Joseph Silva Jr

10/10/21

SNHU

DAT 220: Milestone 3

**Analysis Organization:**

The clusters analyzed the data for customers purchases. The web store purchases, and in-restaurant purchases were compared by using Hierarchical Cluster and a K Means Cluster. These clusters showed visual interpretation for the amount of revenue each type of purchase brought to the company from the information provided by the customers.

Next, the linear regression model explained the web channel expenditure with various characteristics of customers such as age and income. The model showed a visualization graph and a linear line comparing web store spending and the characteristics with data points. The graphs were able showed statistics of the RSquare, which compared the variables to each other and if the variables were related to each other.

Lastly, the logistic regression model explained the variables of the WEB\_PURCH\_YN and WEB\_VISITS. This model showed the relationship between customer web store visitation and if they purchased from the web store. The model showed a visual graph showing a web purchase was most likely to happen when the web visit for the customer was 1+.

**Source of Error:**

When reviewing the data, I was unable to find any errors in the raw data which was used to create the graphs. The only issues I found will be discussed in a few paragraphs below under the heading Inaccurate Depictions of Data because it does not show if a purchase was made during each web visit or if it was only one purchase out of multiple visits.

**Meaningful Patterns:**

A Hierarchical Cluster and a K Means Cluster were created using the variables of Web store spend and Restaurant (Spend). Looking at the patterns from the clusters, I was able to determine customers purchased more using the web store for Bubba Gump than in-restaurant purchases. Next, the patterns of a logistic regression using the variables WEB\_PURCH\_YN and WEB\_VISITS showed the data for customers became less when it came to visiting the web store multiple times.

**Inaccurate Depictions of Data:**

An inaccurate area of data would be comparing the variables using WEB\_PURCH\_YN and WEB\_VISITS because we are unable to determine if each visit resulted in a web purchase of an item. This inaccurate data most likely caused an issue with the percentage for the logistic regression model graph because the model showed a 100% when a customer visited the web store 2+ times. These results are more than likely inaccurate due to the 100% results of purchase from the web store from the data mined from the warehouse.

**Alternative Analytic Methods:**

When it came to the graphs, an alternative method would be to break down the data even further into smaller groups than the 500 data points. This method would be most helpful when it came to the linear regression model because the data points were too much to see if there were any patterns within the data points. The graph was unable to show if any characteristics were related when it came to the web store purchase.