Phase 2: EDA

Joseph Nelson Farrell & Michael Massone
DS 5220 Supervised Machine Learning
Northeastern University
Professor Steven Morin, PhD

Data Overview

Dataset :
Customer Segmentation

Origins :
https://www.kaggle.com/datasets/vetrirah/customer?select=Train.csv

Size : (8068, 11)
Instances: 8068
Attributes: 11

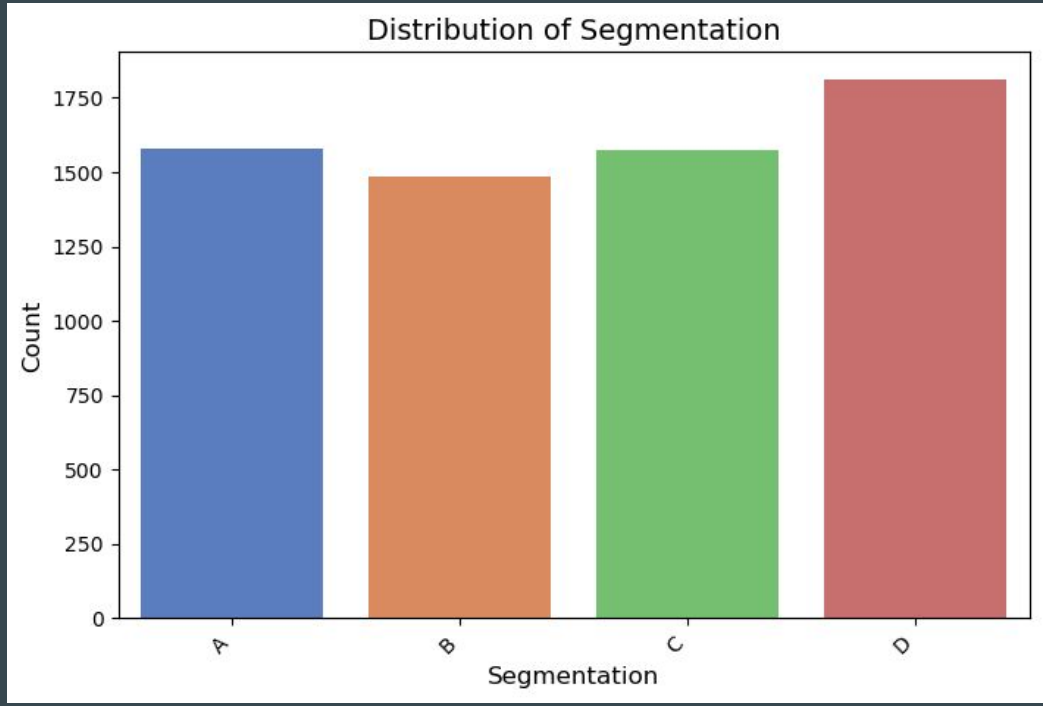
Description :
The dataset contains information about customers of an automobile company segmented into 4 classes (target).

Attribute Name	AttributeType	Percent Missing Values	ML Attribute Designation
index	Numerical - Discrete	0.0%	non_ML
ID	Numerical - Discrete	0.0%	non_ML
Gender	Categorical - Nominal	0.0%	ML
Ever_Married	Categorical - Nominal	1.7%	ML
Age	Numerical - Discrete	0.0%	ML
Graduated	Categorical - Nominal	0.9%	ML
Profession	Categorical - Nominal	1.6%	ML
Work_Experience	Numerical - Discrete	10.0%	ML
Spending_Score	Categorical - Ordinal	0.0%	ML
Family_Size	Numerical - Discrete	4.1%	ML
Var_1	Categorical - Nominal	0.9%	ML

The Target Variable

The column name of the target variable is segmentation.

- There are **0** missing values in the target variable column.
- As a result, **0** observations (rows) are dropped.
- Dataset size: (8068, 11)
- The datatype of the target is a string with 4 values:
  - Value: A ~ Proportion: 0.244422
  - Value: B ~ Proportion: 0.230293
  - Value: C ~ Proportion: 0.244175
  - Value: D ~ Proportion: 0.281111
- The categories appear relatively balanced.
- These values correspond to customer segments used in outreach.

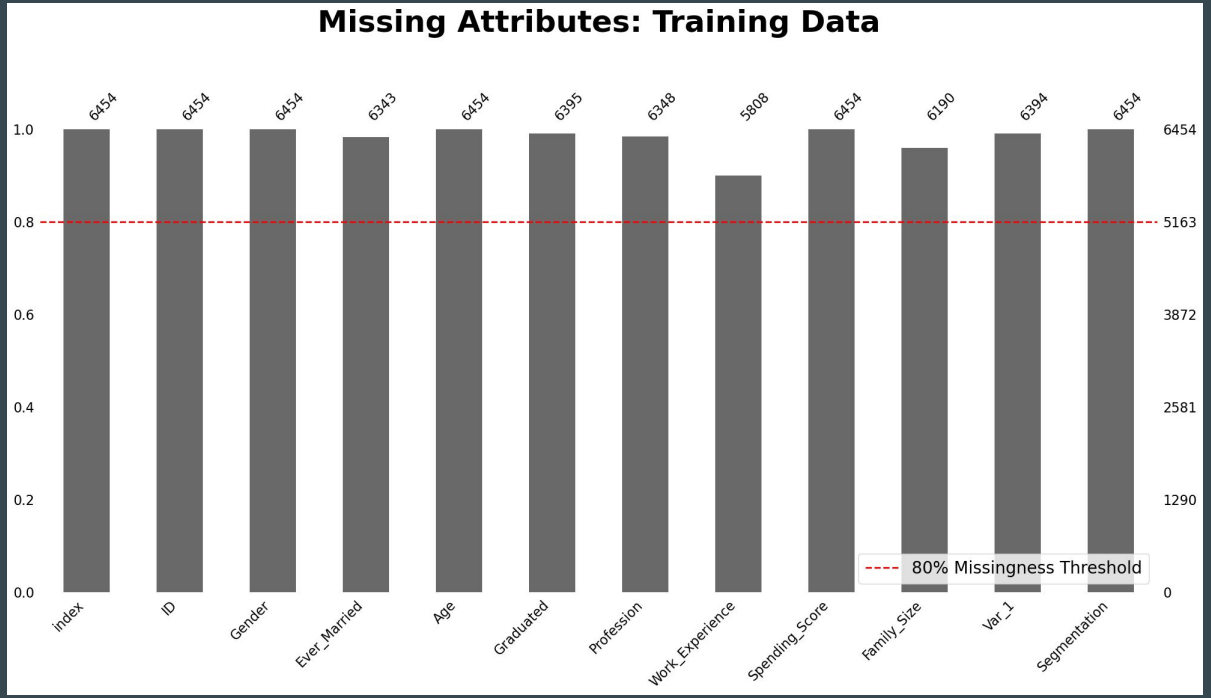


Missingness

The proportion of missing values (NaN ratios) was evaluated for each attribute, including the target variable:

NaN Counts and Ratios: Both the absolute counts and percentages of missing values were displayed directly in the notebook.

Threshold Assessment: Each attribute was compared against a predefined NaN ratio threshold of 20%. No attribute exceeded this threshold, ensuring all features were retained for further analysis.



EDA Approach

Categorical

The categorical attributes were isolated, and the following analyses were conducted:

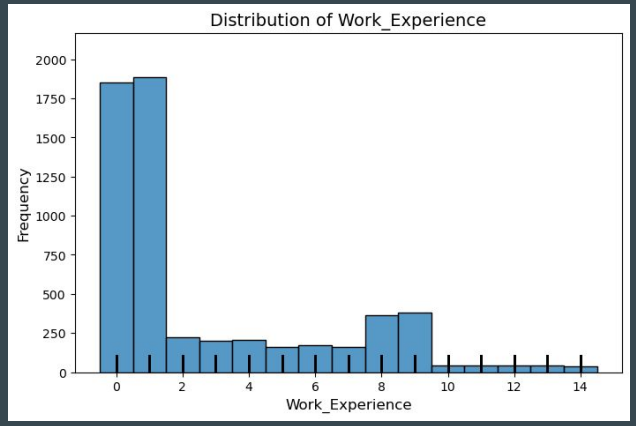
- Cardinality Assessment: Evaluated the number of unique values for each attribute to understand their variability and potential usefulness.
- Class Distribution Overview: Displayed the count and proportion of instances within each class, providing insights into class imbalances.
- Contingency Analysis: Generated a contingency table comparing attribute classes against target classes, visualized through a heatmap to highlight relationships and patterns.
- Bar Plot Visualization: Presented the distribution of each attribute's classes as bar plots, facilitating easy comparison across categories.

Numerical

The numerical attributes were processed through a dedicated pipeline, yielding the following insights:

- Descriptive Statistics: Calculated key metrics such as mean, standard deviation, minimum, maximum, and other summary statistics to understand the overall distribution and central tendencies.
- Outlier Detection: Identified outliers using both the Interquartile Range (IQR) method and Z-score analysis to flag anomalous values.
- Frequency Distribution Visualization:
  - Histogram: Illustrated the frequency distribution of values within each numerical attribute.
  - Boxplot: Provided a visual summary of data spread, highlighting the median, quartiles, and potential outliers.
- Correlation Analysis:
  - Correlation Matrix: Measured and displayed pairwise relationships between numerical attributes.
  - Pairplot: Offered a comprehensive view of attribute relationships through scatterplots and histograms.
- Heatmap of Attribute Means by Target Class: Generated a heatmap to visualize the mean values of numerical attributes across target classes, identifying potential patterns and differences.

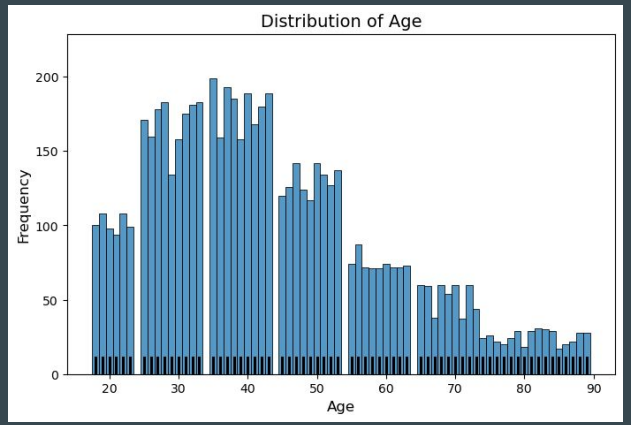
Exploratory Data Analysis - Numerical

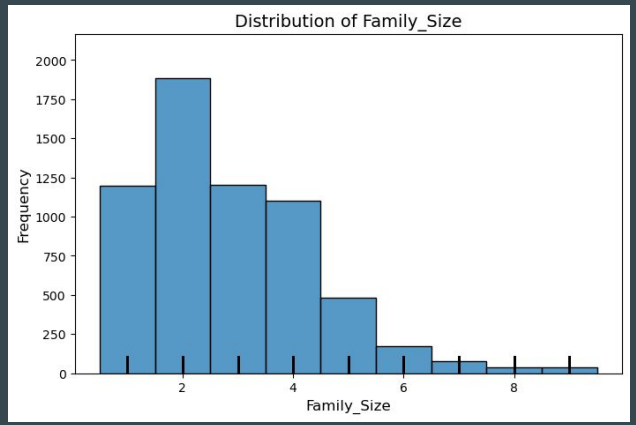


Numerical Feature Distributions:

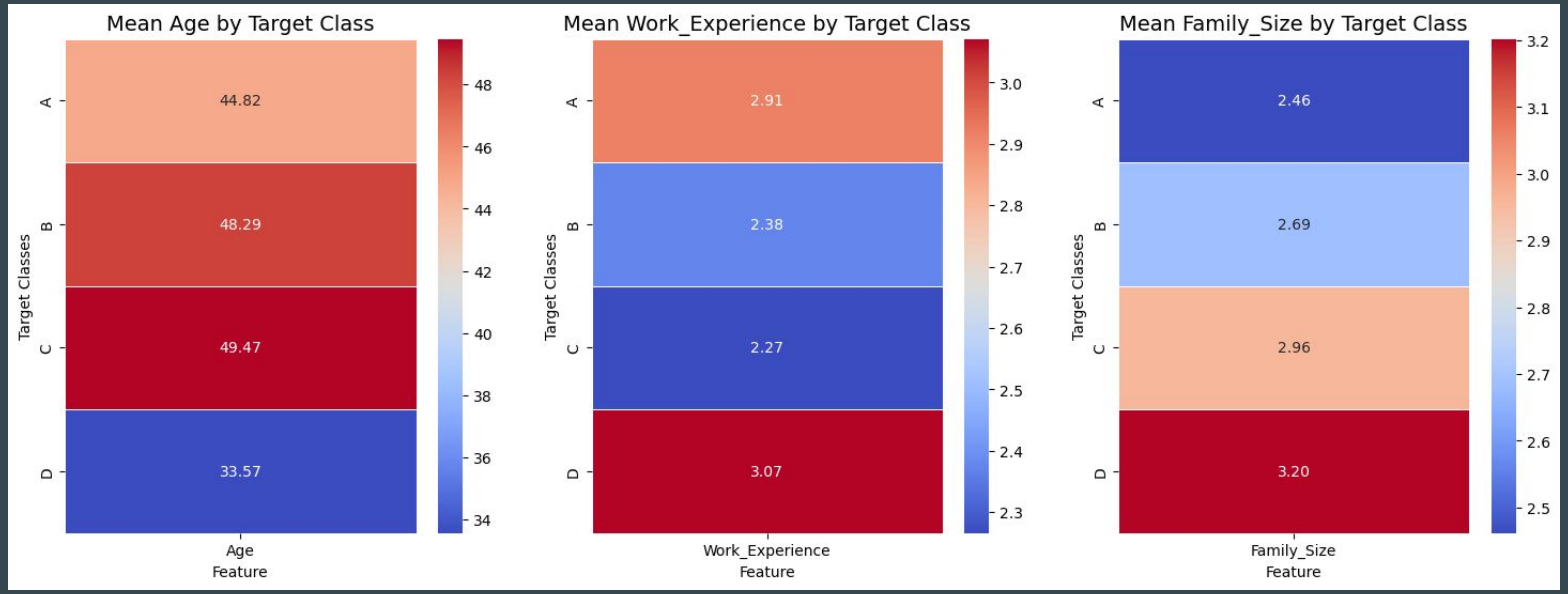
The numerical attributes predominantly exhibited unimodal, right-skewed distributions, with values concentrated toward the lower end and a longer tail extending toward higher values.

There is no clear evidence of clustering within the distributions that could potentially correspond to the four classes in the target variable, with the exception of *Work_Experience* which does have a small cluster around 9 years. This suggests that the variation in these numerical features may not directly align with the distinctions between the target classes.





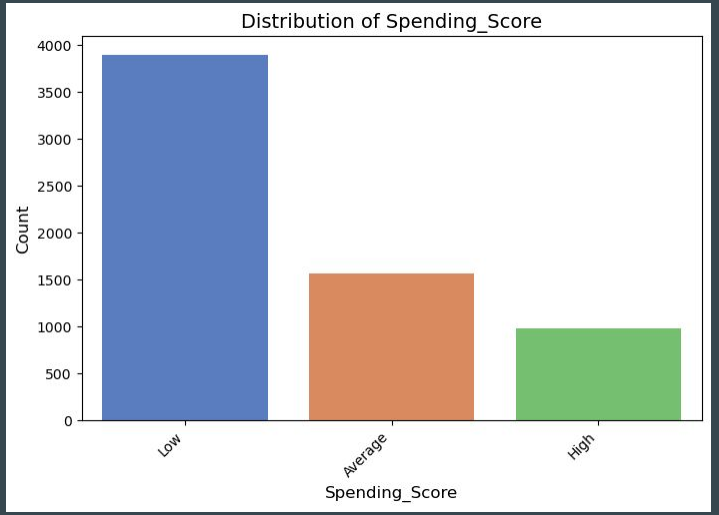
Exploratory Data Analysis - Numerical

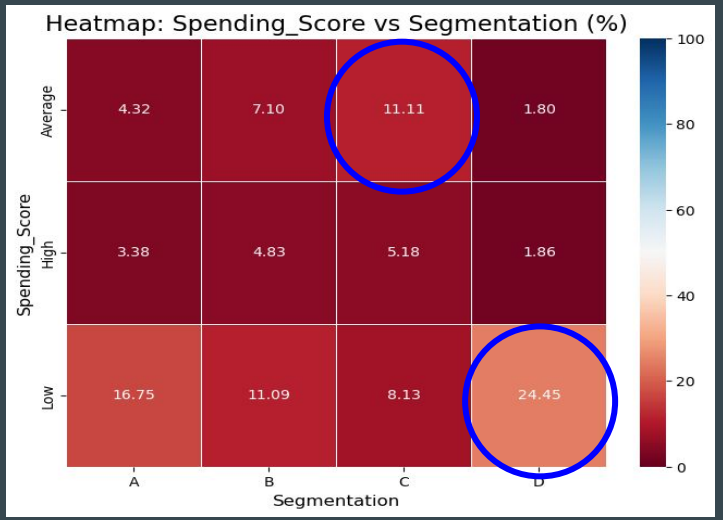


There was some variation in the mean values of numerical attributes when grouped by target classes, with certain features showing more distinct patterns than others. Notably, the Age attribute exhibited the clearest separation, where the mean value for target class D was significantly lower compared to the other three classes. This separation suggests that Age might carry meaningful information for distinguishing class D from the rest, while other numerical attributes did not display as pronounced differentiation across the target classes.

Exploratory Data Analysis - Categorical

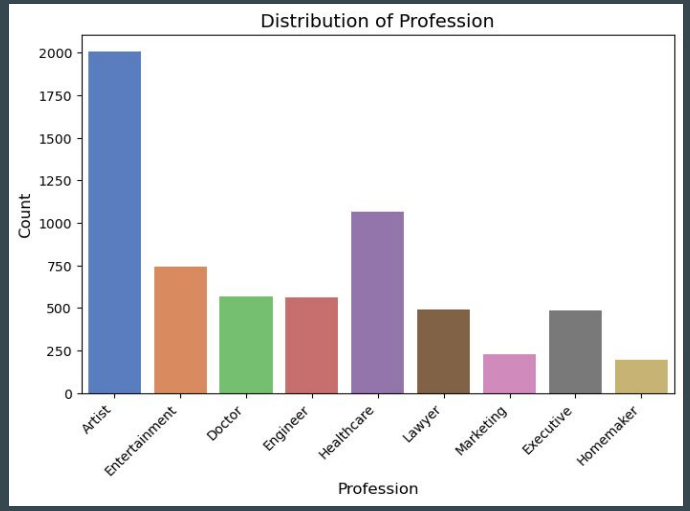
For the Spending Score attribute, the analysis revealed a slightly higher proportion of low scores among target class D. In contrast, class C displayed a greater concentration of average spending scores. This subtle pattern suggests potential behavioral differences between these two classes, with class D individuals tending toward lower spending habits, while class C aligns more closely with moderate spending patterns. However, the signal is weak and the distribution of Spending Scores attribute is unbalanced, with a skew that limits the attribute's discriminatory power across the target classes.

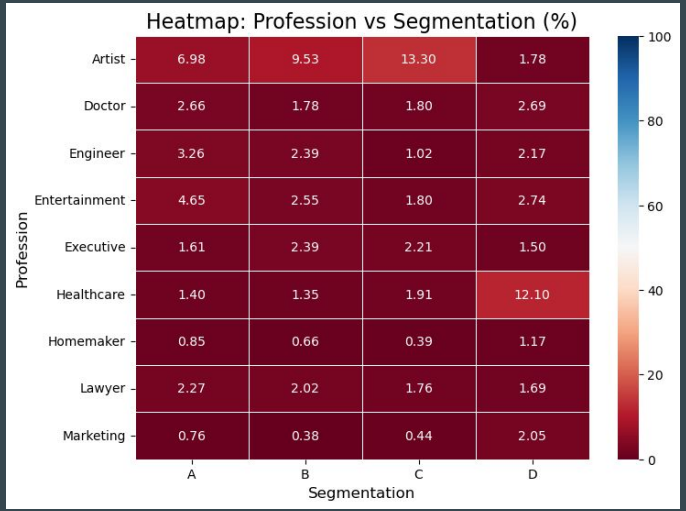




Exploratory Data Analysis - Categorical

For the Profession attribute, the distribution is heavily skewed toward artists, indicating that individuals in artistic professions make up the majority of the dataset. However, target class D shows a notably higher proportion of healthcare workers compared to the other classes. This suggests that profession type could provide some distinguishing power for class D, although its predictive strength may be limited due to the overall dominance of artists in the data.





Clustering - DBSCAN & KMeans

Clustering (unsupervised) was performed in a effort to identify a latent number of target classes. This had two motivations:

- Reduce model complexity:
  - If we could reduce the number of target classes to 2 then our complex multi class classification problem could be reduced to a simpler binary classification problem.
- Inform model interpretation:
  - If target classes were found to be very similar in the encoding space than error rates between those similar classes could perhaps be less impactful to the stakeholders. It could indicate that similar classes would be receptive to either reachout strategy.

Methodology:

- Gridsearch over over DBSCAN & KMeans using PCA and UMAP for dimensionality reduction.
- Evaluation using intrinsic metrics, silhouette score (KMeans) and validity index (DBSCAN)..

Results:

- No meaningful results were discovered.

Interpretation:

- Not surprisingly, the human labels applied to the observations do NOT follow a partition that an unsupervised algorithm can identify. This is a result of the arbitrary nature of both encoding space and the human generated labels.

Conclusion

The exploratory data analysis (EDA) did not reveal any strong correlations, linear relationships, or distinct clusters among the attributes, limiting the potential for straightforward feature-based classification of the target classes. Although some attributes—such as Age, Spending Score, and Profession—showed variability across the target classes, these differences were subtle and unlikely to serve as reliable predictors in isolation.

For Class D, the analysis identified minor class-specific patterns that could provide limited discriminatory power. In particular, Class D was associated with a higher proportion of healthcare workers and a tendency toward lower spending scores compared to the other classes. However, these signals were weak, and the overall attribute distributions were often skewed or unbalanced, which could negatively impact model performance.

Acknowledgments

1. Steve Morin, PhD
2. Seaborn Documentation