ECON 607 Assignment 5

Andrew Girgis

October 2023

Contents

List of Figures

1 SKLearn K-means

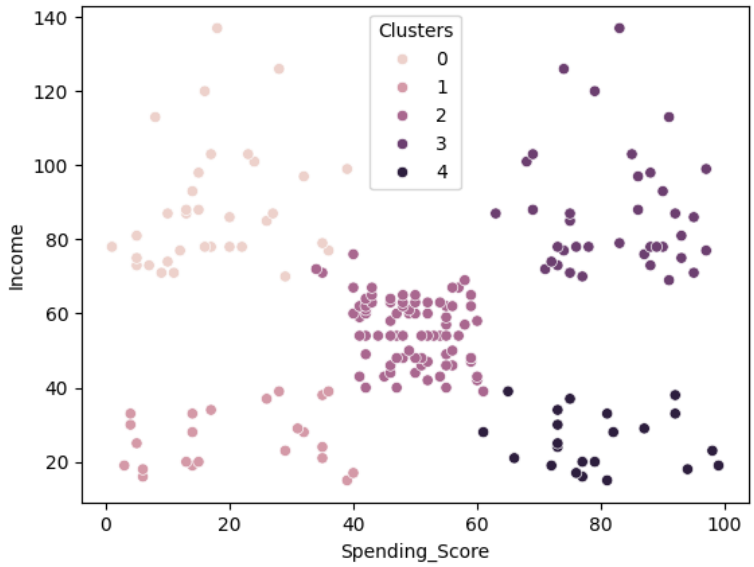
1.1 Mall Customer Data set

The mall customer segmentation data set is created for the learning purpose of the customer segmentation concepts, also known as market basket analysis. This analysis holds importance for owners of supermarkets or grocery stores where, through membership cards, they obtain basic data about their customers. This data set is composed by the following five features:

- **CustomerID**: Unique ID assigned to the customer
- **Gender**: Gender of the customer
- **Age**: Age of the customer
- **Annual Income (k$)**: Annual Income of the customer
- **Spending Score (1-100)**: Score assigned by the mall based on customer behavior and spending nature.

1.2 Cluster scatterplot of Spending Score and Income

Figure 1: K-means clusters scatterplot on Spending Score and Income



1.3 K-means clusters labels

Looking at figure 1 above we can observe 5 clusters of customers. The x-axis represents the customers spending score, the spending score is a score assigned by the mall based on customer behavior and spending nature. The y-axis is the customers annual income in thousands of dollars. We can see 5 separate clusters. Currently they are labeled by number where:

- **Label 0**: Customers who have a low spending score and a high income. We will call these customers *holders*. I am calling these customers holders because these customers have a lot of income however they are frugal in their behavior and spending nature.
- **Label 1**: Customers who have a low spending score and a low income. We will call these customers *rationals*. I am calling these customers rationals because these customers have low income and they are frugal in their behavior and spending nature, which is rational.
- **Label 2**: Customers who have a median spending score and a low to median income. We will call these customers *medians*. I am calling these customers medians because these customers have a low to median income and they have a median spending score which tells they are about average in their behavior and spending nature.
- **Label 3**: Customers who have a high spending score and a low income. We will call these customers *irrationals*. I am calling these customers irrationals because these customers have low income and they are lavish in their behavior and spending nature, which is irrational.
- **Label 4**: Customers who have a high spending score and a high income. We will call these customers *spenders*. I am calling these customers spenders because these customers have high income and they are lavish in their behavior and spending nature.

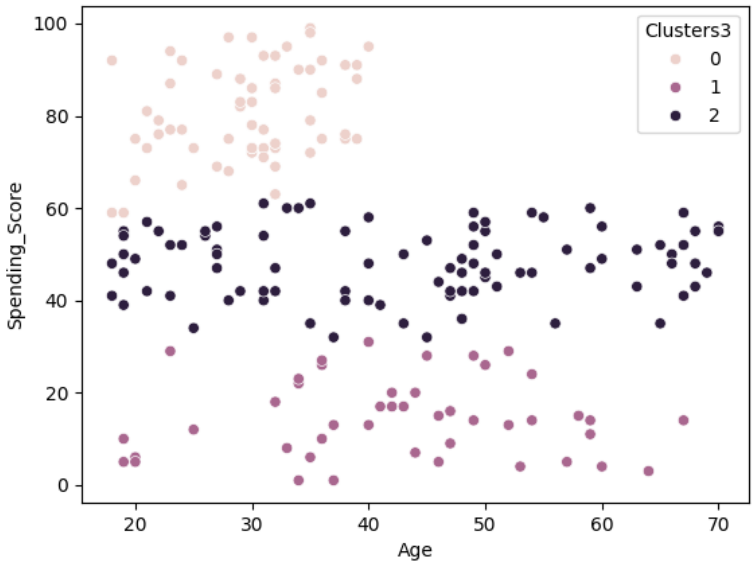
1.4 Insights of Spending Score and Income Clusters

The reason we use k-means clustering is to use machine learning to identify underlying patterns in data. By clustering we are creating groups in our data were each point or individual in a specific group are (in some way)similar and the points or individuals in separate groups are (in some way)different. In figure 1 we create a scatterplot of of Spending Score and Income and differentiate the clusters visually through the use of colours which are defined in the legend. We outline the labels of these clusters in section 1.3. After using machine learning to create the clusters the challenge still remains of how the clustering can be utilized to make informed business decisions. Using the value counts function we find that there are 81 *medians*, 39 *spenders*, 35 *holders*, 23 *rationals*, and 22 *irrationals*. Based on this we can see that a clear majority of individuals are in the *median* group and differences in individuals between the other four groups

aren't as large. This result always us to determine our market segmentation, since a large majority have a median spending score and a low to median income and the rest of the individuals are fairly evenly distributed between the other four groups it is beneficial for the supermarket business to cater to a large audience with a wide range of products and services. This is because as a business, we want to maximize the spending score for all individuals especially those individuals with a high income.

1.5 Clusters scatterplot of Age and Spending score

Figure 2: K-means clusters scatterplot on Age and Spending Score



1.6 Insights of Age and Spending Score Clusters

The first thing I noticed about the clusters scatterplot on age and spending score is the lack of individuals in the high age and high spending score range. Based on this we can see that the older demographic of customers have a lower spending score than the young to middle aged demographic. We know that the spending score captures the customer behavior and spending nature of an individual. According to Radu Valentin the article *Consumer behavior in marketing – patterns, types, segmentation* customer behaviour is defined as the study of how people make purchase decisions to satisfy their needs, wants, or desires

and how their emotional, mental, and behavioral responses influence the buying decision (Radu, 2023). Since we know that customer behaviour includes ideas explaining a consumers susceptibility to marketing, brand loyalty, trends, and more. With this understanding, we can infer that the older generation is less likely to be susceptible to customer behaviour specific marketing and are likely to spend less (since the spending score captures both variables). Based on this inference the supermarket should cater its marketing for the age group of 18-40 year olds. Another interesting visual we see in the graph is that the clusters are divided horizontally based on the spending scores. We see group 1 are individuals with a spending score below 40 of all ages, group 2 are individuals with a spending score between 35 and 60 of all ages and group 0 are individuals with a spending score of 60 and greater, and as mentioned above, contain only ages 18 to 40. Based on these groups being clustered in this manner we can infer that the supermarket should cater to the needs of the separate groups of spending scores, this can be done through methods like including big brand name items and alternatives that include non brand name products with a lower price tag. By catering to the different spending scores in this manner we can ensure that individuals in all groups in our clustering are satisfied and have a positive shopping experience increasing the likelihood they return.

References

Radu, V. (2023, October). Consumer behavior in marketing - patterns, types, segmentation - omniconvert blog. https://www.omniconvert.com/blog/consumer-behavior-in-marketing-patterns-types-segmentation/

2 Code

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import sklearn.cluster as cluster
import warnings

warnings.filterwarnings('ignore')

# Load the dataset
df = pd.read_csv('/Users/andrew/Downloads/UW courses/ECON 607/
Assignment 5/Mall_Customers.csv')

# Display the first few rows of the dataset
print(df.head())

# Rename the columns for easier access
df.rename(columns={'Annual Income (k$)': 'Income',
'Spending Score (1-100)': 'Spending_Score'}, inplace=True)
print(df.head())

# Display descriptive statistics of the dataset
print(df.describe())

# Visualize the pairwise relationships in the dataset
sns.pairplot(df[['Age', 'Income', 'Spending_Score']])
plt.show()

# Perform K-means clustering with 5 clusters using
'Spending_Score' and 'Income' variables
kmeans = cluster.KMeans(n_clusters=5, init="k-means++")
kmeans = kmeans.fit(df[['Spending_Score', 'Income']])
kmeans.cluster_centers_

# Assign the cluster labels to the dataframe
df['Clusters'] = kmeans.labels_

# Display the first few rows of the dataframe with cluster labels
print(df.head())

# Save the dataframe with cluster labels to a CSV file
#df.to_csv('mallClusters.csv', index=False)

# Visualize the clusters using a scatterplot
```

```
sns.scatterplot(x="Spending_Score", y="Income",
hue='Clusters', data=df)
plt.show()


# Perform K-means clustering with 3 clusters using
'Age' and 'Spending_Score' variables
kmeans3 = cluster.KMeans(n_clusters=3, init="k-means++")
kmeans3 = kmeans3.fit(df[['Age', 'Spending_Score']])
kmeans3.cluster_centers_

# Assign the cluster labels to the dataframe
df['Clusters3'] = kmeans3.labels_

# Display the first few rows of the dataframe with cluster labels
print(df.head())

# Save the dataframe with cluster labels to a CSV file
df.to_csv('mallClusters_with_3_clusters.csv', index=False)

# Visualize the clusters using a scatterplot
sns.scatterplot(x="Age", y="Spending_Score", hue='Clusters3',
data=df)
plt.show()
```