Problem Statement- The mall Customer data is given, here we need to perform the market segmentation to analyze the customer groups.

The given dataset is Mall_customers.csv.

1) **Exploring and understanding the dataset**: The different features and the characteristics of the dataset as follows:

Customer Id: The primary key of the dataset by which each individual customer or row can be uniquely identified.

Gender: Male or Female.

Age: Age of the customer.

Annual Income: The annual income is given In thousands of dollars.

Spending Score(1-100):Score assigned by the mall based on the customer behaviour and spending nature.

2) **Data Preprocessing:**

1) Data Cleaning:

a) Handling the missing values:There are no missing values in the dataset based on the sample provided.by which there is no necessity to fill any data

b) ,Noisy data or outliers: based on the given data , we are gonna identify if there exists outliers or not by visualizing using the box plot by using the matplotlib library.

The implementation as follows:

In box plot the whiskers are the lines extending from the box would be having the upper and lower bounds which would be indicating the range of the data, anything beyond these are called outliers.

The range is termed to be (IQR) abbreviation Interquartile Range:

Considering the data we would be calculating the IQR for both Annual income and spending score:

How IQR is calculated (Sample Example):

IQR=q3-q1

Where q1 is the 25th percentile and q3 is the 75th percentile

Formula to determine the lower and upper bounds for outliers:

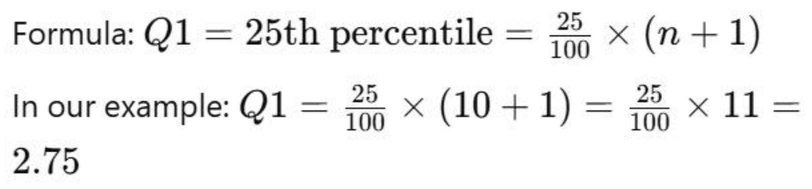
Lower Bound = q1 - 1.5*IQR

Upper Bound= q3 + 1.5*IQR

Small Scale Example how it is done internally:

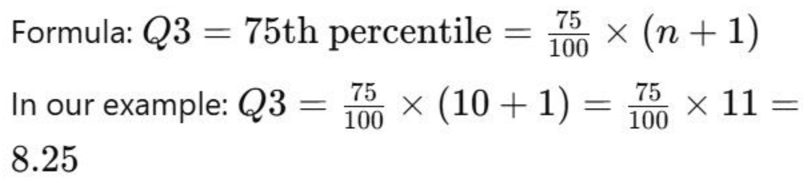
Consier this dataset {18,20,22,24,25,27,30,32,35,40}

25th percentile would be:



Formula: $Q1 = \text{25th percentile} = \frac{25}{100} \times (n+1)$

In our example: $Q1 = \frac{25}{100} \times (10+1) = \frac{25}{100} \times 11 = 2.75$



Formula: $Q3 = \text{75th percentile} = \frac{75}{100} \times (n+1)$

In our example: $Q3 = \frac{75}{100} \times (10+1) = \frac{75}{100} \times 11 = 8.25$

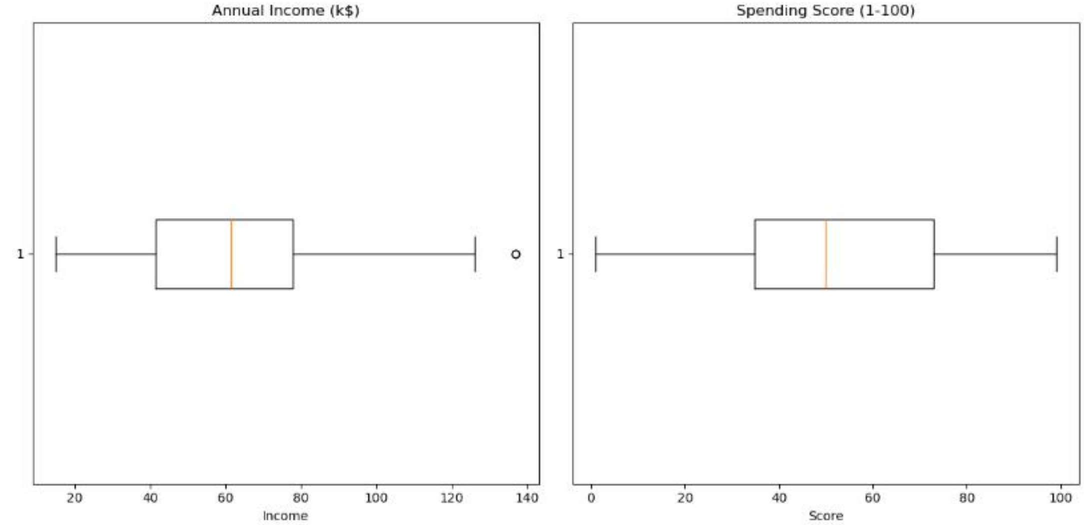
So by rounding of the q1 and the q3 : q1=22 and q2=35

Here the 22 is the lower limit and the 35 would be the upper limit and all the other values lying out side this range would be considered as the outliers.

The python code as follows:

```
import pandas as pd
import matplotlib.pyplot as plt

# Assuming your data is already in a DataFrame named 'df'
# If not, load your data into a DataFrame first

# Extract the columns of interest
annual_income = df['Annual Income (k$)']
spending_score = df['Spending Score (1-100)']

# Create a figure with two subplots
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(12, 6))

axes[0].boxplot(annual_income, vert=False)
axes[0].set_title('Annual Income (k$)')
axes[0].set_xlabel('Income')

axes[1].boxplot(spending_score, vert=False)
axes[1].set_title('Spending Score (1-100)')
axes[1].set_xlabel('Score')

plt.tight_layout()

# Show the plot
plt.show()
```

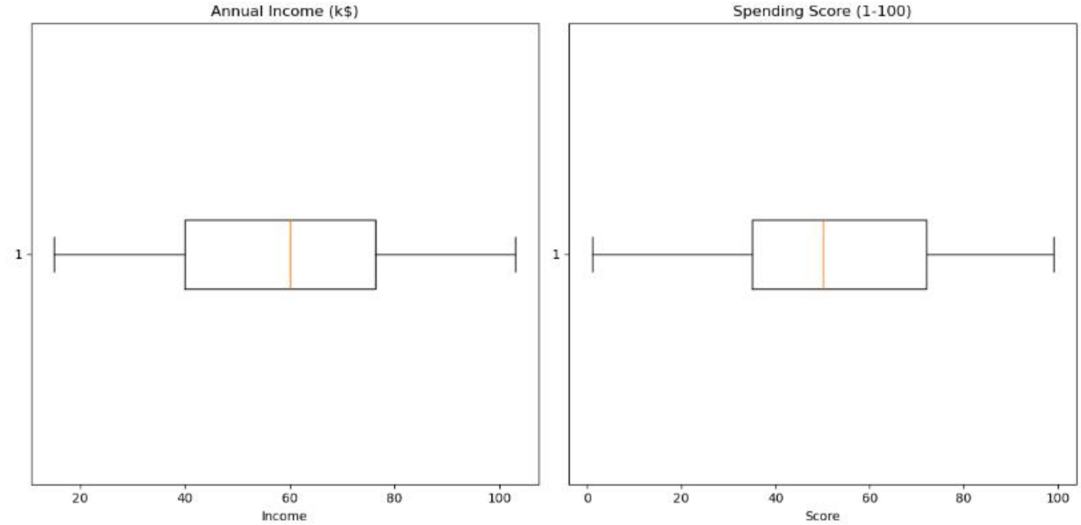
The given dataset has the outlier in one of the feauture, The annual income has one outlier which need to be removed.



Since we have identified an outlier here we would directly be removing the outlier:

Simple Range-Based Method: Remove data that are above a certain percentile or below a certain percentile. This can be effective if the outliers are clear do not represent critical data.

```
max_threshold = df['Annual Income'].quantile(0.95)
df = df[df['Annual Income (k$)'] <= max_threshold]
```



The ouliers are removed.