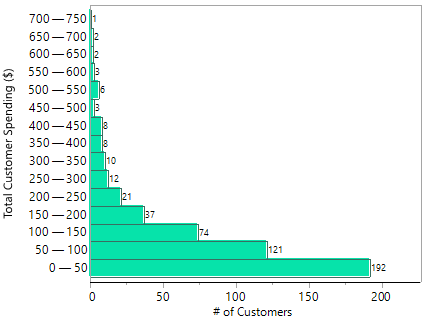
**Website Purchasing by Income** (Logistic Scatterplot)



Histograms were used to determine if there is a pattern within the variables regarding the amount spent at Bubba Gump. Figure 9 places the total spending, online and restaurant combined, into groups and counts the number of customers in each group. The range of the total amount spent at Bubba Gump is $0 to $750 at intervals of $50. Though the histogram is displayed vertically, it is considered left-skewed. As the total amount customers have spent (online and restaurant) at Bubba Gump decreases, the number of customers decreases. Figure 10 visualizes the amount spent in Bubba Gump’s restaurant and the number of Bubba Gump customers in each restaurant spending group. The range of restaurant spending is $0 to $225 at intervals of $25. As the restaurant spending decreases, the amount of customers decreases with a peak at $25 to $50. Figure 11 visualizes the amount spent on Bubba Gump’s website number and the number of Bubba Gump customers in each online spending group. The range of website spending is $ 0 to $500 at intervals of $50. Due to the large number of customers, $0 was created an independent spending group. As the spending on website purchases decreases, the amount of customers decreases.

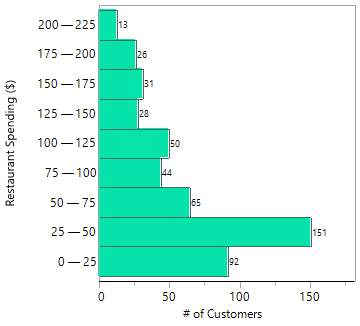
**Figure 9**

**Total Bubba Gump Spending of Customers** (Histogram)

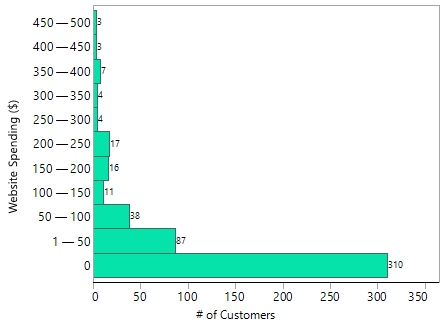


**Figure 10**

**Restaurant Spending of Customers** (Histogram)



**Figure 11**

**Website Spending of Customers** (Histogram)

## Validity, Reliability, Limitations

Validity refers to the accuracy that the analysis researched the intended variables or subject (Price et al., 2015). The data came from internal sources. Based on the information provided, the full dataset come from data within Bubba Gump. The data was not full set pulled from public sources. It maintains accuracy because it allows the research to remain focuses on Bubba Gump instead random businesses or population sets. The variables were selected that connected to the research questions. However, an argument is plausible that the variables differed slightly from the research question. At one point, the relationship between customer purchase amounts and income was examined. A complete summary of restaurant and website spending was not used. Instead, restaurant spending and website spending were examined independently. From certain point-of-views, that method may be considered partially inaccurate. However, to unify the analysis, the same income variables used in each examination, using the exact same intervals. The analysis may be considered partially inaccurate because it did not explore enough variables or values to make solid business decisions. No relationship was found between restaurant variables and income. Since no pattern or relationship was found, it has little to no usability outside of the analysis. The analysis of website and spending was obscure, so it is most likely not applicable to generalize outside of the study. The analysis did not reveal any patterns to apply to broader data. However, the data did provide possible target markets to perform further research. For the relationship between customer spending and the number of customer, the pattern can be applied outside of the analysis. It showed a clear relationship between spending the number of customers. It could applied to regions with different, including larger, number of customers. Bubba Gump could also use it adjust prices, assess how it compare to their revenue, such as ratios or percentages to determine profit, sales, or revenue. One limitation is the finding specifically focus on current Bubba Gump customers. The findings do not research to the general population to gain new customers. The dataset had no information regarding customer satisfaction to find ways to help Bubba Gump understand their customers. No period was available to research. Time periods are useful to find sales trends, discover how sales affected Bubba Gump, research customer loss or changes in demographics, gain new customers and/or increase customer loyalty.

Reliability refers to the consistency of the research (Price et al., 2015). The question to address is: Are the finding repeatable when applied to different data samples or performed by another researcher (Stephanie, 2016)? The analysis does have issues with reliability. The labelling of the variables is one issue that affects reliability. If another researcher performed the test independently, the researcher may have different interpretation of what the variables represent, such as if “Restaurant” represents the restaurant number instead of the amount spent at the restaurant. In this analysis, the “Restaurant” variable was compared to the location of customers. The restaurant variable did not demonstrate any relationship to customer locations. With a review of how the restaurant spending related to the values in other relationships, the restaurant variable was deemed as the amount spent at Bubba Gump restaurant visits. With analysis involving spending at Bubba Gump, the results of another analysis or sampling are likely to show the similar findings. However, because intervals can be grouped in different intervals, the results may have some variation in how prominent the patterns are. To address this issue, multiple variations of the intervals (for each variable used in the spending histograms) were tested to review consistency in the patterns. With some slightly variations due to the size of the patterns were consistent across the different intervals of each variable, allowing for practical application of generalization. For the scatterplots analyzing the relationship of purchases and income, the results have a limitation for practical usage. Outliers may have affect the interpretation of the results. Other sampling or a different dataset may have revealed a relationship between restaurant purchases and income or just the opposite for website purchases and income. Different samples were used to test out the relationships between spending and income. Some slight variation did occur, but overall the relationship stayed the same. Colorization had a dual purpose for the scatterplots. The first purpose was for viewers to have a better visualization. However, the second purpose was to help find outliers that may have been missed with all black dots. The scatterplot are applicable for generalization, but additional dataset would be advisable to further valid the relationship and lack of relationships. The limitation is sampling is only available on this dataset, which is somewhat unreliable. Another limitation is the trend may not be evident because of the small size of the data set and lack of a time frame. It is unknown is the values are from recent transactions, over a larger time span, or when Bubba Gump was very successful. For the cluster bars, the trend is restaurant visits are generalizable. Trends are evident with each age group and overall . Similar to the Bubba Gump spending histograms, different intervals were test to explore variations. The trends were consistent among the tests. The website testing has a lower level of generalization and consistency. The limitation is that the website visits and age groups produced randomization. Other visuals were tried, such as a standard bar graph and scatterplots, but did not provide any more effect means of interpreting the values for possible trends. Another researcher may use a more efficient chart/graph that better displays that produces more stable findings. Though it is consistent in showing target groups, it is too randomization for general application spread across the age groups.

## Resulting Decision Influence

To note, in this section, Instead of repeating “client and team”, “client” is referring to any client, superior, or team involved in the review and decision making.

The analyzer/writer of the report may fully understand all the material within the analysis, but it does not mean the client can understand it (Brown, 2017). However, it is better to assume that the client prefers some explanation and guidance in properly interpreting the analysis (Brown, 2017). This project would not be closed just because the paper/digital report is sent to the client. Decisions should be made based on a solid understanding of the material that applies to the situation. Clear methods of communication are available for the client and analyst to discussion the analysis, such as meetings, emails, phone calls, and chats. The analyst can make clear and simple statements on how the variables and analysis relates to real-life application (Brown, 2017). For example, the analyst can give suggestions on areas that the analysis is applicable with, such as marketing, understand clients, or addressing finances (Valcheva, n.d.). The analyst can always use the communication to discover if additional steps need are needed. Some examples are finding and pairing outside information or possibly redoing parts (such as visuals or reports on findings) that the client is still confused on after the discussions (Brown, 2017). If external information needs to be paired with the current data, the analyst can find the proper and reliable source of data. The analyst can review how the data relates, format it to match with the database, combined the data, and provide the findings (Brown, 2017). Depending on the goal or question at hand, the pairing of data can lead to more stable decision. The company or individual can view how the outside information compares with Bubba Gump’s dataset. Bubba Gump could find out how competitors compare to them and adapt their sales accordingly (Valcheva, n.d.). If desired, they could also find out more about possible target markets and their preferences (Valcheva, n.d.). Overall, decision should supported by facts, not assumptions made by the analyst or client. The analysis is a reliable method that test states, produces the facts, and offers ideas for additional paths to explore (Brown, 2017).

The Bubba Gump analysis report generates possible recommendations for the restaurant to consider. However, Bubba Gump should be aware that a business analyst and data analyst are not necessarily same (UMD, 2020). A business analyst is in transforming the findings into valuable suggestion and solutions with a background or specific training in business affairs (UMD, 2020). A data analyst defines business goals to focus the analysis on, produces the findings, and translates the data into non-technical terms (UMD, 2020). Data Analysts are still capable of assisting in business decisions (UMD, 2020). Although supplementary business knowledge is highly limited, the analysis provides at least a few recommends for Bubba Gump. Bubba Gump can either broaden their online market or focus on high-level target market. Based on website visits per age group, Bubba Gump can focus on customers from 37 to 47 and 58 to 67. Out of all the age groups, customers from the age groups have already started to visit the website two to three times. If they want to expand their online market, Bubba Gump can target the age groups of 18 to 27 and 48 to 57. The age group of 18 to 27 has the largest amount of zero website, but some first to second visits customers. The age group of 48 to 57 has a medium level of zero visits with a medium number of first visits also. With a small amount of customers visiting the website a second time, Bubba Gump can try to increase the number of website visits of 48 to 57, leading to a wider market.

Bubba Gump needs to increase customer loyalty online and in-store. The restaurant visits per age group shows a sharp drop from one visit to two visits per age group. Also, 30% to 50% page of each age group have not visited the website. Combined with the histograms regarding spending, many customers only spend $0 to $100 dollars. If Bubba Gump’s prices are higher, Bubba Gump needs to offer deals or lower prices. If Bubba Gump’s prices are lower, Bubba Gump should consider raising prices. To help appeal to higher income levels, Bubba Gump should add special entrees of a higher price. According to the scatterplots, higher income customers purchase less and lower amounts than other customers.. To help increase loyalty, Bubba Gump should consider implementing a customer rewards system that benefits both the customers and Bubba Gump. The marketing department, sales, and financing should review the system before implementation to make sure it benefits both sides.

Visual Evaluation

Clustered bar charts were used for Figure 3, Figure 4, and Figure 5. Clustered bars starts examined the relationship between Bubba Gump visits and the age group of customers. In all three charts, the visits were clustered per age group. The age is grouped together in the exact same manner across all three figures. It starts at 18 years old and continues at 10-year intervals. The intervals may be slightly confusing at a quick glance. Intervals starting at 10 or 20 would cause a misrepresentation. No values were under 18 years old. Starting at 20 years old would have ignore the values of 18 and 19 years old. In presentations to others, it is important that values are correctly represented. Figure 3 examines the relationship between restaurant visits and the age groups of customers. Each visit total had a separate bar, colored independently. The relationships can be viewed within age groups and the overall trend across. A legend was added to associate the colors with the total number of restaurant visits. Overall, Figure 3 was presented properly. Figure 4 and Figure 5 examine the relationship between Bubba Gump website visits and the age groups of customers. Figure 4 is set up the same way as Figure 3. Figure 4 faces the same advantages and issues to viewers as Figure 4. However, it is challenging to find any relationships among the values. Figure 5 is somewhat different from Figure 4. Figure 5 examines the percentages rather than the number of customers. It may be hard to understand since it turns the counts into percentages. However, it gives a better view on what age groups are more likely to visit Bubba Gump website.

Standard scatterplots examined the relationship between the income of customers and spending at Bubba Gump. Figure 6 examined restaurant spending and income of customers. The cause-and-effect was to see how changes in incomes affected restaurant spending. Income was the factor, correctly placed on the x-axis. The income scale may cause a visual misconception. The income scale starts at $20,000 instead of $0. Among the data sample, no customers were listed with an income under $20,000. The computer software automatically sent the range based on the values. The income was labelled according to scale, but it may be overlooked that it starts at $20,000. The scale could have started at $0, but may have caused a larger confusion with the large gap in values at the start of the scatterplot. Instead of box of all black dots, the dots were converted to a color scale to create a better visualization for the restaurant purchasing scale. A legend added for viewer to associate the colors with the amount spent at the restaurant. Figure 7 examined website spending and the income of customers. The cause-and-effect was to see how changes in incomes affected restaurant spending. The income of the customers Income was the factor, placed on the x-axis. Figure 7 has the same issue with the income scale as Figure 6. Due to the software’s application of the values, the income scale starts at $20,000 instead of $0, which may be overlooked. Among the data sample, no customers were listed with an income under $20,000. Like in Figure 6, starting the scale at $0 may have caused more confusion rather than clarity. The misleading part is that the scale is larger for the website spending than the restaurant spending. However, a legend was created clearly stating the values associated with the colors in both Figure 6 and Figure 7.

The logistic scatterplot about the likelihood of website (Figure 8) was useful for visualizing the increase in probability. The yes and no responses were colorized (blue and red) to understand what the line in the graph was representing, which was the cutline between yes and no. Labels were placed to help comprehend the scatterplot. However, without proper explanation, the logistic scatterplot would be challenging to comprehend, especially when association with the overall meaning and how the visualization relates to each label. It may have been better represented using a statistics table, but would require carefully selection in how to display the statistics, such groupings, labelling, or numbers versus percentages.

Histograms were used in two sections of the analysis. Figure 1 counted the number of customers per restaurant visits. Figure 2 counted the number of customers per website visit. Figure 9, 10, and 11 counted the number of customers per spending group, including total spent, restaurant purchases and website purchases. The axis included the scale, grouping, and labels. The histogram bars remained uniform in color because they were representing the same variable. Histograms are used to display the frequency of continuous variables. Since restaurant visits and spending are technically infinite, the histogram was the proper selection. It clearly categorized the values into groups (as needed) and provided the customer count, which is the frequency, in the chronological order of the visits or spending. The number of customers was added to the end of each bar/group. With its similarities to the bar chart, the visualization of the customer count is clear. Most patterns are fairly clearly and easier to understand visually than by descriptive. The challenge is understanding the official name of the pattern according to histogram terminology. Overall, a short explanation usually clarifies any confusion or misunderstanding of the histogram.

## Next Steps

Before making additional decisions, a new analysis should be considered. With the goals of Bubba Gump already established, additional hypothesis are the new starting point of the research. However, the dataset needs better preparation with more information available, such as the time ranges, better labelling, or explanations on what the data is representing. Data regarding time would be especially useful. The restaurant may be more popular in the summer. However, customers may prefer shopping online during the winter, but cannot find any good deals that appeal to them. Information related to time can help pinpoint issues or develop new hypothesis to explore.

Since the business is considered with sales and not just customer demographics, a dataset focusing specifically on marketing, advertising, and their relation to sales is highly recommended. It can be compared with demographic, visits, and spending. Bubba Gump will likely find relationships in their advertising and marketing compared to customer purchases and possibly demographics. It may also be helpful to gather data regarding customer satisfaction, rating, and the reputation of Bubba Gump. Analysts and businesses including restaurants should no automatically assume they know what the problem is or what the root cause is. It is better to have information that provides a solid foundation for future decisions and addressing issues that the company is experiencing. For example, the current information can be used to select target markets for Bubba Gump. However, Bubba Gump needs to hopefully find out what is appealing to the target markets, their purchasing and visiting habits at Bubba Gump, and what the target market’s view/opinion is about Bubba Gump.

# References

Ahlemeyer-Stubbe, A., & Coleman, S. (2014). *A practical guide to data mining for business and industry*. Wiley. https://mbsdirect.vitalsource.com/#/books/9781118981863/

AnswerMiner. (2017, August 1). *Histogram – The Ultimate Guide of Binning - AnswerMiner*. AnswerMiner. https://www.answerminer.com/blog/binning-guide-ideal-histogram

Australian Bureau of Statistics. (2013, July 3). *Statistical language - census and sample*. Australian Bureau of Statistics. https://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+census+and+sample

Biddix, J. P. (2018). *Research methods and applications for student affairs*. Jossey-Bass, A Wiley Brand.

Brown, M. S. (2017, May). *If You Want To Succeed With Analytics, Effective Communication Is A Must*. Forbes. https://www.forbes.com/sites/metabrown/2017/05/30/if-you-want-to-succeed-with-analytics-effective-communication-is-a-must/#748e11f83765

Cocchi, M. (2019, May 14). *Backward Elimination - an overview | ScienceDirect Topics*. ScienceDirect. https://www.sciencedirect.com/topics/computer-science/backward-elimination

Crane, K. (2018, June 21). *14 Must-Read Restaurant Trade Magazines*. Restaurant Insider. https://upserve.com/restaurant-insider/restaurant-industry-magazines-put-reading-list/

Enago Academy. (2017, August 17). *How to Identify a Meaningful Research Question*. Enago Academy. https://www.enago.com/academy/things-need-know-writing-good-research-question/

Gibson, P. (2018). *Types of Data Analysis*. Chartio; Chartio. https://chartio.com/learn/data-analytics/types-of-data-analysis/

Harmath, L. (2014, May 13). *Line Graph - When to Use It?* ClickDimensions. https://www.edrawsoft.com/chart/when-to-use-line-graph.html

Import.io. (2019, October 29). *What is data visualization and why is it important?* Import.Io. https://www.import.io/post/what-is-data-visualization/

Investopedia. (2019). *Stepwise Regression*. Investopedia. https://www.investopedia.com/terms/s/stepwise-regression.asp

JMP. (n.d.). *Key Features of JMP Pro*. Www.Jmp.Com. Retrieved June 28, 2020, from https://www.jmp.com/en\_us/software/predictive-analytics-software/key-features-of-jmp-pro.html

Klipfolio. (2019). *What is Data Visualization? Definitions, Graph Types and How to Use Them*. Klipfolio. https://www.klipfolio.com/resources/articles/what-is-data-visualization

Mehta, A. (2017, October 13). *Analytics Insight*. Analytics Insight. https://www.analyticsinsight.net/four-types-of-business-analytics-to-know/

Mujawar, S., & Joshi, A. (2015). Data Analytics Types, Tools and their Comparison. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(2). https://doi.org/10.17148/IJARCCE.2015.42110

Price, P., Rajiv Jhangiani, & Chiang, I.-C. A. (2015). *Research methods in psychology*. Saylor.Org. https://opentextbc.ca/researchmethods/chapter/reliability-and-validity-of-measurement/

Rakesh Kumar. (2018). Impact of Demographic Factors on Consumer Behaviour - A Consumer Behaviour Survey in Himachal Pradesh. *Global Journal of Enterprise Information System*, *6*(2), 35–47. https://doi.org/10.18311/gjeis/2014/3084

Ribecca, S. (n.d.). *Line graph*. Data Visualisation Catalogue. Retrieved May 30, 2020, from https://datavizcatalogue.com/methods/line\_graph.html

Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, *5*(1). https://doi.org/10.1186/s40537-018-0143-6

Soni, D. (2018, March 22). *Supervised vs. unsupervised learning*. Towards Data Science. https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

Stephanie. (2016, July 1). *Reliability and Validity in Research*. StatisticsHowTo. https://www.statisticshowto.com/reliability-validity-definitions-examples/

TEC. (n.d.). *JMP: Reviews, Pricing, Alternatives & Ratings | TEC*. Www3.Technologyevaluation.Com. Retrieved June 28, 2020, from https://www3.technologyevaluation.com/sd/solutions/jmp-68310

Thibodeaux, W. (2019). *Advantages & Disadvantages of a Demographic Environment*. Chron. https://smallbusiness.chron.com/advantages-disadvantages-demographic-environment-22283.html

ThoughtCo. (2019). *What Is a Histogram and How Is This Graph Used in Statistics?* ThoughtCo. https://www.thoughtco.com/what-is-a-histogram-3126359

TrustRadius. (n.d.). *JMP Pro Reviews & Ratings 2020*. TrustRadius. Retrieved June 28, 2020, from https://www.trustradius.com/products/jmp-pro/reviews

University of Maryland (UMD). (2020, March 18). *Comparing Analytics Careers: Business Analyst vs. Data Analyst*. Online Business UMD. https://onlinebusiness.umd.edu/blog/business-analyst-vs-data-analyst/

Uzialko, A. (2019, June 23). *What Is B2B?* Business News Daily. https://www.businessnewsdaily.com/5000-what-is-b2b.html

Valcheva, S. (n.d.). *7 Real-World Examples Of Data Mining In Business, Marketing, Retail*. Http://Www.Intellspot.Com/Data-Mining-Examples/.

Valcheva, S. (2017, December 1). *What Does a Scatter Plot Show? What is The Purpose Of It?* http://www.intellspot.com/scatter-plot/

Vessel, D. (2018, May 9). *Analytics 101: 5 Steps of Discussion Support*. IBM Business Analytics. https://www.ibm.com/blogs/business-analytics/planning-analytics-101-what-is-our-plan/

Walker, B. (2017, October 17). *The Pros and Cons of Consumer Segmentation*. https://insights.patientbond.com/blog/the-pros-and-cons-of-consumer-segmentation

‌

‌