Mall Dataset

Ryan Meston

Bellevue University

Abstract

For this project I am going to look into a mall dataset with information on customers annual income and there spending habits. In this dataset there spending habits are represented as a spending score to determine how willing they are to spend. Looking through this dataset I will be able to find insights on who I can market towards based on different spending groups.

Introduction and Background

CustomerID – customer identification number
Genre – gender of customer.
Age – age of customer.
Annual Income (k$) – annual income of customer.
Spending Score (1-100) – spending score, how much the customer is willing to spend.

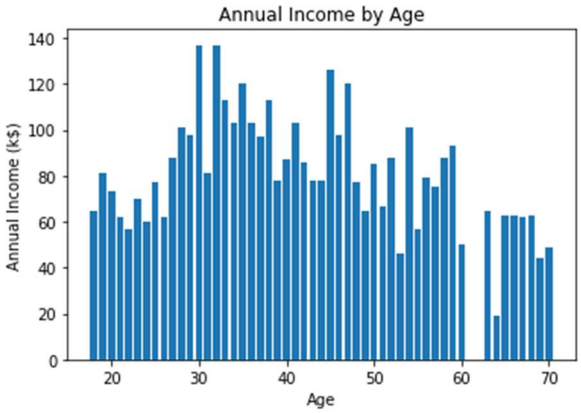
For this project I want to look into data that can help with identifying customers and targeting customers through marketing strategies. I am going to go through a dataset that has age, gender income, and spending likelihood to determine who would be a good customer to target adds towards while shopping. The questions I am going to ask are:

- Which customers are low, medium, and high value customers?
- Which gender and age group are more willing to spend money?
- Does income determine how much one will spend?
- What groups are the best target audience for ads and coupons?

Methodology

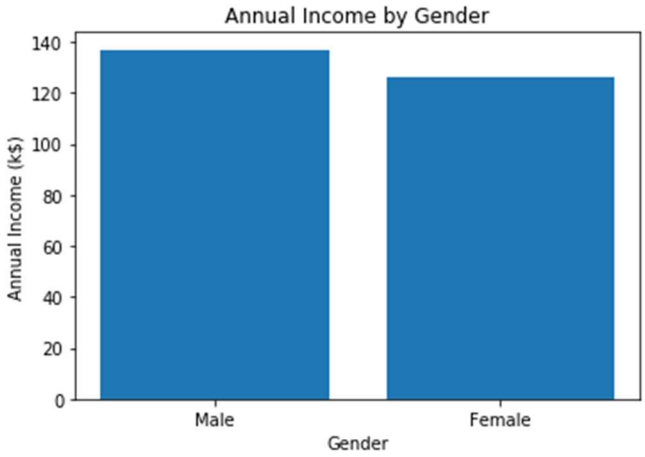
For this project I looked into K-means and customer segmentation to analyze this dataset. I want to use K-means to cluster information to find out how I can market towards different groups of people. I was able to use K-means to determine how I can market towards different groups within the dataset.

The below visual shows annual income by age.



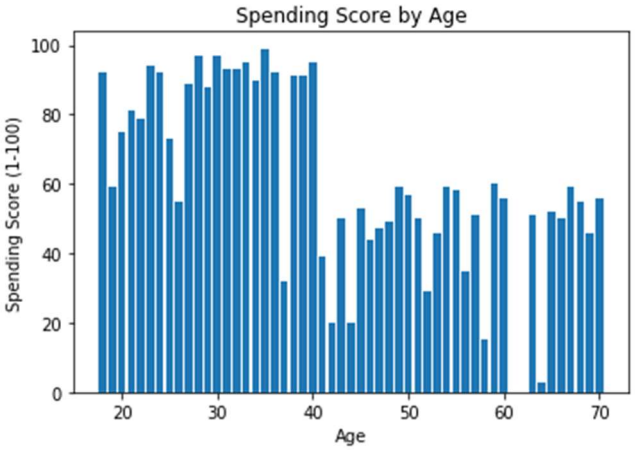
Here we can see that that the 30-50 range is what accounts for the most income for their age group.

The below visual represents annual Income by gender.



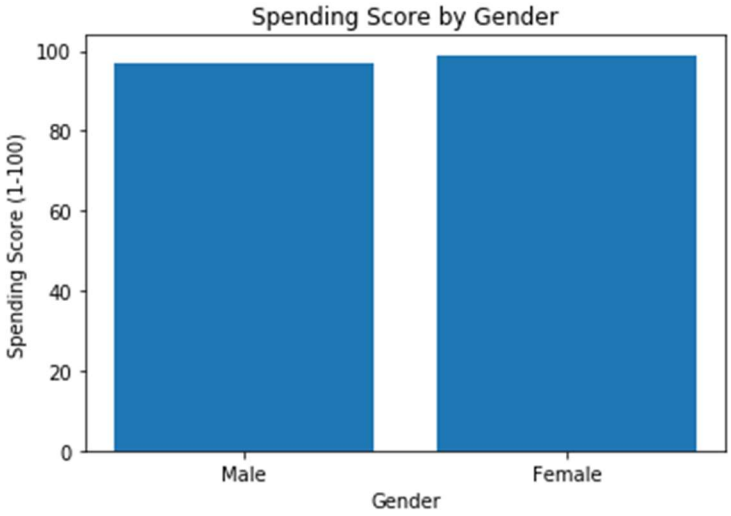
In this visual we can see that males in this dataset have a slightly higher income overall than the females in this dataset.

The below visual is a representation of spending score by age.



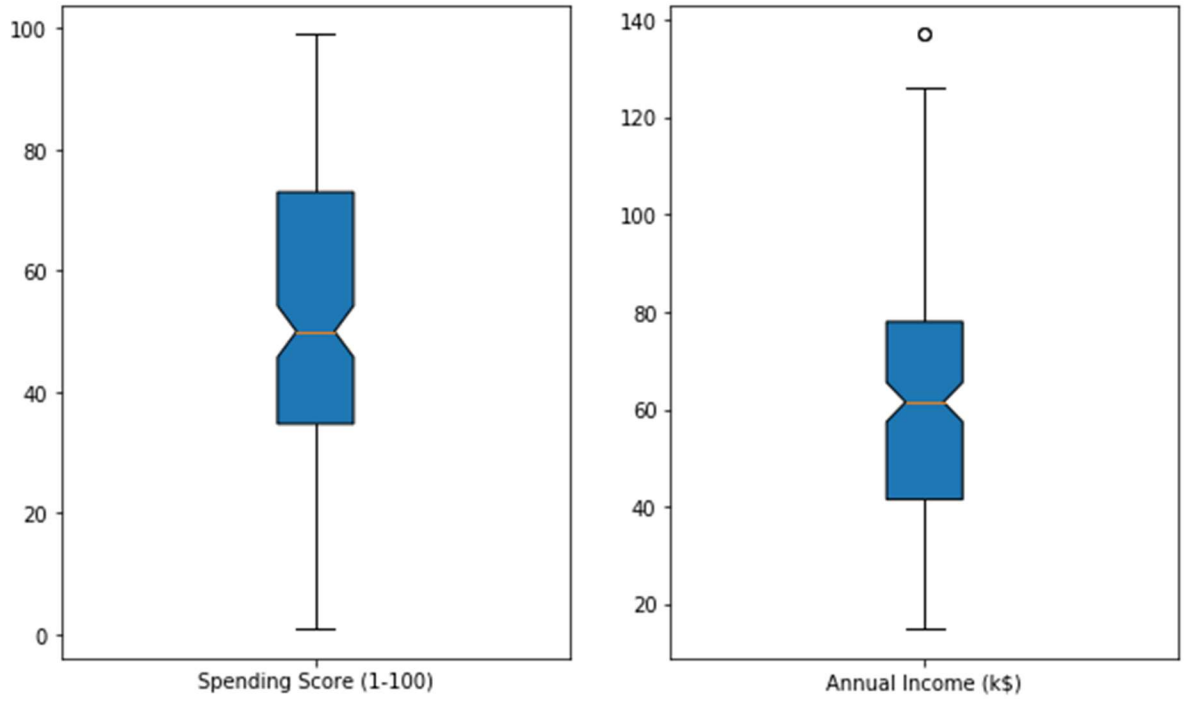
Here we can see that the spending score is much higher in the lower age group of this data. In our visual with Income by age group we saw that the most income was between 30-50, and in this visual we can see that the groups that spend the most are 19-40. There is some overlap between ages in both visuals, but the younger group tends to spend more.

The below visual is showing us the comparison between spending score by gender.



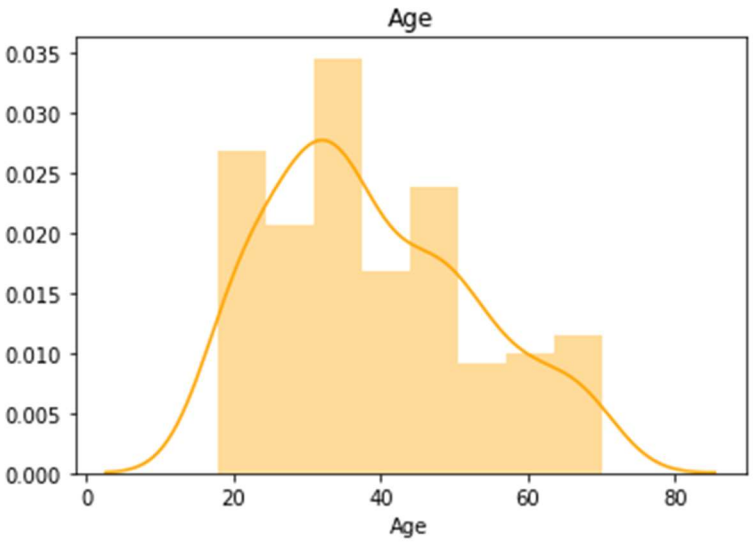
In this visual it really looks very close in spending habits between both males and females in this dataset. Both genders seem to have similar spending scores within the dataset.

Below is a boxplot to compare income and spending score.



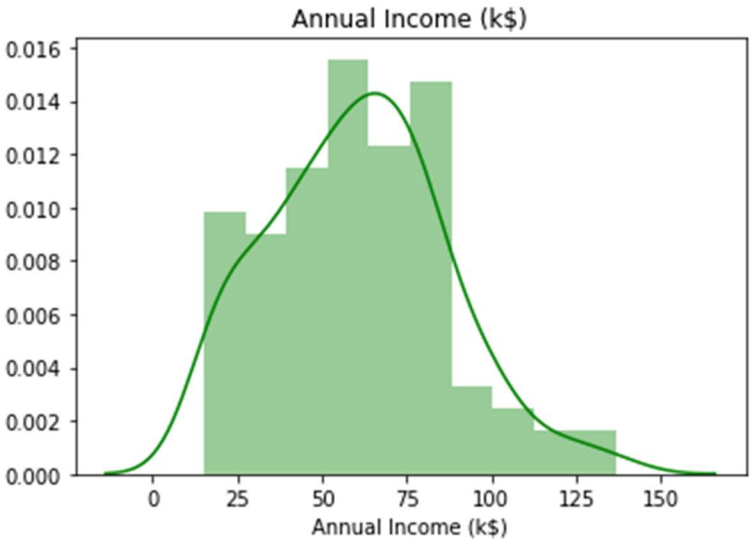
This visual is showing us more in terms of income versus spending. It looks like spending outpaces income. People are more willing to spend even if they do not have the money.

This visual shows us age distribution.



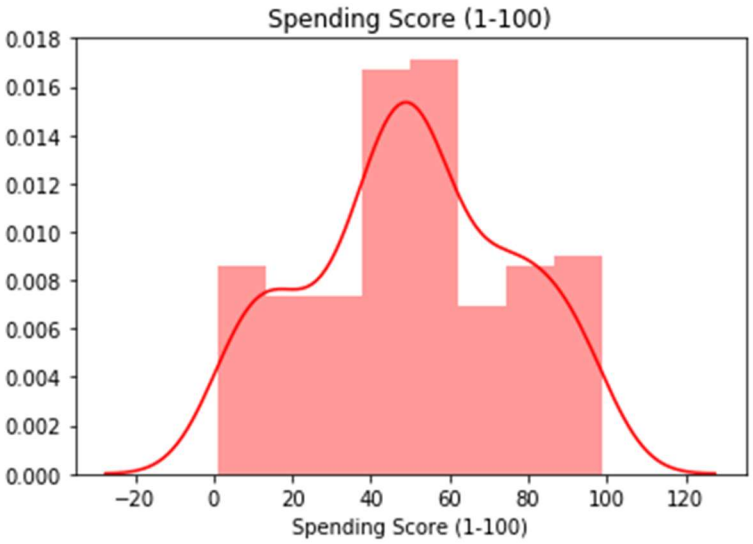
Here we can see that the primary age group in this dataset is right around 30-35 years of age.

Below we can see a visual showing annual income distribution.



Here we can see that most people's income within the dataset is right around 50,000-60,000 a year.

In the spending score distribution visual below we can see that the majority of the spending scores is in the 40-60 range. The majority of the people in this dataset our roughly 50/50 on whether they will spend money or not.

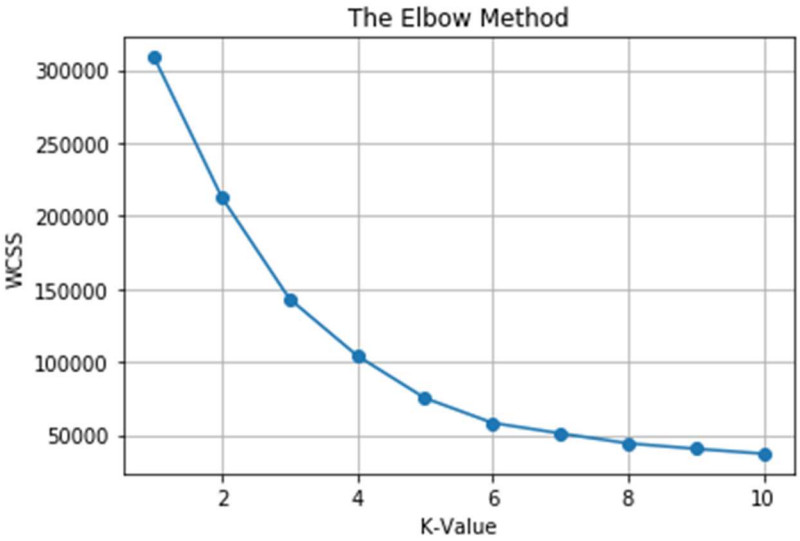


I also looked in trying a heatmap style visual based off of annual income and the spending score with age being my heatmap factor.

This visual is when I noticed the grouping and decide on going the K-means route. When we look at the above visual, we can see that we are starting to develop 5 groups. In the visual above I noticed the middle group had a concentration of many of the same ages within the dataset.

Results

I went ahead and used the elbow method to determine what my K-value would be.



I didn't think there was a distinct elbow in this visual, it looked very close to 4 or 5. I ended up using 5 as my K-value for this project.

Below is the visual of groupings after utilizing K-means algorithm with my K-value as 5. We can see we have 5 different groupings or patterns in this visual. The red group or group 0 can be seen as our lowest income and least likely to spend group. The orange group or group 1 is the group that is the most likely to spend and has the most annual income.

The questions I wanted to answer with this project are as follows:

- Which customers are low, medium, and high value customers?
    - Red Group – lowest annual income and spending score.
    - Green Group – above average annual income with a low spending score.
    - Yellow Group – average annual income and average spending score.
    - Blue Group – low annual income and high spending score.
        - This group loves to spend money, but how much do they have to spend?
    - Orange Group – high annual income and high spending score.
        - I would target the orange clients with marketing ads to increase consumer spending at stores.
- Which gender and age group are more willing to spend money?
    - Females spend slightly more than males do in this dataset.
    - 19–40-year-olds are the most willing to spend money.
- Does income determine how much one will spend?
    - With this dataset I do not believe income relates to how much someone will spend. There are different groups that have a high annual income and have a low spending score. There are also groups with low annual income and very high spending scores.
- What groups are the best target audience for ads and coupons?
    - The Orange Group is the best group in this scenario to target for ads and other marketing techniques. This group has the highest annual income and the highest sending score.

References

1. Thakur, Shivashish. 2020. KDnuggets.com. 21 Machine Learning Projects – Datasets Included. https://www.kdnuggets.com/2020/03/20-machine-learning-datasets-project-ideas.html
2. Kaggle.com. Mall Customer Data CSV File. https://www.kaggle.com/shwetabh123/mall-customers
3. Khalid, Irfan Alghani. June 1, 2020. Customer Segmentation in Python. https://towardsdatascience.com/customer-segmentation-in-python-9c15acf6f945
4. Adithyan, Nikhil. November 20, 2020. Customer Segmentation with K-Means Python https://medium.com/codex/customer-segmentation-with-k-means-in-python-18336fb915be
5. Scikit-learn. Package Information for Clustering and K-Means. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
6. Analytica Vidhya. April 26, 2021. K-Means Clustering Simplified in Python. https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/
7. Analytica Vidhya. January 28,2021. Profiling Market Segments Using K-Means Clustering. https://www.analyticsvidhya.com/blog/2021/01/profiling-market-segments-using-k-means-clustering/
8. Karaman, Baris. May 4, 2019. Customer Segmentation. https://towardsdatascience.com/data-driven-growth-with-python-part-2-customer-segmentation-5c019d150444
9. Nemke, Mike. June 5, 2021. Customer Segmentation: A Technical Guide with Python Examples. https://www.mktr.ai/applications-and-methods-in-data-science-customer-segmentation/
10. Sagar, Abhinav. 2019. KDnuggets.com. Customer Segmentation Using K-Means Clustering. https://www.kdnuggets.com/2019/11/customer-segmentation-using-k-means-clustering.html

Appendix

For this project I was able to analyze a dataset and use clustering and K-means to find out what groups would be the best suited for marketing towards. This project helped me learn more about clustering and how useful it can be in finding patterns in your data. I was able to analyze a lot of my data with visualizations to gain insights into my data to determine different outcomes.

One visual I created utilized annual income and spending score with age as a heatmap option. I was hoping to have more time to get a more detailed visual out of Power BI to show the age groups of each 'hotspot' as you scrolled over the visual. I am happy with what I was able to do in this project and I was able to use different resources to find ways to solve problems within the dataset to gain insights and target specific groups for marketing.

10 Questions

1. What age group/age range makes the most money?
2. What age group/age range spends the most money?
3. What age group/age range makes the least money?
4. What age group/age range spends the least money?
5. After we determine clustering, how can we identify customers to market towards?
6. What causes groups with lower incomes to have a higher spending score?
7. What cause groups with higher incomes to have a lower spending score?
8. What types of items our customers spending their money on?
9. Do customers spend more on needs or wants?
10. How can you predict spending score of a customer?