Mall Customer Segmentation Data

Market Basket Analysis

Dataset from kaggle
https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

Introduction

We are owning a supermarket mall and through membership cards, We have some basic data about Wer customers like Customer ID, age, gender, annual income and spending score.

Spending Score is something We assign to the customer based on Wer defined parameters like customer behavior and purchasing data..

Problem Statement

We own the mall and want to understand the customers like who can easily converge **[Target Customers]** so that the sense can be given to the marketing team and plan the strategy accordingly.

Main Objective

Our objective is to create an unsupervised learning model to make a market analysis with a clustering approach.

By generating clustering from the model, it is hoped that the marketing team can use it to develop a marketing strategy that is right on the target.

Data Importing and First Lookup

```
os.chdir("data")
df = pd.read_csv("Mall_Customers.csv")
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

Our data contains 200 observations and 5 variables without target.

It has 1 object type (categorical variables), 4 int types (numerical variables), and it looks like we don't have missing values.

Drop customer id

```
df.drop("CustomerID", axis=1, inplace=True)
df.head()
```

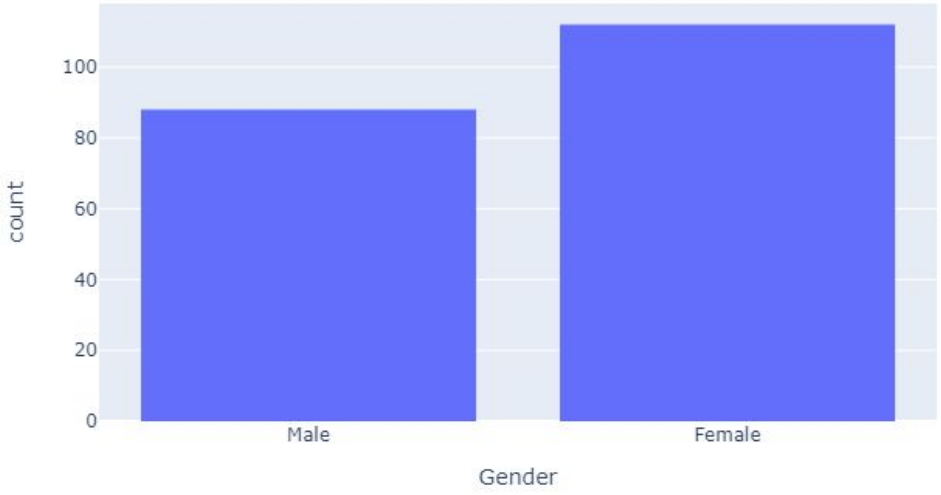
	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40

We drop Customer ID because we don’t need it for further analysis.

Exploratory Data Analysis

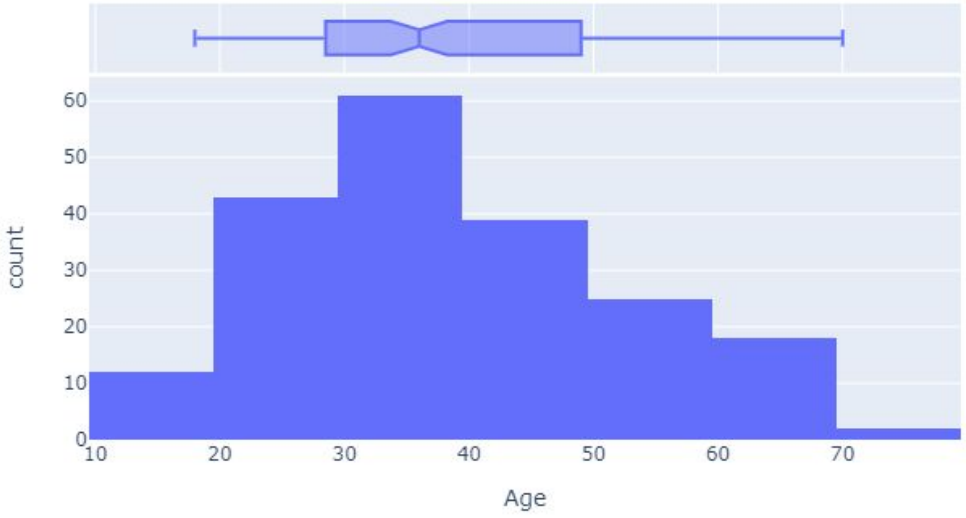
Univariate Analysis

Gender



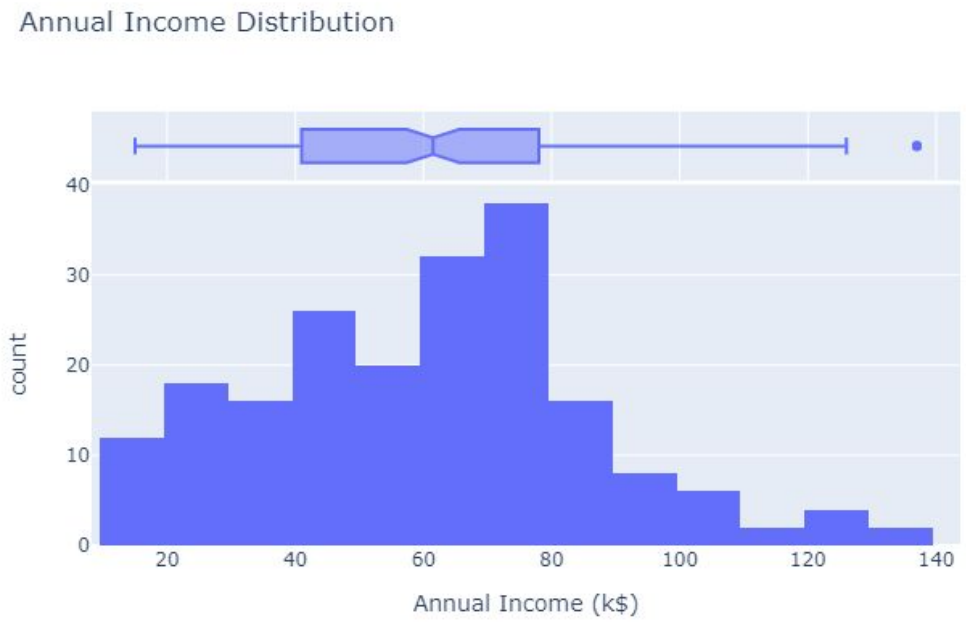
From the graph above, we can conclude that there are more female customers than male customers.

Age



Most of our customers come from the age group of 30 to 40 years, followed by customers aged 20 to 30 years. This indicates that our customers are mostly from the millennial generation.

Annual Income



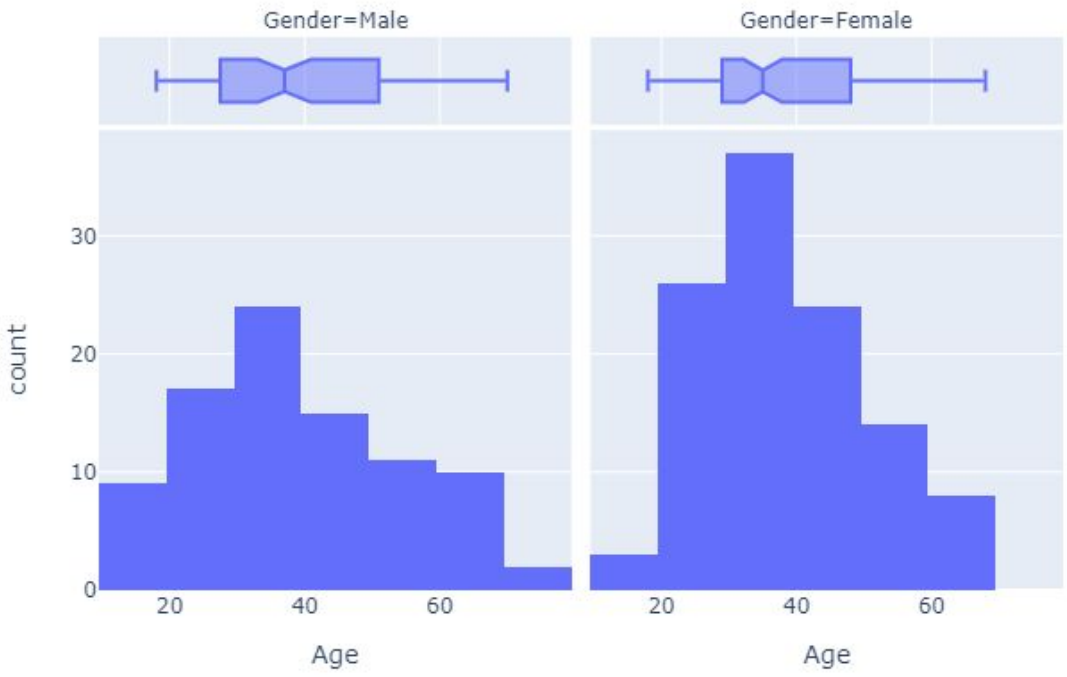
Most customers come from groups with an income of 60 to 80k$, namely the middle class.

Spending Score

Judging from the spending score, the highest intensity is in the score range 40 to 60 (middle) and this is directly proportional to the highest intensity in variable income.

Multivariate Analysis

Age by gender



Both have a similar distribution. However, there are far more female customers in the age range of 20 to 40 when compared to male customers.

T-test
H0: There is no difference in the average age of male and female customers.
H1: there is a difference in the average age of male and female customers.

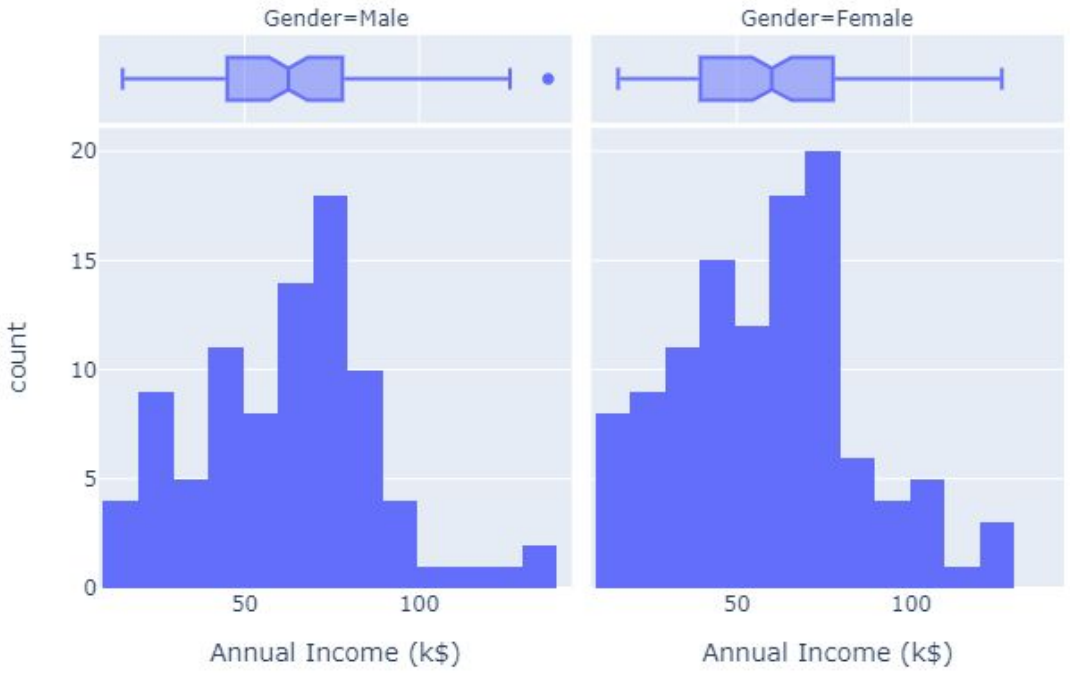
```
import scipy.stats as st

ttest = st.ttest_ind(
    df.loc[df["Gender"]=="Male", "Age"],
    df.loc[df["Gender"]=="Female", "Age"]
)
print("T-test p-value :", ttest.pvalue)
```

```
T-test p-value : 0.391890737788257
```

When viewed from the p-value in the t-test, it can be concluded that there is no difference in the average age of male and female customers (failing to reject H0).

Annual income by gender



Both have a similar distribution. There is an outlier in the annual income of men.

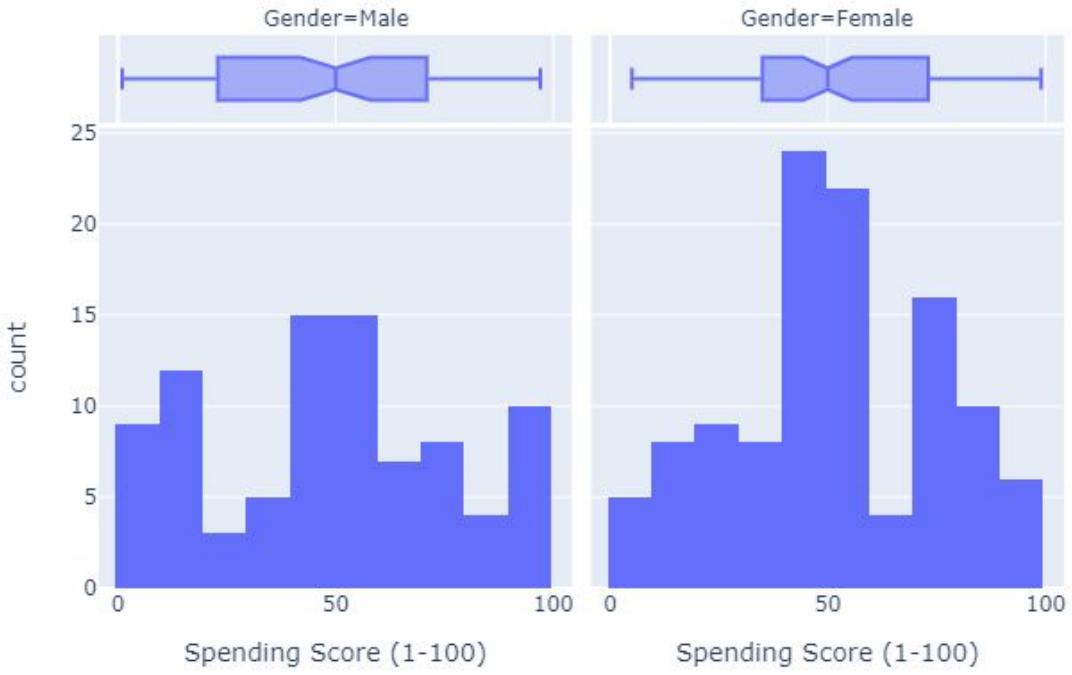
T-test
H0: There is no difference in the average annual income of male and female customers.
H1: There is a difference in the average annual income of male and female customers.

```
ttest = st.ttest_ind(
    df.loc[df["Gender"]=="Male", "Annual Income (k$)"],
    df.loc[df["Gender"]=="Female", "Annual Income (k$)"]
)
print("T-test p-value :", ttest.pvalue)
```

```
T-test p-value : 0.42755249399927864
```

If seen from the p-value in the t-test, it can be concluded that there is no difference in the average annual income of male and female customers (failing to reject H0).

Spending score by gender



There are far more female customers with spending scores in the range of 40 to 60 when compared to male customers.

T-test

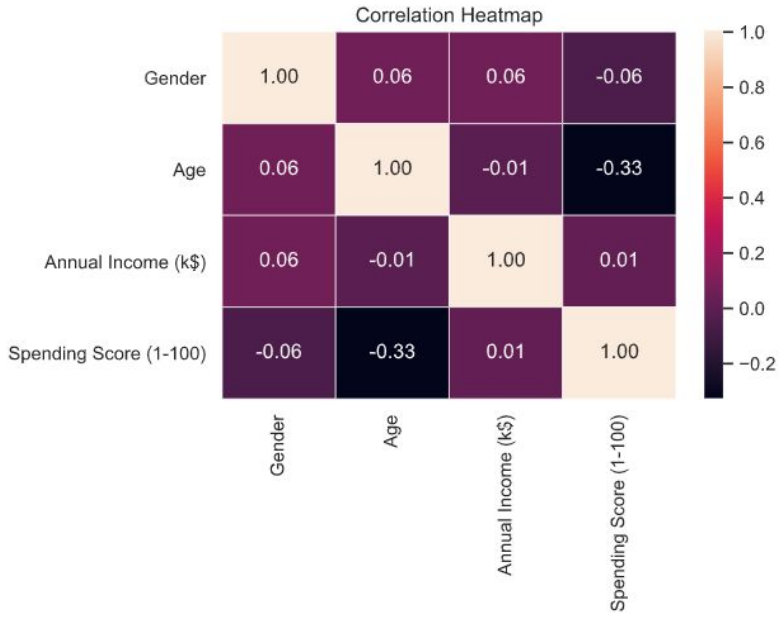
H0: there is no difference in the average spending score of male and female customers.
H1: there is a difference in the average spending score of male and female customers.

```
ttest = st.ttest_ind(
    df.loc[df["Gender"]=="Male", "Spending Score (1-100)"],
    df.loc[df["Gender"]=="Female", "Spending Score (1-100)"]
)
print("T-test p-value :", ttest.pvalue)
```

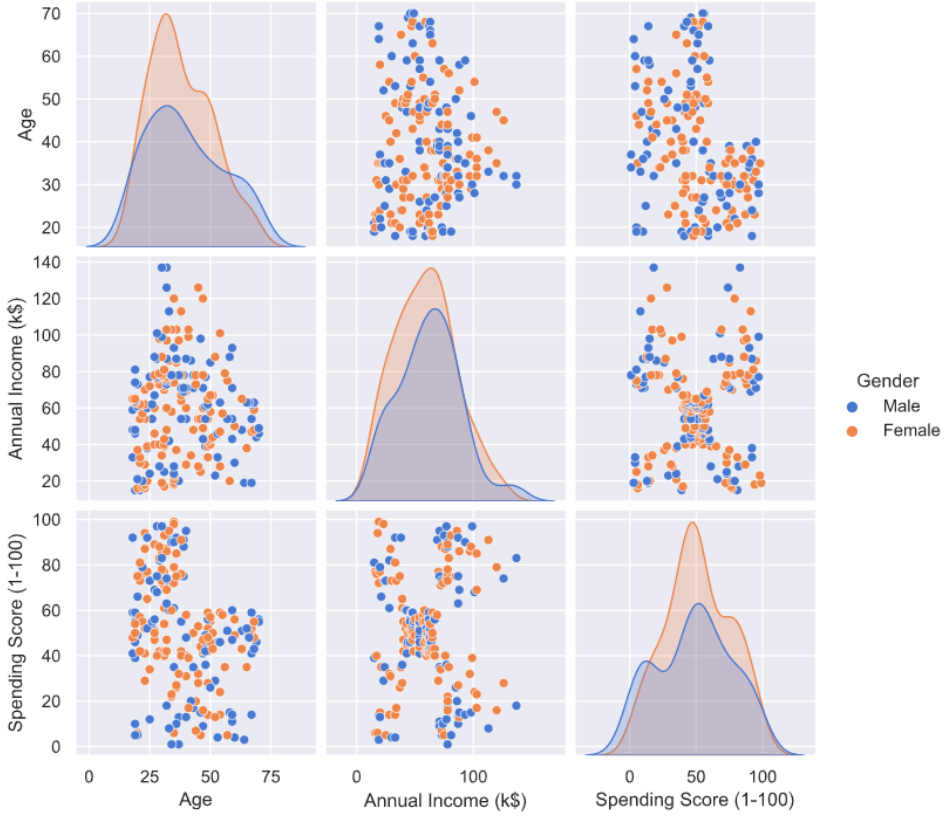
```
T-test p-value : 0.4137446589852176
```

When viewed from the p-value in the t-test, it can be concluded that there is no difference in the average spending score of male and female customers (failed to reject H0).

Features Correlation



The variable age and spending score have a strong enough correlation when compared to the others.



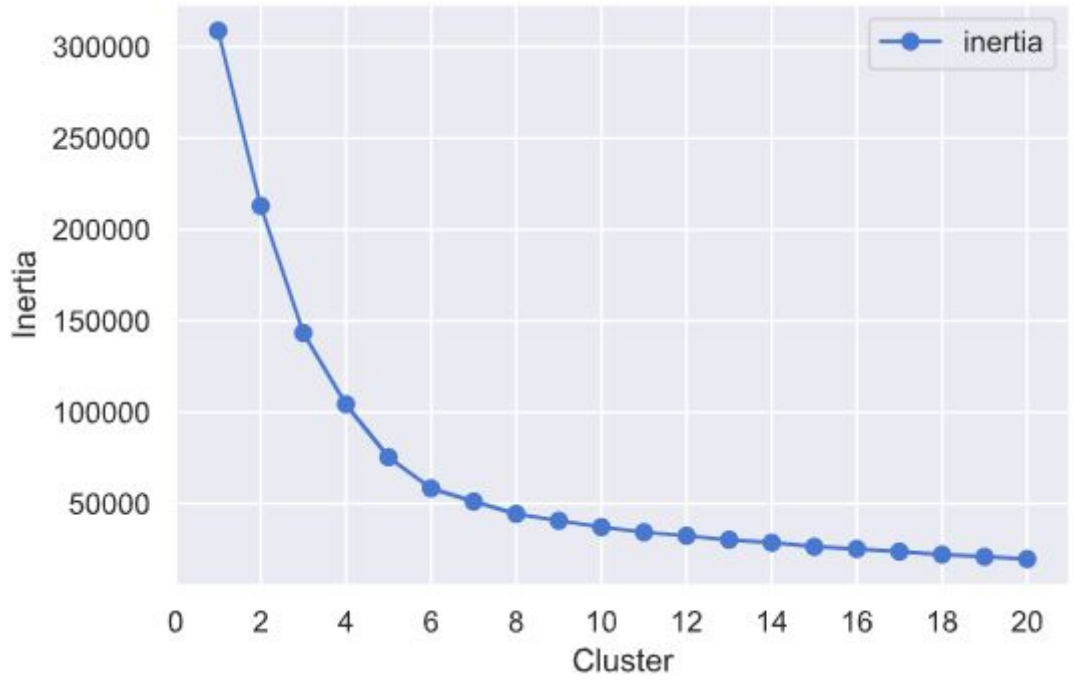
When viewed from the gender distribution, there seems to be no pattern because it tends to be random. Meanwhile, if we look at the scatter plot between income and spending score, we can see that there are 5 clusters depicted.

Modeling

At this modeling step we will use 4 machine learning, namely:

1. K-means
2. Mean Shift
3. Agglomerative Clustering
4. DBSCAN

Inertia

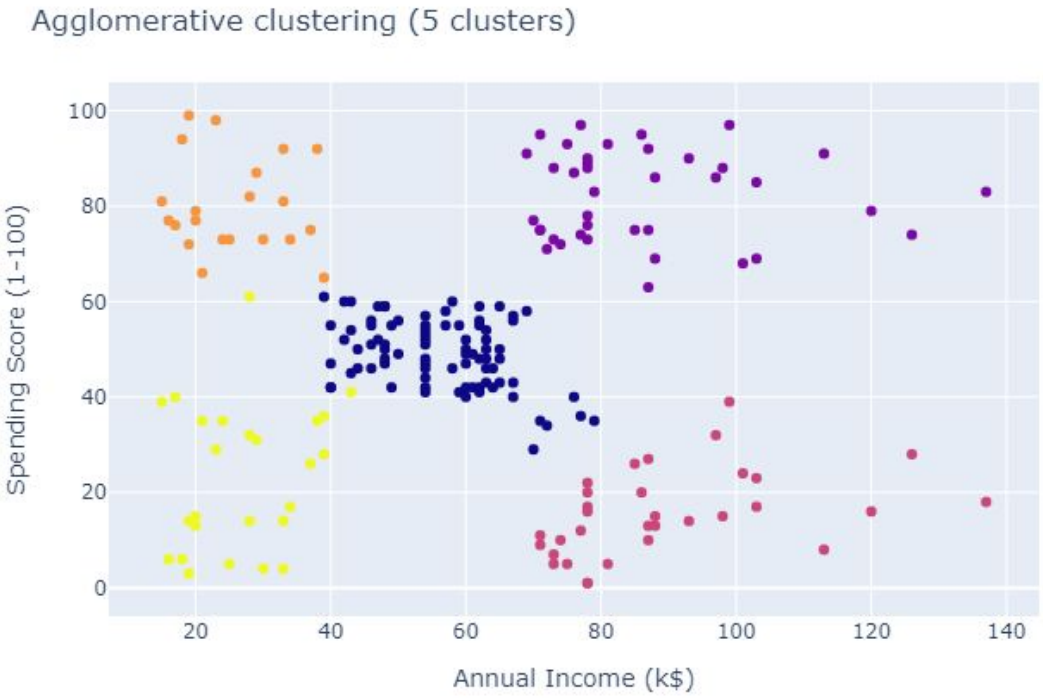


From the inertia value, we will use 5 clusters because after 5 clusters there is no significant decrease in the inertia value.

1. **K-means**

2. **Mean Shift**

3. Agglomerative Clustering



4. DBSCAN

Conclusion

1. K-means and Agglomerative Clustering successfully categorize our data into 5 clusters.
2. Mean Shift and DBSCAN can not work well on our data because our data has a fairly far spread.
3. Models that are feasible to use are K-means and Agglomerative Clustering

Insight

1. There are 5 categories of customers who visit our mall.
2. Category 1: customers with low income and low spending score.
   Category 2: customers with low income but high spending scores.
   Category 3: customers with middle income and middle spending score.
   Category 4: customers with high income but low spending score.
   Category 5: customers with high income and high spending scores.
3. With this categorization, it is hoped that the marketing team can apply different marketing strategies to each customer category.

Suggestions for next steps

- In this analysis, we draw conclusions from the variable Spending Score and Annual Income, for further analysis it can be done on a combination of other variables, for example, the variable Age and Annual Income.
- The results of clustering can be adjusted not only to 5 clusters. It depends on the decision of the management.

.