DATA ANALYTICS AND MACHINE LEARNING

ASSIGNMENT COVER SHEET

Student Name: Siddhi Kelshikar

Course Title: Master's of Science in Financial Technology (MSc)

Lecturer Name: Ciara Feely

Module/Subject Title: B9FT114 Data Analytics and Machine Learning

Assignment Title: (CA1)

Student Number: 10627249

TABLE OF CONTENT

INTRODUCTION

The retail industry is becoming increasingly competitive, and gaining an understanding of customers and their preferences is crucial for success. "Shop Customer Data" provides a comprehensive analysis of a shop's customer base, which can help the owner gain insights into their customers' interests, beliefs, and desires. This information can be collected through membership cards or other means of data collection.

The data set provides valuable insights into customers' spending habits, which can help the owner understand which products are most popular in their business. These insights enable the owner to make informed decisions about their business strategy and identify areas for improvement. By analyzing the data, the owner can also identify the strengths and weaknesses of their business, which can lead to changes that improve the customer experience, boost sales, and drive growth.

To enhance the usefulness of the data set, additional demographic data such as customer ID, gender, age, annual income, spending score, profession, work experience, and family size can be included. This will help identify patterns and trends among different customer groups, which can further inform business decisions. Detailed purchasing information such as which products are purchased together, at what times of day, and whether customers are more likely to make impulse or planned purchases can also be included.

Overall, "Shop Customer Data" is a valuable resource for businesses seeking to optimize their operations and stay competitive in the retail industry.

EXPLORATORY ANALYSIS

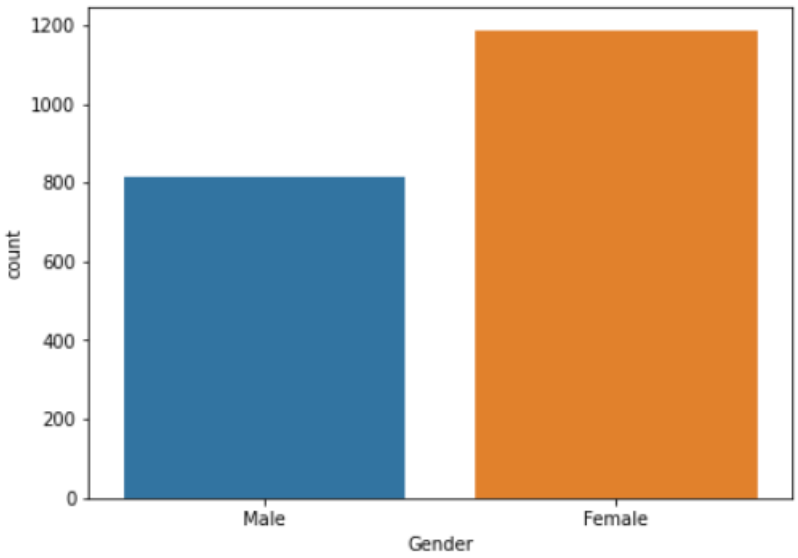
With a sample size of 2000 survey responses from the membership cards, the data set offers powerful insights that could help the owner of the retail shop to more accurately target this demographic. Following are the eight different categories or parameters such as customer ID, gender, age, annual income, spending score, profession, work experience - in years, family size which are further classified into categorical and numerical variables as shown in the notebook.

Firstly we have identified Central and Variational measures through info() function which helped us to know different parameters and there types included in our data. However, describe () function has provided us the count, mean, standard deviation, min and max function for the data for different parameters.

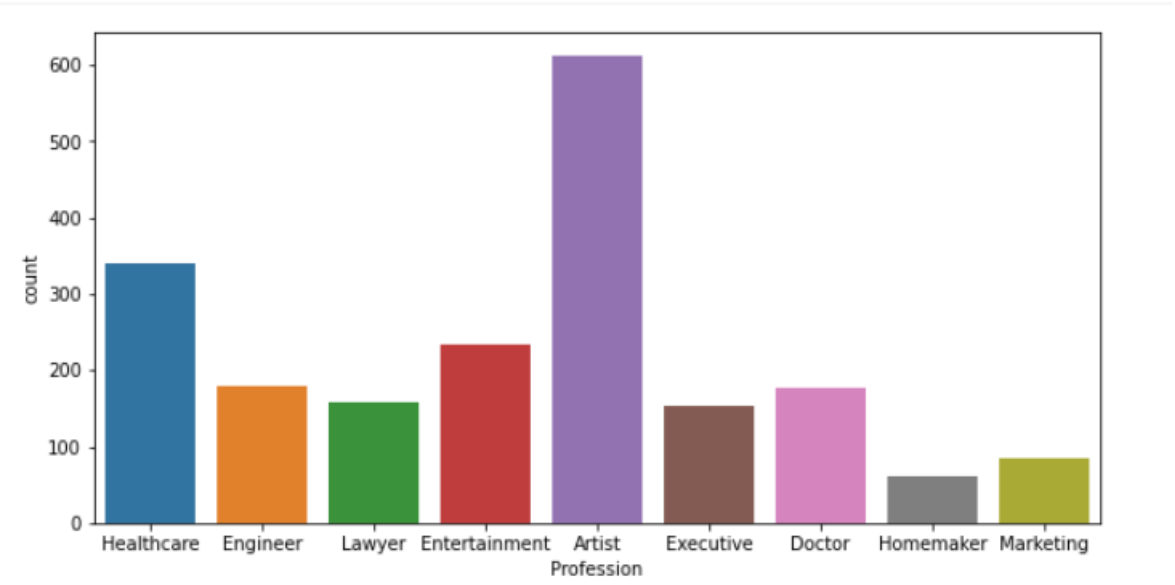
Additionally, we have dropped the “Customer ID” column as it doesn’t provide any useful information for the analysis.

Further we have Identified which type of segment is most responsive to engaging shopping experiences by analyzing the relationship between all of the features to understand customer’s shopping habits using various graphs such as histplot, boxplot and scatterplot to know the interesting trends or correlations between different segments.

1.Visualise Distribution of Categorical Features

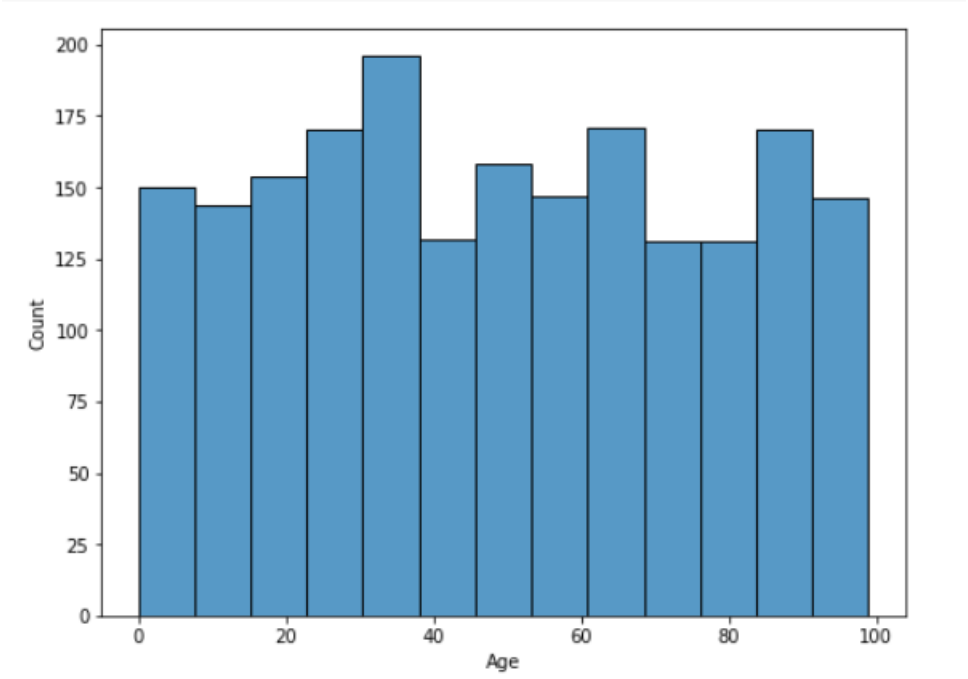


This graph describes that Females customers are more than male customers by 400 points which helps us to know that Females customers use our membership card facility frequently and purchase products from our retail store.

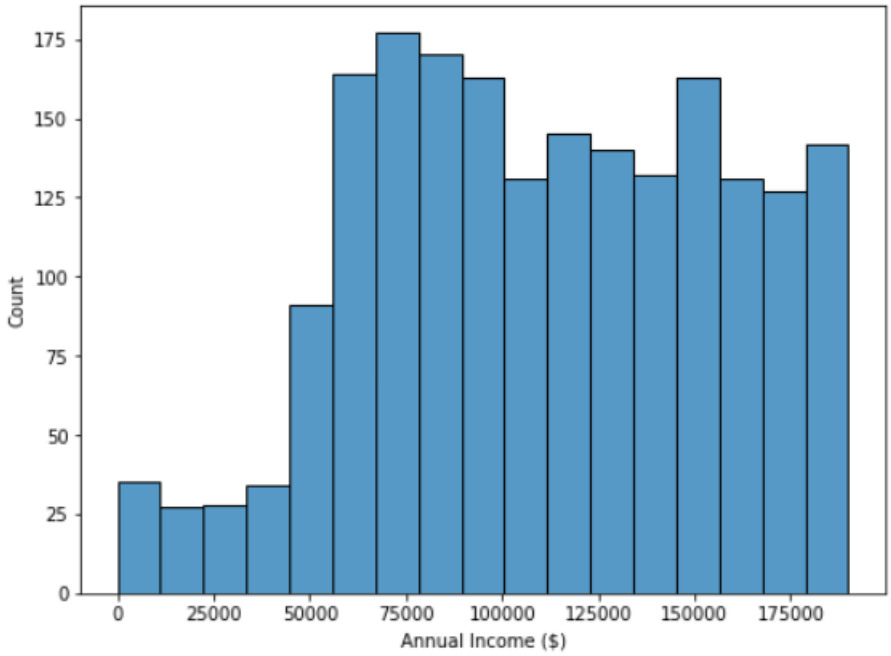


This graph demonstrates through which category our customer belongs from which helps us to identify which sector of products are more sell in our retail store and there spending habits based on there profession. Through this graph we can identify that our customer from “Artist Profession” background are the highest and “Healthcare” sector being the second highest in the graph whereas customers who are “Homemaker” and from “Marketing” sector are the lowest as shown in the figure. However the customers from the remaining sectors such as Engineer, Lawyer, Entertainment, Executive and Doctor sector lies in same range. Through this we can assume that customers from “Artist Profession and Healthcare” sector can be our loyal customers as compared to the rest.

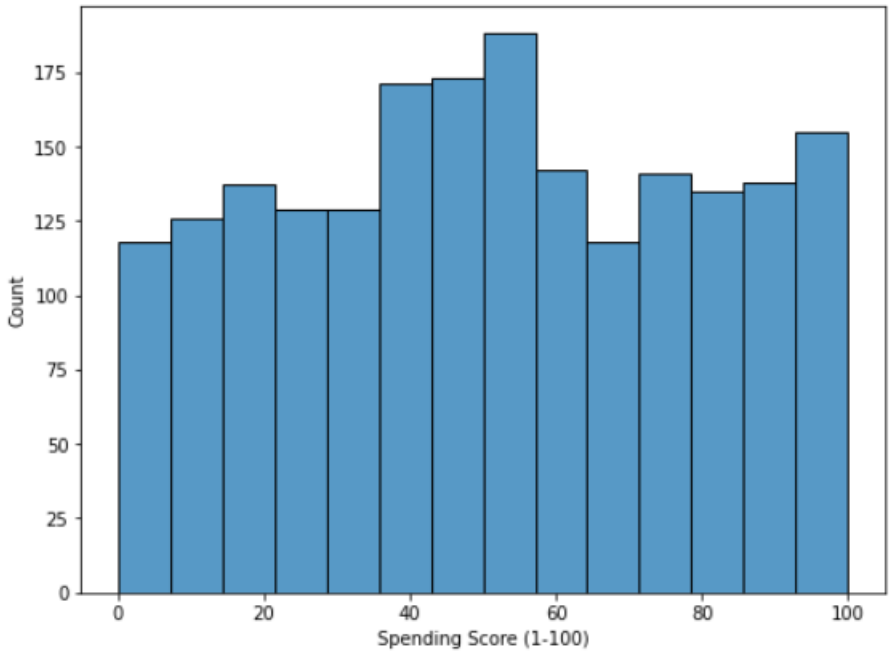
2. Visualise Distribution of Numeric Features



This graph tells us that we have highest number of customers whose age ranges from 20 – 40 and 60 – 65 which helps us to know which products should be kept for their age

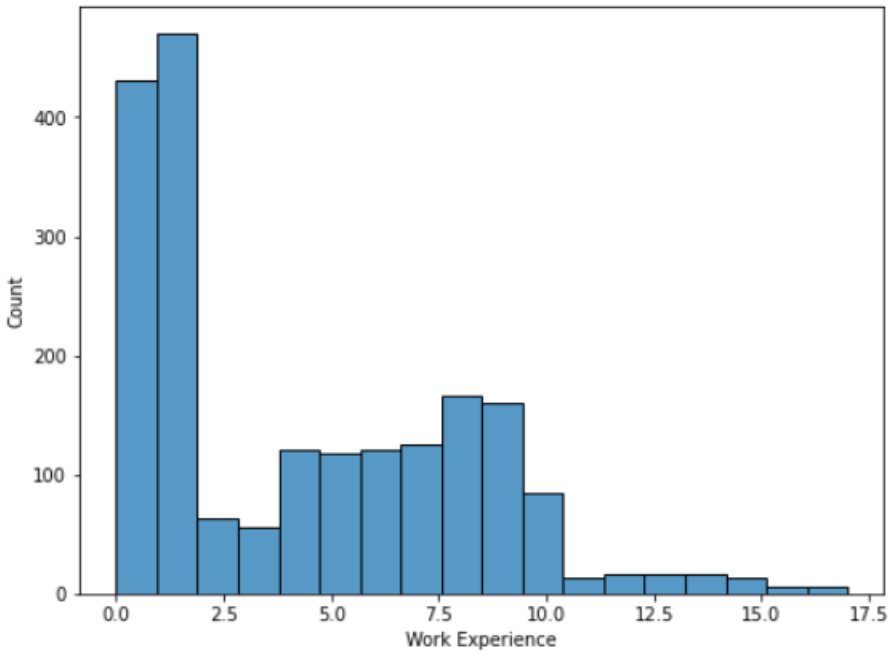


The above graph describes the annual income of our customers which will helps us to know from which category our customers come from through their annual income. As shown in the graph we can tell that 80% of our customers annual income is above 50000 ($) which is helpful for the profitability of the store.

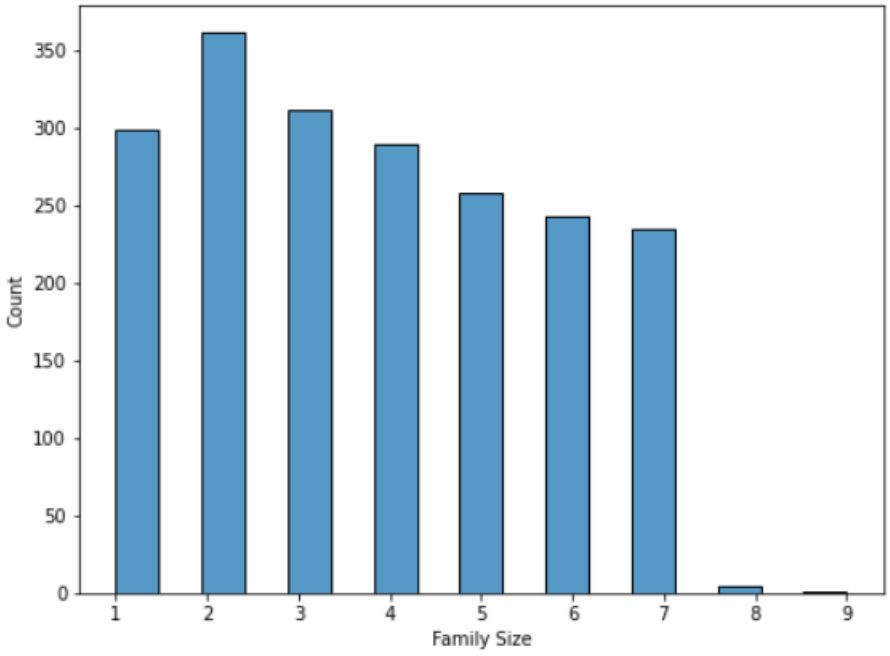


Spending Score graph plays a vital role as this tells as how much our customer spend in our retail store. With the help of our membership card assigned by our

shop, we have traced the spending score of our customers, based on customer behaviour and spending nature. However, from this graph we say that most of our customers spending scores are above 40 which is positive sign for our store as tells us that our product and services are highly scalable.

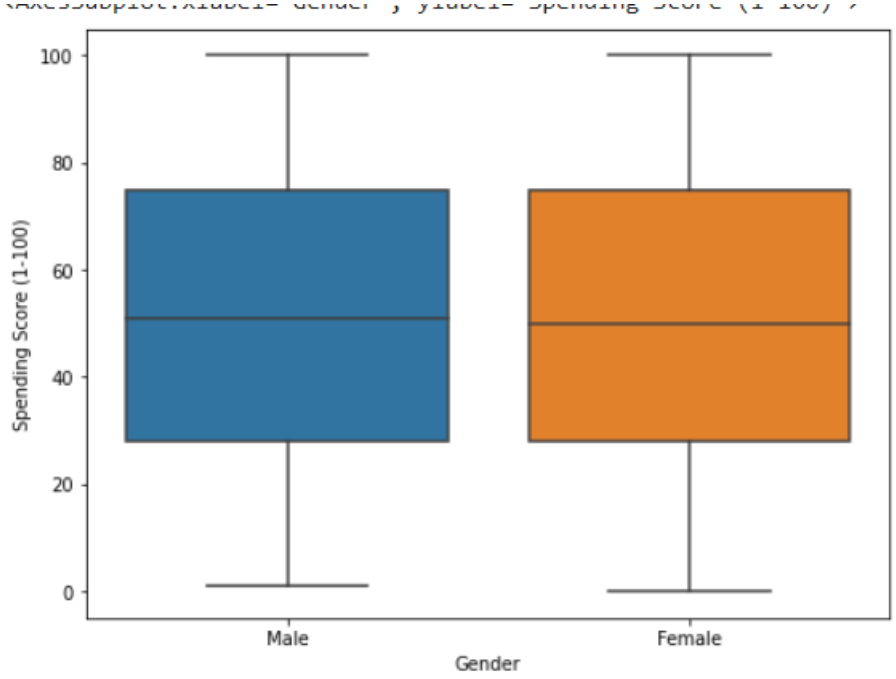


This graph describes that most of our customer are fresher and doesn't have any relevant work experience however many of them have work experience of 5 to 10 years and few of them have more than 10 years of work experience in their professional industry.

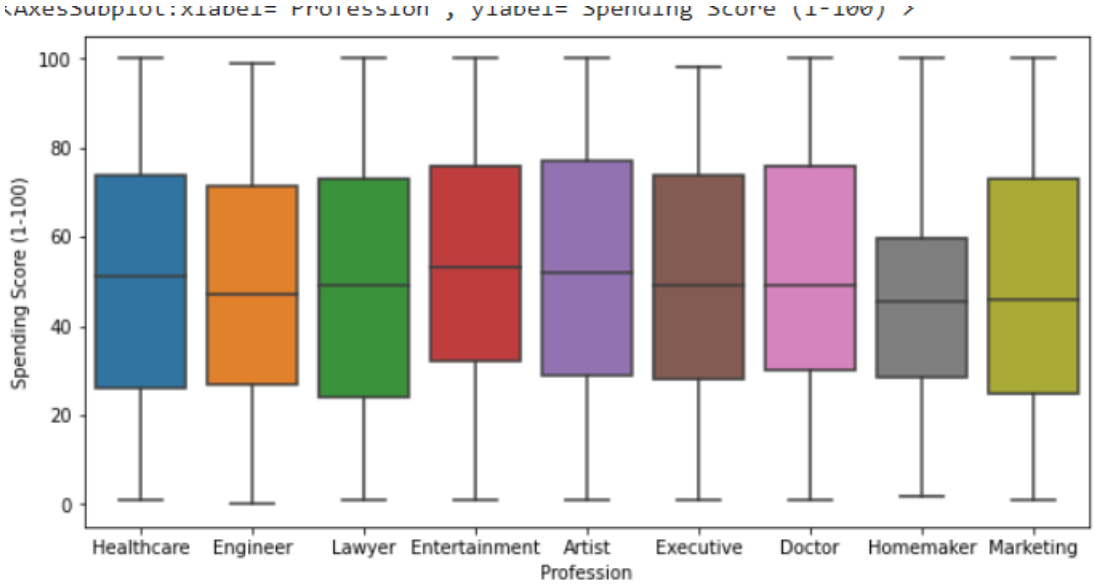


This graph helps us to know the family size of our membership customer which will help us to further analyse their spending score based on their family size. The above graph tell us the most of our customer have 2 members in their family and more of them have 3 and 4 members in their family. Through this we can tell us that more the member more will be the spending score and less the member less will be the spending score.

3. Relationships between continuous and categorical



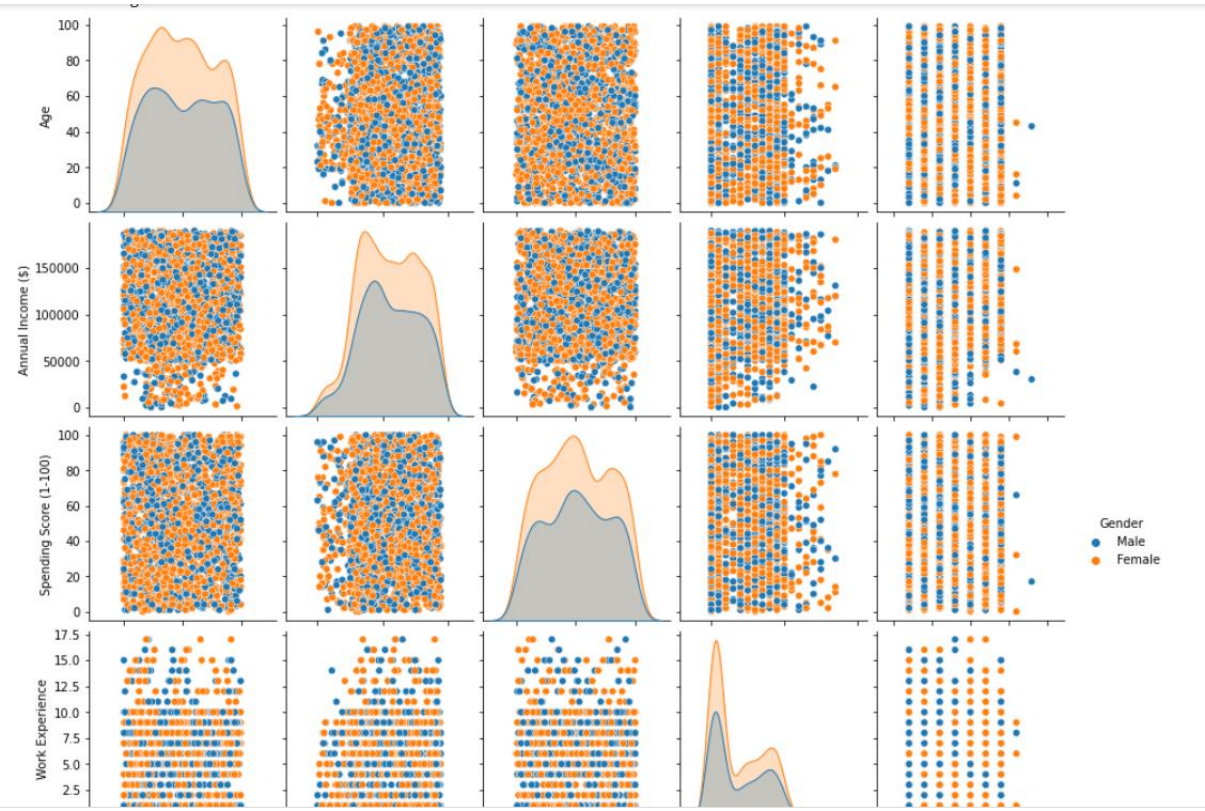
This box plot graph shows the relationship status between the Gender and the Spending score of the customer. The graph depicts that the male and female have the equivalent spending score where the lower quartile is approximately 25, median value is around 50 and the upper quartile is 75 for both male and female.



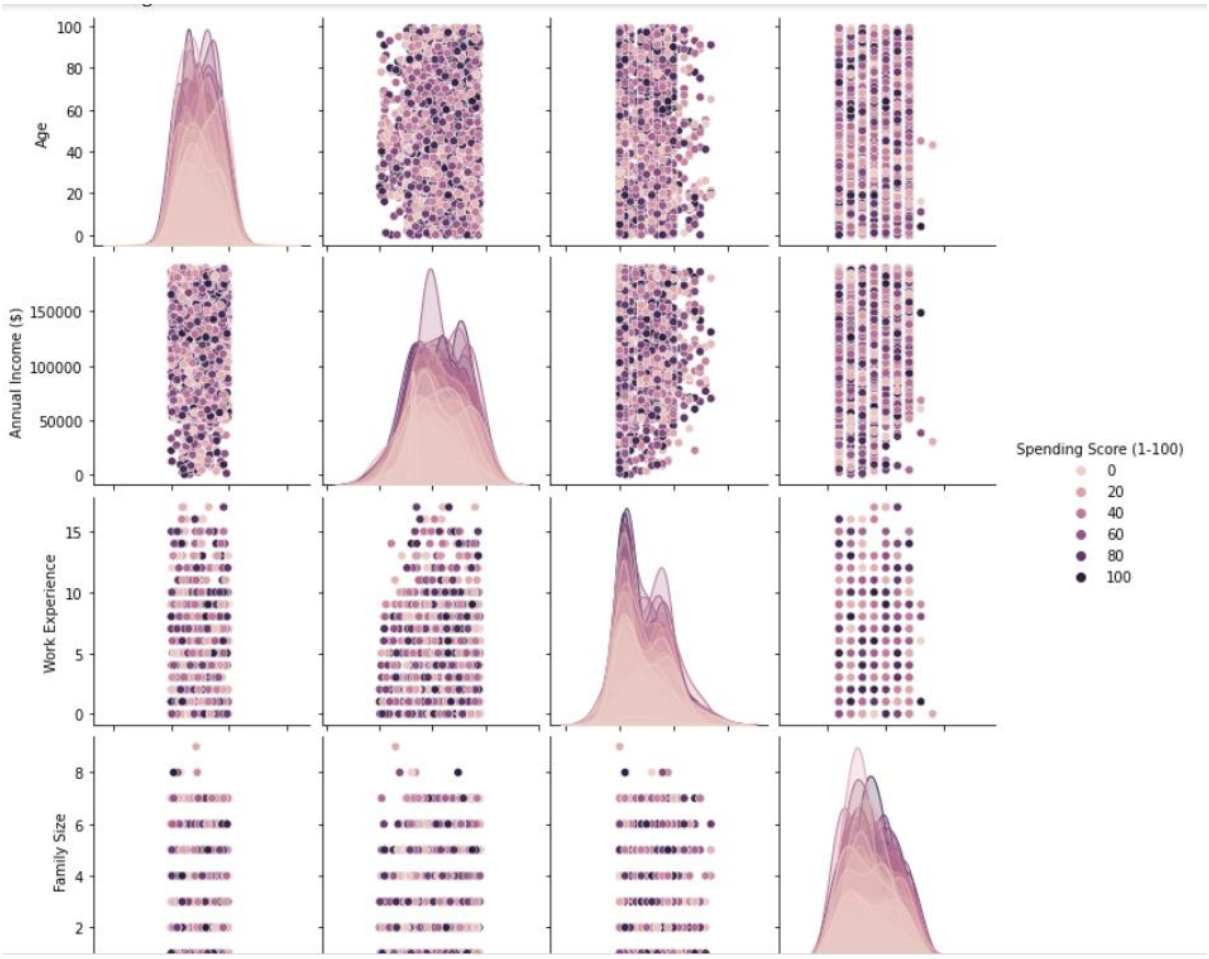
The above box plot graph shows the relationship status between the Profession and the Spending score of the customer. The graph depicts that the customer belonging form healthcare, engineer, lawyer, entertainment, artist, executive, doctor and marketing sectors have the equivalent spending score whereas homemaker has a less spending score.

Furthermore, we have calculated the correlation coefficient and heatmap to analyze the relationship between variables in a dataset and to visualize the strength of the correlation using a color-coded map.

1. Pair plot

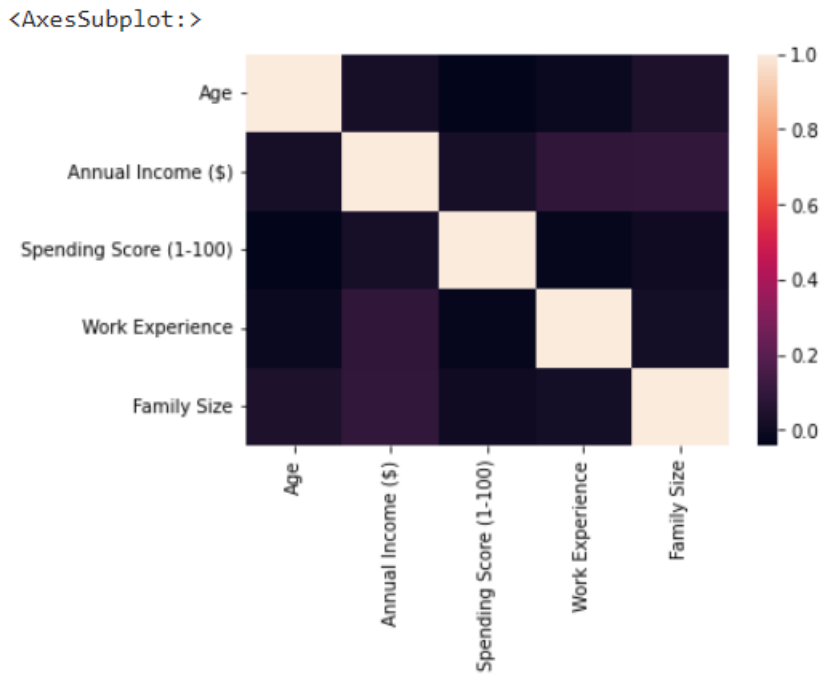


The above pair plot is of Gender which describes the relationship with different factors. The plot for age and spending score is highly condensed, whereas plot for age and annual income, annual income and spending score, age and work experience, work experience and annual income, work experience and spending score are less compact with some scattered points however plot for age and family size, annual income and family size, family size and spending score, work experience and family size is quiet scattered which tells us that female customers are more than male customers.



The above pair plot is of spending score which describes the relationship with different factors. The plot for age and annual income, age and family size, annual income and family size is condensed, whereas plot for age and work experience, annual income and work experience is less compact with some scattered points however plot for work experience and family size its quiet scattered which gives a clear picture that a spending score which ranges from 20-60 are more than others.

2. Heat map



In this visualization, each square represents the correlation between two variables. The correlation value ranges from -1 to +1, where values closer to zero indicate no linear trend between the two variables. A correlation closer to +1 indicates a strong positive relationship, where an increase in one variable is associated with an increase in the other. Similarly, a correlation closer to -1 indicates a strong negative relationship, where an increase in one variable is associated with a decrease in the other.

The diagonals of the visualization are all light pink, indicating a perfect correlation between each variable and itself. The other squares represent the correlation between two different variables. A larger number and lighter colour represent a higher correlation between the two variables, while a smaller number and darker colour indicate a lower correlation.

Overall, this visualization provides a clear understanding of the relationships between different variables and how they are correlated with each other.

DATA PREPROCESSING

To prepare the data for machine learning modelling, we have converted categorical variables such as "Gender" and "Profession" into dummy variables. This is necessary because many machine learning algorithms require numerical inputs rather than categorical ones.

The next step is to split the dataset into training and testing subsets. This is a critical step in machine learning, as it helps to evaluate the performance of a model. The dataset is typically split into two subsets: the training set, which is used to train the model, and the testing set, which is used to evaluate its performance. The purpose of this split is to ensure that the model can generalize well to new data and prevent overfitting.

The 80/20 split is a commonly used ratio, where 80% of the data is used for training and 20% for testing, but the split can vary depending on the size of the dataset and the complexity of the model. By splitting the dataset into two subsets, we can improve the robustness of the model and its ability to make accurate predictions in real-world scenarios. Overall, splitting the dataset into training and testing subsets is a crucial step in the machine learning process and helps to ensure the accuracy and reliability of the model.

1. Handling Missing Values

Our data doesn’t carry any missing data.

2. Outliers

An outlier is a value in the data set that is extremely distinct from most of the other values.

Outliers are further classified into:

1.1 Z score outliers

Usually z-score greater than +3 or less than -3 is considered as outlier which is pretty much similar to standard deviation method.

Through this method we haven’t found any outlier in our data set

1.2 Box plot outliers

Box plot is a popular visualization technique used to display the distribution of a dataset. It is particularly useful for detecting outliers and understanding the spread of

the data. The function takes the data as input and generates a box-and-whisker plot that displays the median, quartiles, and outliers of the data.

By examining the box plot, we identified:

Q3 = 73.25

Q1 = 25

IQR = 48.25

Upper bound = 145.625

Lower bound = -47.375

We can also use Chebyshev's Theorem to identify outlier

Chebyshev's Theorem, also known as Chebyshev's Inequality, is a mathematical concept that was discovered by Pafnuty Chebyshev. It is useful for any dataset, especially those that have non-normal probability distributions. The theorem estimates the minimum proportion of observations that will fall within a specified number of standard deviations from the mean. The equation uses the value of k to limit the number of standard deviations away from the mean, and the probability of a random variable falling more than k standard deviations away from the mean is at most 1/k^2. By applying this theorem, we can estimate the proportion of data points that fall within a certain distance from the mean and gain valuable insights into the dataset.

PREDICTIVE ANALYSIS

Predictive analytics is the task of predicting the output variable given the values of input features.

As our dataset belongs to a numeric feature, we will be using Linear Regression model.

Input variable is X and output variable X = Gender, Age, Annual Income ($), Profession, Work Experience and Family Size

Y = Spending Score

Interpretation of the model parameters:

1. Mean absolute error = 24.09

2. Mean absolute percentage error = 1.65

If Mean absolute percentage error is 1.65 over 100% which is a lot, if spending score is 100 than it is off by 100 or more which is irrelevant.

3. Mean squared error = 28.24

Mean squared error will penalize the error b.y 28.24.

4. R2 score = 0.005

R2 score represents by what proportion of the model of the variance Y is explained by the model however 0.005 is too low score.

Model summary

In our model summary constant is 53.4925.

If value of Age increases by 1 our coefficient value decreases by 0.035 by every unit increase in Age.

If value of Annual Income increases by 1 our coefficient value increases by 3.44 by every unit increase in Annual Income.

If value of Work experience increases by 1 our coefficient value decreases by 0.1055 by every unit increase in Work experience.

If value of Family size increase by 1 our coefficient value increases by 0.22 by every unit increase in Family size.

If value of Sex male increase by 1 our coefficient value increases by 0.27 by every unit increase in sex male.

If value of Prof Doctor by 1 our coefficient value decreases by 2.48 by every unit increase in Prof Doctor size.

If value of Prof Engineer size increase by 1 our coefficient value decreases by 3.46 by every unit increase in Prof Engineer.

If value of Prof Entertainment increase by 1 our coefficient value increases by 0.67 by every unit increase in Prof Entertainment.

If value of Prof Healthcare size increase by 1 our coefficient value decreases by 4.53 by every unit increase in Prof Healthcare.

If value of Healthcare size increase by 1 our coefficient value decreases by 2.97 by every unit increase in Healthcare.

If value of Prof Homemaker size increase by 1 our coefficient value decreases by 7.215 by every unit increase in Prof Homemaker.

If value of Prof lawyer size increase by 1 our coefficient value decreases by 4.46 by every unit increase in Prof lawyer.

If value of Prof marketing size increase by 1 our coefficient value decreases by 4.53 by every unit increase in Prof marketing.

P value

According to our model summary we do not have any P value significant in our data. Therefore, our data is insignificant.

CONCLUSION

Through this dataset we have identified Females are our ideal customer and also the spending score depends on various other factors such as Profession, Age, Gender. True linear regression we can conclude that this model is a moderate or not that well fit for the further analyse.

However, through Data Analytics and Machine Learning model we identified our ideal customers and has made it viable to make predictions about the various choices and spending patterns of our customer which will help to enhance the profitability of a retail store.

Link for the python file.

https://colab.research.google.com/drive/1nF0inzgFH2lIgsvYXJNiMcgY66bpfLFt?usp=sharing