Capstone Project #2 - Mall Customer Segmentation
Milestone Report

Overview

We all go to malls to shop. When said “we”, I meant everyone of us including people of any gender, age, socio-economic status, etc. On the other hand, all mall business owners might try to make any arrangements regarding their customers. If they know who usually visits their shop, customers age, gender, socio-economic status and any other indicators, they would be able to develop better marketing strategies and arrangements. This study aims at bringing a perspective on marketing strategies for mall business owners based on different features, such as age, gender, annual income, and spending scores.

Potential Stakeholders

The potential stakeholders for this study are mall business owners and customers who shop in those malls. This study will help mall business owners to develop better marketing strategies so that the customer will also get positively affected from these strategies. Other business owners might also use the model developed in this study to develop their own marketing strategies.

Data and Deliverables

The was obtained from a Kaggle competition (https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python). There are five features in total with 200 data points. However, CustomerID will not be considered as a feature because it is a label feature for customers. Figure 1 summarizes the first five data points in the dataset. The features are: gender, age, annual income in dollars, and spending score between 1 and 100. According to the Kaggle definition, ‘Spending Score’ is something you assign to the customer based on your defined parameters like customer behavior and purchasing data.

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Figure 1. Summary of the first five data points in the dataset

At the end of this study, I will deliver:

1. A proposal of the project
2. Data wrangling notebook with summaries
3. Data storytelling notebook with summaries
4. Statistical application on data with summaries
5. A milestone report
6. A notebook with applied model
7. A final report
8. Slide deck

Data Wrangling

In order to prepare the dataset for a further analysis, I applied some data cleaning and wrangling techniques. First of all, there was no null values in the dataset which has 200 data points and four analyzable features: gender (female - male), age, annual income (k$), and spending score (1-100).

Females are the 56% of all customers. While the maximum age of the customers is 70, the minimum is 18. I categorized age groups as seen in Figure 2:

```
In [9]: # Customer ages range between 18 and 70. I want to categorize customers based on their ages.
        #I will create labels for different ranges
        dfmallcustomer['Age'] = pd.cut(dfmallcustomer.Age, bins=[0, 25, 35, 60, 70],labels=['Young Adult','Adult',
                                                                                             'Middle Age','Older Adult'])
```

Figure 2. Age categories based on age groups

Doing this categorization, I found that most of the customers are middle age customers (42.5%) who are between 35 and 59 years old. Adults (25-34 years old) are 30%, young adults (17-24) are 19%, and older adults (60-70) are 8.5%.

In terms of the annual income, the maximum income is $137k while the minimum income is $15k. I created different income groups based on customers' annual incomes. Table 1 shows each group with their annual income range and the percentage of them among all customers. As seen in the table, the majority of customers earn between 35k and 84k (64%). The lower income group has the lowest percentage (8%) which makes sense because the lowest income might mean lowest interest in shopping in malls.

Table 1. Annual income groups with their income range and percentage among all customers

Income Group	**Annual Income Range ($)**	**Percentage (%)**
Lower	0-19k	8
Lower Middle	20k-34k	11
Middle	35k-64k	40
Upper Middle	65k-84k	24
Upper	85k and above	17

Finally, I did the same analysis for the spending score. As seen in Table 2, the majority of the customers have either good or very good scores (25-74). On the other hand, the lower and excellent scores are very close to each other (19%).

Table 2. Customer groups based on their spending scores

Spending Score Group	**Spending Score Range**	**Percentage (%)**
Lower	0-24	32
Good	25-49	29.5
Very Good	50-74	19.5
Excellent	75-100	19

As the final step, I saved my new dataset as CSV.

Data Storytelling

For data storytelling, I loaded both the original dataset and the dataset created after the data wrangling process. I wanted to create stories based on different datasets. In addition to some information provided in the data wrangling process above, I created some new stories. For example, in terms of the age of the customers, I found out that the most of the customers were between 18 and 50 years old (see Figure 3).

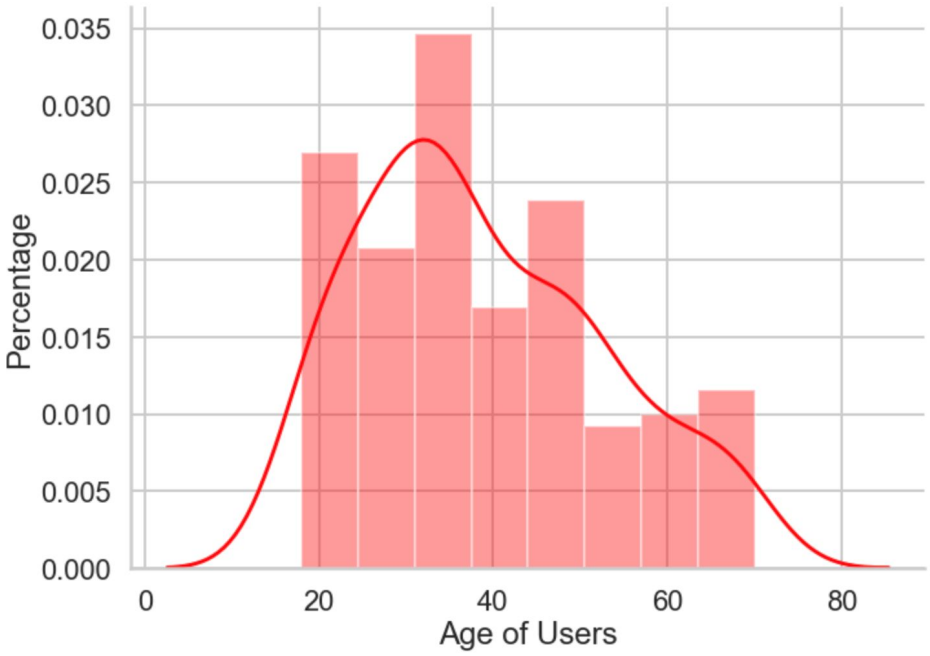

Figure 3. Customer age and percentages

I also did some binomial analysis. For example, I checked age groups based on gender using the dataset I created after the data wrangling. As seen in Figure 4, female customers are the majority for each age group except older adults. Similarly, females are the majority group in all income groups and in spending scores except those who have low spending scores.

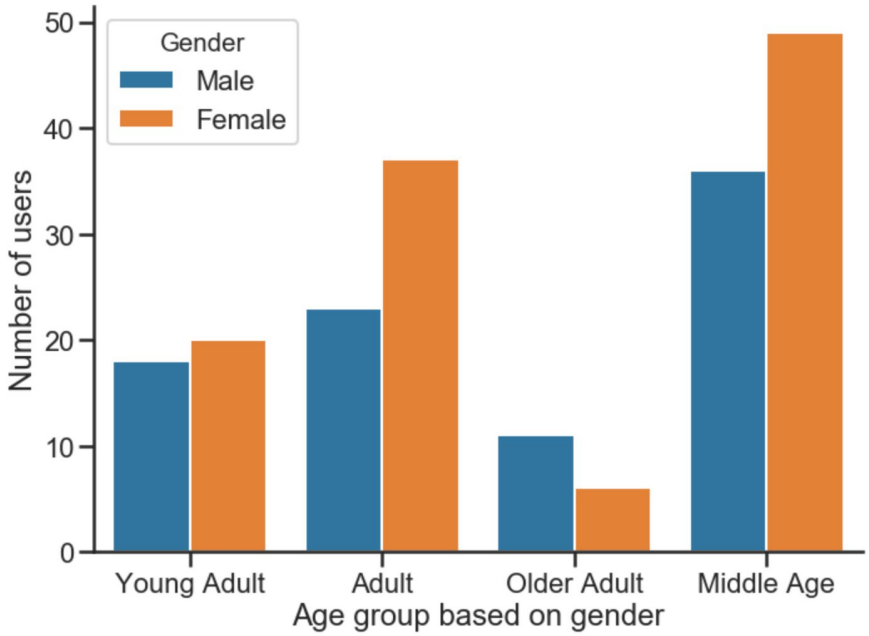

Figure 4. Age groups based on gender

I also checked the distribution of annual income based on age groups. As seen in Figure 5, middle age customers have middle annual income. In the upper group, there are young adults and older adults. This totally makes sense because older adults might probably get retired while young adults either do not have a job or in their early career.

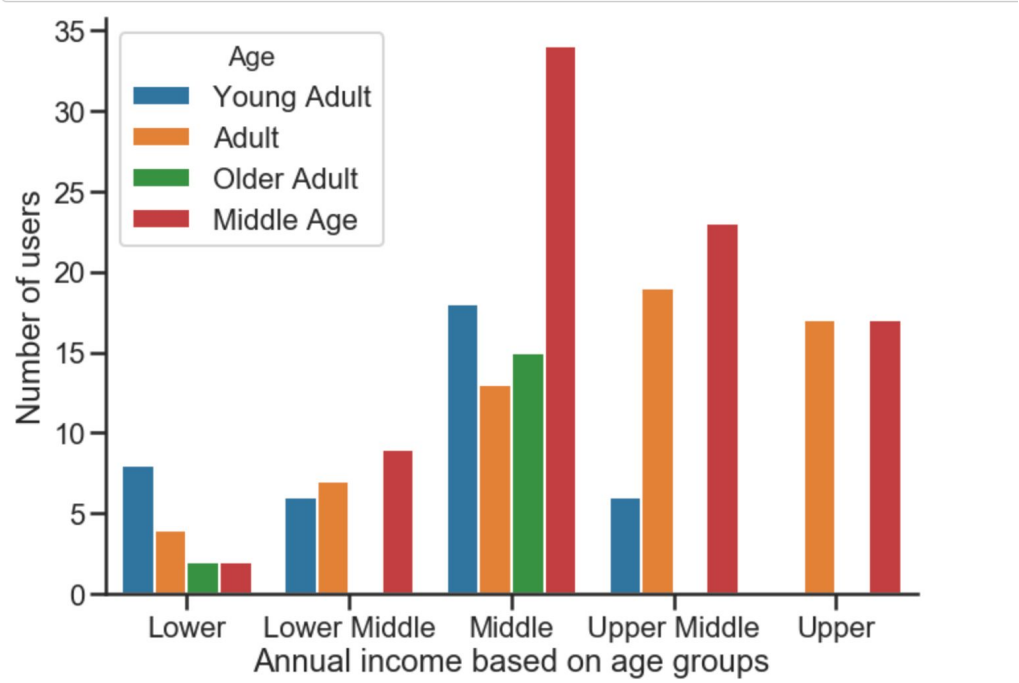

Figure 5. Annual income groups based on age groups

As the final step, I checked the correlation between the features to make a decision on whether or not a feature to include in my analysis. I created a heatmap to see all correlations (see Figure 6). Eliminating the customer ID as the label variable, it can be said that we can include all variables for our analysis because there is not good correlation between variables. The best correlation was seen between age and spending score, which is still not good (-0.33).

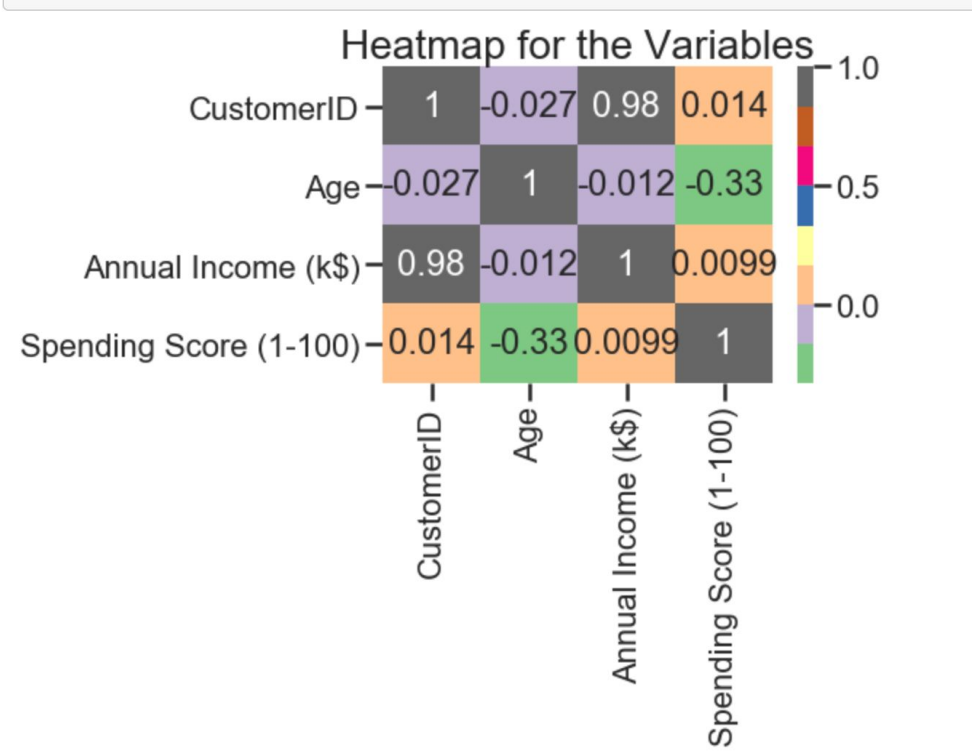

Figure 6. Correlation heatmap of features

Applying Inferential Statistics

In terms of inferential statistics, I used a frequentist approach and bootstrap methods for hypothesis testing. For the student's T-test, the null hypothesis was "there is no significant difference in spending scores between males and females." I found out that the difference between the means of the spending scores of males and females was 3.015422077922082 and the standard error was 0.0565. Figure 7 shows the calculation of t and p values manually and using scipy. As seen in Figure 7, the p-value is higher than 0.05, we reject null hypothesis and say that there is significant difference between the means of males and females ages. This aligns with what I found in data storytelling.

Standard Deviation Female: 24.007052468801657

In [12]:
```
#calculate t value manually
n0 = len(male_ss)
n1= len(female_ss)
std0 = male_ss.std()
std1= female_ss.std()
mean0 = mean_male
mean1= mean_female
sp = np.sqrt( ((n0-1)*(std0)**2 + (n1-1)*(std1)**2)/ (n0+n1-2) )
t_ = (mean1 - mean0)/(sp * np.sqrt(1/n0 + 1/n1))
print(t_)
```

```
0.8190464150660334
```

In [13]:
```
# Use 0.05 Significance level in two sample t-test
t_val=((male_ss_mean - female_ss_mean)-0)/SE
print(t_val)
```

```
-53.34416477675928
```

In [14]:
```
#calculate p value manually
p_value = (1 - t(n0 + n1 - 1).cdf(t_)) * 2
p_value
```

Out[14]:
```
0.4137397159674374
```

In [15]:
```
#calculate t and p values using scipy
ttest_ind(male_ss, female_ss)
```

Out[15]:
```
Ttest_indResult(statistic=-0.8190464150660333, pvalue=0.4137446589852176)
```

Figure 7. Calculation of t and p values manually and using scipy

For bootstrap and confidence interval, I performed 10000 replicates for males and females immediately after setting the random seed to 47, I got the value 47.215. The 95% confidence interval of the difference in standard deviation between females and males was [-0.16439370371649292 , 7.500105521493774]. The null hypothesis was “there is no difference in standard deviations between males and females.” Under the null hypothesis, I got a p-value of 0.7956. Thus it is sufficiently likely that the null hypothesis is true and thus I retain the null hypothesis. There is no significant difference in spending scores between females and males.