Problem Definition:

This dataset is from a mall, and they want to understand their customers better. Their goal is to determine what kind of shops to bring into the mall that would cater to all clusters. This would in turn result in the increased customer lifetime value of all customers. To do this, they would like to determine the hidden segments of their customers. Identification of the underlying segments would result in understanding their customer's behavior. This will help the mall understand by developing labels for their different types of customers and would in turn build the blocks for actionable intelligence which would help with achieving their goal. This would also help with developing and implementing marketing strategies to determine tailor-made marketing techniques to increase customer lifetime value of all clusters as they don't want to lose customers, they want to increase the number of customers and retain existing customers as well as defining their target market. They would like to segment their customers based on the features they proposed in the dataset. Building a model that answers all these questions would help the company bring in new shops of the right size or market existing shops better and increase the size of the shops to accommodate more products and customers.

Model Formulation:

We have no predefined thresholds to determine our segments; therefore, we use unsupervised learning techniques to determine the underlying patterns of our dataset. We use the KMeans algorithm. KMeans algorithm is used for clustering. It groups data points based on similarity, this means that data points that are similar are put into one cluster. It also helps by ensuring that different clusters are separated from each other to show uniqueness. This means that customers from different clusters have different values and perceptions and would need to be targeted differently. KMeans clustering method works only with quantitative data. It measures the Euclidean distance between data points to find the best centroids. Centroids are centers of the cluster, and they represent the average position of all data points within a cluster. This method will help develop marketing strategies catered for the different clusters. Clusters will be developed based on the features provided in the table below:

Feature/Variable	**Explanation**
Age	The age of the customer
Annual Income	How much the customer earns in thousands
Spending Score(k$)	The score ranging from 0-100. 0 meaning they don't spend at all and 100 meaning they spend a lot
Estimated Savings(k$)	The estimated savings the customer has
Credit Score	Refers to how well you are at paying back money that you borrowed from Financial Service Providers
Loyalty Years	An estimation of the customer and mall relationship length
Gender	The gender of the customer
Preferred Category	The category of products the customer tends to buy when they go to the mall

We can see that there is a mix of demographic, financial and behavioral fields. KMeans is good because it will group customers who have similar demographic, financial and behavioral aspects into one cluster. Once the clusters are formed, we can then identify and interpret the unique characteristics. After this we would have a certain number of clusters. These clusters are now our customer segments. Segmentation is beneficial because we now have a deeper understanding of our customers. This understanding will result in tailored marketing approaches for our different segments which will result in attraction of new customers and retaining existing customers and increasing customer satisfaction.

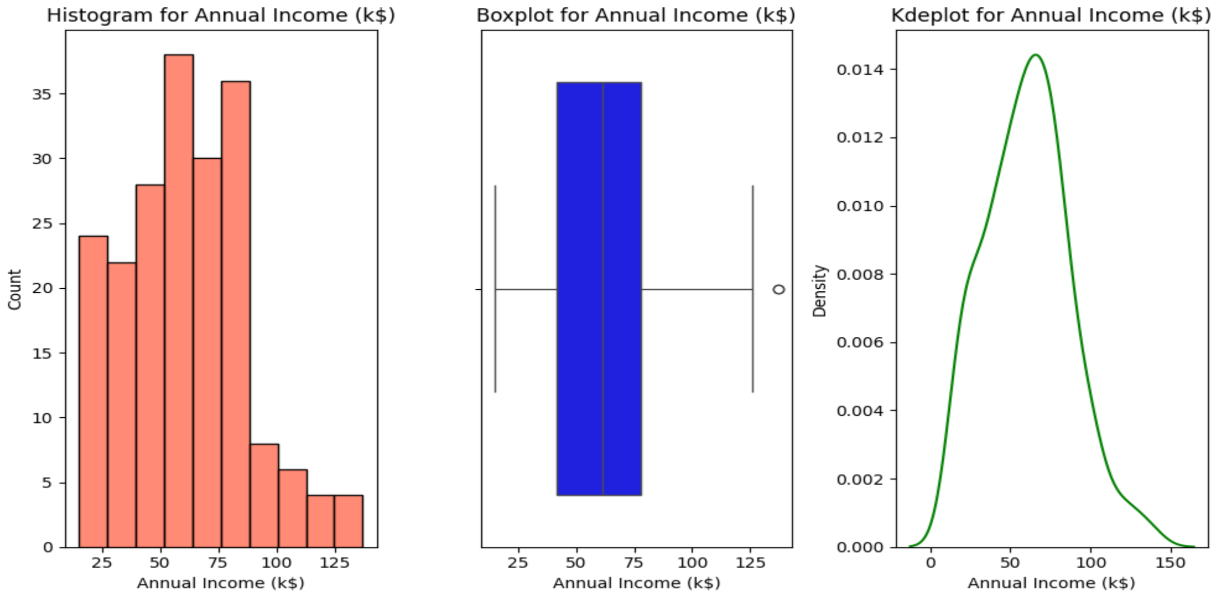
Implementation Summary

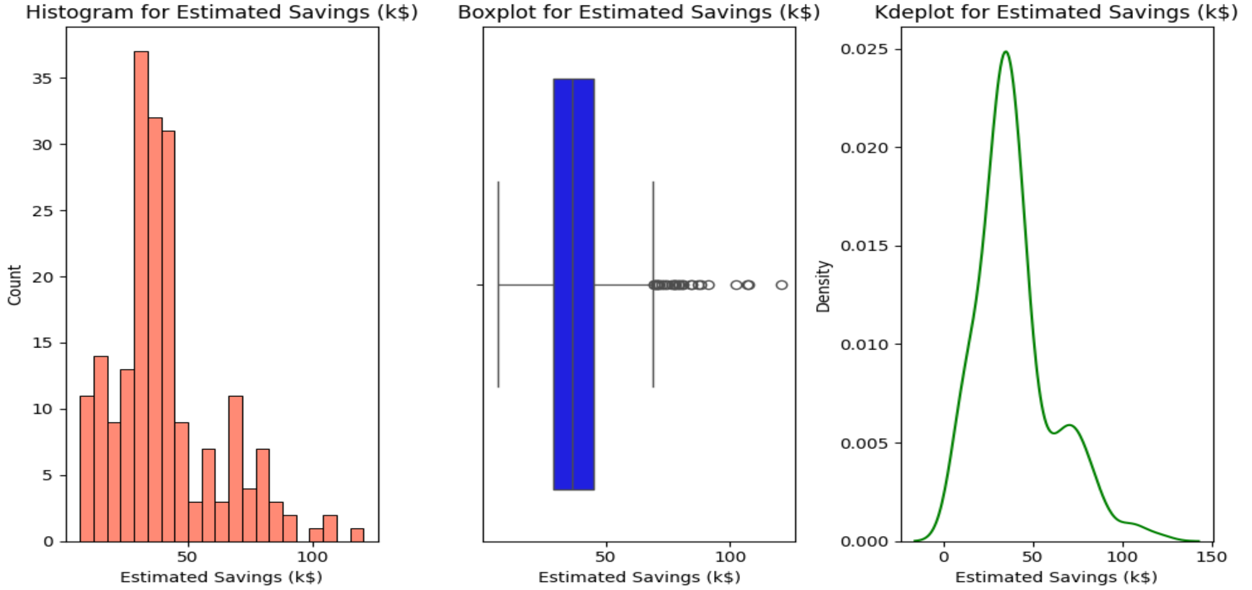
- Data Profiling:

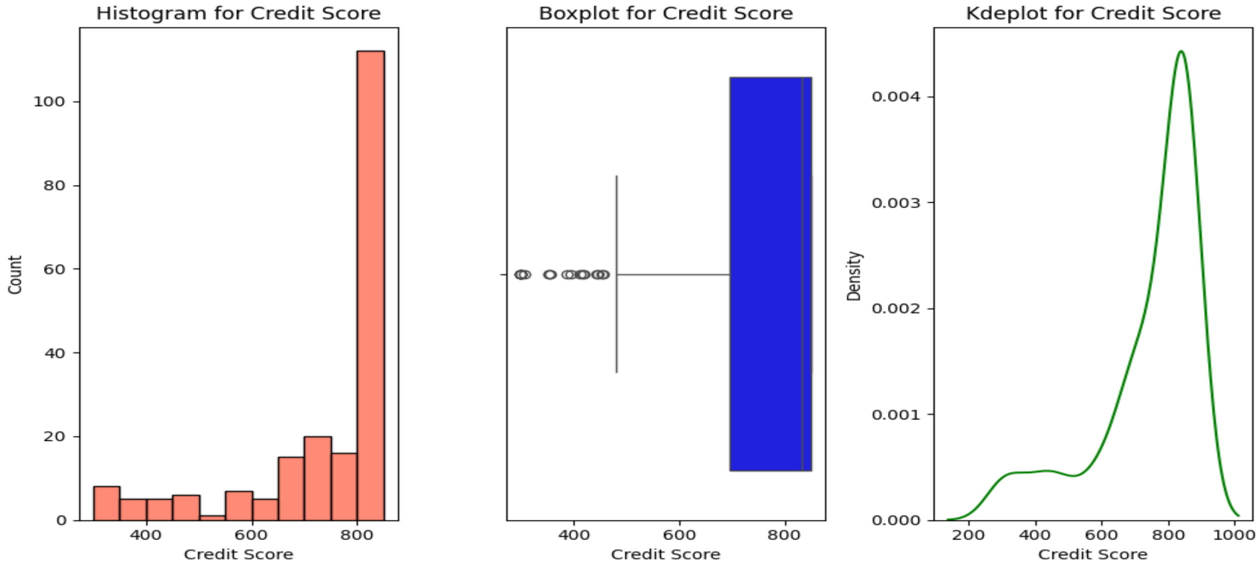
The dataset has 200 rows and 10 columns. There is 4 missing rows in the Age Group column. Since all rows are for age 18, I filled in the missing rows with the correct age group. There are 7 numerical columns and 3 categorical columns. I then decided to drop the CustomerID and Age Group columns for dimension reduction.

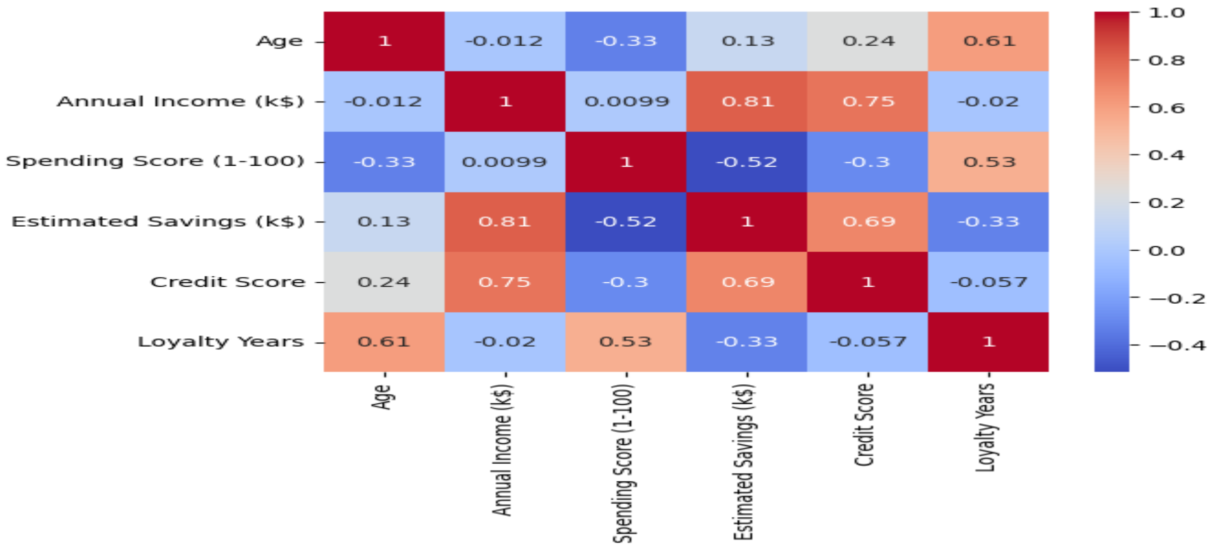
- Exploratory Data Analysis

Numerical Columns are Age, Annual Income, Spending Score, Estimated Savings, Credit Score, Loyalty Years. I built some visualizations that can help with understanding distributions, determining minimum, median and max values as well as detecting outliers.I also plotted a correlation matrix that will help with making sense of clusters. I then identified that there’s outliers in 3 of the numerical columns using the boxplot. I removed these outliers by using the IQR method because KMeans is sensitive to outliers. Below you’ll see the columns with outliers.

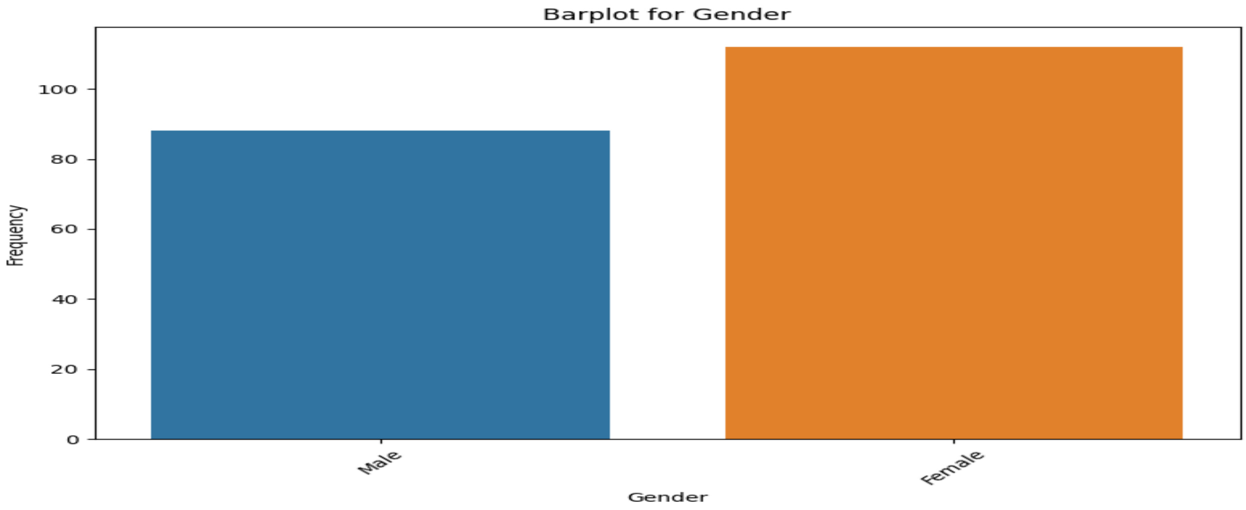


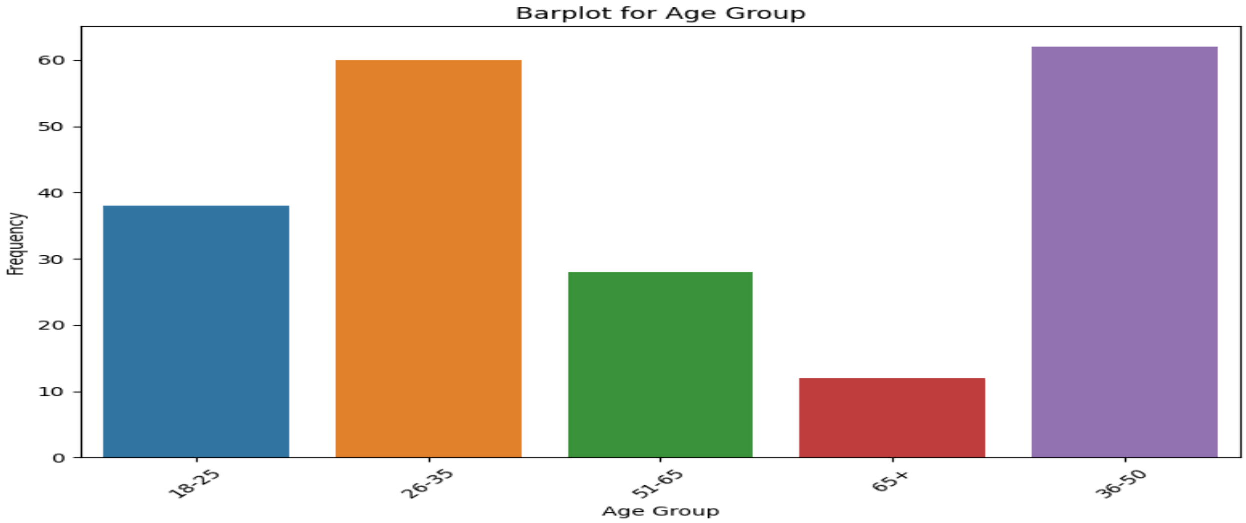

Histogram for Estimated Savings (k$)
Count
0
5
10
15
20
25
30
35
50
100
Estimated Savings (k$)
Boxplot for Estimated Savings (k$)
50
100
Estimated Savings (k$)
Kdeplot for Estimated Savings (k$)
Density
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
Estimated Savings (k$)

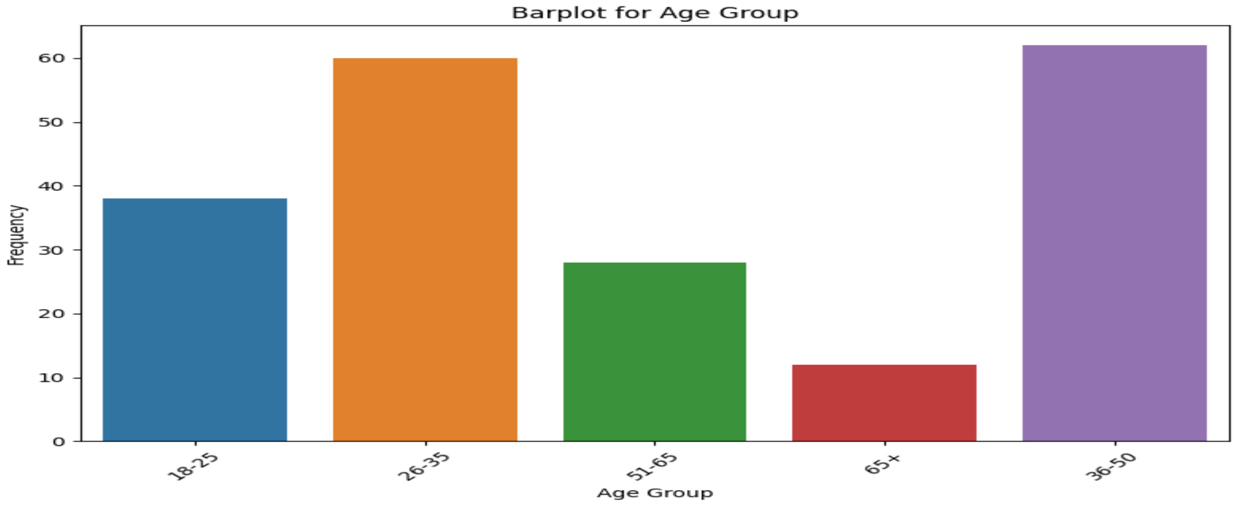

Histogram for Credit Score
Count
0
20
40
60
80
100
400
600
800
Credit Score
Boxplot for Credit Score
400
600
800
Credit Score
Kdeplot for Credit Score
Density
0.000
0.001
0.002
0.003
0.004
200
400
600
800
1000
Credit Score


Age
Annual Income (k$)
Spending Score (1-100)
Estimated Savings (k$)
Credit Score
Loyalty Years
1
-0.012
-0.33
0.13
0.24
0.61
-0.012
1
0.0099
0.81
0.75
-0.02
-0.33
0.0099
1
-0.52
-0.3
0.53
0.13
0.81
-0.52
1
0.69
-0.33
0.24
0.75
-0.3
0.69
1
-0.057
0.61
-0.02
0.53
-0.33
-0.057
1
Age
Annual Income (k$)
Spending Score (1-100)
Estimated Savings (k$)
Credit Score
Loyalty Years
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Categorical Columns are Gender and Preferred Category. The dataset consists mostly of females whilst the most frequently bought category is Electronics while the 36-50 years age group is the most frequent.

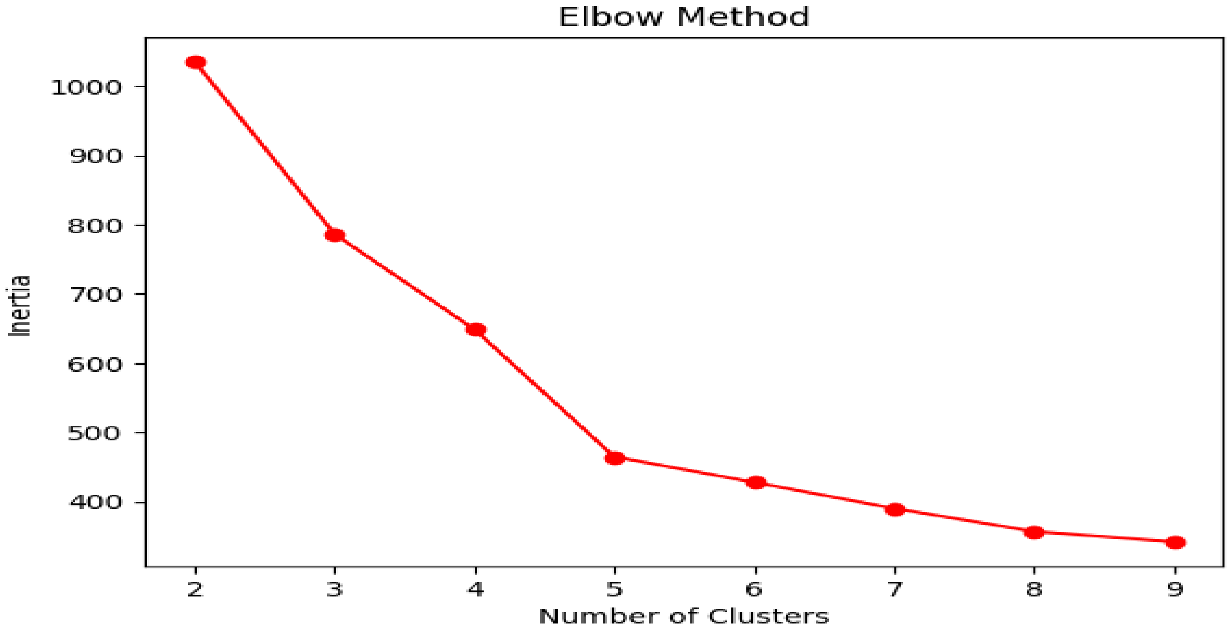




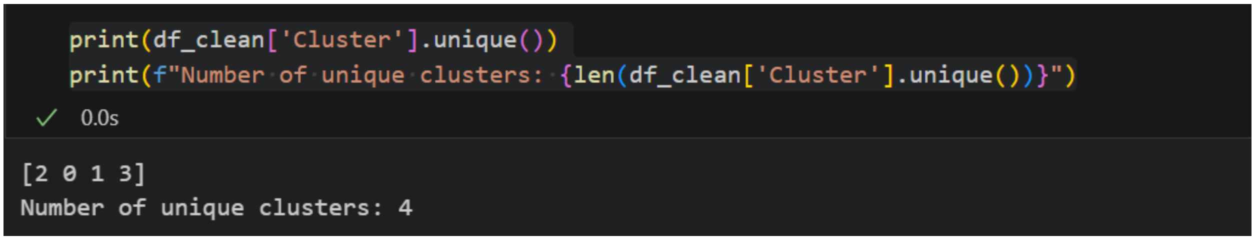


- KMeans Implementation

KMeans calculates the Euclidean distance between data points therefore we need to encode our relevant categorical variables into numerical data. I use one-hot encoding to do this. Our numerical data also needs to be scaled therefore I scaled my data using StandardScaler. KMeans is sensitive to magnitudes of data which can impact the Euclidean distances calculation. After scaling and encoding I then created a function that will take the scaled dataset as an argument and then produce a line plot with the number of clusters that will tell us the best number of clusters to select when running KMeans. This is known as the elbow method. It calculates the inertia per cluster. The moment the inertia gradually decreases(elbow) that is number of clusters we select for our KMeans model.



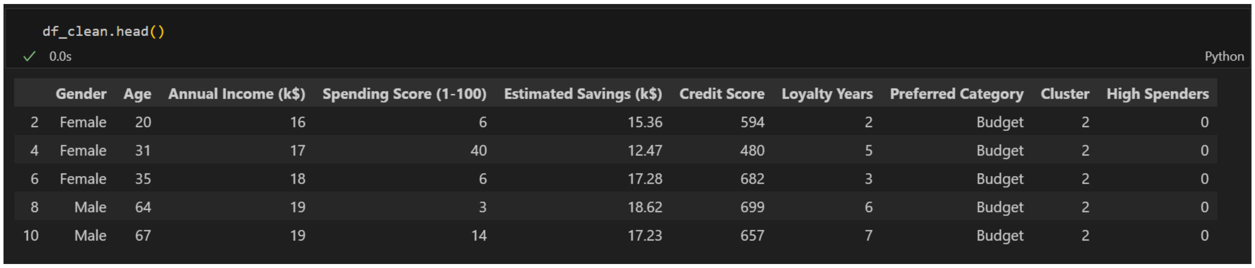
We have now identified our optimal number of clusters (k = 4) and now we can run the KMeans model. To run the KMeans model, I created a pipeline that has two steps. The first step scales the numerical values and encodes the categorical columns, and the second step is that it runs the KMeans model. The model runs and now we need to create a new column on our original, clean dataset with no outliers called 'Cluster' and add the relevant value to the row, I also added a new column that works as a threshold for High Spenders. This new field will also provide information on High Spenders in different clusters. This will give us the cluster for each row. We do sanity checks to see that unique clusters align with the number of clusters we specified. Below you'll see that the number of clusters aligns with the specified number of clusters and our KMeans model works. You will also see a snapshot of the dataset with the two new columns we added.



```
print(df_clean['Cluster'].unique())
print(f"Number of unique clusters: {len(df_clean['Cluster'].unique())}")
```

✓ 0.0s

```
[2 0 1 3]
Number of unique clusters: 4
```



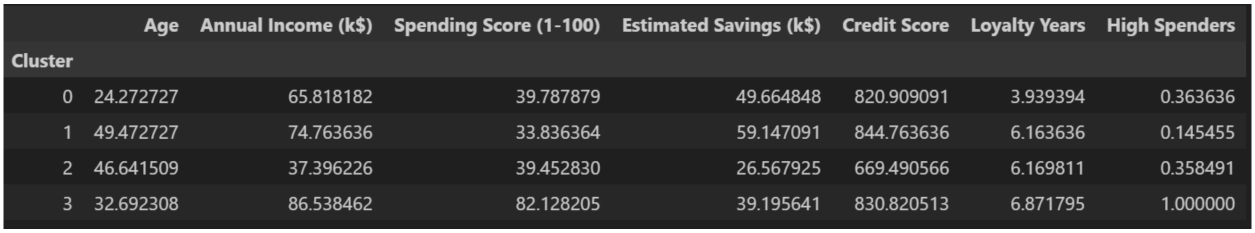
```
df_clean.head()
✓ 0.0s                                                                                  Python
```

	Gender	Age	Annual Income (k$)	Spending Score (1-100)	Estimated Savings (k$)	Credit Score	Loyalty Years	Preferred Category	Cluster	High Spenders
2	Female	20	16	6	15.36	594	2	Budget	2	0
4	Female	31	17	40	12.47	480	5	Budget	2	0
6	Female	35	18	6	17.28	682	3	Budget	2	0
8	Male	64	19	3	18.62	699	6	Budget	2	0
10	Male	67	19	14	17.23	657	7	Budget	2	0

Results and Interpretation:

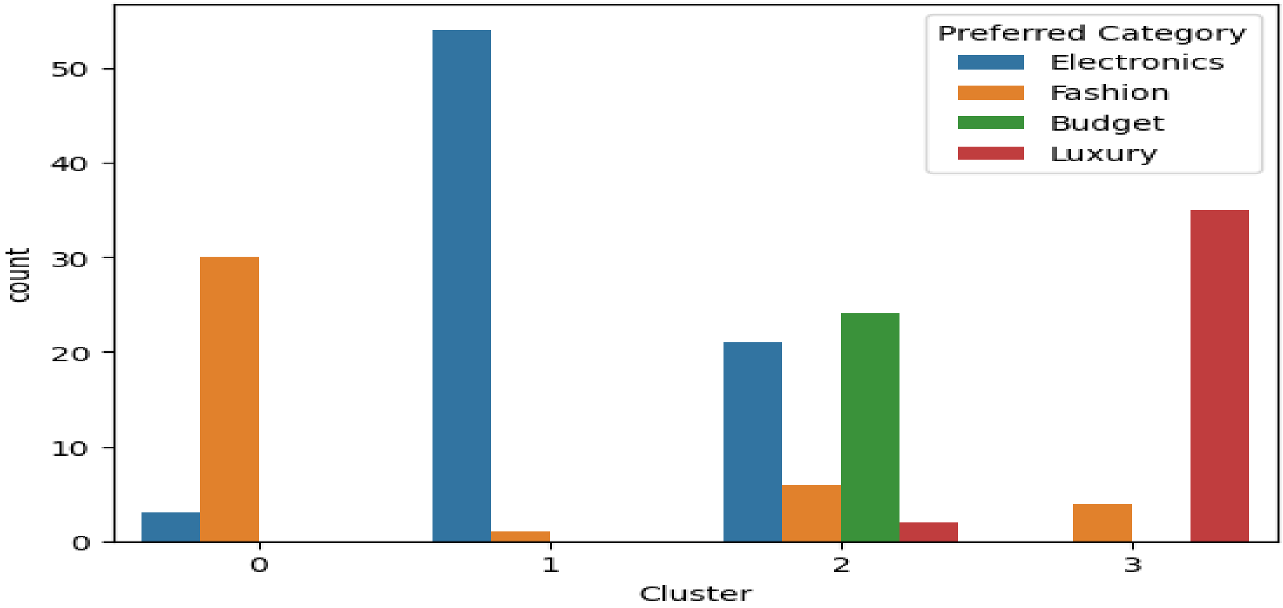
- Cluster Profile:

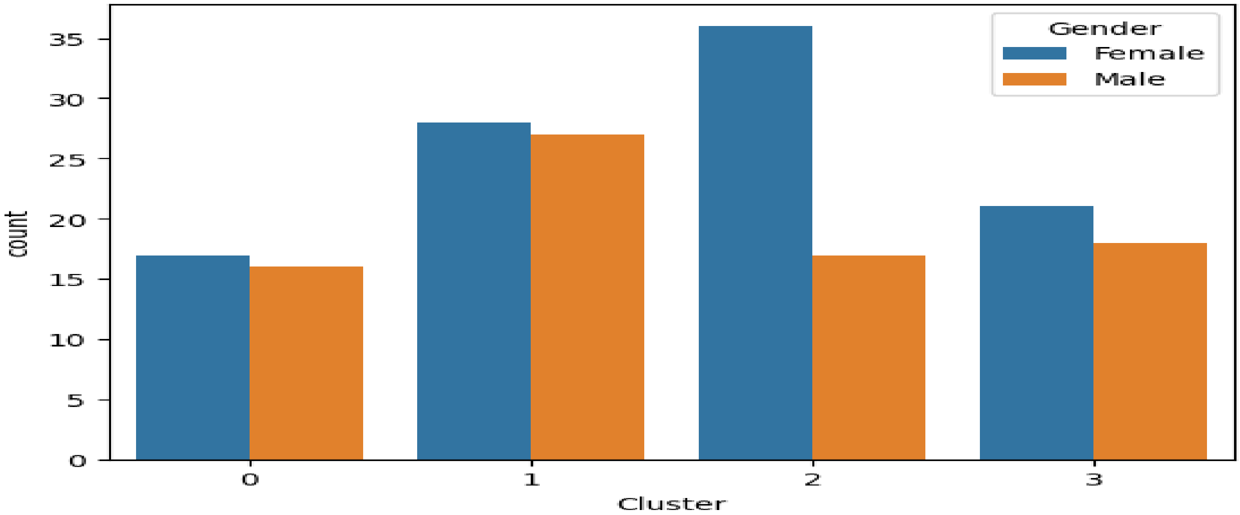
The first table is based on the numerical columns of the dataset.



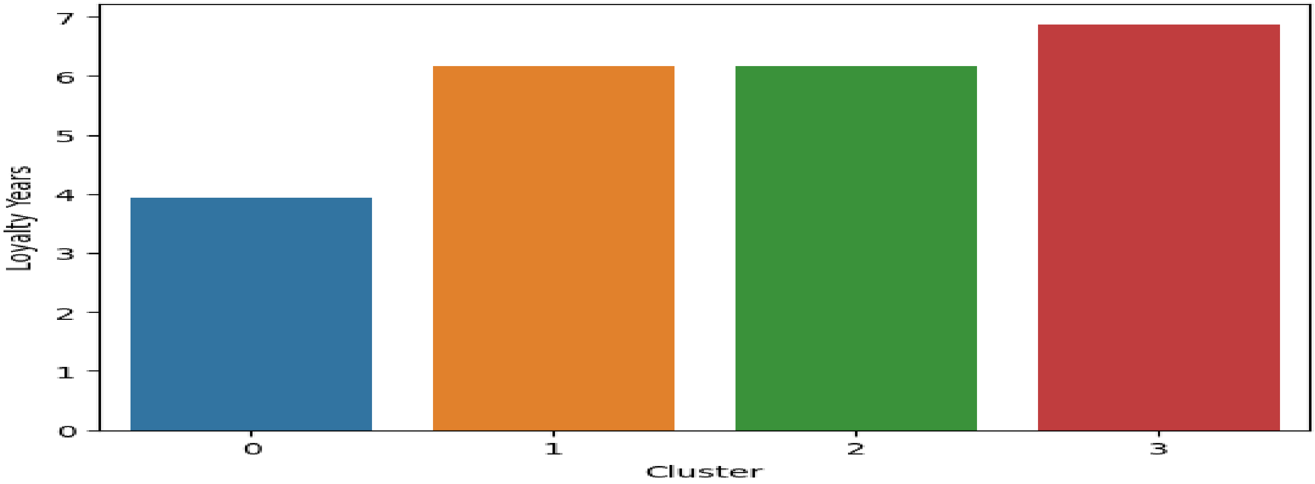
	Age	Annual Income (k$)	Spending Score (1-100)	Estimated Savings (k$)	Credit Score	Loyalty Years	High Spenders
Cluster							
0	24.272727	65.818182	39.787879	49.664848	820.909091	3.939394	0.363636
1	49.472727	74.763636	33.836364	59.147091	844.763636	6.163636	0.145455
2	46.641509	37.396226	39.452830	26.567925	669.490566	6.169811	0.358491
3	32.692308	86.538462	82.128205	39.195641	830.820513	6.871795	1.000000

The visualizations are based on the categorical columns





This shows the population of each cluster in the dataset



Cluster	Explanation	Marketing Strategy
Cluster:0	Youngest group, on average earn 65 000 dollars annually, low spending score, have customer relationship with the mall close to 4 years and 36% of them are high spenders	Because they young we can focus on growth and future engagement. We can develop social media ads and discounts tailored to this group. This will also attract other young individuals and increase this cluster.
Cluster:1	Oldest age group, on average, earn about 74 000 dollars annually, low spending score and have a customer relationship with the mall for 6 years and 14% are High spenders	We can implement the buy-now pay later with these individuals, they are conservative with their money but good with credit. We tailor marketing strategies whereby they can buy on credit distribute their funds. Doing this can potentially increase their spending scores and will migrate to other clusters
Cluster:2	Old age group, on average earn 37 000 dollars annually, low spending score like cluster:0. Also have a customer relationship	These individuals are budget friendly and therefore we need to aggressively advertise discounts for them, show them why they can benefit the most for loyalty programs and provide evidence on the value for

		money. The good thing about this cluster as well is they also buy all types of products and they dominate the population of the dataset with Cluster:1
Cluster:3	Young adults, who earn around 86 000 dollars annually, with a high spending score and around a have a customer relationship with the mall that spans to 6 years and all of them are high spenders.	They like luxurious products and are not afraid of paying for them. We need to market luxurious products to them and give them a sense of exclusivity such as early access to luxurious products before the general population.

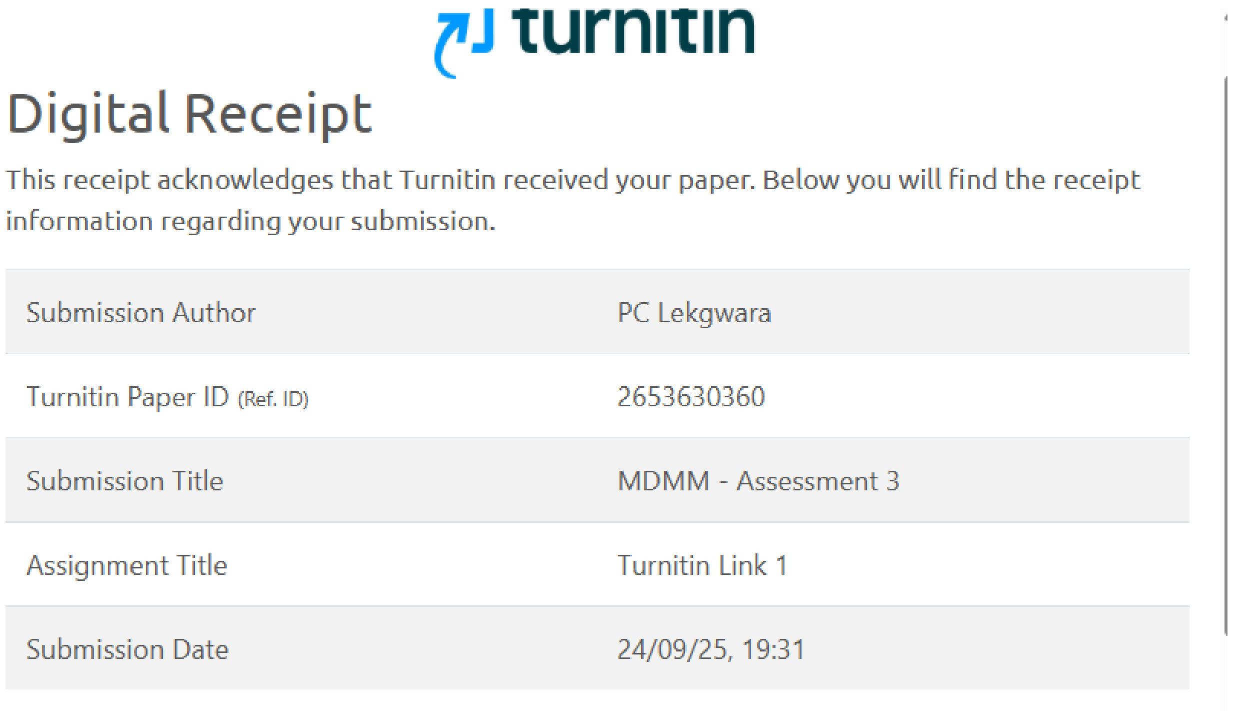
This scatter coincides with the table of summary statistics it shows that Cluster:1 and Cluster:3 is very similar.

Conclusion

We identified 4 distinct clusters, each of them with their own different traits. The understanding of the clusters helps in having a deeper understanding of the customers. Understanding of these clusters also provide evidence of who the mall's target market is. They can now build marketing strategies to ensure retainment of this target market. From a business perspective they also understand the dynamics of the mall and can now determine what types of shops to bring into the mall and increase all cluster sizes.

References

KartikSethi000, n.d. *prediction and clustering*. [Online]
Available at: https://www.kaggle.com/code/kartiksethi000/prediction-and-clustering/input
[Accessed 24 September 2025].



turnitin

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

Submission Author	PC Lekgwara
Turnitin Paper ID (Ref. ID)	2653630360
Submission Title	MDMM - Assessment 3
Assignment Title	Turnitin Link 1
Submission Date	24/09/25, 19:31