STAT 450 Project

**Customer Personality Analysis**

**Submitted to**

Dr. Rebecca Le

**Prepared by**

Evan Cabrera

Sean Cunniff

Gregory Lent

Luis Osorio

May 7, 2023

1. **Background and Introduction**

Businesses serve a large and varied demographic of customers. As data science expands, more information is being collected by businesses about customers every day. To maintain customer happiness and increase profitability, it is useful to provide targeted advertisements to customers. Targeted advertisements allow customers to not have their time wasted, and for businesses to not waste money on meaningless promotions. The decision about how to best advertise to customers can be complicated, but being able to identify specific groups and needs is a good first step.

This leads to our main research question of interest: How can customer personalities be used to divide customers into groups? If there are no prior assumptions being made about group numbers or structures, an excellent technique to answer this question is via Cluster Analysis. Cluster analysis is a form of unsupervised classification, in which variables are grouped into clusters based upon similarities or distances. Furthermore through our process of data exploration we developed a secondary research question: What is the impact of education and children on the total purchase amount and income?

1. **Data and Exploratory Analysis**

To answer our research question, we utilized the Customer Personality Analysis [dataset](https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis) found on Kaggle.com. This dataset contains information about 2240 customers. Each row of data includes several types of information including demographic information, purchase history, purchase format, and how the customer has responded to marketing outreach.

We began our investigation into the data with exploratory data analysis, and basic visualization of the data set which can provide further insights. We examined the overall shape of the data, investigated outliers in numeric variables, and calculated new variables that were believed to be helpful for future analysis.

Our initial exploration found that the data was composed of 2240 customers across 27 variables. We examined the data for absent or missing values, and we found 24 customers without income data; these observations were dropped from the data set. Furthermore, we plotted several numeric variables (see Appendix A**)** which we suspected might contain outliers, and through this process determined that one customer’s income was significantly larger than other customers, and three customers had ages believed to be outliers. These four customers were removed from the dataset. Leaving the final number of observations at 2212.

Age and “time as customer” are provided in the dataset as dates, so they have been given an appropriately calculated value in years. Two additional variables were calculated due to suspected usefulness: total spent by a customer, and total number of advertisements responded to. The final part of our data exploration included constructing a correlation plot of our numeric variables to receive a general understanding of how our data might be related (see Appendix A).

1. **Model and Results**

**Cluster Analysis**

Clustering is an unsupervised learning technique which classifies observations of an unlabeled dataset into groups. There are several different types of Clustering algorithms which all use different distance metrics. For our specific dataset we used K-Means clustering, which is a non-hierarchical clustering method that uses the smallest centroid distance to classify the observation. A notable feature of K-Means clustering is that it is very sensitive to the number of clusters chosen.

In order to determine the optimal number of clusters, we considered the Elbow Method and the Silhouette Score. The Elbow method uses Inertia, or within-cluster sum of squares, which calculates the sum of squared distances between each data point and the centroid of its assigned cluster. As the number of clusters increases, the within-cluster sum of squares will decrease. These points are plotted in a graph, and the elbow of the graph is chosen, hence the name. However, for this dataset, the Elbow Method proved to be an impractical choice.

Instead, we used the Silhouette Score, a criterion which takes on a value between -1 and 1. This score provides a measure of how well each data point fits into its assigned cluster based on both how close it is to other points in the cluster and how far it is from points in the neighboring clusters. A score close to 1 indicates that the data point is well-matched to its own cluster, while a score close to -1 indicates a poor fit. A score close to 0 indicates that the point is roughly equally close to points in its own cluster and those in neighboring clusters. Using the Silhouette Score metric we discovered that the optimal number of clusters for our model was two with a Silhouette Score of 0.286.

In these two clusters, the first cluster contained 864 observations and the second cluster contained 1344 observations. To investigate what separates these two clusters, we examined the criteria used by the K-Means algorithm to find the difference between these two groups. To do this, we examined the differences between the means of each variable between the two clusters. In order to quantify these differences, we conducted a t-test using a significance level of 0.05/20 (the number of predictors) = 0.0025 testing the differences between the means of each predictor (See Appendix A). Since the t-test only works for continuous variables, we used a Chi-Squared Test of Homogeneity for the categorical variables in our dataset. These tests showed that income, children, amount spent on meat in the last 2 years, amount spent on fish in the last 2 years, number of purchases made using a catalog, and total amount spent differed significantly between the two clusters.

**MANOVA**

In order to quantify the effect of education and children on income and spending, a two-way MANOVA (Multivariate Analysis of Variance) was performed. The customers were grouped according to education level (high school, college degree, or postgraduate degree), as well as whether they have any children at home, for a total of six groups. The response variable was a vector containing yearly income and total spent in the store in the last two years. Additionally, a post-hoc Tukey's HSD test was performed to test the individual interactions between the grouping variables.

When testing for the multivariate normality assumption of the MANOVA model, it was found that income does not follow a normal distribution, so the assumption of multivariate normality was not satisfied (see Appendix A). Additionally, the assumption of homogeneous variance-covariance was not satisfied for education level or number of children (see Appendix A).

Both income and total spent were shown to be significantly affected by both education and presence of children (*p* < 0.0001, see Appendix A). Additionally, most interactions between variables were significant in the post-hoc tests. For income, insignificant interactions include the presence of children for those with a high school or college education, and the difference between a graduate and postgraduate college degree for customers with and without children. For total spent, insignificant interactions include the presence of children for those with a high school or college education, the difference between those without children and a high school education and those with children and a postgraduate degree, and the difference between a graduate and postgraduate college degree for customers with and without children (see Appendix A). All other differences between groups were significant.

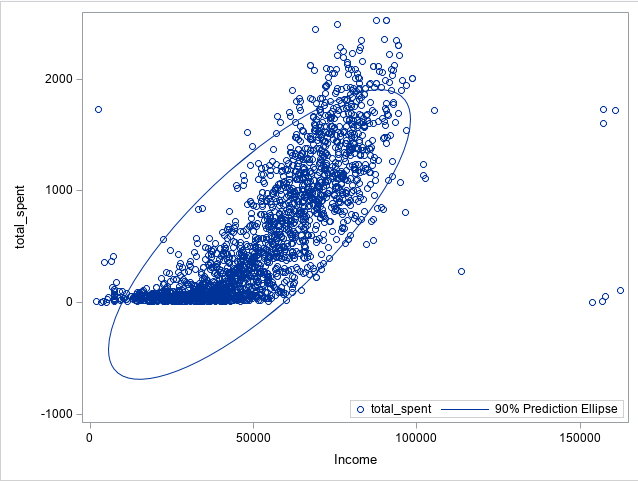
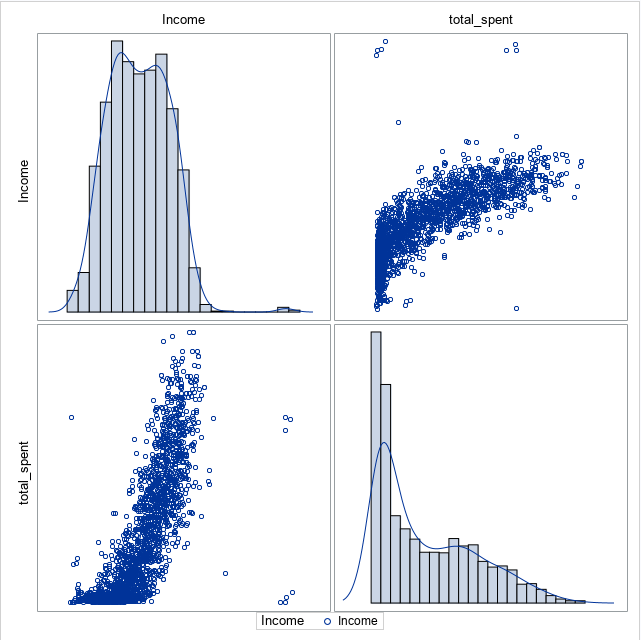
1. **Discussion and Conclusion**

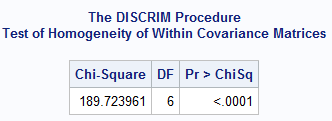
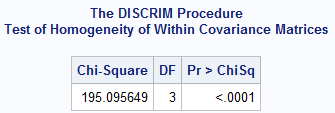
After cleaning the data and exploring the surface level associations, we used a K-means clustering method to group customers into clusters. The number of clusters to be used within our K-means method was specifically picked following consideration of the Elbow Method and Silhouette Score for the number of clusters to use. Comparing the column means between our clusters we were able to eliminate insignificant variables from the clusters and develop a characterization of the clusters.

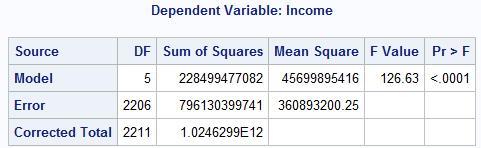
Considering the significance of children in the cluster, and our previous expectation of the importance of education, we decided to quantify the influence of these two variables on total amount spent and income. Our multivariate normality assumptions failed, deeming our MANOVA findings unreliable. However, we were able to determine that income and total spending were shown to be significantly affected by the presence of children and education level.

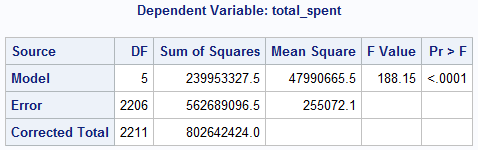
In conclusion, we have found that our multivariate statistical methods of Clustering and MANOVA to be useful in the analysis of customer data. These methods were able to break customers into groups, and further define relationships and interactions between variables. The results from using either of these methods could potentially help businesses to better provide services to customers based upon customer shopping personalities.

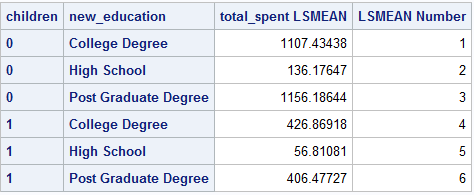
**Appendix A: Graphs and Results**

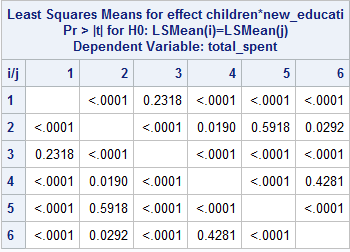
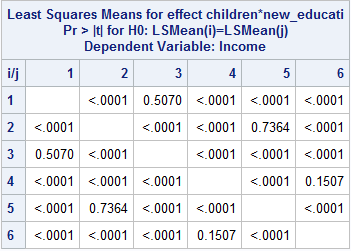
****

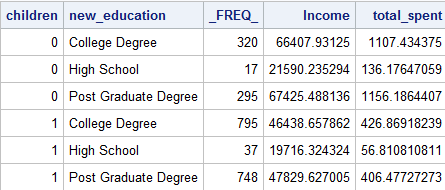
****

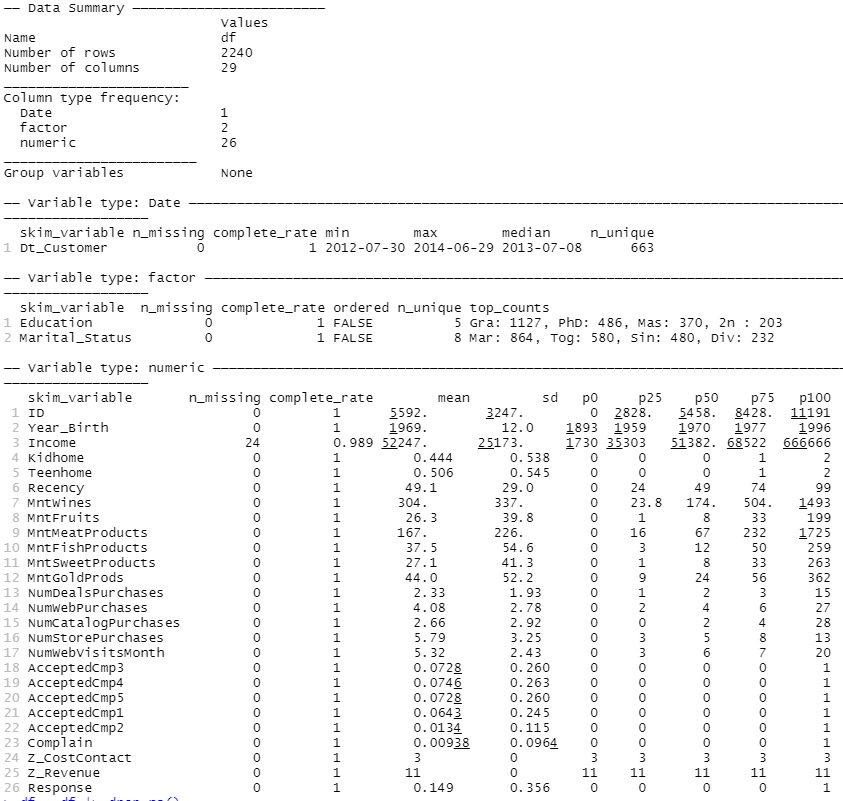
****

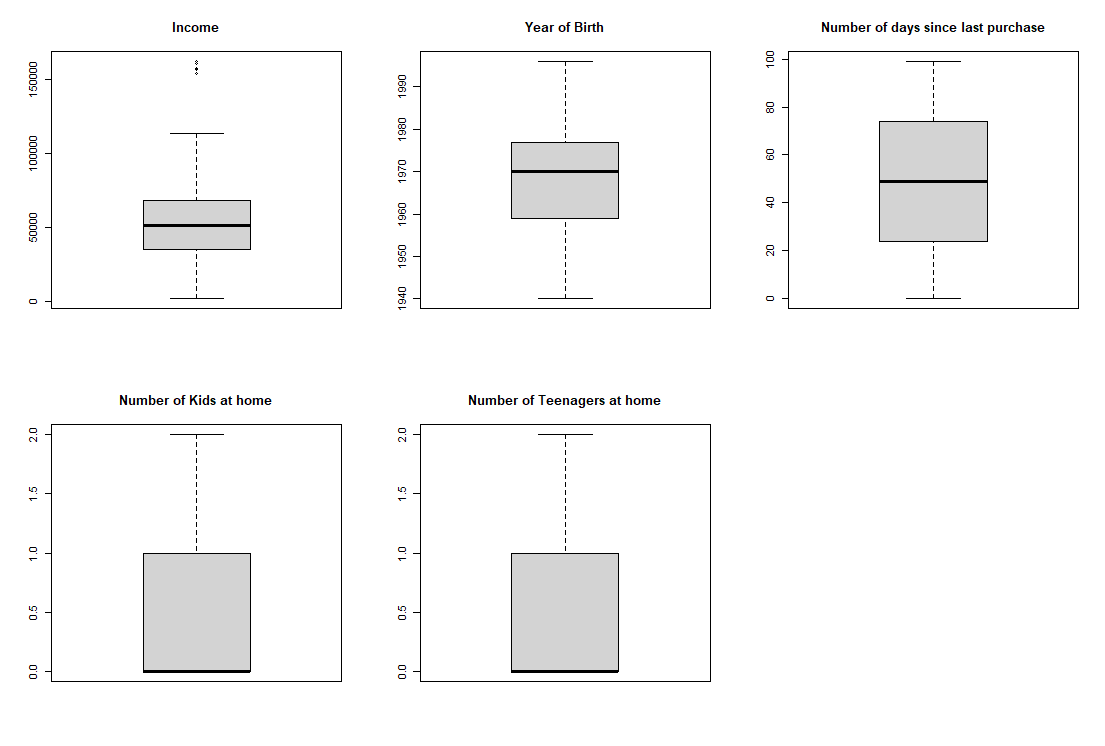
****

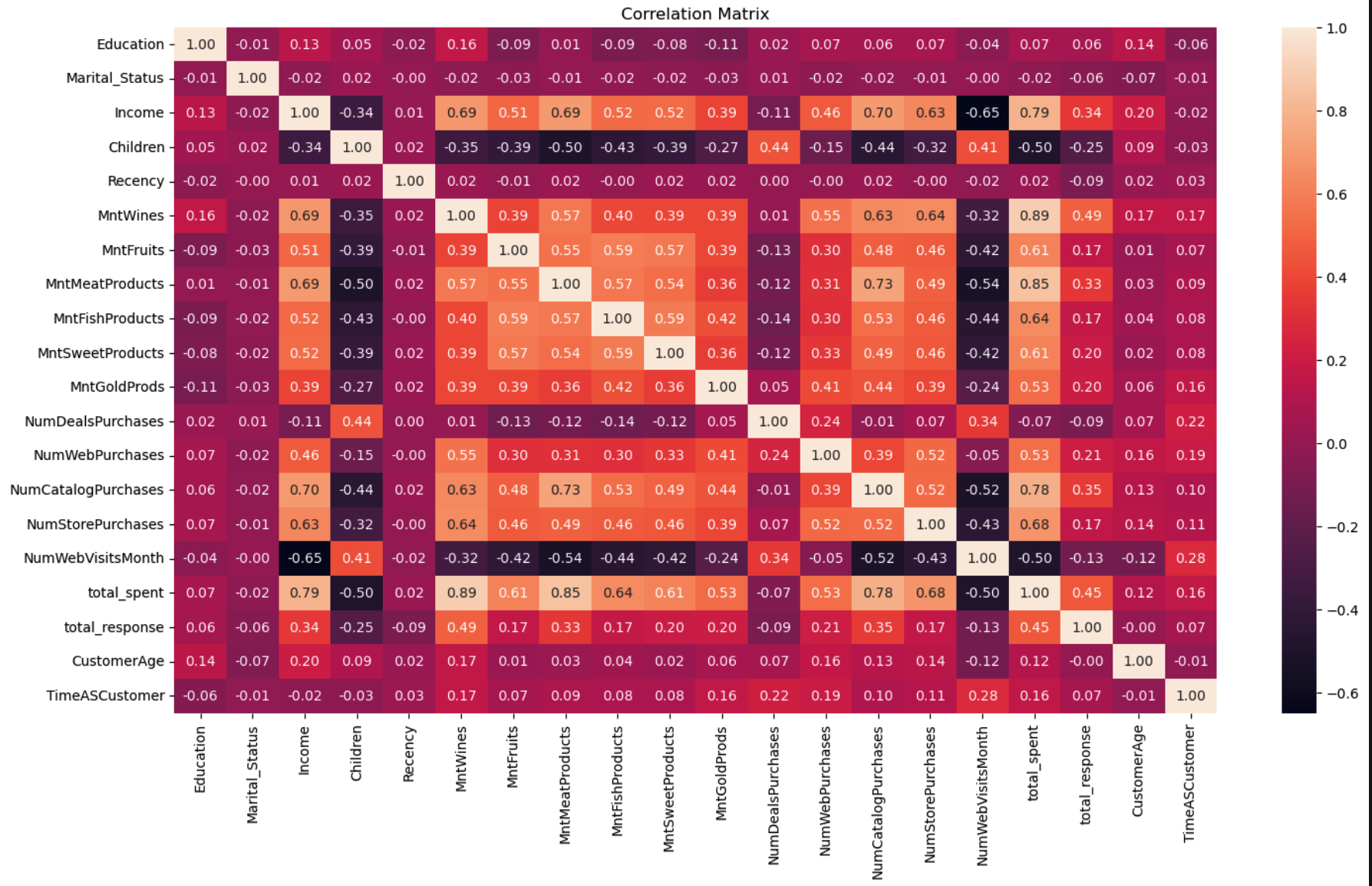
****

****

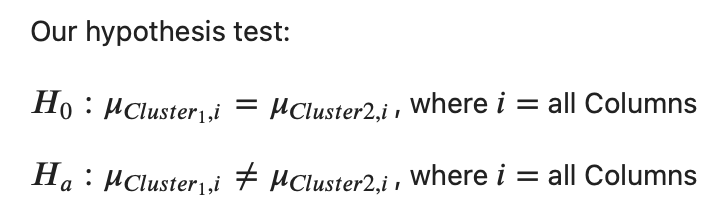
****

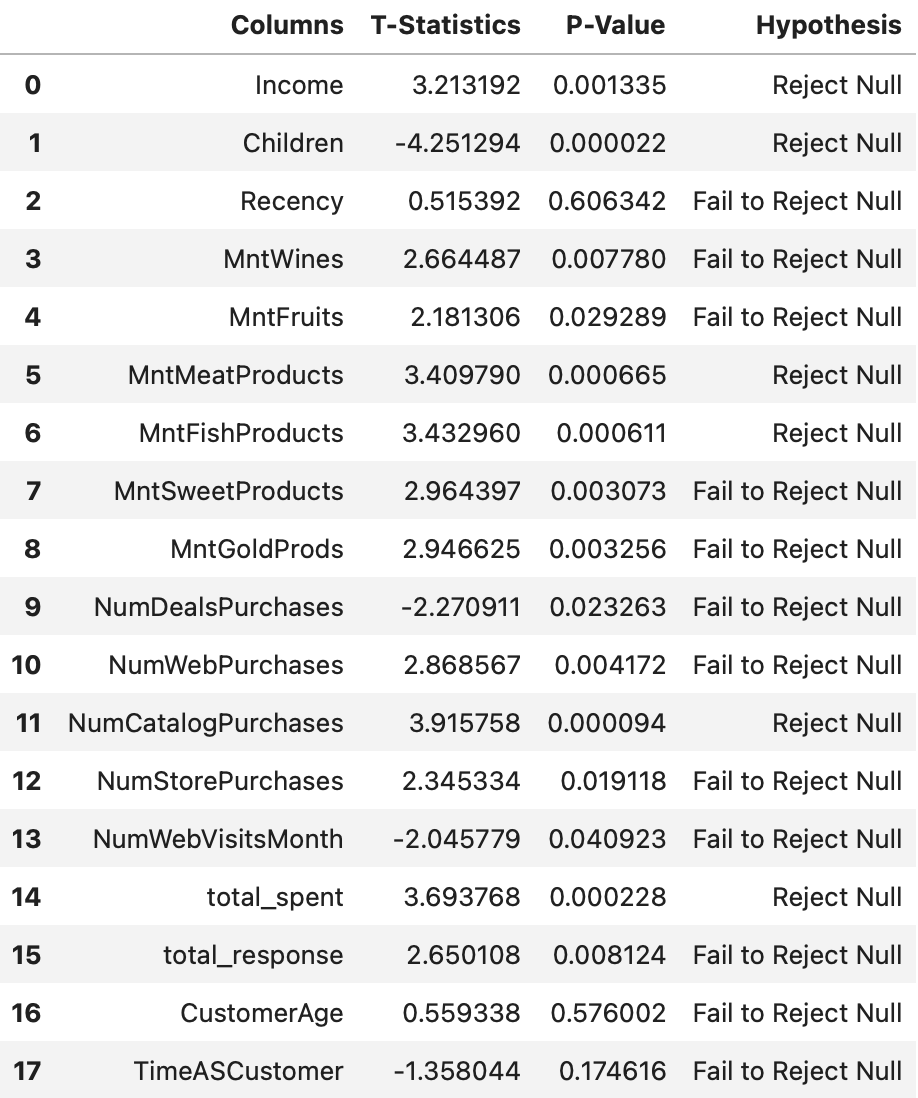
****

****

****

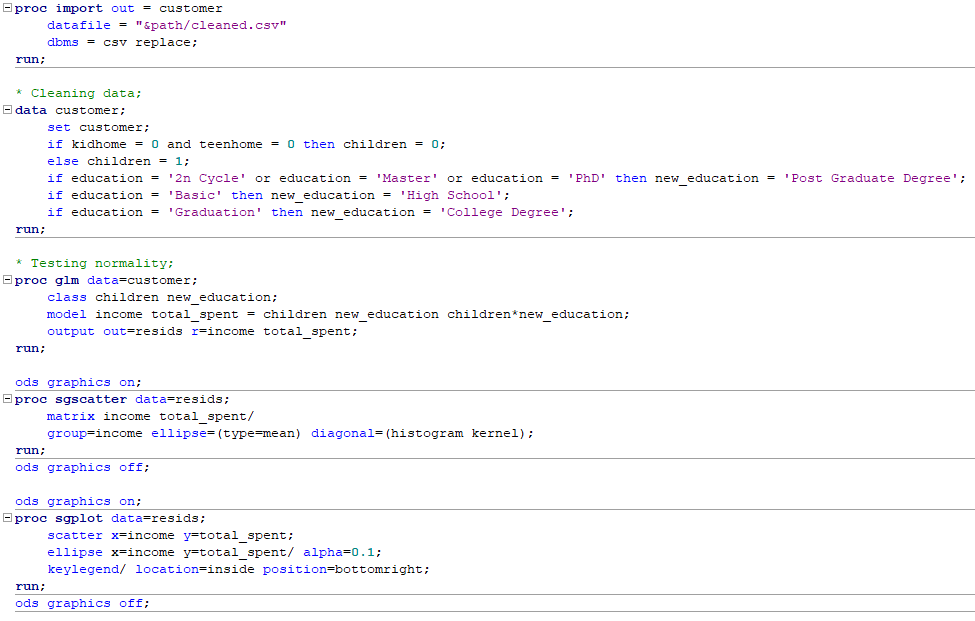
**Unpaired T-Test Results for our 2 Clusters:**

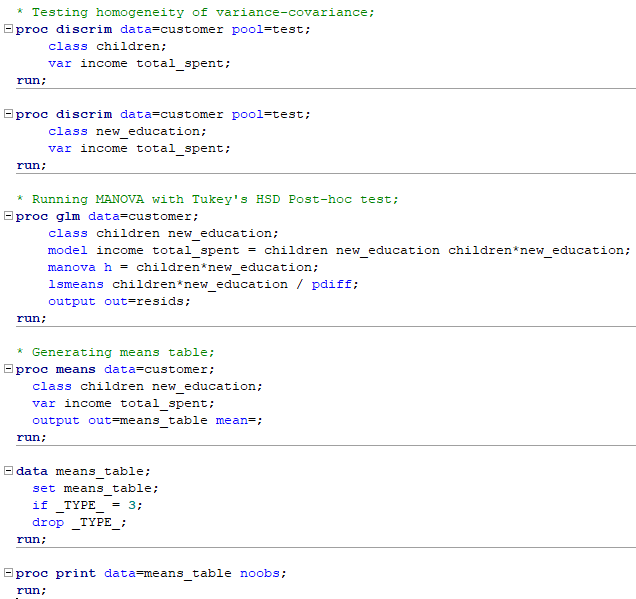
****

****

**Appendix B: Code**

[Python Code](https://drive.google.com/drive/folders/1x5jJMkjv-4p9-o1SV-b4Q7X2p0MPgla7?usp=share_link) can be found here in ipynb and pdf format along with the data sets.

****

****

****