# Which Domain?

*What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you’re doing with these data.*

The domain comes from a Kaggle page where data has been collected for segmentation project purposes. The dataset can be accessed at this URL: <https://bit.ly/2RXienx>

1. Choudhary, V. (2018, August 11). Mall Customer Segmentation Data. Retrieved from https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python
   1. This is our main data source.
2. Team, D. F. (2019, May 3). Data Science K-means Clustering - In-depth Tutorial with Example. Retrieved from <https://data-flair.training/blogs/k-means-clustering-tutorial/>
   1. This article from Data Flair discusses K-means clustering as an unsupervised learning algorithm.
3. Data Novia. (n.d.). K-Means Clustering in R: Algorithm and Practical Examples. Retrieved from <https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/>
   1. A much more in-depth look at K-means clustering use and practical examples.
4. Nguyen, Tien Anh. (2018, July 3). Customer Segmentation: A Step By Step Guide For B2B. Retrieved from https://openviewpartners.com/blog/customer-segmentation/#.Xp9P1IhKgmI
   1. This is a great general overview of customer segmentation as a business concept, without getting into the programming aspect.
5. Shopify. (n.d.). Customer Segmentation. Retrieved from <https://www.shopify.com/encyclopedia/customer-segmentation>
   1. Another resource for understanding customer engagement from a business perspective.
6. Hanlon, Annmarie. (2020, April 10). The segmentation, targeting and positioning model. Retrieved from <https://www.smartinsights.com/digital-marketing-strategy/customer-segmentation-targeting/segmentation-targeting-and-positioning/>
   1. A useful review of how to translate consumer segmentation into a marketing strategy.
7. Yadav, Jyoti. (2019, Aug 16). Selecting optimal number of clusters in KMeans Algorithm(Silhouette Score). Retrieved from <https://medium.com/@jyotiyadav99111/selecting-optimal-number-of-clusters-in-kmeans-algorithm-silhouette-score-c0d9ebb11308>
   1. Using the silhouette method to visualize with k-means clustering.
8. Fonseca, Luiz. (2019, Aug 15). Estimating the number of clusters in a data set via the gap statistic. Retrieved from <https://statweb.stanford.edu/~gwalther/gap>
   1. This is a longer article discussing the use of the Gap Statistic Method for compare intracluster variation with expected values.
9. Tibshirani, Robert, Walther, Guenther, and Hastie, Trevor. (2000, Nov). Clustering Analysis in R using K-means. Retrieved from <https://towardsdatascience.com/clustering-analysis-in-r-using-k-means-73eca4fb7967>
   1. Another resource for K-means usage and a real-world example.
10. DataCamp. (n.d.). Cluster Analysis. Retrieved from <https://www.statmethods.net/advstats/cluster.html>
    1. Various cluster methods using the R language specifically.

# Which Data?

*What is the dataset you’ll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.*

The data being used consists of customer information for a shopping mall business. The data features included are age, gender, annual income, and spending score. Each customer is assigned a Customer ID for identification purposes. The spending score is a value provided for each customer which helps us to define their spending behavior. A customer who spends more has a higher score than one who spends less. There are a total of 200 observations in our data set.

# Research Questions? Benefits? Why analyze these data?

*How are you proposing to analyze this dataset? This is about your approach. Here, you’ll be proposing your research questions as well as justifications for why you’d offer these data in this way.*

Customer segmentation is a process which allows a business to divide their customer base into specific groups based on similarities such as gender, age, shopping habits, and more. Using machine learning algorithms, I plan to segment customers of a shopping mall in this manner. With segmentation, companies can gain a deeper understanding of their user base and meet the specific needs of those groups.

# What Method?

*What methods will you be using? What will those methods provide in terms of analysis? How is this useful?*

Customer segmentation is a method of unsupervised learning. We want to utilize a clustering technique for this data to identify our targets customers from the user base. Specifically, I plan to enlist K-means clustering for this project. K-mean clustering is best when you are working with an unlabeled dataset, which we are in this scenario. With this method we can group our data points in terms of their similarity.

# Potential Issues?

*What challenges do you anticipate having? What could cause this project to go off schedule?*

Having worked with K-means clustering in multiple assignments and a previous project, I feel mostly comfortable with its use. I do not anticipate having trouble with this method as its applied to my data set. As with any new data set, once I get into the data exploration phase it is possible that I could be presented with a surprise or two that may slow my analysis down. However, I feel pretty confident in my ability to work through the roadblocks that could be presented. I also feel slightly more comfortable in R than Python at this point in my program, so I look forward to completing this project in R.

# Concluding Remarks

*Tie it all together. Think of this section as your final report’s abstract.*

For my third project I plan to apply the technique of customer segmentation to a data set consisting of mall shopper behavior. Customer segmentation is an incredibly valuable business tool in today’s world. This method enables businesses to target those groups in order to make their marketing efforts most effective. With our results, we can take a look at target customers and ways that this marketing method translates into the real world.