ABOUT DATASET

Automobile Customer Segmentation Dataset:

An automobile company has plans to enter new markets with their existing products (P1, P2, P3, P4, and P5). After intensive market research, they've deduced that the behaviour of the new market is similar to their existing market.

In their existing market, the sales team has classified all customers into 4 segments (A, B, C, D). Then, they performed segmented outreach and communication for a different segment of customers. This strategy has work exceptionally well for them.

Acknowledgements

The dataset was acquired from the Analytics Vidhya hackathon.

Content

Variable	**Definition**
ID	Unique ID
Gender	Gender of the customer
Ever_Married	Marital status of the customer
Age	Age of the customer
Graduated	Is the customer a graduate?
Profession	Profession of the customer
Work_Experience	Work Experience in years
Spending_Score	Spending score of the customer
Family_Size	Number of family members for the customer (including the customer)
Var_1	Anonymised Category for the customer
Segmentation	(target) Customer Segment of the customer

DATA PRE-PROCESSING

Data Cleaning

- Dropping the fields with null values.
- Dropping the unnecessary columns.
- Dropping all duplicate values.

Data Transformation

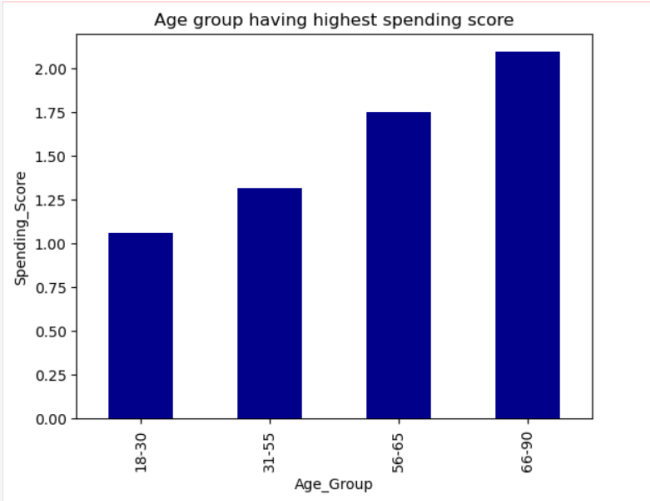
- Label encoding the categorical columns.
- Adding that columns for visualization purposes.

DATA INSIGHTS

- Statistical Data Analysis:
  1. How age affects the spending score of the customer?
     -The correlation between age and spending score is positive correlation. (0.45)

  2. What is the average age of consumer?
     -The average age of customer is 43. This indicates that middle aged customers are more frequent customers.

  3. Which age group has the highest spending score?



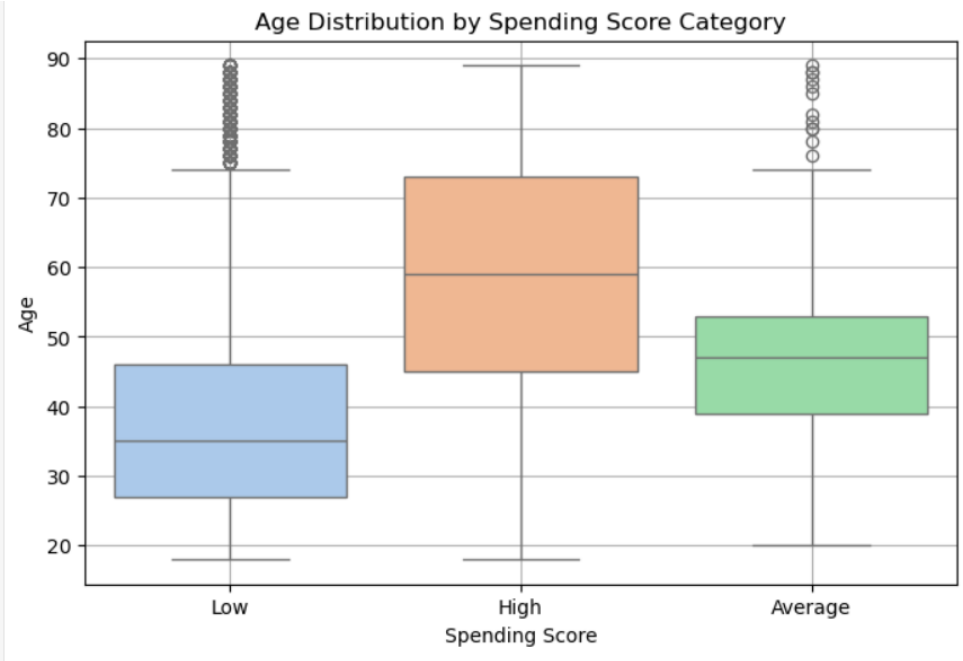
-The oldest age group i.e from 66-90 has the highest spending score with the mean around 2.2.

- Insights from the visuals:

1. Does the spending score vary significantly between genders?

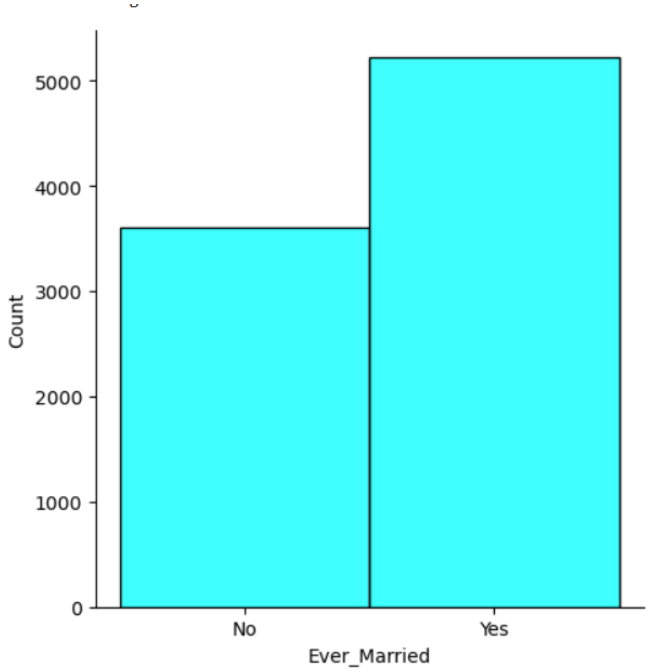
-Men have comparatively a little higher average of spending score than females standing around 1.4 whereas females have a average spending score of 1.37.

2. Age Distribution by Spending Score Category:



-High Spending Score = Avg age – 60 years
-Average Spending Score = avg age – 48 years
-Low Spending Score = Avg age – 35 years

3. What proportion of customers are married versus unmarried?
   -Out of around total 8819 customer's data more than 5000 customers are married and less than 4500 customers are unmarried.



4. What are the most common professions among the customers?

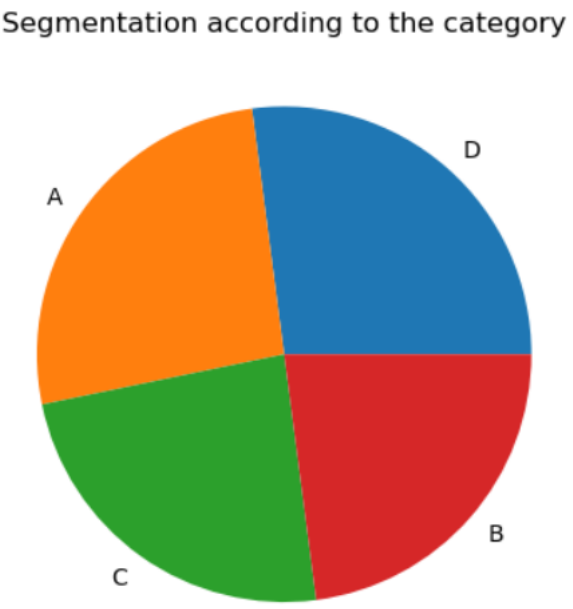
```
Most Common Professions:
Profession
Artist           2888
Healthcare       1414
Entertainment    1063
Doctor            798
Engineer          777
Lawyer            673
Executive         652
Marketing         325
Homemaker         229
Name: count, dtype: int64
```

The highest customer base profession is Artists i.e. 2888.

The lowest customers base has homemakers i.e. 229

5. How many customers fall into each segmentation category (A, B, C, D)?

```
Segmentation category-wise customer count:
Segmentation
D    2388
A    2308
C    2101
B    2022
Name: count, dtype: int64
```



Segmentation category-wise customer count:

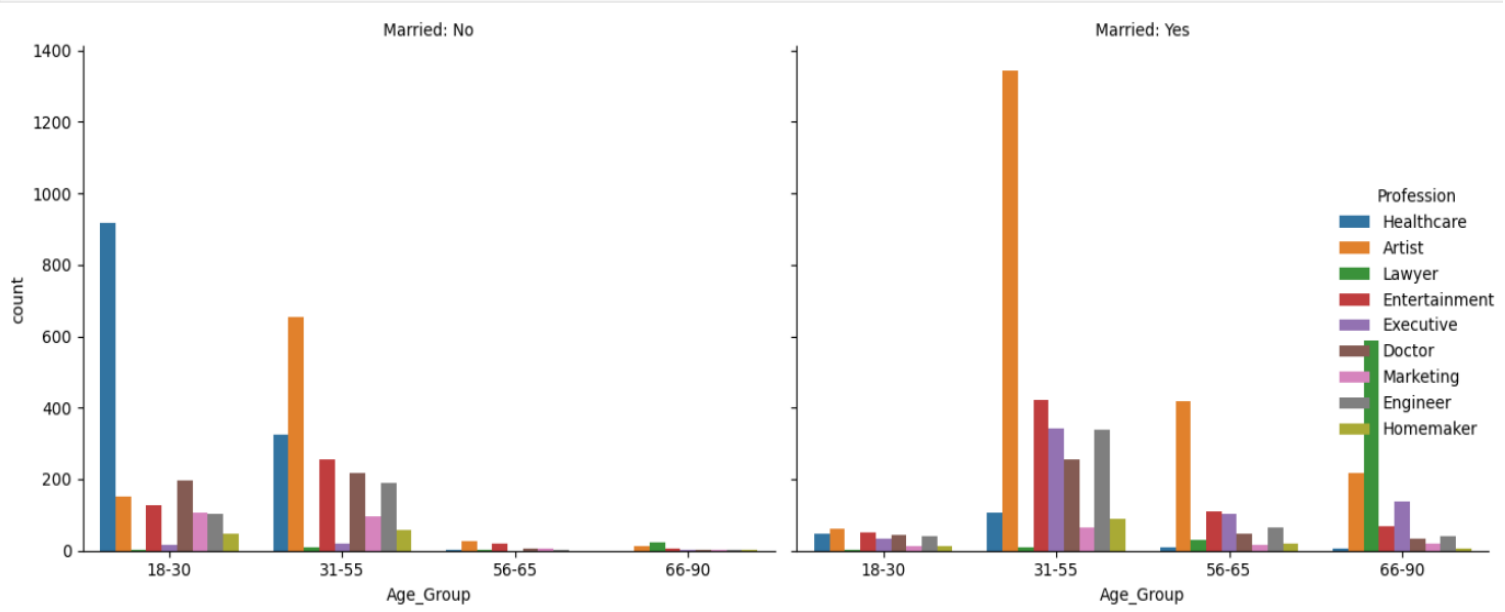
The D segment has highest count of customers. Hence, the company has to apply more marketing efforts over this segment.

6. How does family size affect spending score?
   The correlation between family size and spending score is positive. Hence, the family size affects directly affects the spending score but since the correlation is 0.0694. The influence will be lesser.

7. Number of graduates and non-graduates with their spending score.

```
Graduated
Yes     5594
No      3225
Name: count, dtype: int64
```

8. Multi-Panel Comparison by Marital Status:
   For side-by-side plotting of profession vs age group for married vs unmarried.



-

PREDICTIVE MODELLING

For Modelling:

Algorithm Used

Logistic Regression (Multinomial)

Reason: Interpretable baseline model for multiclass classification.

Model Training

- Data was split into 80% training and 20% testing sets.
- Logistic Regression was trained using lbfgs solver with multi_class='multinomial'.

Model Evaluation:

Classification Report Output:

```
Classification Report:
              precision    recall  f1-score   support

           A       0.41      0.46      0.43       482
           B       0.36      0.12      0.18       398
           C       0.41      0.58      0.48       399
           D       0.59      0.63      0.61       485

    accuracy                           0.46      1764
   macro avg       0.44      0.45      0.42      1764
weighted avg       0.45      0.46      0.43      1764
```

Confusion Matrix:

Shows how many customers in each actual segment were correctly or incorrectly classified.

Key Insights:

- Segment D had the highest prediction accuracy.
- Segment A and B had more misclassifications, suggesting class overlap or underrepresentation.
- Overall model accuracy: ~45%

TECHSTACK

Programming Language: Python

For data pre-processing and data transformation: Pandas

For statistical operations: Numpy

For visualizations: Matplotlib and Seaborn