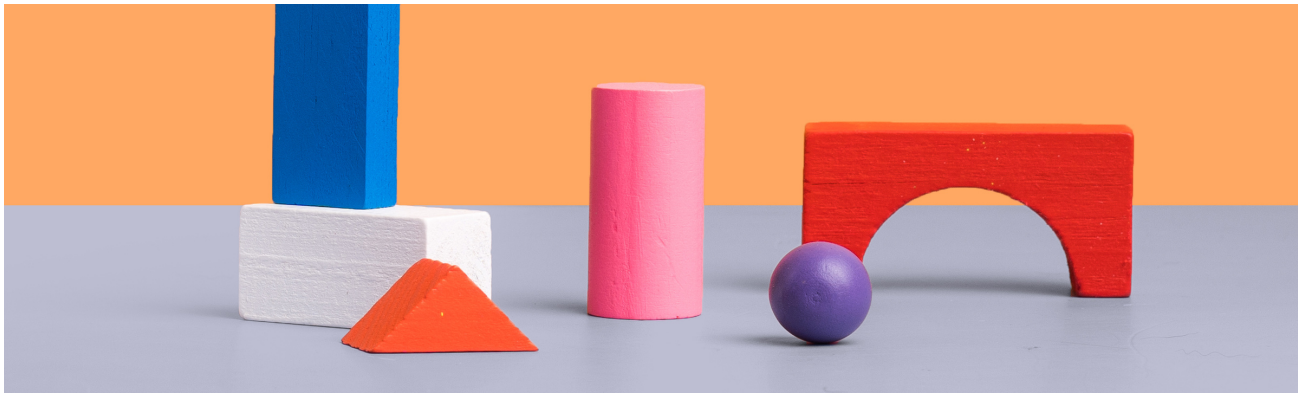

Specification Doc

Context

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 behavior 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).

We tackle the task of cleaning the customers' dataset, ask some relevant questions and visualize their answers using a visualization tool.

Dataset

This dataset is said to be acquired from the Analytics Vidhya hackathon.

Link https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation

it has the following features :

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	# of family members for the customer (including himself)
Var_1	Anonymised Category for the customer
Segmentation	Segment of the customer

Exploratory Data Analysis

As the title suggests, we will perform some EDA techniques such as :

- Identifying anomalies in our dataset.
- Getting rid of them for a more representative data.
- visualize the outcome in a BI tool of our choice.

Methodology

A quick warmup for the data cleaning exercice shows us that this dataset has mainly categorical data, at first glance we see missing values in different columns, for this matter we might use median or linear regression to fill in the nan cells corresponding to numerical features, and the most used value for categorical ones, a quick checking of box plots shows some outliers laying low in our dataset, so we might get rid of them, since they will impact our dashboards.

Once we are sure of the cleanliness of our dataset we will then move to visualizing the results in a responsive dashboard.

Analytical Questions

Analysis Type	Questions
Demographic analysis	• What is the gender distribution within the dataset ? • How many indivisuals are married vs unmarried ? • What is the age distribution of the dataset ? • What is the proportion of individuals who have graduated ? • What is the distribution of family sizes in the dataset ? • Which professions are most common among the dataset ? • What is the distribution of spending scores among individuals ?
Work Experience and Age	• How does work experience vary with age?

Analysis	• What is the average work experience for different age groups? • How does work experience relate to the family size?
Spending score Analysis	• How does the spending score vary across different segments? • What is the spending behavior based on profession? • How does the spending score vary between married and unmarried individuals?
Customer Segmentation Analysis	• How are customers segmented in the dataset? • What are the characteristics of each segment based on age, and profession? • How does the segmentation differ between those who have graduated and those who haven't? • Are there any notable differences in spending behavior across segments? • How does the segmentation vary by gender? • How does the segmentation differ based on the Var_1 category?
Family Size and spending score Analysis	• How does the family size impact the spending score?
Var_1 Analysis	• What are the different categories and their frequencies in Var_1?

Tools

The tools we will be working with are primarly power BI for most of our dashboards and matplotlib or other python library for some graphs that unfortunately don’t exist in Power BI's community edition.

Our Progress So Far

- [x] ~~Data cleaning.~~
- [x] ~~Data visualisation part 1 (python).~~
- [ ] Data visualisation part 2 (Power BI).

Conclusion

Our dataset is intended for modeling purposes, which means the cleaning will highly depend on the end goal, since in our case we won't be performing machine learning on our DS, we won't be doing some advanced data preprocessing such as encoding or generating dummies etc. Nevertheless we hope that our final product will please our teacher Miss Ourdou.