** Kinnaird College for Women Lahore**

**Project.**

Customer Segmentation [Visualization]

**Submitted By:**

**Group Members**

**Mishaal Arbab**

**Hibba Mehmood**

Major: BSCS (Semester VI)

Course Name: Data Science with Python

Department: Computer Science

Kinnaird College for Women Lahore

**Submitted To:**

**Mam Nimra Zaheer**

Department: Computer Science

Date: January 1, 2020

**Table of Contents**

[Dataset 3](#_Toc60336305)

[*Customer id:* 4](#_Toc60336306)

[*Gender:* 4](#_Toc60336307)

[*Age:* 4](#_Toc60336308)

[*Annual Income (k$):* 4](#_Toc60336309)

[*Spending score (1-100):* 5](#_Toc60336310)

[*Class (poor, middle, rich):* 5](#_Toc60336311)

[*Box plots of individuals features* 6](#_Toc60336312)

[*Box plots of two feature* 7](#_Toc60336313)

[*When a customer belongs to poor class effect on annual income* 7](#_Toc60336314)

[*Class and spending score (1-100) box plot:* 8](#_Toc60336315)

[*Box plots with multiple features along y* 9](#_Toc60336316)

[*Gender:* 10](#_Toc60336317)

[*Age:* 10](#_Toc60336318)

[*Annual Income (k$):* 11](#_Toc60336319)

[*Spending Score (1-100):* 11](#_Toc60336320)

[*Class:* 12](#_Toc60336321)

[*Customer ID:* 12](#_Toc60336322)

[*Class and Annual Income:* 13](#_Toc60336323)

[*Class, annual income, spending score:* 13](#_Toc60336324)

[*Class, Age:* 14](#_Toc60336325)

[*Customer ID:* 15](#_Toc60336326)

[*Age:* 15](#_Toc60336327)

[*Annual Income (k$):* 16](#_Toc60336328)

[*Spending Score (1-100):* 16](#_Toc60336329)

**CUSTOMER SEGMENTATION**

Visualization and plots

# Dataset

Mall customer is our dataset which have 6 features,

* Customer ID (Id of customer)
* Gender
* Age (age of customer)
* Annual Income (k$) [annual income of customer]
* Spending score (1-100) [points that the customer earn by spending money on particular brand or shop]
* Class [poor,middle,rich on the bases of annual income]

Statistics

In statistics I explain the min, max, counts, mean, and standard deviation of each features. Table given below,

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Stats* | *Customer ID* | *Gender* | *Age* | *Annual Income (k$)* | *Spending Score (1-100)* | *Class* |
| Count | 200.00 | 200.00 | 200.00 | 200.00 | 200.00 | 200.00 |
| Mean | 100.50 | 0.44 | 38.85 | 60.56 | 50.20 | 0.68 |
| Std | 57.88 | 0.50 | 13.97 | 26.26 | 25.82 | 0.66 |
| Min | 1.00 | 0.00 | 18.00 | 15.00 | 1.00 | 0.00 |
| 25% | 50.75 | 0.00 | 28.75 | 41.50 | 34.75 | 0.00 |
| 50% | 100.50 | 0.00 | 36.00 | 61.50 | 50.00 | 1.00 |
| 75% | 150.25 | 1.00 | 49.00 | 78.00 | 73.00 | 1.00 |
| Max | 200.00 | 1.00 | 70.00 | 137.00 | 99.00 | 2.00 |

This table shows the values of mean, median, quartiles, minimum, maximum of each features, I take the value till 2 decimal places.

25% 🡪 quartile 1 (q1)

50% 🡪 median (q2)

75% 🡪 quartile 3 (q3)

Mean 🡪 average of data

Std 🡪 standard deviation (the measure of dispersion of a set of data from its mean)

## ***Customer id:***

In customer id we have id’s of 200 customer the maximum id of customer is 200 because we have 200 customer and the minimum id is 1 because id of customer start from 1.standard deviation is 57.88. Average of customer id is 100.50 because total customers are 200, first 25% customer have id till 50.75 and 50% of customer have id between (1-100) and 75% of customer have id between (1-150).

## ***Gender:***

For gender we encode male and female in to zero or 1 maximum value is 1 and minimum is 0. Mean /average of gender is 0.44. . The standard deviation of gender is 0.50. Quartile 1 (25%) of customer are females. Median (50%) middle point of gender is 0.00 which means middle gender of customer is female. Quartile 3 (75%) or we can say that middle from second half of data is 1.00(male), 75% of customer gender are either male or female.

## ***Age:***

For age maximum age of customer is 70 and minimum age of customer is 18 and mean /average age of customer is 38.85, the half of 70 is same 38.85 that’s why average is 38.85. The standard deviation of age is 13.97 means how data disperse from 38.85 upward or downward, it is downward disperse .quartile 1 (25% of age) we can say that middle value in first half of age is 28.75 that means 25 % of customer age are less than or equal to 28.75.median (50% of age) middle point of age is 36.00 which means middle age of customer is 36. Quartile 3 (75% of age) or we can say that middle from second half of age is 49.00, 75% of customer age are less than 49.00 or equal to 49.00.

## ***Annual Income (k$):***

Maximum annual income of customer is 137.00 (k$) which means maximum annual earning of customer is 137.00 (k$) and minimum annual income of customer is 15.00(k$) and mean /average annual income of customer is 60.56(k$), the half of 137.00(k$) is same 60.56(k$) that’s why average is 60.56(k$). The standard deviation of annual income is 26.36(k$) means how annual income disperse from 60.56(k$) upward or downward, it is downward disperse .quartile 1 (25% of annual income) we can say that middle value in first half of annual income is 41.50(k$) that means .median (50% of annual income) middle point of annual income is 61.50(k$). Which means middle value of annual income is 61.50(k$). Quartile 3 (75% of annual income) or we can say that middle from second half of annual income is 78.00(k$), 75% of customer have income less than or equal to 78.00(k$).

## ***Spending score (1-100):***

Maximum spending score of customer is 99.00 which means maximum point’s customer earn is 99.00. And minimum points that the customer earn is 1.00. And mean /average spending score of customer is 50.20. The standard deviation of spending score is 25.82.quartile 1 (25% of spending score) we can say that middle value in first half of spending score is 34.75 which means 25 % of customer earn 34.75 or less than 34.75 score. Median (50% of spending score) middle point of spending score is 50.00. Which means middle value of spending score is 50.00. Quartile 3 (75% of spending score) or we can say that middle from second half of spending score is 73.00, 75% of customer earn score less than or equal to 73.00.

## ***Class (poor, middle, rich):***

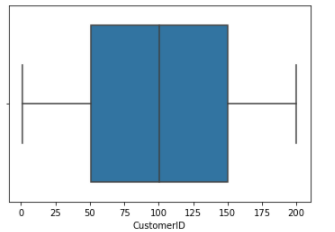
For class we encode poor, middle, rich class in to 1, 0, and 2. Maximum value is 2 and minimum is 0. Mean /average of class is 0.68. The standard deviation of age is 0.66. Quartile 1 (25%) customers belong to middle class on the bases of annual income. Median (50%) customers belong to poor class or middle class. Quartile 3 (75%) customer belongs to either poor or middle class.

Visualizations

**Box plots**

## ***Box plots of individuals features***

#### **Customer ID:**



* Minimum customer id is 0
* Maximum customer id is 200
* Quartile 1 of customer id is 50
* Median of customer id is 100
* Quartile 3 of customer id is 150(75% of customer have id between (1-150))

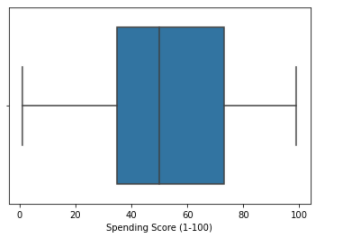
#### **Age:**

* Minimum age of customer 18
* Maximum age of customer 70
* Quartile 1 of customer age is 28(25 % of customer age are less than or equal to 28)
* Median of customer age 38
* Quartile 3 of customer age is 49.( 75% of customer age are less than 49 or equal to 49)

#### **Annual Income (k$):**

* Maximum annual income is 137.
* Minimum annual income of customer is 15.
* Quartile 1 of customer annual income is 41(25 % of customer have annual income less than or equal to 41(k$)).
* Median of annual income is 61
* Quartile 3 of customer annual income is 78(k$),( 75% of customer have income less than or equal to 78(k$))

#### **Spending Score (1-100):**

*  Maximum spending score is 99
* Minimum spending score is 0
* Quartile 1 of spending score is 34(25 % of customer earn 34 or less than 34 score)
* Median of spending score is 50
* Quartile 3 of spending score is 73(75% of customer earn score less than or equal to 73).

## 

## ***Box plots of two feature***

#### **Class and annual income (k$) box plot:**

## ***When a customer belongs to poor class effect on annual income***

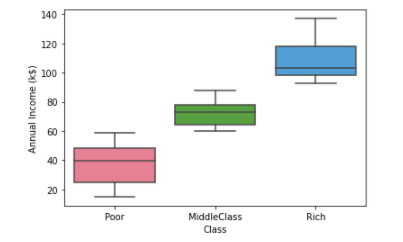
The maximum annual income of customer 60 and Minimum annual income of customer which belongs to poor class is 18. quartile 1 of annual income of customer is 25(k$) which means 25% customer of poor class have annual income 25(k$) or less than 25(k$).median of annual income 39(k$) which means the customer belongs to poor class have median of annual income is 39(k$).quartile 3 of annual income of customer is 49(k$) which means 75% customer of poor class have annual income 49(k$) or less than 49(k$).

#### **When a customer belongs to middle class effect on annual income**

The maximum annual income of customer 89 and Minimum annual income of customer is 59 which belongs to middle class. quartile 1 of annual income of customer is 64(k$) which means 25% customer of middle class have annual income 64(k$) or less than 64(k$).median of annual income 115(k$) which means the customer belongs to rich class have median of annual income is 115(k$).quartile 3 of annual income of customer is 125(k$) which means 75% customer of rich class have annual income 125(k$) or less than 125(k$).

#### **When a customer belongs to Rich class effect on annual income**

The maximum annual income of customer 138(k$) and Minimum annual income of customer is 87(k$) which belongs to rich class. quartile 1 of annual income of customer is 97(k$) which means 25% customer of rich class have annual income 97(k$) or less than 97(k$).median of annual income 73 which means the customer belongs to middle class have median of annual income is 73(k$).quartile 3 of annual income of customer is 78(k$) which means 75% customer of middle class have annual income 78(k$) or less than 78(k$).



## ***Class and spending score (1-100) box plot:***

#### **When a customer belongs to poor class effect on spending score**

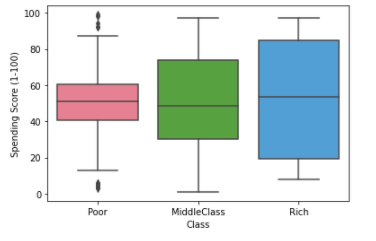
The maximum spending score of customer 85 and Minimum spending score of customer which belongs to poor class is 15. quartile 1 of spending score of customer is 41 which means 25% customer of poor class earn score 41 or less than 41.median of spending score is 52 which means the customer belongs to poor class have median of earning score is 53.quartile 3 of spending score of customer is 59 which means 75% customer of poor class earn score 59 or less than 59.

#### **When a customer belongs to middle class effect on spending score**

The maximum spending score of customer 97 and Minimum spending score of customer which belongs to middle class is 0. quartile 1 of spending score of customer is 32 which means 25% customer of middle class earn score 32 or less than 32.median of spending score is 50 which means the customer belongs to middle class have median of earning score is 50.quartile 3 of spending score of customer is 71 which means 75% customer of middle class earn score 71 or less than 71.

#### **When a customer belongs to rich class effect on spending score**

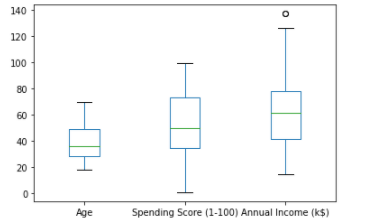
The maximum spending score of customer 97 and Minimum spending score of customer which belongs to rich class is 10. quartile 1 of spending score of customer is 41 which means 25% customer of rich class earn score 18 or less than 18.median of spending score is 54 which means the customer belongs to rich class have median of earning score is 54.quartile 3 of spending score of customer is 82 which means 75% customer of rich class earn score 82 or less than 82.



## ***Box plots with multiple features along y***

#### **Class [along x axis] and spending score (1-100), annual income (k$), age [along y axis] box plot:**

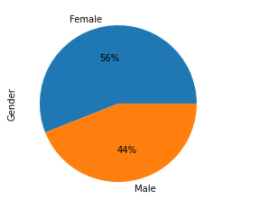
Here I show the box plots of age, spending score, annual income, with respect to class means how the class affect this features that are along y axis .how we can interpret these features through rich, middle, poor class. Minimum age of customer is 18 and maximum age of customer is 70 median of age is 38 and quartile of age is q1 is 28 and q3 is 50 .for annual income minimum is 18 maximum is 125 dollar and median is 60 dollar or quartiles are q1 is 42 dollar and q3 is 79 dollar .for spending score minimum is 0 maximum is 100 median is 50 and quartiles are q1 is 38 and q3 is 71.



**INDIVIDUALS PLOTS:**

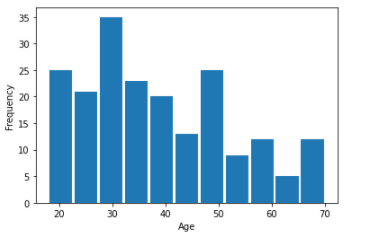
## ***Gender:***

I choose pie chart for gender because gender has two value male and female or we can easily predict these through pie chart. This pie chart shows 56% are female and 44% are males in mall customer segmentation dataset.



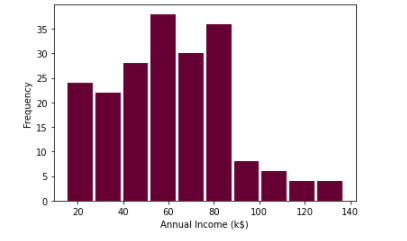
## ***Age:***

I choose histogram for age of the customer because through histogram we can easily see the distribution of ages of customer. We can see that which age group people are more. This histogram shows that the customer of age 26 to 32 are exist more in mall customer dataset, and the people of age 61 to 65 are less in numbers.



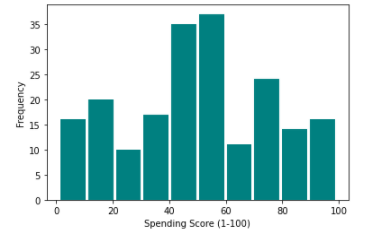
## ***Annual Income (k$):***

I choose histogram for annual income of the customer because histogram give good distribution of income as compared to bar plot, line plot or pie chart. Through binning in histogram we can easily see that highest annual income or lowest annual income of customer. This histogram shows that the customer whose annual income lies between (55(k$)-95(k$)) are exist more in mall customer dataset, and the people whose annual income lies between (95(k$)-138(k$)) are less in numbers.



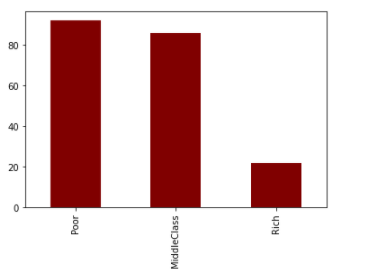
## ***Spending Score (1-100):***

I choose histogram for Spending Score of the customer because spending score have random data points which are not in sequence and histogram give good distribution of score as compared to bar plot, line plot or pie chart. Through binning in histogram we can easily see that highest highest or lowest spending score of customer. This histogram shows that the customer whose spending score lies between (40-60)are exist more in mall customer dataset, and the people whose spending score lies between (20-35)are less in numbers.



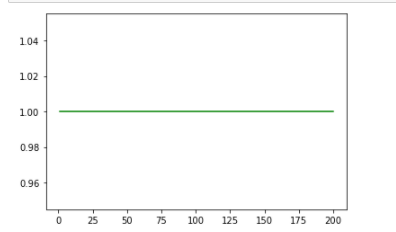
## ***Class:***

We choose bar plot for class because through bars we can easily see class’s rich, poor, middle class. Class is categorical data for categorical data bar plot shows better visiualizations.this bar graph shows that 83% customer lies in middle or poor class which means 83% people have middle or low income and 17% people have high income lies in rich class.



## ***Customer ID:***

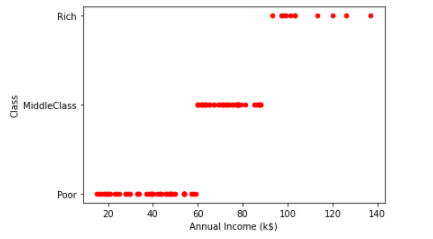
For customer id we choose line plot because id’s are in sequence 0 to 200 that why it shows in single line unique id assign to each customer.



**COMPARATIVE PLOTS:**

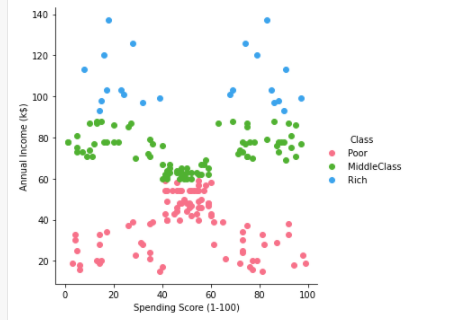
## ***Class and Annual Income:***

We draw scatter plot between annual income and class .annual income is along x axis and class is along y axis .this graph interpret that the customer whose income lies between (15k$ to 59k$) are belongs to poor class and the customer whose income lies between (60k$ to 85k$) belongs to middle class and the customers whose income lies between (85k$ to 138k$) belongs to rich class.



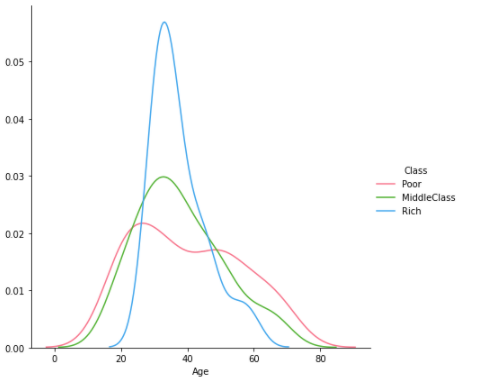
## ***Class, annual income, spending score:***

We draw scatter plot between annual income, spending score and class .annual income is along y axis and spending score is along x axis .this graph interpret that the spending score and annual income of customer according to class. Pink color shows poor class, blue color shows rich class and green color shows middle class. Pink dots shows that the customer belongs to poor class earn how much annual income and earn how much score. Green dots shows that the customer belongs to middle class earn how annual income or earn how much spending score. Same case for blue dots/rich class. Poor class customer annual income lies between 0 dollar to 59 dollar but they earn score between 0 to 100,middle class customer have annual income between 59 dollar to 85 dollar and they earn spending score same between 0 to 100,rich class people have annual income between 85 dollar to 139 dollar and they earn score between (0-40,80-100).



## ***Class, Age:***

Draw a facet grid graph between class and age this graph shows that the person belong to poor class have age factor positively skewed more because it spread toward right side more and the customer belongs to middle class have age factor normal not positively skewed or not negatively skewed, and the customer belongs to rich class are highly skewed upward because its values going upward. Age factor of rich class is between (20-70), age factor for middle class is between (0-80) same for poor class.



**PAIRPLOT:**

From the picture below, we can observe the variations in each plot. The plots are in matrix format where the row name represents x axis and column name represents the y axis. The main-diagonal subplots are the univariate (distributions) for each attribute.

## ***Customer ID:***

Customer who have id between (0-70) belong to poor class and the customer who have (50-140) id are normal skewed or the persons who have (140-200) id are highly skewed or belong to rich class.

## ***Age:***

Age of poor class customer lies between (0-80) right skewed means spread toward right. For middle class age lies (10-70) it is normally skewed. For rich class age lies between (20-65) it is highly skewed.

## ***Annual Income (k$):***

Poor class people have (0k$-60k$) it is right skewed and middle class people have annual income (50k$-110k$) highly skewed graph values going upwards and rich class people have annual income (110k$-150k$) it is also highly skewed.

## ***Spending Score (1-100):***

Poor class people have earn score (-50-100) it is inward graph and middle class people have earn score between (50-130) normally skewed graph and rich class people earn score (-30-120) it is also highly left skewed.

Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our data-set.

