Exploring Marketing Strategies for Mall Business Owners

Springboard Data Science Intensive Program
Capstone Project #2

Remzi Kizilboga, 2020

Problem Statement

- Consider that you want to or already own a business in a mall
- **Goal:** Better serve customers and make more benefit
- **Solution:** Developing marketing strategies
- Marketing strategies can vary based on the type of business, work culture, the type of customers, systems, location, etc.
- This study aimed 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

- **Primary:** Current and future business owners
- **Secondary**: Customers

Data and Deliverables

- Kaggle competition

	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

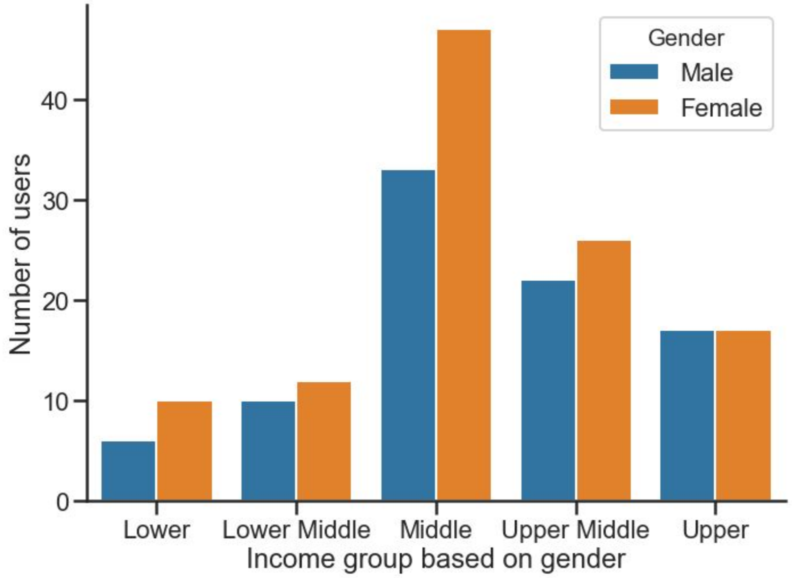
- Spending Score’ is something you assign to the customer based on your defined parameters like customer behavior and purchasing data.

Data Wrangling

- Explore each feature seperately
- **56% Female, 44% Male**
- Categorize and label age, annual income, and spending score
- **Age (18-70):** Young adult, Adult, **Middle age(42.5%),** Older adult
- **Annual Income(15k-137k):** Lower, Lower middle, **Middle (40%)**, Upper middle, Upper
- **Spending score (1-100)**: Low, **Good (32%),** Very good, Excellent

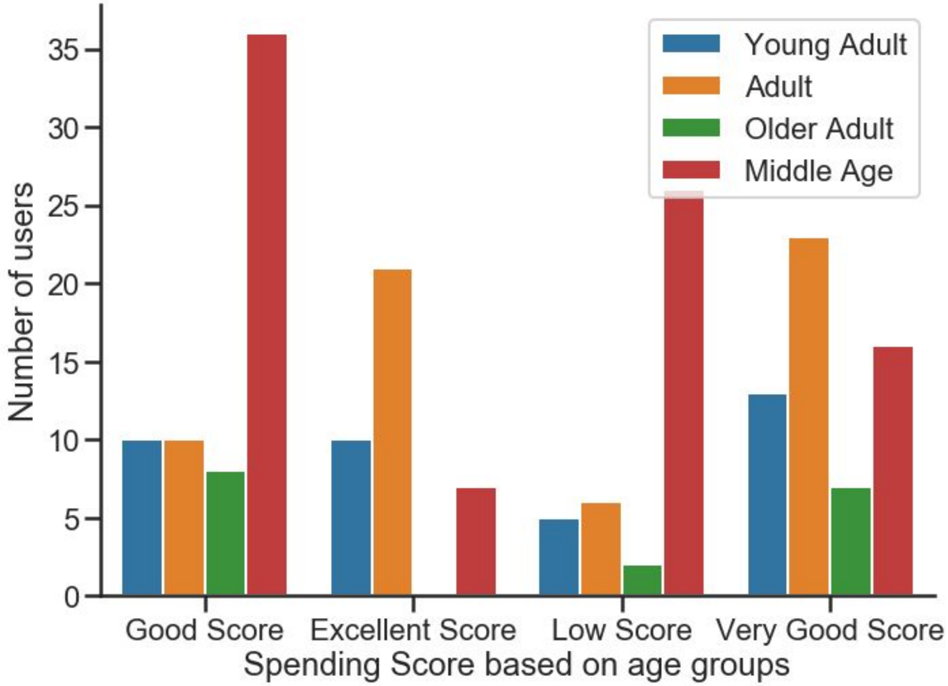
Data Storytelling

- Univariate analysis for each feature
- Bivariate analysis



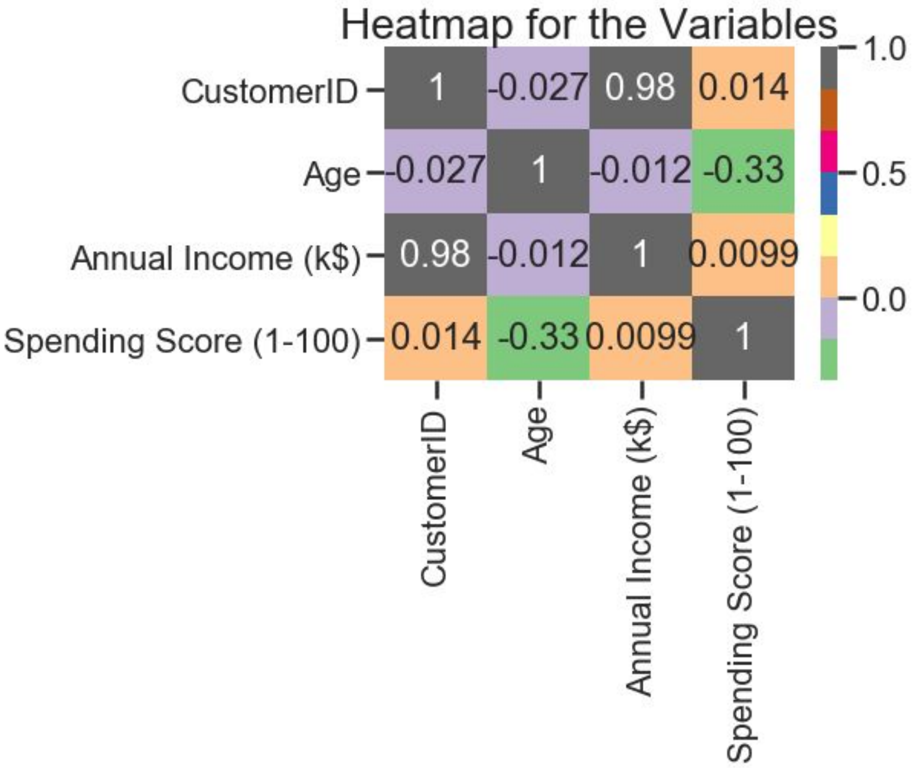
Data Storytelling

- Bivariate analysis
- Older adults do not have excellent score. Adults have the best scores (very good and excellent score)



Data Storytelling

- Correlation heatmap
- No correlation between features



Inferential Statistics - Student T-test

- Student-t test and Bootstrap
- Ho: there is no significant difference in spending scores between males and females Ha: there is a significant difference in the spending scores between males and females.
- Reject Ho

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)
```

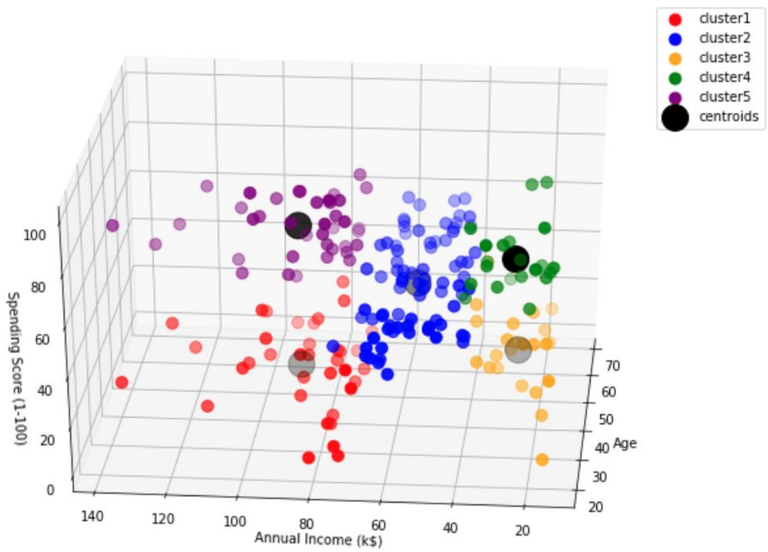
Inferential Statistics - Bootstrap

- Bootstrap
- Ho: there is no significant difference in spending scores between males and females Ha: there is a significant difference in the spending scores between males and females.
- Reject Ho

```
In [25]: # # Shifting the Dataset so that the two groups have equal means

         # First calculating the combined mean
         combined_mean = np.mean(np.concatenate((male_bts, female_bts)))

         # Generate the shifted dataset
         male_shifted = male_bts - np.mean(male_bts) + combined_mean
         female_shifted =  female_bts - np.mean(female_bts) + combined_mean

In [27]:
         # Draw the bootstrap replicates from the shifted dataset
         bs_replicates_male = draw_bs_reps(male_shifted, np.mean, size=10000)
         bs_replicates_female = draw_bs_reps(female_shifted, np.mean, size=10000)

In [28]: # Get the differences for the bootstrap simulated sample
         bs_differences = bs_replicates_male - bs_replicates_female


         # Get the observed difference from the actual dataset
         obs_diff = np.mean(male_bts) - np.mean(female_bts)
         obs_diff

Out[28]: -3.015422077922082

In [29]: # Calculate the p-value by comparing the bootstrap replicates against the observed difference of the means
         # The fraction of values WITHIN bootstrap replicates array that meet a certain criteria against the obs_diff

         p = np.sum(bs_differences >= obs_diff)/ len(bs_differences)
         print('p-value =', p)

         p-value = 0.7956
```

Model Application
KMeans

- Elbow method (k-5)
- Misclustered data points



INTERPRETATION: **Cluster1 (red cluster):** Generally people older than 40 years old with lower annual income and lower spending score.

Cluster2 (blue cluster): Generally people between 20-50 years old with more than 80k annual income and moderate spending score.

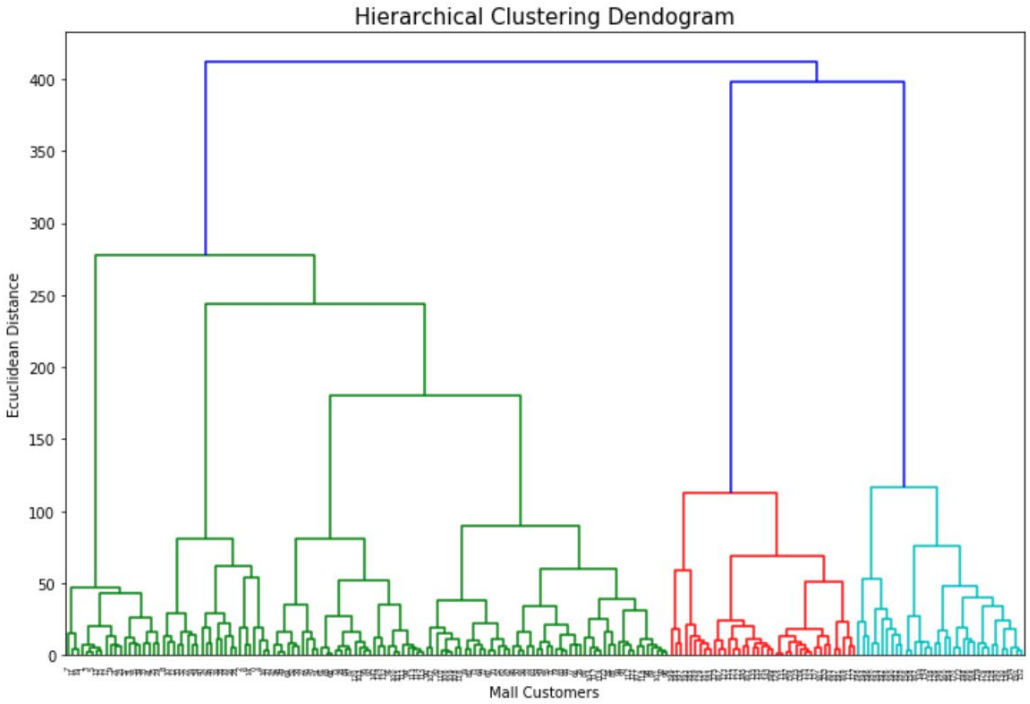
Cluster3 (orange cluster): These are the people who are spreaded in terms of age. It seems that their centroids are about 60 years old. But they have higher annual income (80k and more) but don't spend too much.

Cluster4 (green cluster): These customers do not have either very high or low annual income. They are spreaded to all ages. They also do not have either very high or low spending scores.

Cluster5 (purple cluster): These customers have the lowest population. They are younger with lower annual income. Their spending score is moderate.

Model Application Hierarchical

- Hierarchical Clustering Dendogram
- 3 clusters



Model Application KMeans

- 3 clusters
- More clear results
- Target younger people with higher annual income
- Explored each cluster one-by-one