Week 2 Assignment: K-Means Clustering

Candidate: Sneha Santha Prabakar

Coding platform: Jupyter Notebook

Please Note:

The python notebook included in the submission contains all the charts and steps mentioned in the grading criteria. This PDF document only provides a summary for each section executed in the python notebook. Hence, please refer to the python notebook for the execution steps.

Part 1: Data summary

The dataset consists of 200 rows, 6 columns.

From the preliminary analysis, we can see make the following observations:

1. There is **no missing data** in any column.
2. **Age** of the customers is in the range: 18 - 69. This shows that the data includes customers from young adults to senior citizens. This allows for age-related behavioral segmentation. The median (43.5) is almost the same as the mean (43.4), which shows that there is a uniform distribution of all age groups in the dataset.
3. The **annual income** of the customer spans a wide range (15k - 119k). This may influence the customers purchasing power. The median (66.5) is close to the mean (67.1), so its not heavily skewed distribution.
4. **Spending score** is also widely distributed (1-100), suggesting the presence of both frugal and generous spenders. The Median (52.35) is close to the mean (50.8), so its not heavily skewed distribution.
5. **Purchase Frequency** also has a wide range (1-10) but uniformly distributed, since the median (5.45) is close to the mean (5.33). It suggests that some customers shop frequently whereas other shop rarely.
6. The **Avg Purchase Value** ranges from 10-137, with a slightly right-skewed distribution as the median (36) is lower than the mean (43). The 75% percentile is 59.6, while the max is 137.6 - this may suggest that while most customers spend modestly, there are some outliers which push the mean higher.

Part 2: Data Pre-Processing

Exploratory Data Analysis

We can make the following observations about the individual distribution of the variables and the relationship between the variables:

- Individual distribution:
  - Age and Annual income: Roughly uniform distribution.
  - Spending score: Faily normal distribution since it is bell-shaped.
  - Purchase frequency: Looks relatively even, possibly slightly skewed towards more frequent shoppers.

    - Avg purchase value: Right-skewed which suggests that most customers make smaller-value purchases. The high-spenders are a minority (some may even be outliers).

- Relationship between the variables:
    - Most scatterplots show no strong correlation between the variables.
    - The only variables that show a *slightly positive* correlation are:
        - Spending Score vs Annual Income
        - Spending Score vs Purchase Frequency
        - Avg Purchase Value vs Annual Income

Heatmap for Correlation Analysis

This heatmap confirms that most of the variables not linearly correlated, except Annual income and Spending score. The correlation coefficient between Annual Income and Spending score is 0.6 which is greater than 0.5, which suggests a slightly positive correlation between the two (which was also observed on the Scatterplots above). However, it is still not strong enough to indicate a positive dependency as it is lesser than 0.75.

Hence, we can conclude that the variables are not linearly correlated to each other.

Handling Missing Values and Outliers

There are no missing values in the data.

After plotting the box plot for all variables, we notice that Outliers are present only Avg Purchase value:

- Outlier 1 (CustomerID 48): 137.6 is only approximately 1.1x more than the upper bound
- Outlier 2 (CustomerID 133): 123.4 is very close to the upper bound

This tells us that:
1. These outlier values are valid, since they are not largely off the charts
2. Since they are not too far from the upper bound (for the reasons stated above), they most probably won't influence the creation of K means clusters much.

CustomerID 48 has the highest Avg Purchase Value, though his purchase frequency is relatively low (2.6 on a scale of 10), his Annual Income is in the third quartile, and his spending score is also relatively low (23.6 on a scale of 100).

CustomerID 133 shows a relatively high annual income, spending score, purchase frequency and close to normal Avg Purchase value (i.e. close to the upper bound).

We see that removing CustomerID 48 does not really have a major impact on the mean/median/std deviation of the Avg Purchase Value column. Since CustomerID 133 has an even lower Avg Purchase Value than CustomerID 48, it is unlikely that removing CustomerID 133 is also going to have any significant impact on the column. Hence, we will simply retain these two rows seeing that they do not siginficantly impact the distribution of Avg Purchase Value.

Scaling the features

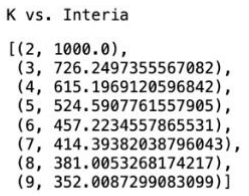
We need to standardize the features to make sure the clustering algorithm is not biased (since it is a distance-based algorithm). We choose standardization over normalization because:

- our variables differ in range
- most of them are normally/uniformly distributed
- we have two outliers

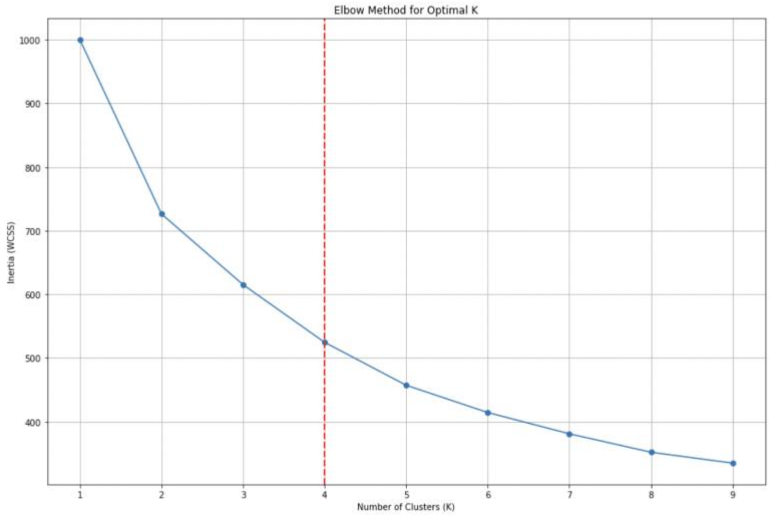
The transformed data has all features with mean approx. 0 and standard deviation approx. 1.

Part 3: Determining the optimal number of clusters

Elbow Test

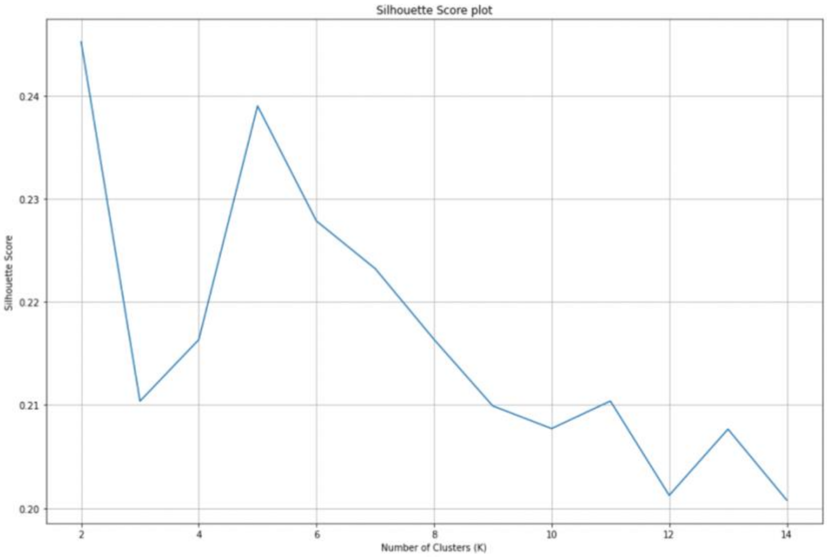


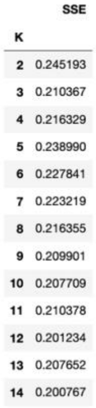
```
K vs. Interia

[(2, 1000.0),
 (3, 726.2497355567082),
 (4, 615.1969120596842),
 (5, 524.5907761557905),
 (6, 457.2234557865531),
 (7, 414.39382038796043),
 (8, 381.0053268174217),
 (9, 352.0087299083099)]
```



The Elbow Test gives an optimal value of K = 4 as the rate of drop of Inertia decreases after that.

Silhouette Score





	SSE
K	
2	0.245193
3	0.210367
4	0.216329
5	0.238990
6	0.227841
7	0.223219
8	0.216355
9	0.209901
10	0.207709
11	0.210378
12	0.201234
13	0.207652
14	0.200767

The Silhouette Score peak is at K=5, hence the optimal value of K (by Silhouette method) is 5.

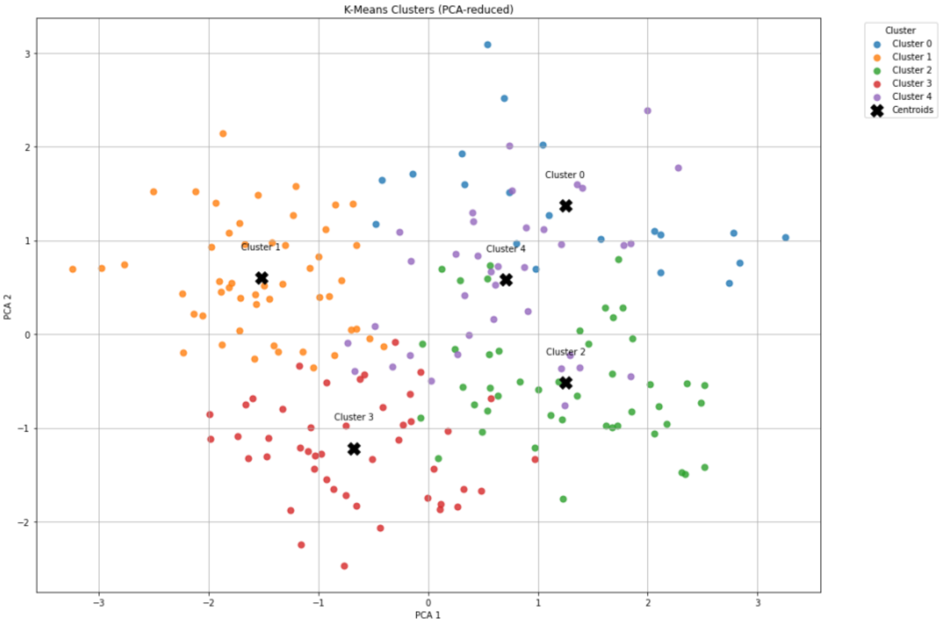
We see that the optimal K determined by both the methods are very close. The Elbow method focuses on the compactness within each cluster whereas the Silhouette score focuses on the seperation and cohesion of the clusters.

We can choose K = 5, since the clusters are better seperated at this point (as suggested by Silhouette score) and it still has good elbow.

Part 4: K-means cluster implementation

PCA Plot

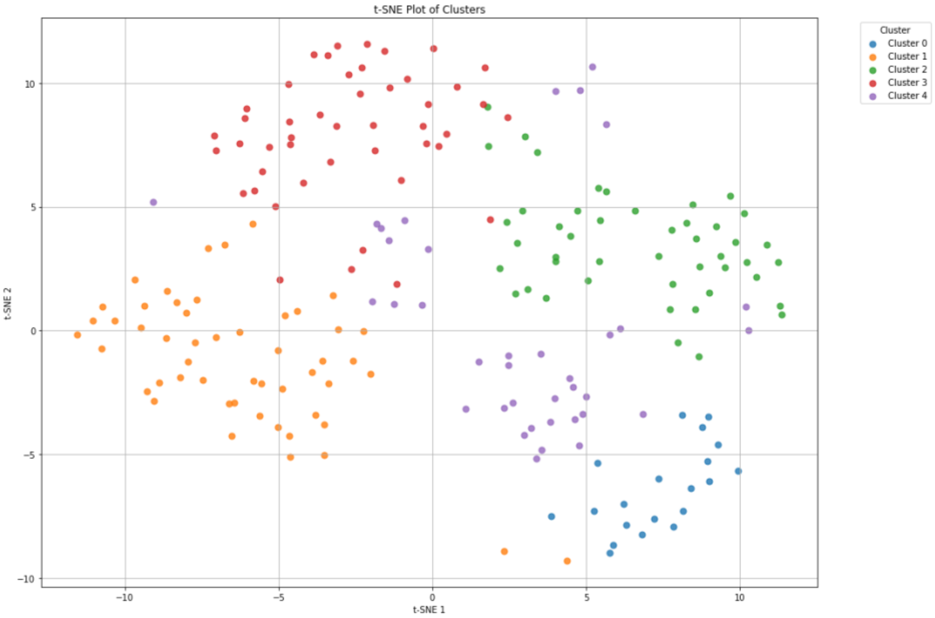
The plot below showcases PCA (2 dimensions) of the clustered data points, with the centroid of each cluster marked by “X”:



In this 2D PCA space, we can see that:

- The clusters are mostly well-seperated with some overlaps.
- The centroids (marked as red X on the plot) are visibly distinct and not tightly packed - this implies that K-means algorithm found meaningful partitions in the data.

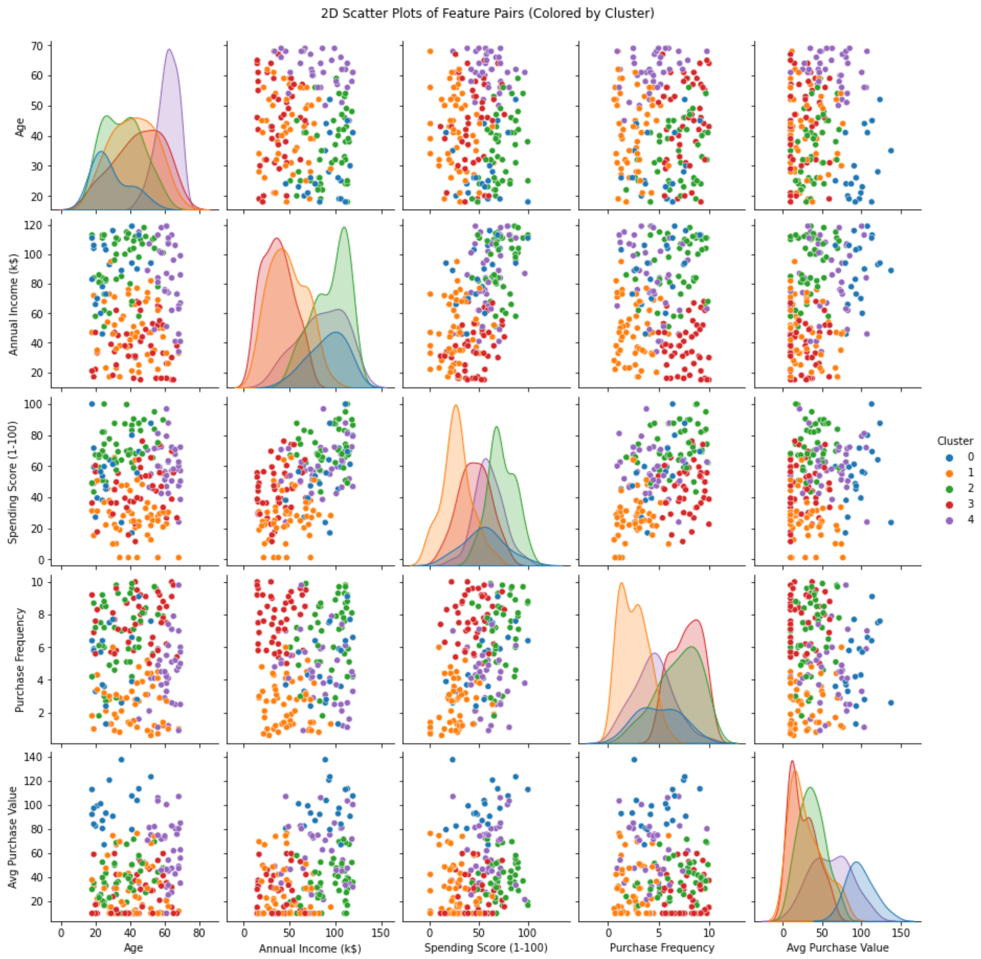
t-SNE Plot



- The clusters are very well-separated with sharper boundaries in the t-SNE plot than in the PCA plot.

- There is still some overlap between the clusters, but the cluster cohesion is visually stronger here.
- Overall, both PCA and t-SNE plots confirm that the K-Means algorithm successfully grouped the customers into 5 distinct clusters.

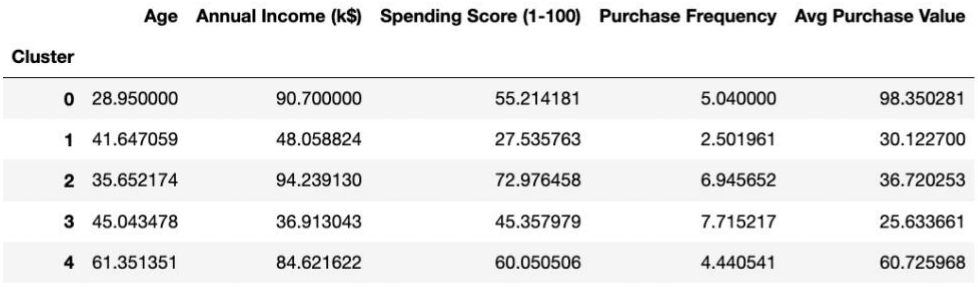
2D plot of Feature combinations across clusters



The pairwise features that show better seperation between the clusters are:

- Annual Income vs Spending Score
- Purchase Frequency vs Avg Purchase Value
- Spending Score vs Purchase Frequency

We can also look at how much the average value of each feature differs across clusters. This helps us identify which features contribute the most to distinguishing one customer group from another.

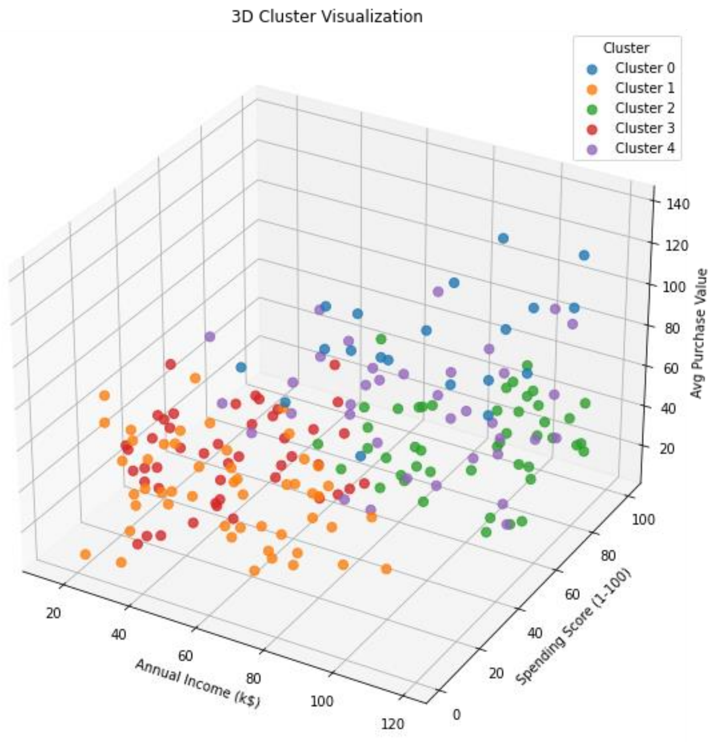


	Age	Annual Income (k$)	Spending Score (1-100)	Purchase Frequency	Avg Purchase Value
Cluster					
0	28.950000	90.700000	55.214181	5.040000	98.350281
1	41.647059	48.058824	27.535763	2.501961	30.122700
2	35.652174	94.239130	72.976458	6.945652	36.720253
3	45.043478	36.913043	45.357979	7.715217	25.633661
4	61.351351	84.621622	60.050506	4.440541	60.725968

Based on the scatterplot visualization and the means table, we can conclude that the following features are the top 3 key variables for customer segmentation:

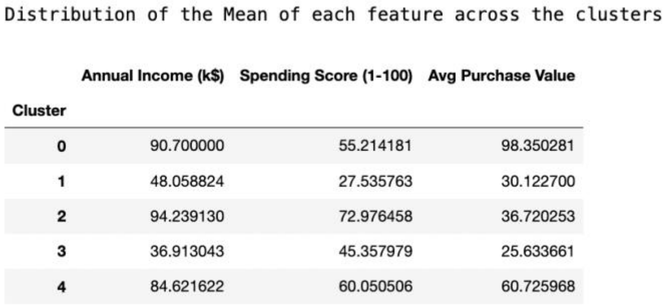
- Annual Income:
    - Shows strong variation across the clusters and distinguishes them based on the income level.
    - Going by the mean of Annual income in each cluster, we can say that:
        - Cluster 0 and 2: High-income group (Cluster 2 is the highest income group with mean=94.2)
        - Cluster 4: Medium-income group
        - Cluster 1 and 3: Low-income group (Cluster 3 is the lowest income group with mean=36.9)
- Spending Score:
    - Helps differentiate engagement levels.
    - Going by the mean of Spending score in each cluster, we can say that:
        - Cluster 2 and 4: High-spending group (Cluster 2 has the highest spending with mean=72.9)
        - Cluster 0 and 3: Medium-spending group
        - Cluster 1: Low-spending group (mean=27.5)
- Avg Purchase Value:
    - Differentiates between premium and budget-concious customers.
    - Going by the mean of the Avg Purchase Value, we can say that:
        - Cluster 0: High-spending per transaction (mean=98.35)
        - Cluster 4: Medium-spending per transaction (mean=60.7)
        - Cluster 1,2 and 3: Low-spending (budget) per transaction (Cluster 3 having the most budget-concious customers, with mean=25.6)

3D plot of Key features



The clusters seem to be more distinct in high-value and low-value groups, while there are some overlaps in the mid-value groups.

Part 5: Customer segmentation



```
Distribution of the Mean of each feature across the clusters
```

Cluster	Annual Income (k$)	Spending Score (1-100)	Avg Purchase Value
0	90.700000	55.214181	98.350281
1	48.058824	27.535763	30.122700
2	94.239130	72.976458	36.720253
3	36.913043	45.357979	25.633661
4	84.621622	60.050506	60.725968

We can make the following observations about the customer segmentation:

- Cluster 0: High-income, high-spending per transaction. Likely loyal, premium buyers.
- Cluster 1: Medium-income, low spenders on all fronts. Likely price-sensitive or low engagement customers.
- Cluster 2: Very wealthy & active spenders. Strong candidates for elite targeting.
- Cluster 3: Lower income, moderate behaviour. Baseline customers.
- Cluster 4: Upper-mid income, moderately engaged. Responsive to smart promotions.