Business Problem:

This project will take in customer data and group customers based on their demographic information. This will allow us to gain insights on customers for the business's future targeted advertisements. A company is planning on cutting their ad costs and advertise only to those who are more likely to purchase their products. To do this they will need models to inform them which ads should be used on different groups of customers.

Background/History:

Customer segmentation is used to place customers into subcategories (Twillo, n.d.). In our case we will be splitting the customers into subcategories for targeted advertisement, and this will be a demographic segmentation which will allow us to send personalized ads/offers. Also, according to Twillo this will be a powerful segmentation and the first place that business should start when completing customer segmentation.

To get a better understanding of customer segmentation this is the process of splitting customers into groups based on shared characteristics (Vetrivel-PS, 2020). We will need to complete a demographic segmentation instead of behavioral because we don't have purchase history, website usage and so on. However, we have age, gender, education, family size, and profession, which are all demographic characteristics. There are many segmentation models that can be used for this and according to an article from Optimove, the best method for customer segmentation is through clustering, this allows for us to identify patterns in the data (Wyse, 2024). Clustering has been around for a while as well. This was first introduced in 1932 and allows us to discover the "unknown" about our data (Sharma, 2025).

Data Explanation:

The dataset that will be used is the Customer Segmentation dataset from kaggle.com: https://www.kaggle.com/datasets/vetrirah/customer?select=Train.csv
This dataset consists of 10 columns:

1. ID: Unique identifier for each customer.
2. Gender: Identifier for if the customer is male or female.
3. Ever Married: Identifies if the customer was ever married.
4. Age: The age of the customer.
5. Graduated: Identifies if the customer graduated or not.
6. Professional: The profession of the customer.
7. Work Experience: The work experience by customer in years.
8. Spending Score: A score used to identify how the customer spends
9. Family Size: The size of the customer's family, including the customer
10. Var_1: This is an anonymized category for the customer.

The first step completed in the data cleaning process was to review the dataset for null values. Nearly every column had null values. To preserve the integrity of the dataset, I chose to remove these rows instead of filling them in. Also, I couldn't be sure that if I filled in the missing values that the new values would be close to what they should be for that customer. The second step was to review histograms for their distributions. While I was able to fix the left skew of the age category with the square root function, I was unable to fix the skew for any of the remaining columns as square root and log were too powerful of a transformation. Next, I removed the outliers on the integer-based columns (age, work experience, family size) based on the interquartile length of their box plots. Finally, for all the columns that contained strings, I converted the string value into categories spanning from 1 to 10.

Methods:

The first method is to explore the data through a traditional exploratory analysis to see if this would provide any powerful insights into the business problem. As we do not have a specific target column to review for in the data, we will also be building an unsupervised model to gain additional insights into the data. The first model will be a K-Means clustering model built from scratch and the second will be the K-Means model provided by Scikit Learn. K-Means was chosen for the model based on the use cases provided on the Scikit Learn website. This being that the model works as a general-purpose model and can work with many samples.

Analysis:

From the first method, I was to complete an exploratory data analysis to gain insights on the data. However, this method didn't provide many insights into the data set. Based on the scatter plots for spending score by age and gender below, there are no correlations between the data.

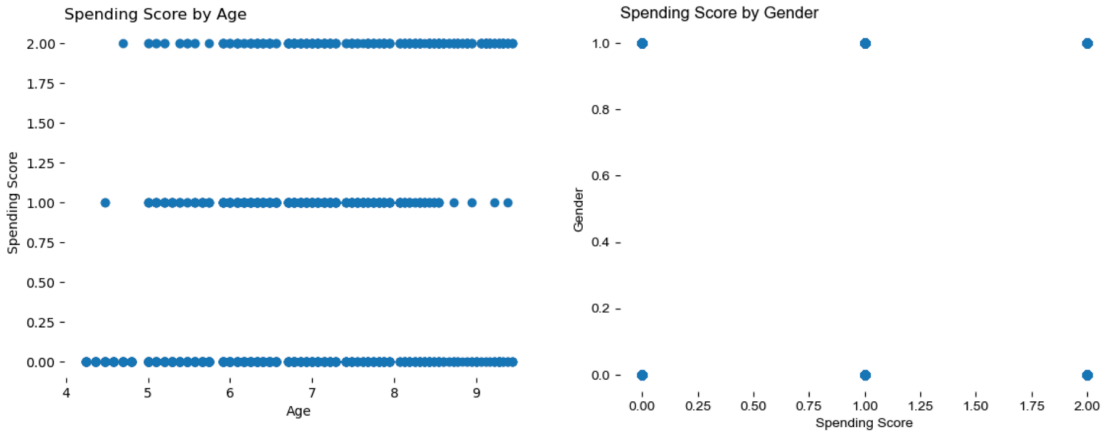


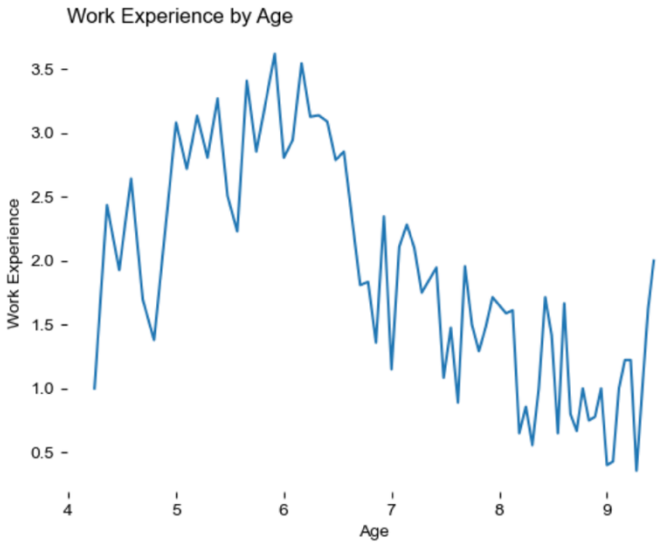
Figure 1-2: Scatter plot for age and spending score (left), scatter plot for gender and spending score (Right)

Also, with this approach we were able to uncover two items from the data. The first being that there is an even split for the spending score between genders and that the population decreases as the spending score increases.

Figure 3: Spending score split among Genders.

This could imply that most of the customers this business receives is buying lower priced items or that the shoppers may be primarily spending based on sales. Another observation that we're able to find out about the data, seems to be a bit odd. The older customers have less work experience that the younger half of the data. This could provide insights that the groups with more work experience may be willing to spend more on the business's products.

Figure 4: Work experience by customer age.



Since I could not uncover much information from the exploratory data analysis, I decided to move onto the K-Means modeling. To choose the best number of clusters, I decided to us the elbow method based on the model's distortion. Using this method, the

best number of clusters for the model was 2. I did apply the number of clusters to both the K-Means model made from scratch and the K-Means model from the Scikit Learn package and received results that were close to one another. However, when checking the model for accuracy, I was only able to receive a silhouette score of 0.426 for the Scikit Learn model and 0.1987 for the model built from scratch which indicates that models were weak, and this can also be seen in the image of the model below.

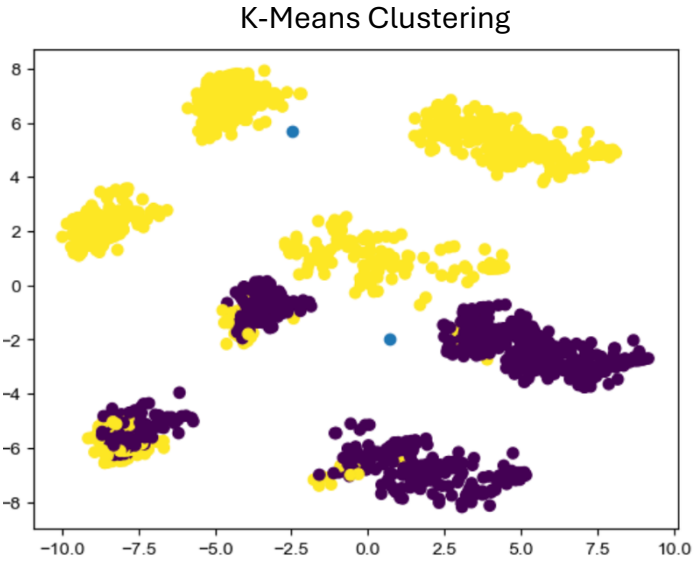


Figure 5: A Visual of the K-Means clustering model.

The different colored data points represent the two clusters, while the blue dots represent the centroids. With this we're able to come up with 2 different categories that the customer fall into.

Category 0: This consists of customers that are slightly more male than female but 62% of the customers were married, customers are slightly older, more than half likley graduated, are doctors, have a median spending score, low family size with an average of 2-3.

Category 1: This consists of an even split between male and female customers, less likely to have been married compared to category 0, are of an older age group, about 70%

have graduated, their profession is more than likely going to be a doctor, have a low amount of work experience, have a lower spending score with a family size of at least 2.

Conclusion:

The traditional exploratory data analysis didn't seem to yield meaningful information for the business problem. From this the K-means Method was able to separate customers into 2 different categories for the targeted advertisements. The K-Means model has a weak silhouette score of 0.426, which means that the model didn't work well with the dataset that was provided. However, this does provide information for the company to base their targeted advertisements on, but before the company invests in this, we should obtain additional data from customers to confirm if the model is a bad fit for the data or if there might have been something wrong in data that was provided.

Assumptions:

I assumed that there was no multicollinearity among the features of the dataset and that each feature column was independent of another. I also assumed that it was best to drop the missing values as I thought there was no guarantee the values that we're filling in for the customer would be accurate. For example, the occupation. I could have used the most occurring occupation to fill in the missing values but there were 8 categories to choose from, and I could be filling in lawyer while the occupation may be entertainment.

Limitations:

I do not feel that the data was of the best quality for this model. There weren't many features, and the data appears to have been randomly generated. This can be seen in the work experience visual. It doesn't seem to make sense that the younger customers have more work experience than the older customers. Also, the spending score based on

genders seemed to be all evenly spilt, but this doesn't seem to be realistic as one would expect to see some variance in the categories. This could have also impacted the silhouette scores making them appear to be weak.

Challenges:

The initial data set I chose to use for this project was too clean and it appeared that all the data was evenly split, leading to an uninteresting and unrealistic project. I switched to this data set after seeing that there was spread for clusters after being visualized for K-Means. However, this data too may have been generated in such a way that the first data set was, leading to almost even splits within the features of the data. For the silhouette score, I did check each point for its accuracy with a loop and 2 was the best as per the elbow plot. However, I was initially comparing the wrong data within the models and wasn't saving the random state when building the model which caused errors in the silhouette score.

Future Uses:

K-means is used for unsupervised learning, which means that the use cases in the future should have no target variable. Other cases that can use this method aside form customer segmentation would be models that take in a client's workout regimen and food diary to gain insights about their weight loss. Also, this can span to other cases like AI for image and document classification.

Recommendations:

Additional information and data should be provided to better improve the model. For example, the different categories of data. We could gain further insights into the business, and we can even split up the model to gain further insights into the customers for

each category. If the K-Means model doesn't improve after this then we can determine that the model is a bad fit for the data and move onto a model such as DBSCAN or Gaussian models.

Implementation Plan:

At this point in time the model should not be implemented until explored with the additional data as suggested in the recommendations section. We would need to ensure that the model is a good fit for the data as the current model has a weak silhouette score. Once a model is obtained that has a satisfactory accuracy score either through the data or a new model, the model can be deployed slowly with a test launch to see if the targeted advertisements suggested by the model have a significant impact on profits. If this does then we can fully deploy the model, but as always, we should ensure data is properly stored, secured, and consistently updated.

Ethical Assessment:

As the data was obtained from Kaggle, there's no restrictions for use with this project. The data is anonymized, and the customers were only identifiable by the ID number assigned to them. This makes it so that the data is safe from leaking customer information in case of data breaches. The model results may include bias as the categories seemed to favor the middle-aged customer groups, but this could be the target audience for the business as well. There seemed to be no bias in the model otherwise regarding gender or profession as each has a different profession and the genders of the customers were nearly equal.

References

2.3. clustering. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/clustering.html

Shakhti, D. (2024). 6 real-world customer segments examples that drive results. Retrieved from https://campaignrefinery.com/customer-segments-examples/#google_vignette

Sharma, N. (2025). K-means clustering explained. Retrieved from https://neptune.ai/blog/k-means-clustering#:~:text=Clustering%20was%20introduced%20in%201932,on%20their%20similarities%20and%20dissimilarities.

twillo. (n.d.). Customer segmentation models: The what, why & how. Retrieved from https://segment.com/growth-center/customer-segmentation/model/

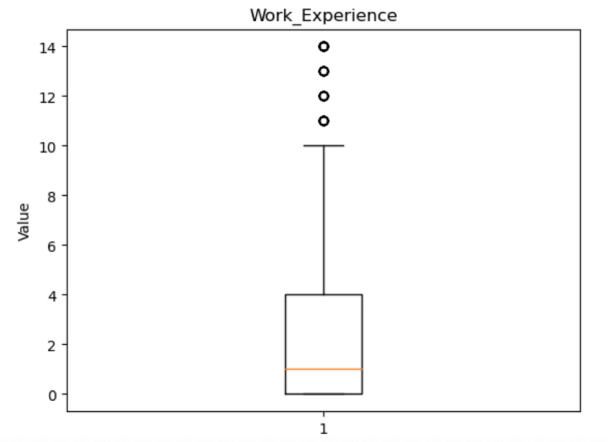
Vetrivel-PS. (2020). Customer segmentation. Retrieved from https://www.kaggle.com/datasets/vetrirah/customer?select=Train.csv

Wyse, R. (2024). 8 customer segmentation models: Unlock the power of Precision Marketing . Retrieved from https://www.optimove.com/blog/types-of-customer-segmentation-models

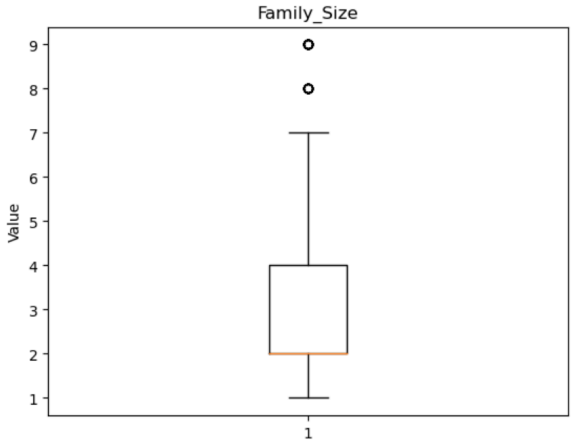
Appendix

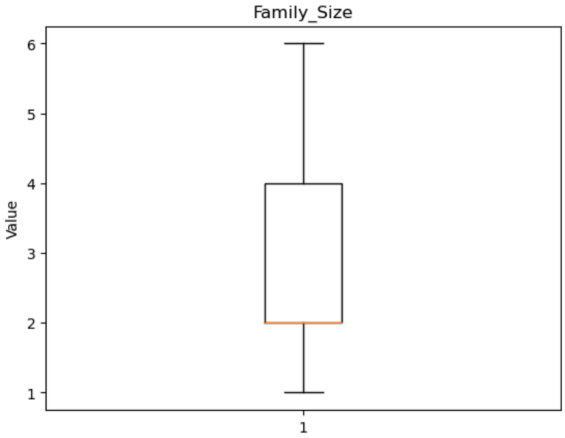
Box plots for removing outliers:

Work Experience before and after:

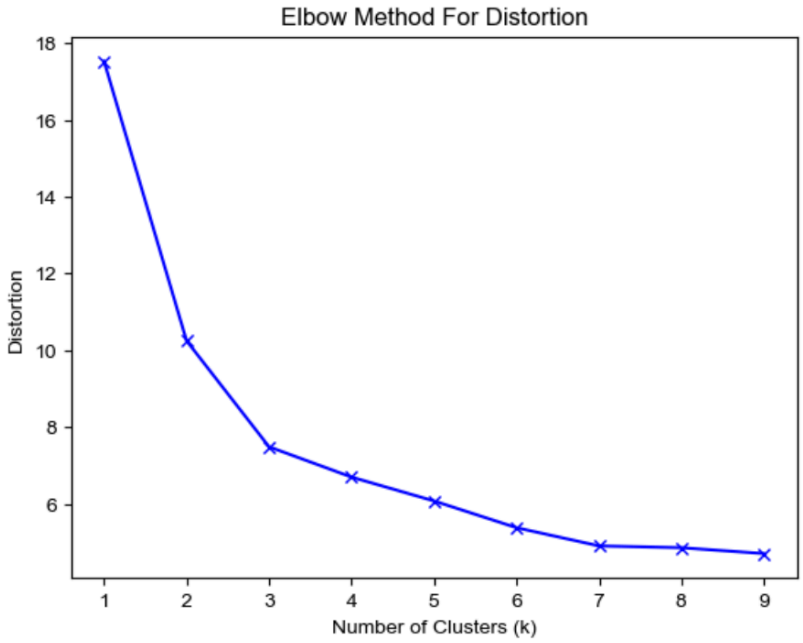


Family Size Before and after:





Elbow Method for K-Means:



Categories for the Scikit Learn model:

	0	1
Gender	0.434357	0.491159
Ever_Married	0.627694	0.530452
Age	6.623126	6.139788
Graduated	0.615284	0.683694
Profession	3.476813	3.532417
Work_Experience	0.702155	6.913556
Spending_Score	0.562378	0.465619
Family_Size	2.778576	2.589391

Silouette Score for this model:

```
Silhouette Score: 0.42635604013380296
```

Categories for the K-means model from scratch:

	0	1
Gender	1.000000	10.000000
Ever_Married	1.000000	1.000000
Age	4.609324	4.331694
Graduated	10.000000	10.000000
Profession	4.375000	5.500000
Work_Experience	1.000000	2.000000
Spending_Score	10.000000	5.500000
Family_Size	2.800000	6.400000

Silhouette score for this model:

```
Silhouette Score: 0.19873023374642873
```