Otilio Alvarado-Cruz

Data Analytics Graduate Capstone

D214 Task 2

**A.**

Utilizing the customer and sales data from 10 malls in California, can ANOVA be utilized to identify the mean difference in spending of customer groups found using K-Means clustering? The analysis aims to identify if customer characteristics play a role in how much they spend. The sales dataset contains transaction details from shopping malls in California, while the customer dataset contains customer characteristics. The null hypothesis is that there is no statistically significant difference between the groups, while the alternative hypothesis is that there is a statistically significant difference between the groups.

**B.**

The customer dataset was obtained by downloading it from Kaggle as two separate XLSX files containing the customer and sales data, which are both joined together through the customer\_id column. The advantage of getting the dataset from Kaggle is that it came prepared specifically for analyzing customer spending behaviors based on their characteristics across different malls. The downside of the dataset is that it contains columns that aren’t relevant to customer characteristics and had some null values. To overcome this, those columns were removed and any rows containing null values were removed.

**C.**



The customer and sales data are extracted from the excel files and merged together through customer\_id using pandas. Pandas is great for the data cleaning process as it “offers a rich set of functions for data transformation and manipulation (Patil, 2023).” The downside to using pandas is that it "can be memory-intensive, especially when working with large datasets (Patil, 2023).”





To identify how many rows and columns the merged dataset has, df.shape is used and it is found to have 99457 rows and 10 columns.





Duplicated.value\_counts() is used to identify if there are any duplicate rows. If a duplicate exists, it will be counted as true. The output gives 99457 false, which matches the number of rows in the dataframe, indicating that duplicates rows aren’t present in the dataset.



A screen shot of a computer

Description automatically generated

To check for nulls in the dataset, isnull().sum() is run, which will print out the total null values present in each column in the dataset. The output shows that there are 119 null values in the age column.

A black background with white text

Description automatically generated

A screen shot of a computer code

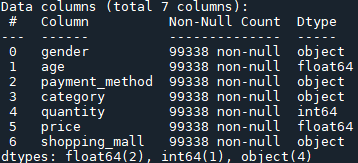
Description automatically generated



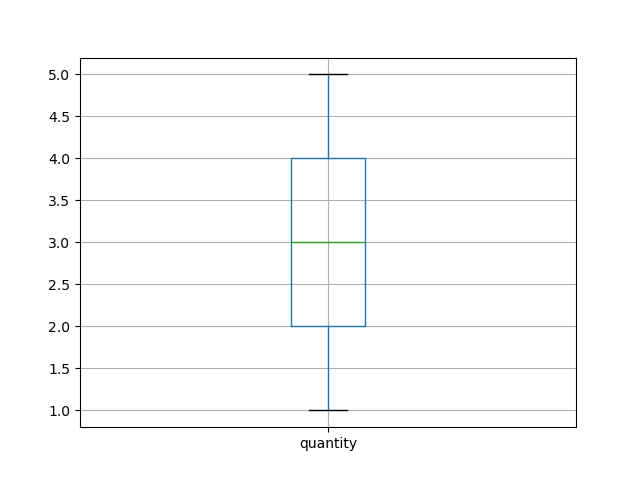
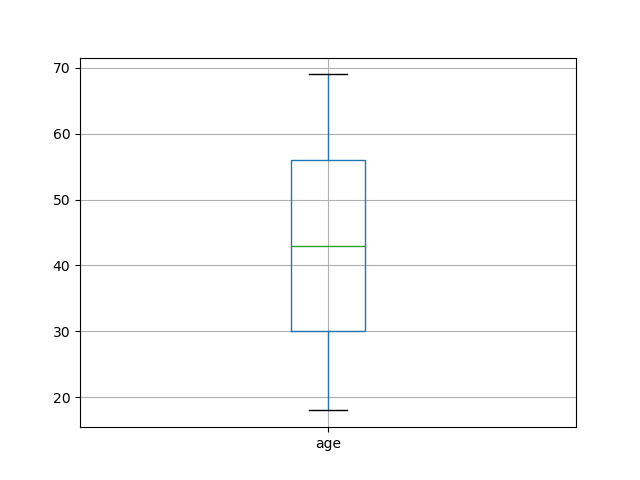
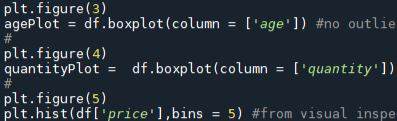
To clean the nulls from the dataset, dropna is used to remove the 119 rows that contain a null value. To verify that the nulls are removed, isnull().sum() is run, which shows that all columns now contain 0 null values, and .shape is used to verify that the 119 rows are removed, and there are now 99338 rows, which verifies that the rows where removed.

A computer screen shot of a code

Description automatically generated



The customer\_id, invoice\_no, and invoice date columns are dropped since these columns function as transaction identifiers. Df.info is used to verify that these columns have been removed from the dataframe.



To identify if outliers are present in the dataset, a boxplot is made for the price, quantity, and age columns. Any value that is outside the whiskers is considered to be an outlier. From the boxplots that were produced, the age and quantity columns don’t contain outliers, but the price column contains outlier values.



A graph with lines and a rectangle

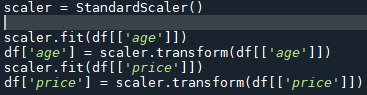
Description automatically generated

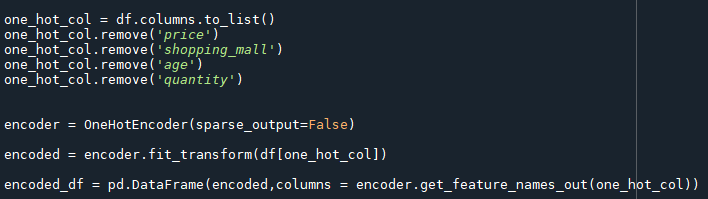
To treat the outliers in the price column, any values above the whisker are imputed with the median price as the “mean value is highly influenced by the outlier treatment, it is advised to replace the outliers with the median value. (Harika, 2024)”. After examining the boxplot post-imputation, the data is now much closer together.

A blue and white graph

Description automatically generated

To identify the normality of the dataset, a histogram is made of the price column. The result graph does not have a bell shape curve, indicating that the data does not have a normal distribution.



The code above is used to standardize the age and price columns into their z-scores. The purpose of standardizing these columns is to “to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance (Firmin, 2024)”.



A screenshot of a computer

Description automatically generated

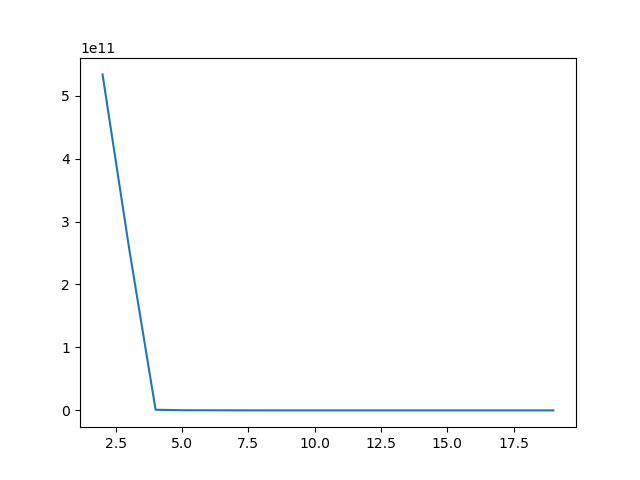
K-means clustering requires that all variables be converted to numerical data. Since there are customer characteristics that aren’t numerical, they are made into numerical data through one-hot encoding, except for the cities column, where frequency encoding is used instead. The advantage to using one hot encoding is that it “preserves the uniqueness of each category. It ensures that the algorithm does not assume any ordinal relationship among the categories (Micheal, 2023)”, but the downside to one-hot encoding is that it “can significantly increase the dimensionality of the dataset, especially when dealing with categorical variables with many unique categories. This can lead to the curse of dimensionality and negatively impact model performance. (Micheal, 2023)”.

It is due to this that frequency encoding is utilized to convert the mall names to numerical data, as there are 10 different mall names, which would lead to a significant increase in dimensionality. Frequency encoding converts each category in a column to the frequency in which that category appears in the column. The advantage of frequency coding over one-hot encoding is that “frequency encoding does not increase the dimensionality of the dataset (Ninja, 2024).” The downside of frequency encoding is that if “two categories have the same frequency in the dataset, they will be represented by the same number after frequency encoding. This can lead to a loss of valuable information, as the model might treat these two distinct categories as the same (Ninja, 2024).” In order to prevent this, the value count of each mall in the shopping\_mall column is printed, and it’s found that no mall has the same value count, which ensures that each mall will have a unique value. To perform frequency encoding, the value count of each mall is mapped to its respective mall.

**D.**

A computer screen shot of a number

Description automatically generated



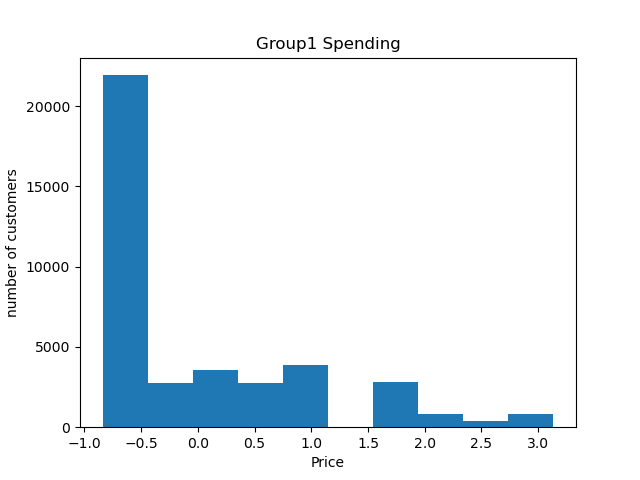
To identify the optimal number of clusters for the model, the elbow method is utilized in which a for loop is run which fits the data into a KMeans model for values of K from 2 to 20. Then the distortion of the model is appended to the distortion array, and then a line plot is made showing the relationship between the number of clusters and the distortion for each value of K. From the line plot shown above, the optimal value of K is 4 as there is a very clearly defined elbow at K=4. The elbow method is an easy and effective way to identify the best value of K for the model, but the main drawback of utilizing the elbow method is that “we do not always have such clearly clustered data. This means that the elbow may not be clear and sharp (Clustering)” which can lead to ambiguity as to which value of K is optimal for the model.



The dataset is then fit into a KMeans model with k set to 4.

A screen shot of a computer code

Description automatically generated

A graph with blue bars

Description automatically generated

A graph with blue bars

Description automatically generated

A graph with blue squares

Description automatically generated

Each cluster is isolated and a histogram is made of each group showing how much customers spent in each group.





To analyze the quality of the clusters that were made, the silhouette score is used. The advantage of using the silhouette score for analysis is that “the score is bounded between -1 for incorrect clustering and +1 for highly dense clustering (Clustering)”, which makes its result easy to interpret. The downside to silhouette score is that it “is generally higher for convex clusters than other concepts of clusters (Clustering).”

A screenshot of a computer

Description automatically generated

The cluster centroids of each group are printed out so that the differences between each variable in each group can be inspected.



Then ANOVA is performed on the price of each of the groups to obtain the p-value. The advantage of utilizing ANOVA on groups found with k-means clustering is that it “can help you know whether or not there are significant differences between the means of your independent variable (Carpenter, 2024).” The downside to utilizing ANOVA is that it “relies on three main assumptions that must be met for the test results to be valid (Carpenter, 2024),” which are normality, homogeneity, and independence, and not meeting these assumptions can affect the accuracy of ANOVA.

**E.**

After separating the customers into 4 distinct groups, a histogram is made on each group showing how much customers spend, and from a visual look, there is very little variation in how much customers spend in each group. The silhouette score of the model is 0.98, which is very close to 1, indicating that the clusters are well separated from each other. When ANOVA is run on the price variable of these clusters, the p-value was found to be 0.99, which is greater than 0.05, indicating that there isn’t a statistically significant difference between the mean value of price between the groups. The cluster centroids of each group are printed out to inspect what makes each group different, and it’s found that the groups are mainly different from the shopping mall column, which represents where the customers made their purchases in. This indicates that across the 10 malls in this dataset, customer spending habits are about the same. A recommended course of action would be to obtain data from different regions in order to obtain a more holistic view on customer spending habits.

One way to do future study on the dataset would be to combine it with other datasets containing customer spending habits from their regions so that customer characteristics of those regions can be represented in the data. An alternative way to do future study on the dataset would be to take one group that was found and create new groups of customers from that group and see if there are statistically significant differences in the spending habits between customers from those groups and identify what kinds of characteristics are present in customers of those groups.

**F.**

Carpenter, Aaron. “What Is ANOVA (Analysis of Variance) Testing?” *Qualtrics*, 26 Sept. 2024, www.qualtrics.com/experience-management/research/anova/#:~:text=Like%20other%20types%20of%20statistical,Why%20is%20that%20useful?

“Clustering.” *Scikit*, scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient. Accessed 14 Dec. 2024.

“Clustering Algorithms: K-Means.” *AI Planet (Formerly DPhi)*, aiplanet.com/learn/unsupervised-learning/clustering-analysis-and-techniques/927/clustering-algorithms-k-means. Accessed 13 Dec. 2024.

Firmin, Sydney. “Standardization in Cluster Analysis.” *Knowledge Center*, 6 Dec. 2024, knowledge.alteryx.com/index/s/article/Standardization-in-Cluster-Analysis-1583461087248#:~:text=Standardization%20prevents%20variables%20with%20larger,the%20algorithm%20with%20equal%20importance.

*Harika. “Detecting and Treating Outliers: Treating the Odd One Out!” Analytics Vidhya, 18 Oct. 2024,* [*http://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/*](http://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/)*. Accessed 18 Nov. 2024*

*Micheal, Oyebamiji. “A Comprehensive Comparison between One-Hot and Ordinal Encoding.” Medium, Medium, 27 Aug. 2023, medium.com/@oyebamijimicheal10/a-comprehensive-comparison-between-one-hot-and-ordinal-encoding-6f899c4f08b3#:~:text=Lack%20of%20Bias:%20Since%20each,second%20or%20a%20third%20class.*

*Ninja, Neural. “Frequency Encoding: Counting Categories for Representation.” Let’s Data Science, 1 Apr. 2024, letsdatascience.com/frequency-encoding/.*

Patil, Rohan. “Exploring the Pros and Cons of Pandas.” *Medium*, Medium, 19 June 2023, rohanjpatil63.medium.com/exploring-the-pros-and-cons-of-pandas-1675463971d4.