

# EDA Project in Customer Segmentation\_Yanshuo Li\_ITP449

[Yanshuo Li](#)

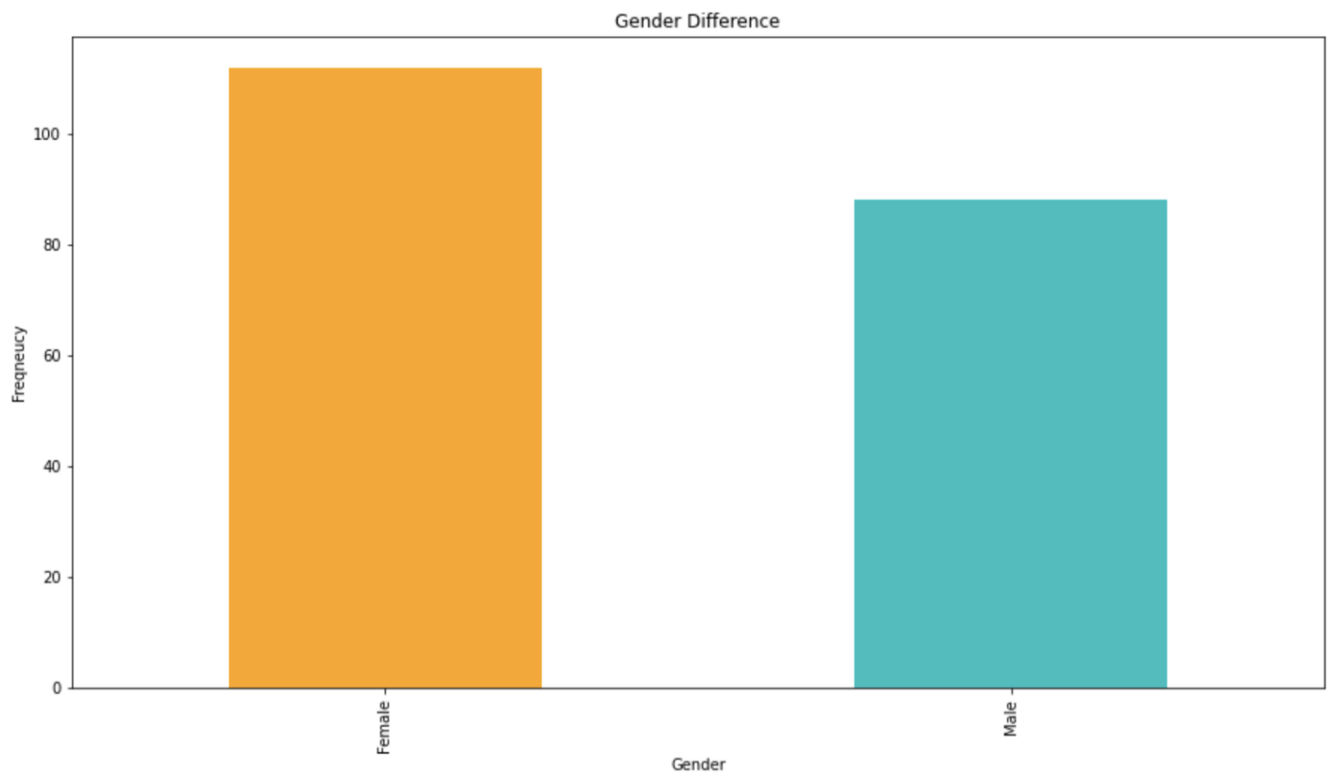
For this project, I am aiming to analyze what's the factor that influences a consumer's purchase power and consumption. In order to achieve my goal, I choose a shopping mall's dataset about its customer identity. It includes variables relevant to marketing such as gender, age, annual income, etc. By doing a machine learning project for this dataset, I would be able to figure out which of them are determinants of consumer spending and market behavior.

Here are few questions I would like to explore:

1. Which gender has higher purchase power?
2. Is there a specific age group that has higher purchase power than any other age groups?
3. Will annual income of a individual influence his or hers purchase power or desire of consumption?

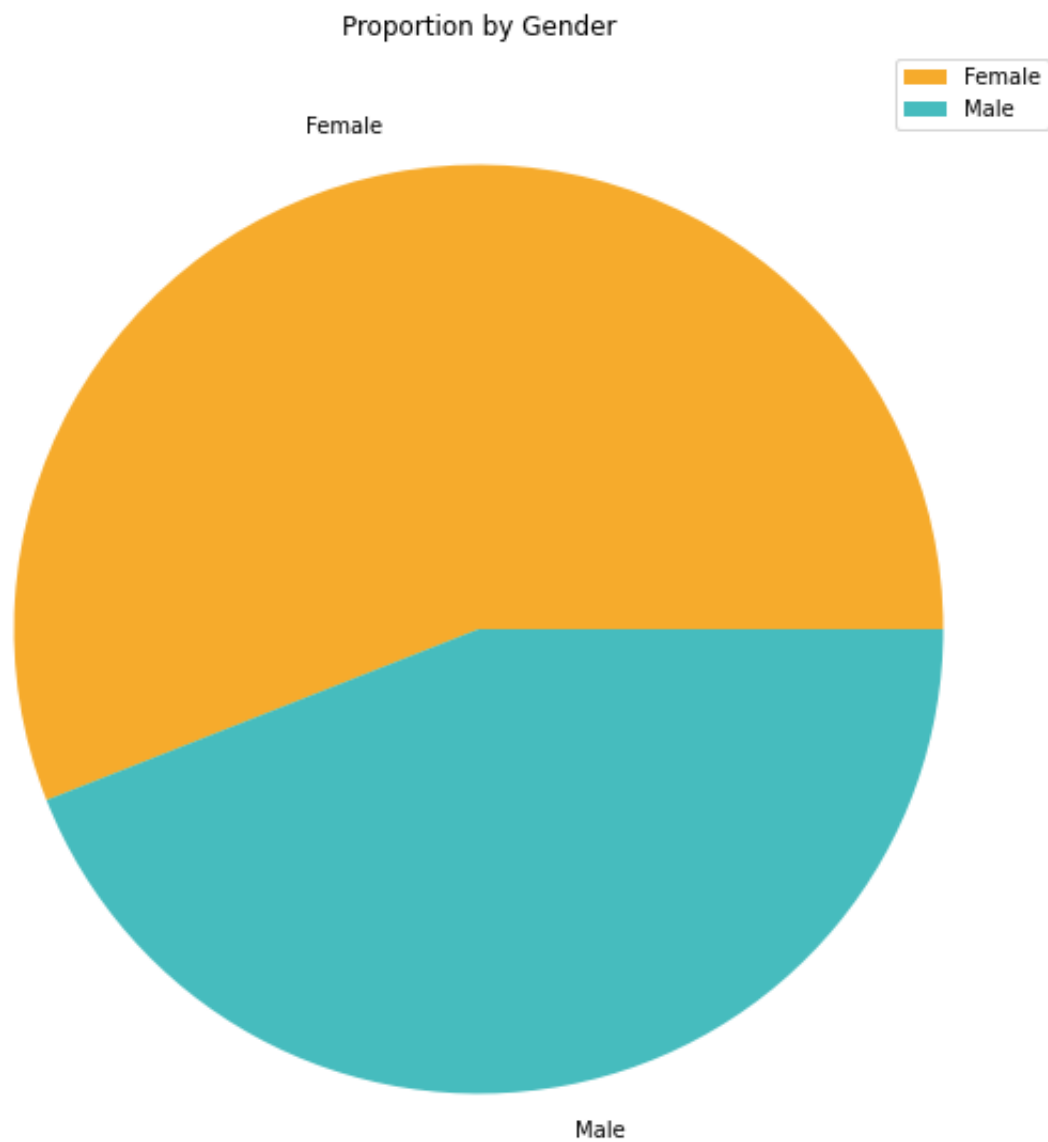
I will focus on solving these problems in my project. Meanwhile, I will also try to rise more new questions with the progress of the project.

## Analysis:



Graph 1

First, from the above diagram (Graph 1), we can conclude that there are more female customers compared to male customers. Females have a higher tendency to go shopping than males.



Graph 2

From the pie chart, we can see that there are 56% female customers, and only 44% male customers; the difference is 12%. This suggests that the difference between the number of customers with different gender is significant.



Image for post

Graph 3

This Histogram of Age shows that the age of all customers is positively skewed. And the range of customer ages is between 15 and 70. This shows that people between the ages of 20 to 40 have comparatively high purchase power and needs of shopping, and with age gradually increase after 40, the purchasing power is negatively related to the customer ages. Therefore, we can conclude that age is one of the determinate that could influence customer spending.

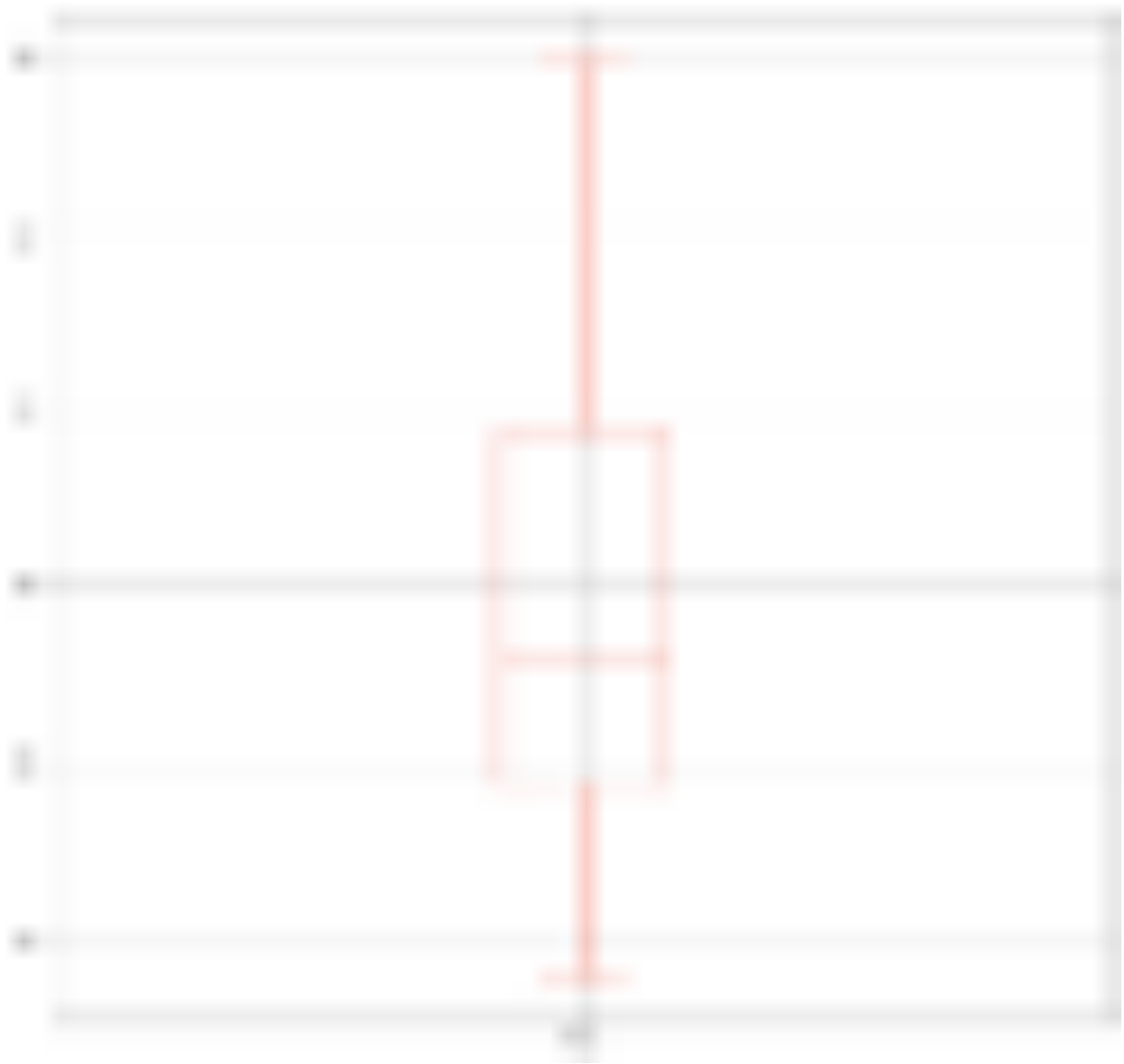


Image for post

Boxplot of Age — Graph 4

From this box-plot visualization, we can see that the maximum customer ages are between 30 and 35, and the 25-percentile and 75-percentile are close to 30 and 50. Therefore, we can conclude that the majority of customers of the mall are age between ages of 30 and 50, and age is one of the determinate of the spending scores.



Image for post

Graph 5

This Histogram of Annual Income shows that the annual income of all customers is positively skewed. And the range of customer annual income is between 15k and 140k. This shows that people with an annual income of 60k to 80k are the majority of customers, and people with higher income (>85) are not the major target for this shopping mall.



Image for post

Graph 6

This Histogram of Spending Score shows that the annual income of all customers is slightly positively skewed, and the maximum spending scores are between 40 and 60. This suggests that middle-class customers are a major source of income for this mall, which is similar to the result we conclude using the Histogram of Annual Income.

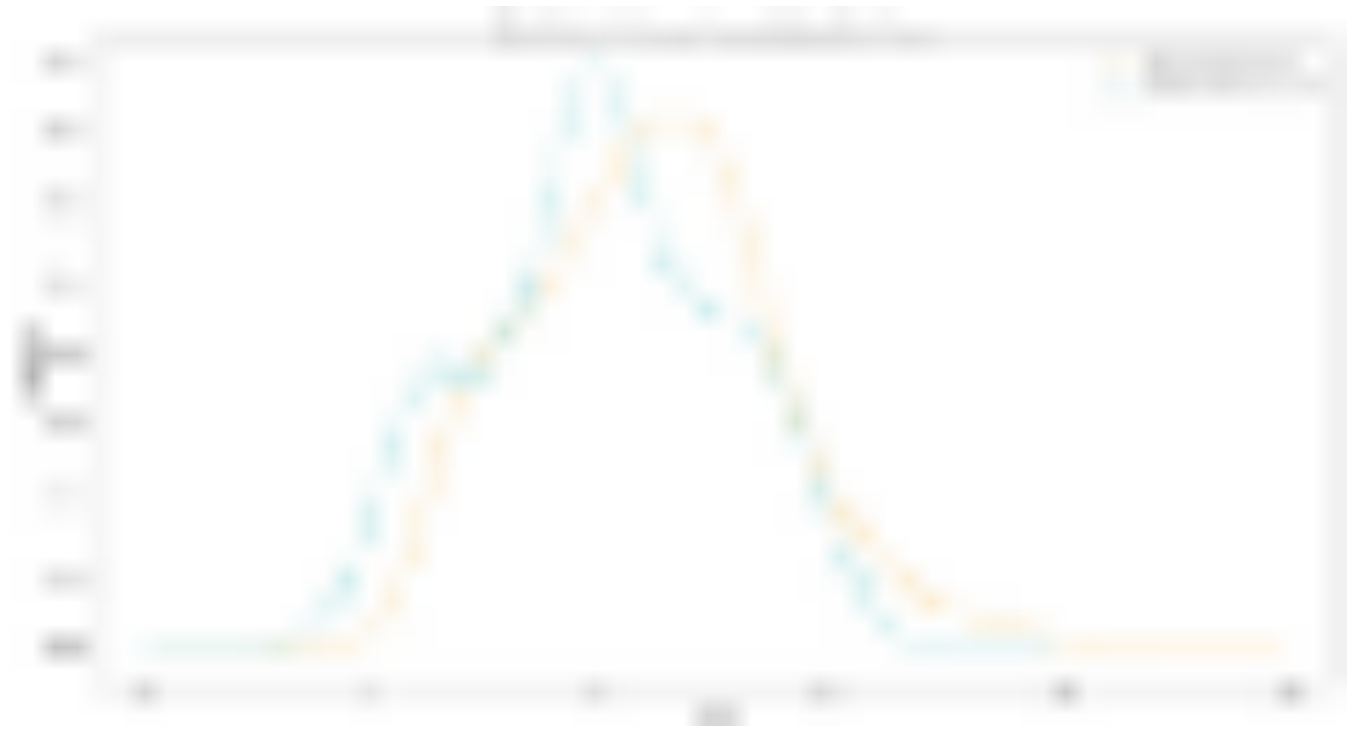


Image for post

Graph 7

From this graph, we can see that the maximum spending score is distributed in a lower class rank compared to the maximum annual income, but the shape of their distribution is generally the same. Therefore, we can conclude that the annual income of customers is a variable that directly influences their spending in the mall, and therefore positively relate to the mall's income and profit. Meanwhile, people tend to spend less than their income.





Image for post

Graph 8

This box-plot of annual income with different genders shows that the variance and standard deviation of males are smaller than females. Meanwhile, male customers have a slightly higher average annual income than female customers; whereas the range of annual incomes of these two groups is generally the same.

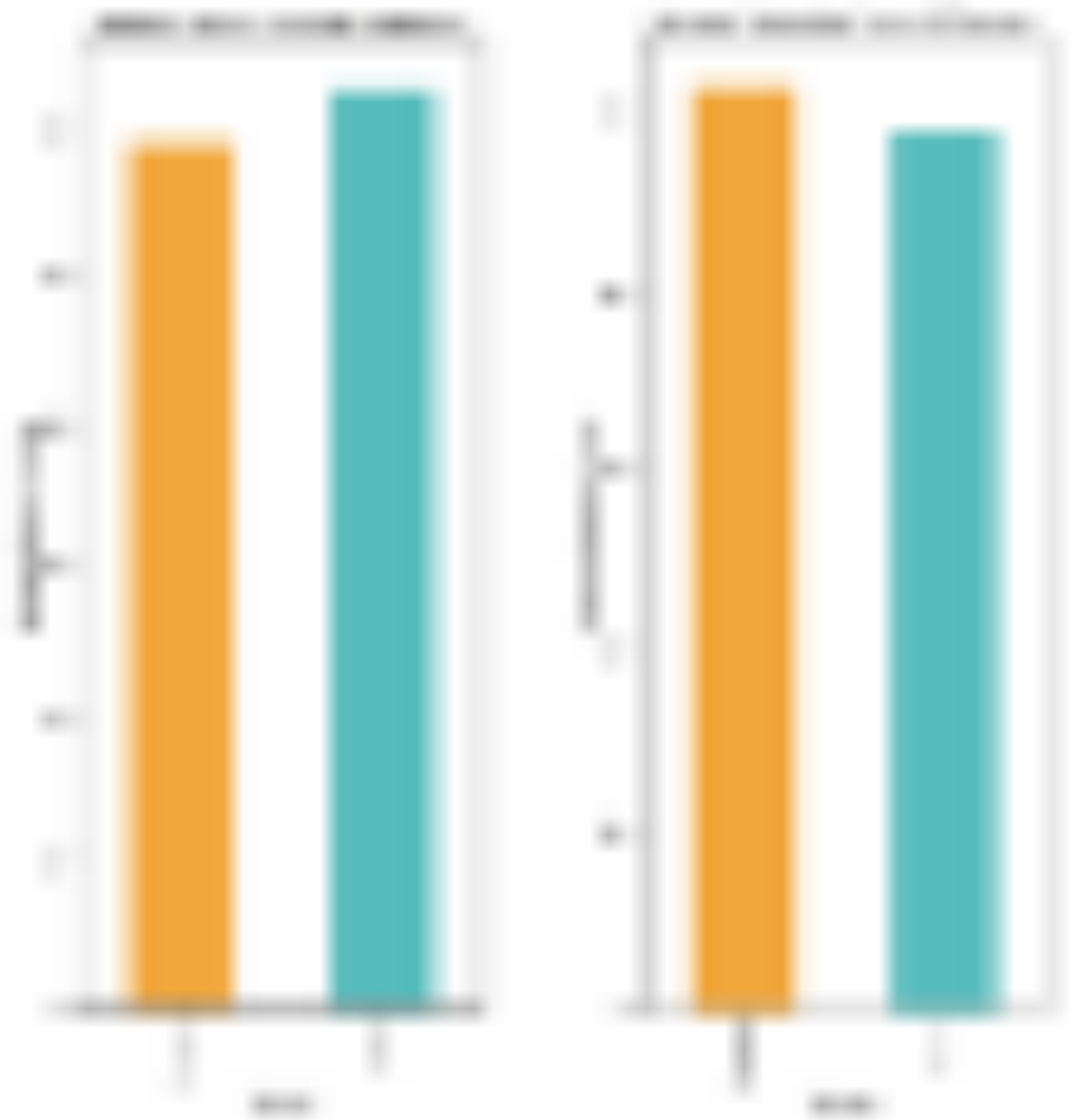


Image for post

Graph 9 & Graph 10

From these two graphs, we can see female customers have comparatively lower annual income compared to male customers. However, female customers have higher average spending scores compared to male customers. Therefore, we can conclude that gender is also one of the

determinants of the spending score, and it has a strong correlation with the income and profit of the mall.

## **Conclusion:**

In conclusion, gender, age, and annual income are three factors that influence the income of the shopping mall. Base on my project, I can conclude that female age between 25 to 40 who has an annual income of 60k to 90k is the group of customers with highest spending score, and contribute most profit to the shopping mall.

## **Work Cited:**

Dataset: Mall Customers Dataset,

<https://www.kaggle.com/shwetabh123/mall-customers>