Customer Segmentation: K-Means

Introduction

- Customer segmentation forms clusters from customer data based on their shared characteristics (Vetrivel-PS, 2020).
- According to Twillo, these clusters are subcategories of of customers.
- K-means clustering has been around since 1932 and helps us uncover the unknown from data (Sharma, 2025).
- As we’re making clusters of customers and K-means analyze clusters, K-means would be the best model in this situation.

Project Goal

- A Company is planning on cutting their advertising costs and would like to focus more on targeted advertisements.
- To do this they will need a clustering model to make groups of customers based on their demographics.

About The Data

- The Data set is the Customer Segmentation data set and was obtained from Kaggle.com.
- This data set contains 10 columns and 2,627 rows of data.
- The ID column and the VAR_1 column will be removed as the ID is to identify the customer and Var_1 is an anonymized category for the customer.

Data Preparation

- The dataset was reviewed for null values and rows with null values were removed.
- Histograms of the data were reviewed, and the right skew of Age was transformed to be normally distributed.
- Outliers were removed from age, work experience, and family size.
- String values were converted to integer values.

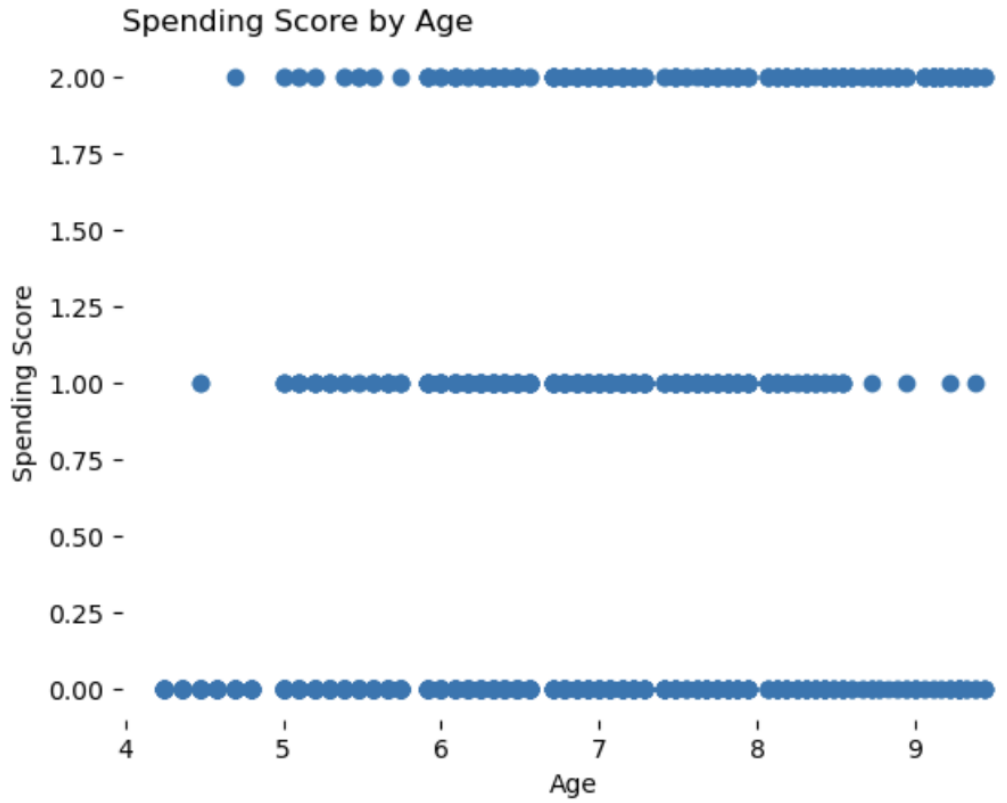
Model Used

EXPLORATORY DATA ANALYSIS.

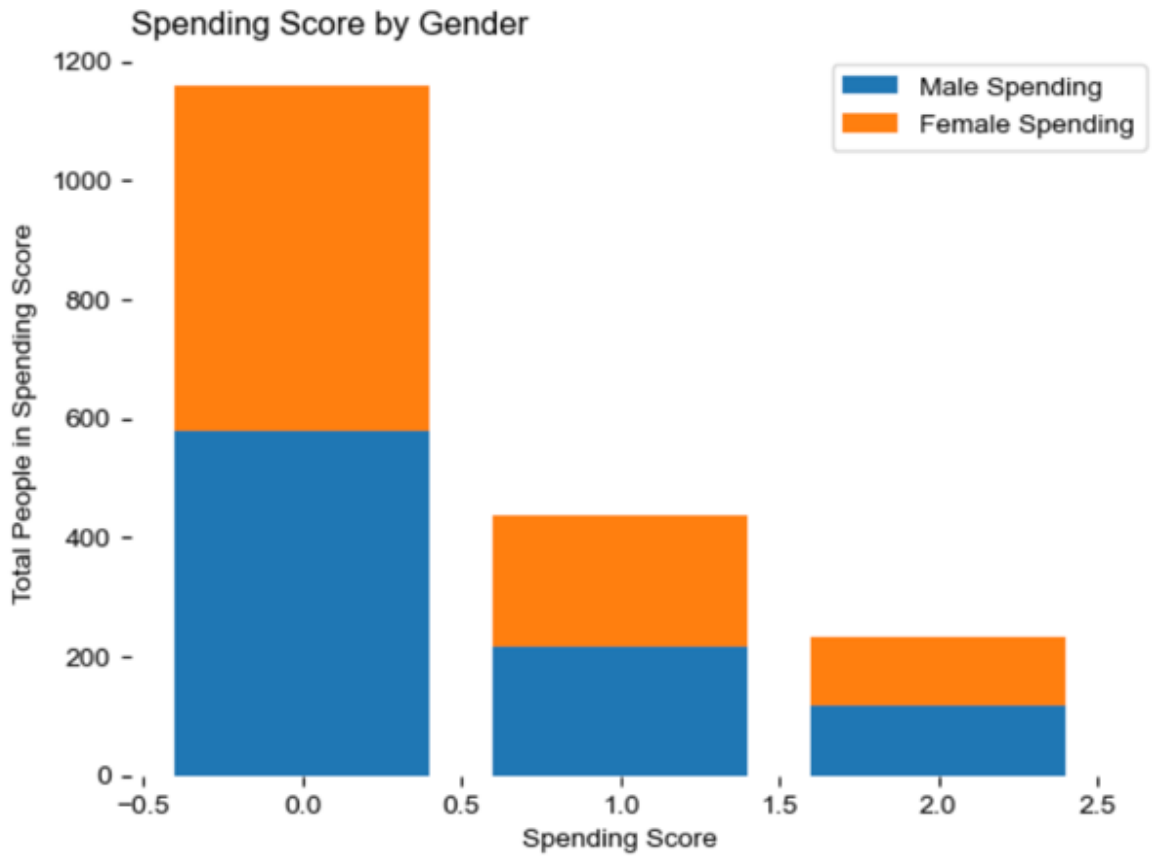
K-MEANS MODEL BUILT FROM SCRATCH.

K-MEANS MODEL FROM THE SCIKIT LEARN PACKAGE.

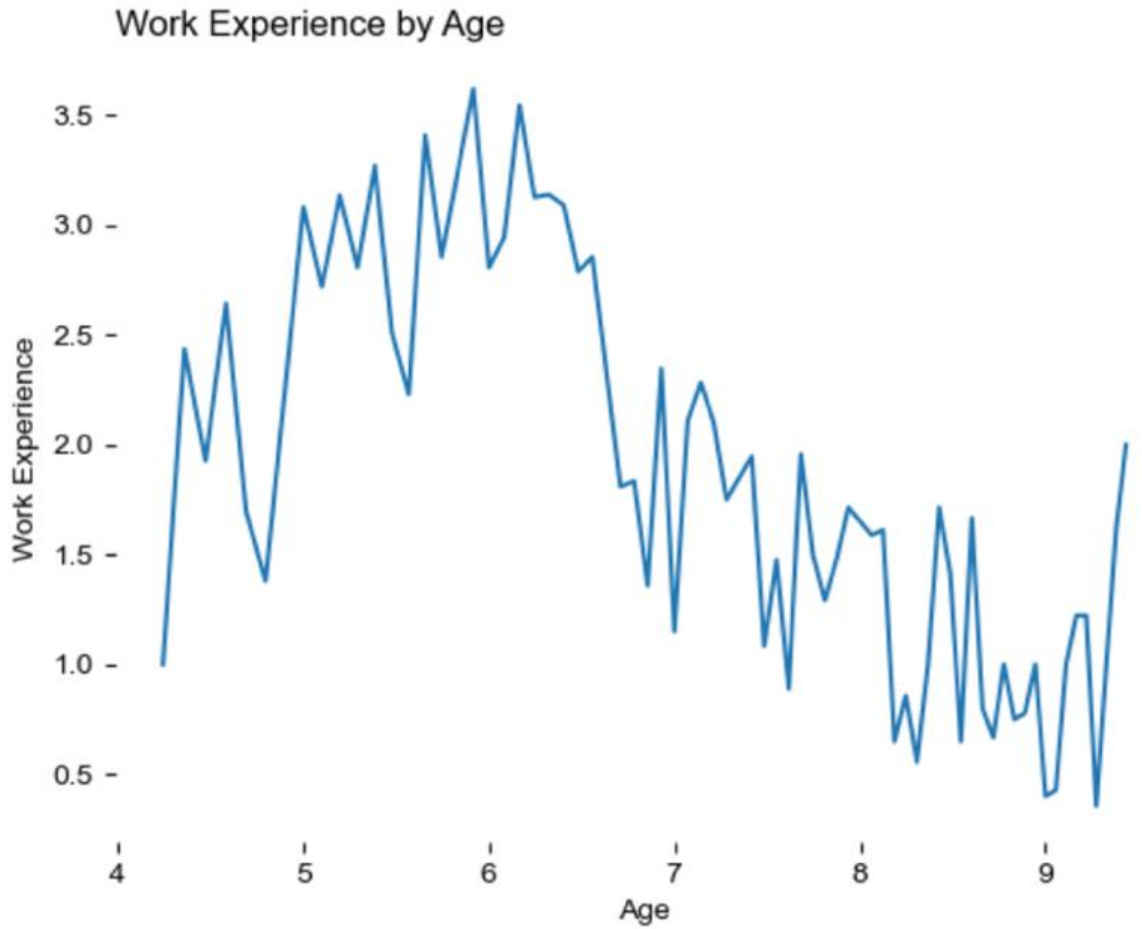
Exploratory Data Analysis (EDA)



EDA: Spending Score By Gender



EDA: Work Experience by Age

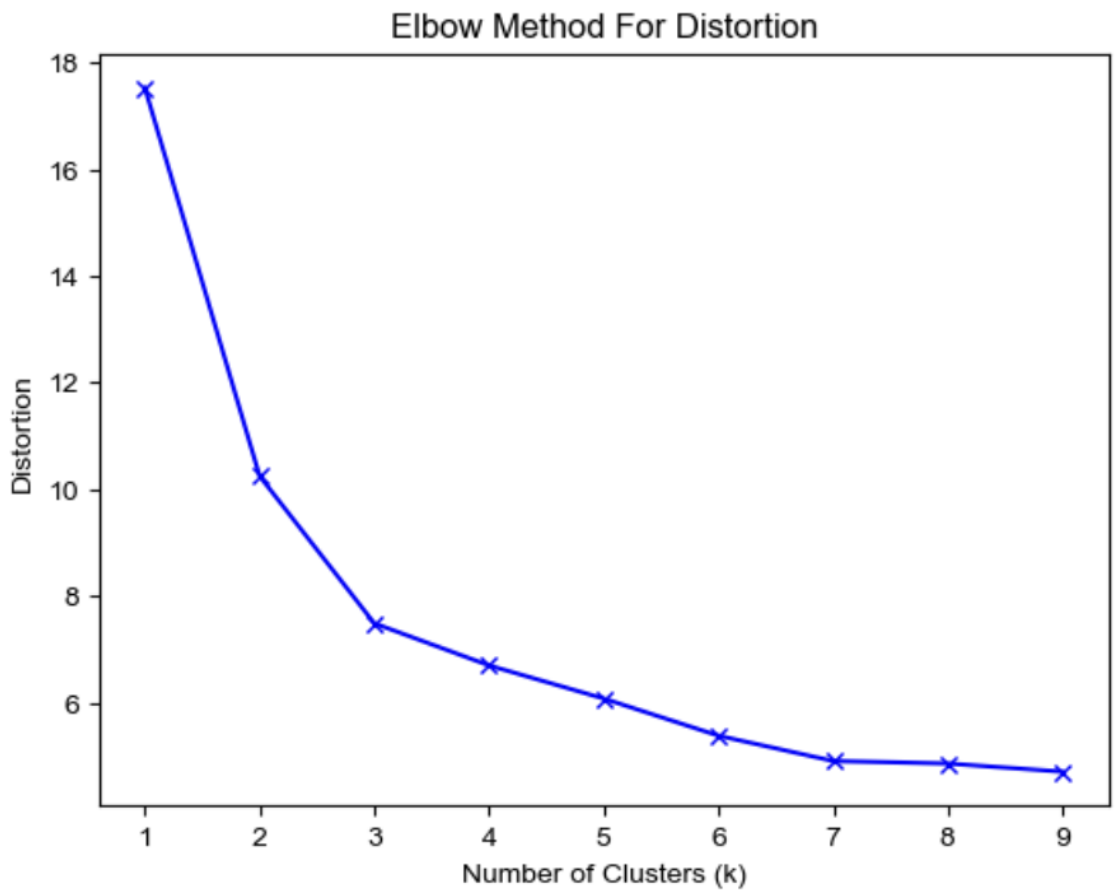


Evaluation method for K-Means

- Silhouette score using distortions.
- Distortion values:

Cluster	Distortion
1	17.52749466966208
2	10.246898580730525
3	7.4845532018542755
4	6.69547142166002
5	6.07409365617229
6	5.3775895110788
7	4.90381734527993
8	4.856261103350349
9	4.704021996903216

Choosing the amount of Clusters



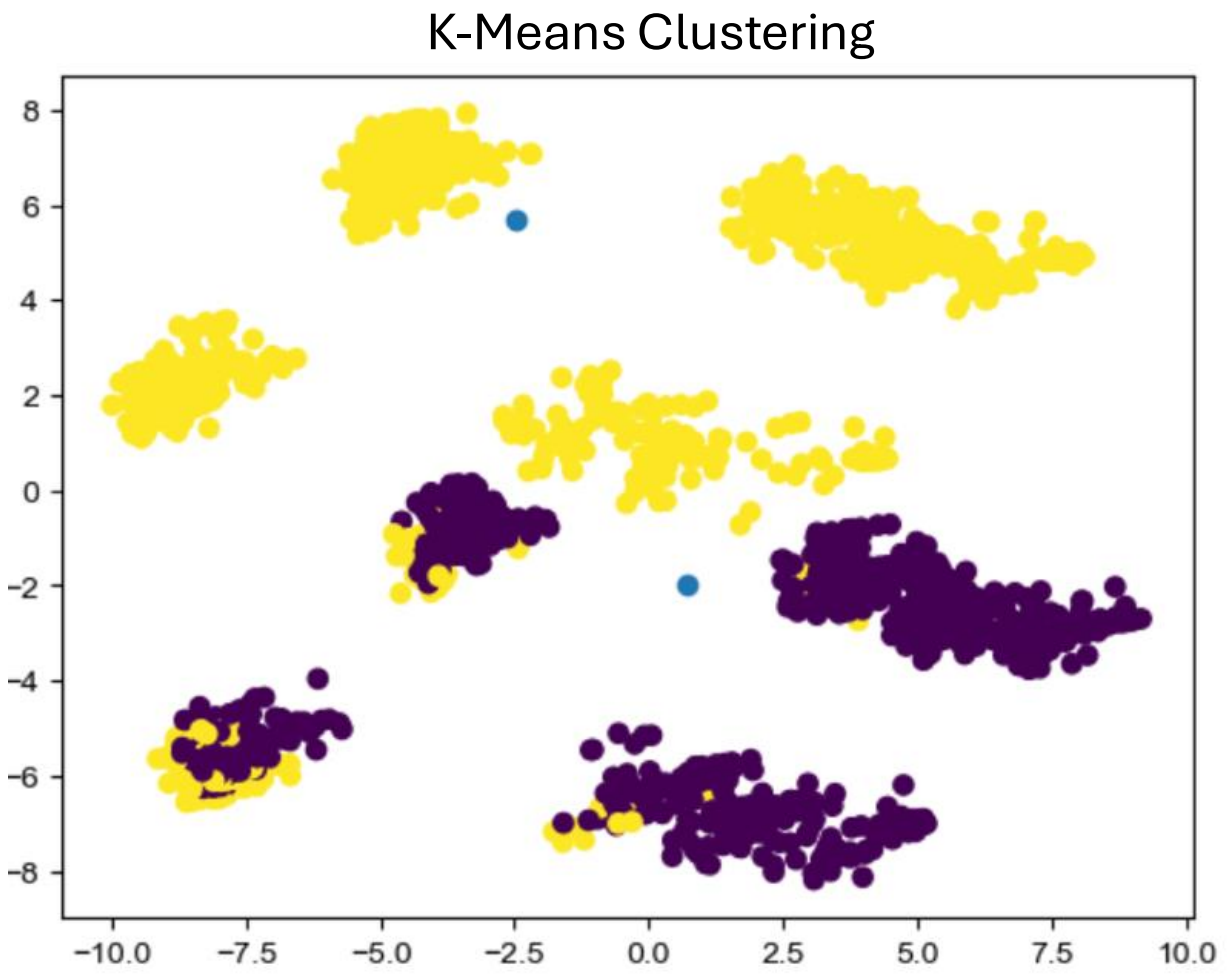
Choosing Clusters Cont.

Clusters	Silhouette Score
2	0.4263560401338029 6
3	0.332269595535650 6
4	0.283387482450383 2
5	0.259506411239969 6
6	0.275636232333462 4
7	0.245809987893256 95
8	0.219977932011180 8
9	0.215989663489771 8

Model Results

Category	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_Siz	2.778576	2.589391

K-Means Visualized



Conclusion

- The exploratory data analysis didn’t yield meaningful results for the company's business problem.
- K-Means was able to create 2 categories for the targeted advertisement with a silhouette score of 0.426.
- While the company can move forward with this model, it’s best that the hold off at this time. The current model is close to being equivalent for random guessing.

Recommendations

- Additional customer information would be needed for model improvements. This can be:
    - Additional Customer data.
    - Additional categories of data. (The actual products being sold).

- If model results do not improve with additional data, other models should be tested such as DBSCAN and Gaussian models.

Future Use cases

- For future use cases we would need to keep in mind that K-Means is for unsupervised learning meaning that there isn’t a specific target column for the model.
- Aside from customer segmentation this can be used for:
  - Monitoring weight loss logs.
  - Image classification.
  - Document classification.

References

- 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