Black Friday Sales Purchase Analysis

51 Siddhant Jain


Black Friday!

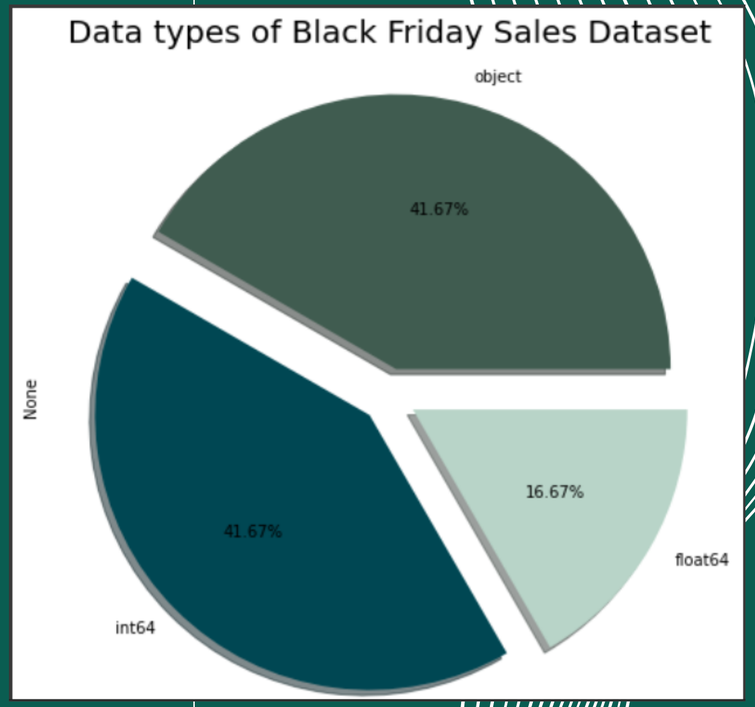

BLACK
FRIDAY
DOORBUSTER
$10

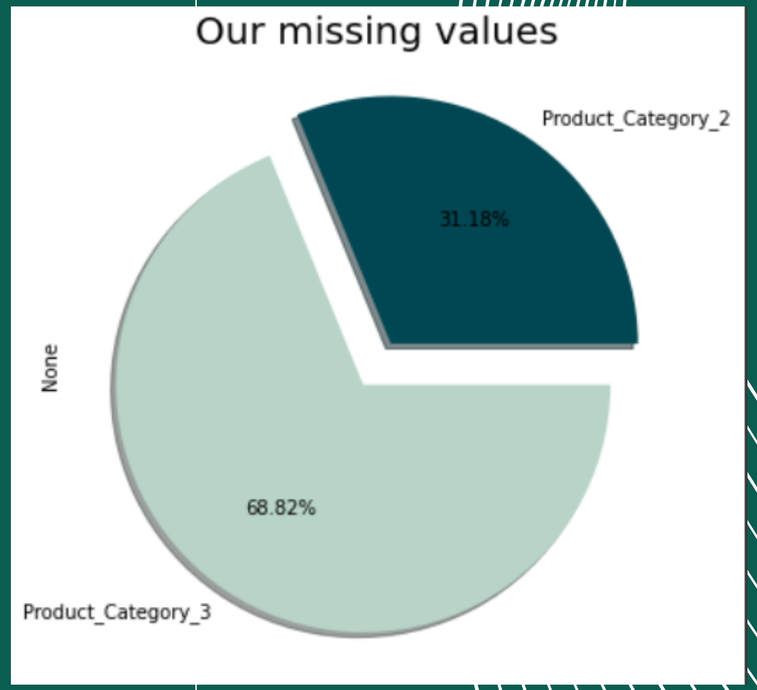

IT'S ON!
BLACK
FRIDAY
SALE

INTRODUCTION ABOUT DATA SET

- A retail company “ABC Private Limited” wants to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high-volume products from last month.
- The data set also contains customer demographics (age, gender, marital status, city_type, stay_in_current_city), product details (product_id and product category), and Total purchase_amount from last month.
- The data set has 12 features of 5891 users accounting for total data of 550068.

Exploratory Data Analysis



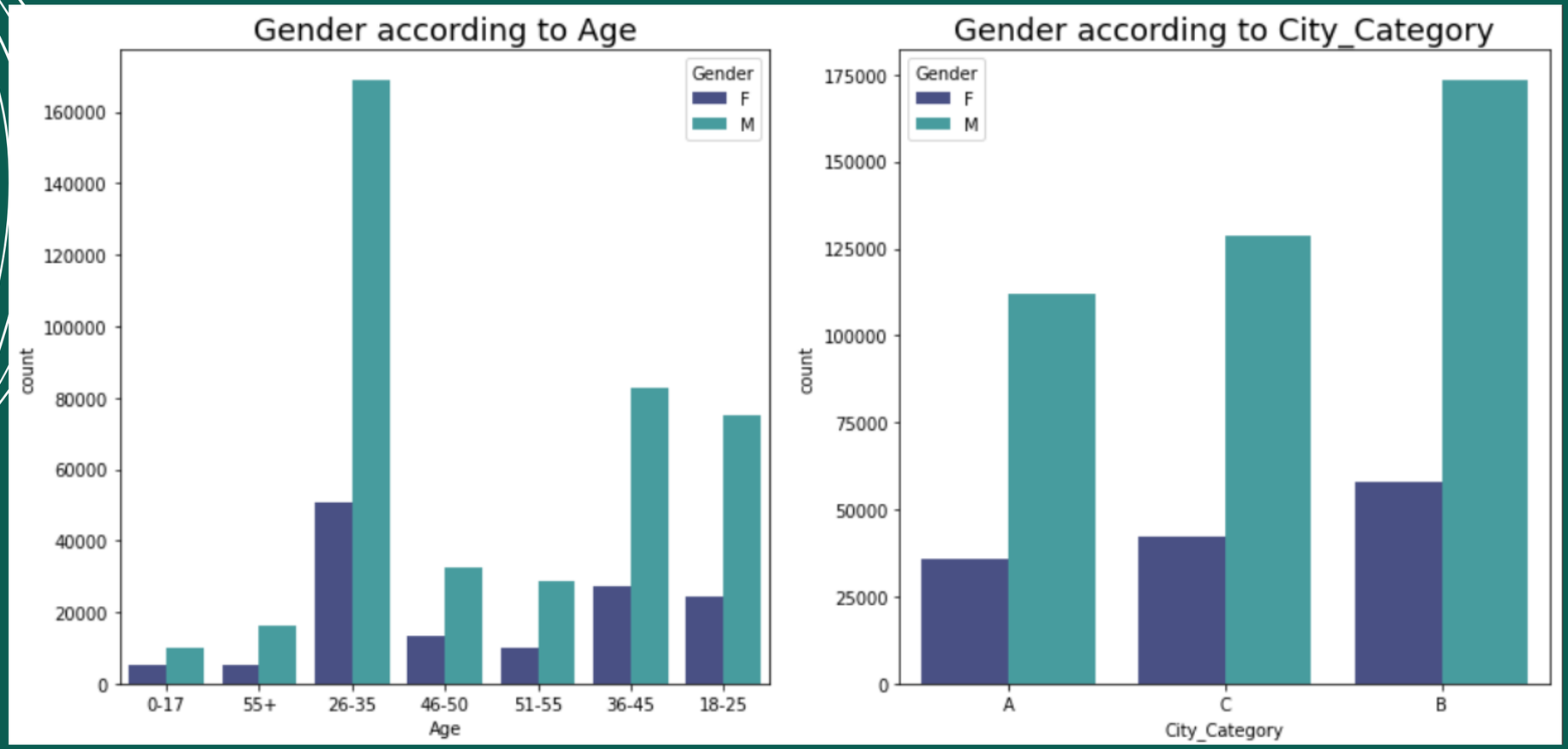


Data Types in RAW data -

- We are having 16.67 percent of float, 41.67 percent of integers, and 41.67 percent of object

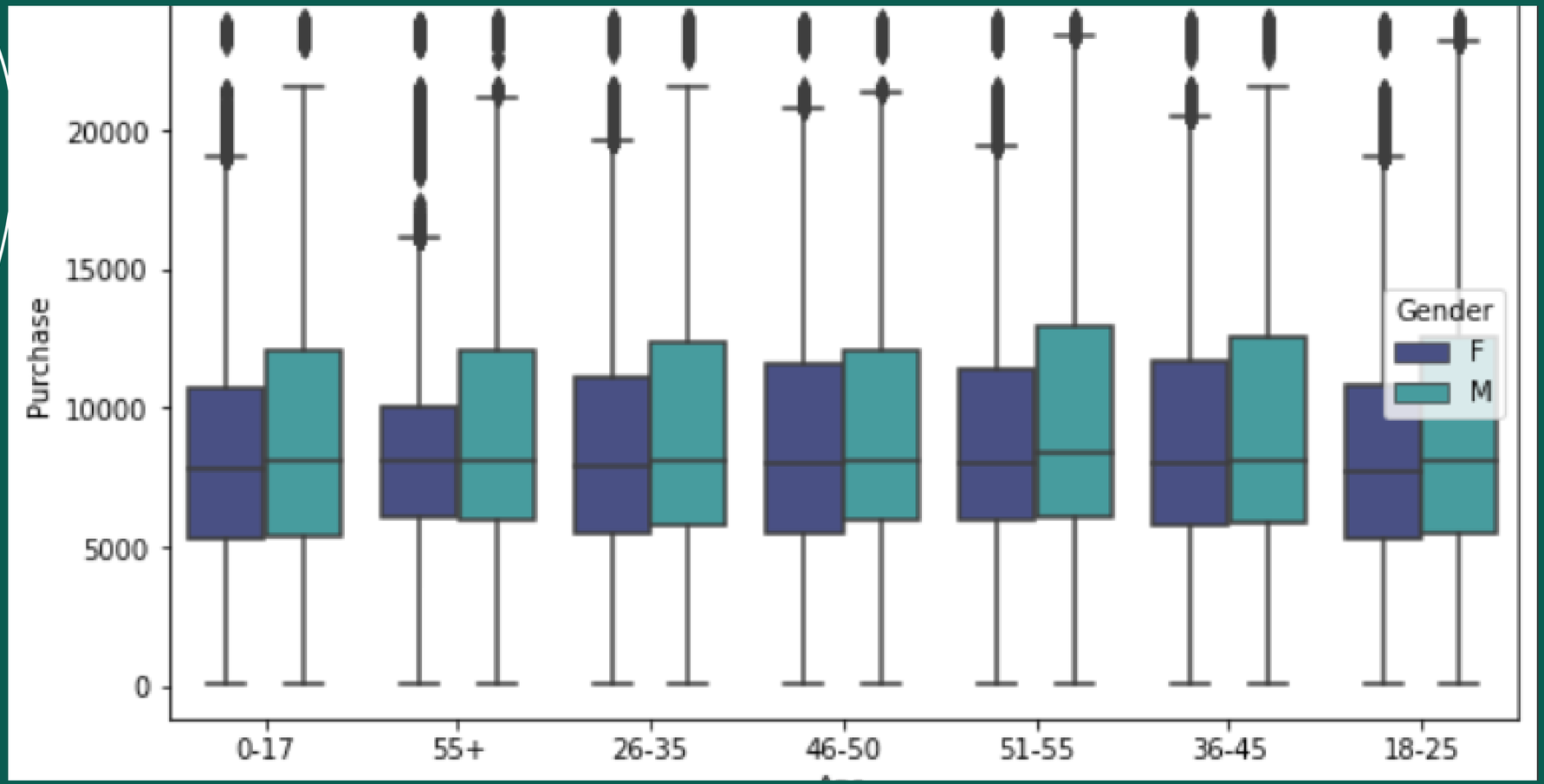
Missing Values -

- We are having 31.18 percent missing values in " Product_Category_2 " column & 68.82 percent missing in " Product_Category_3 "



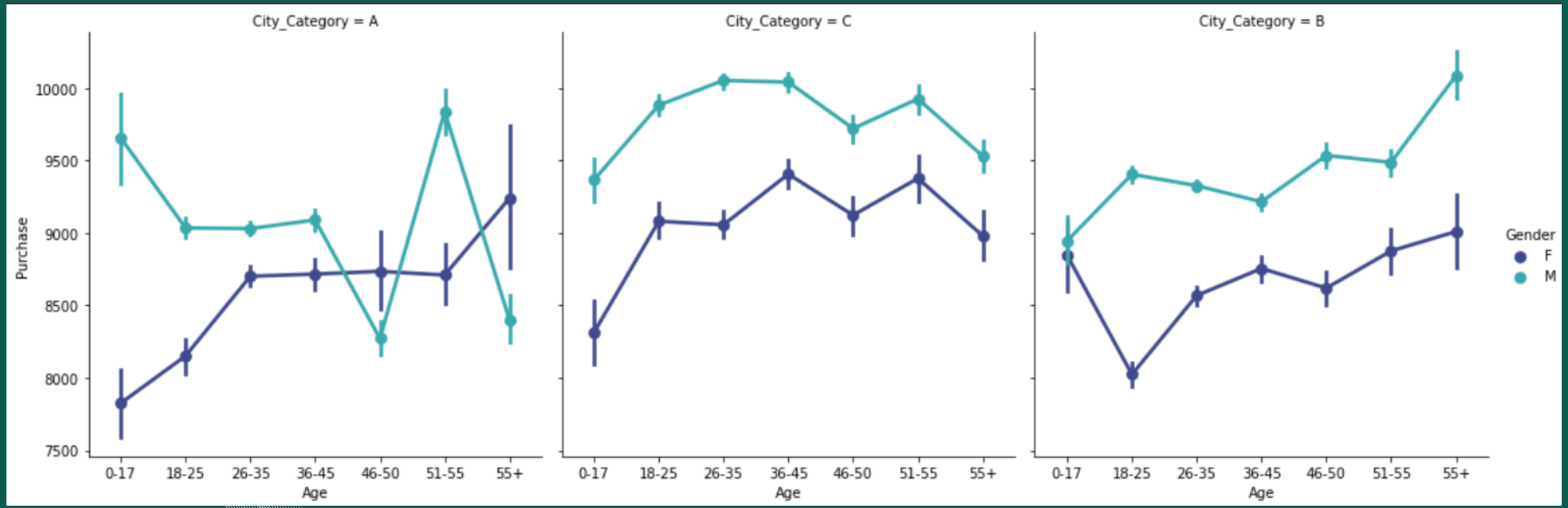
Total Count

- We have maximum people in the age bracket 0f 26-35 even with respect to gender that are part of this analysis.
- Their are maximum participants for both male and female from City Category B.



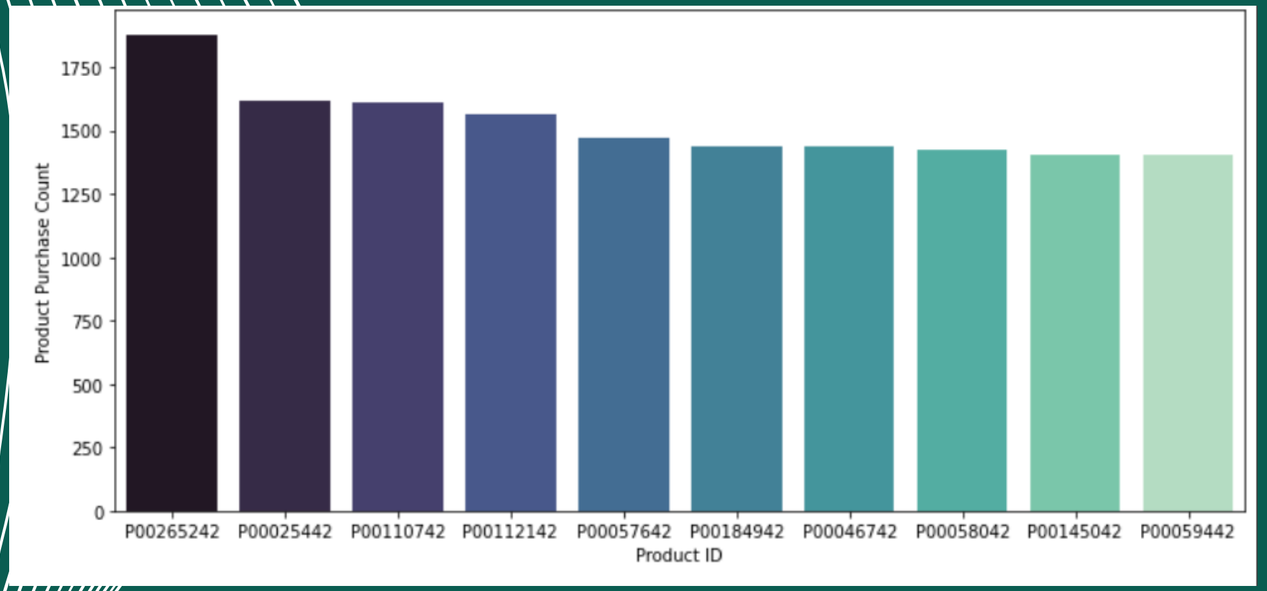
Outliers

- Figure is depicting outliers with respect to purchase across different age groups as well as gender.



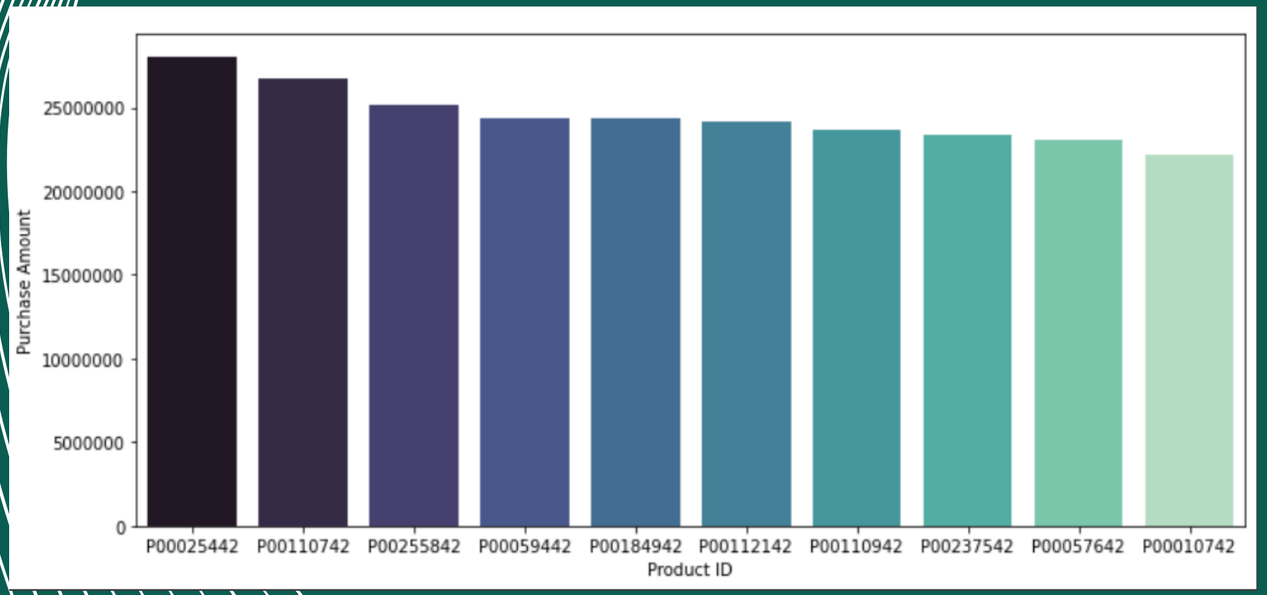
Average Purchase

- Graph is depicting average purchase across different city categories.
- It is also representing average purchase is always more of male w.r.t to female across all age groups for city - B & C.



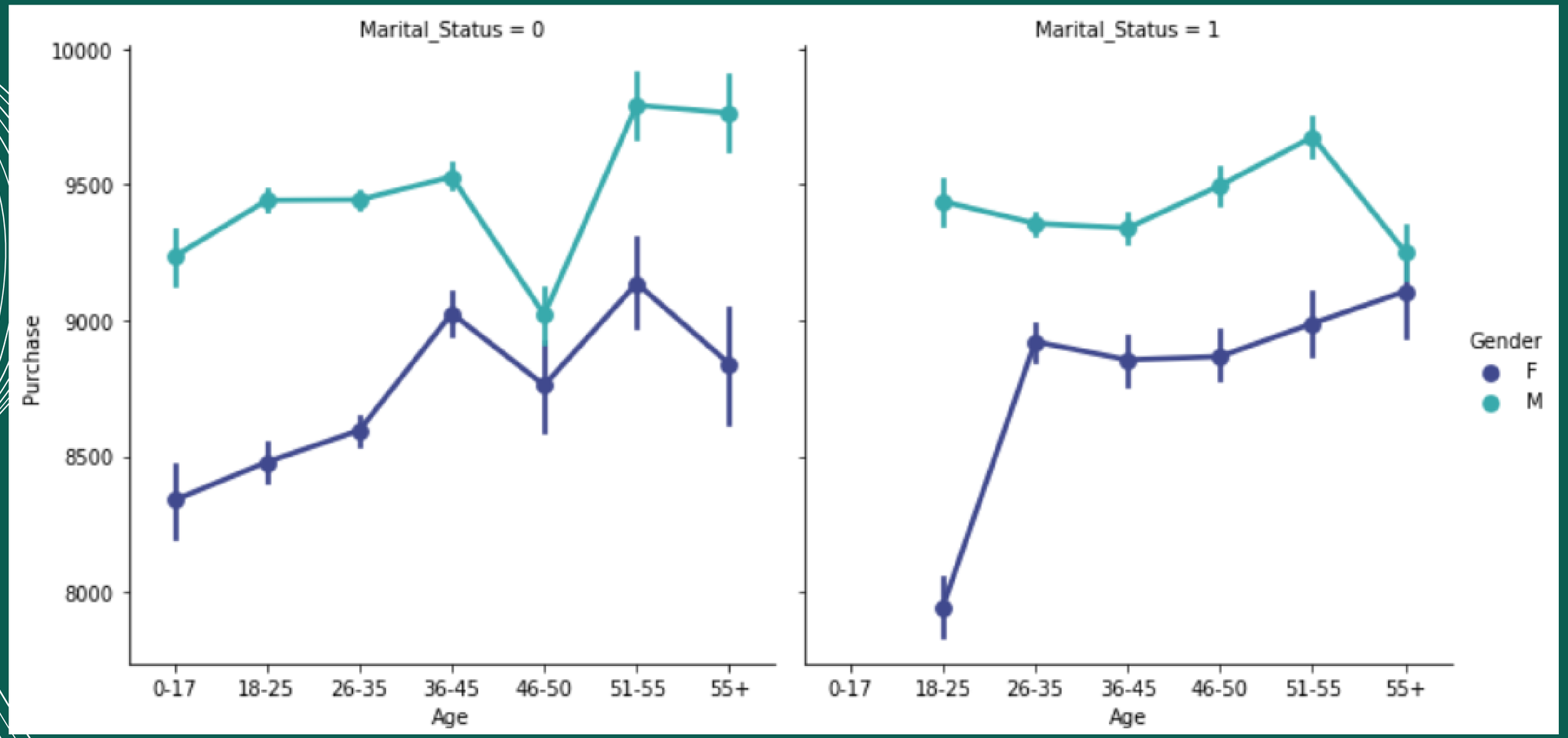
Top Selling Product

- P00265242 - product is the top selling followed by P0025442 and P0011742 ans second and third respectively.



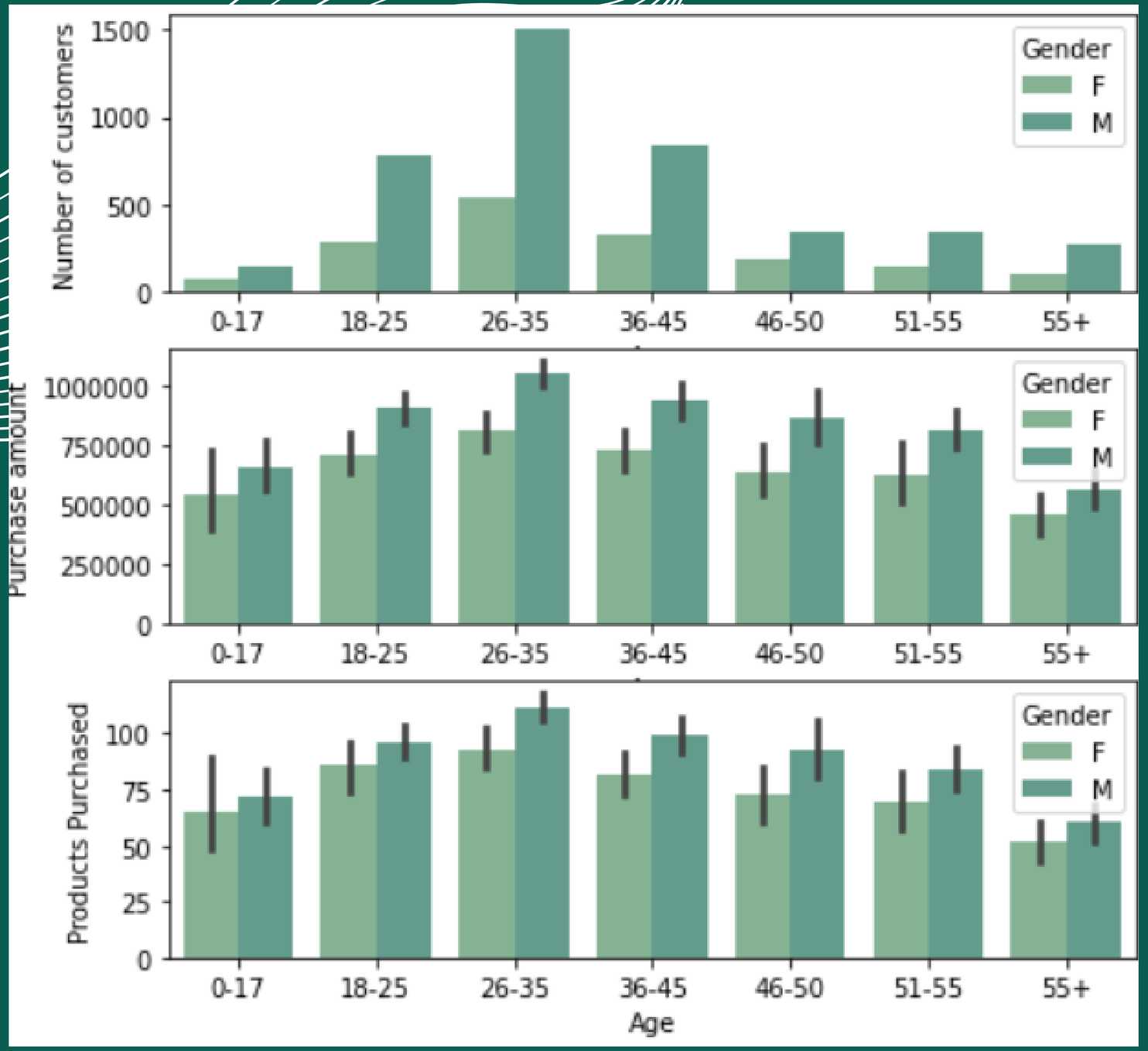
Top Total Purchase Amount Product Wise

- P00025442- product is having maximum total purchase amount contribution followed by P00110742 and P00255842 and second and third respectively.



Average Purchase - Marital Status

- Irrespective of the marital status mens tends to purchase more.



Customer Distribution

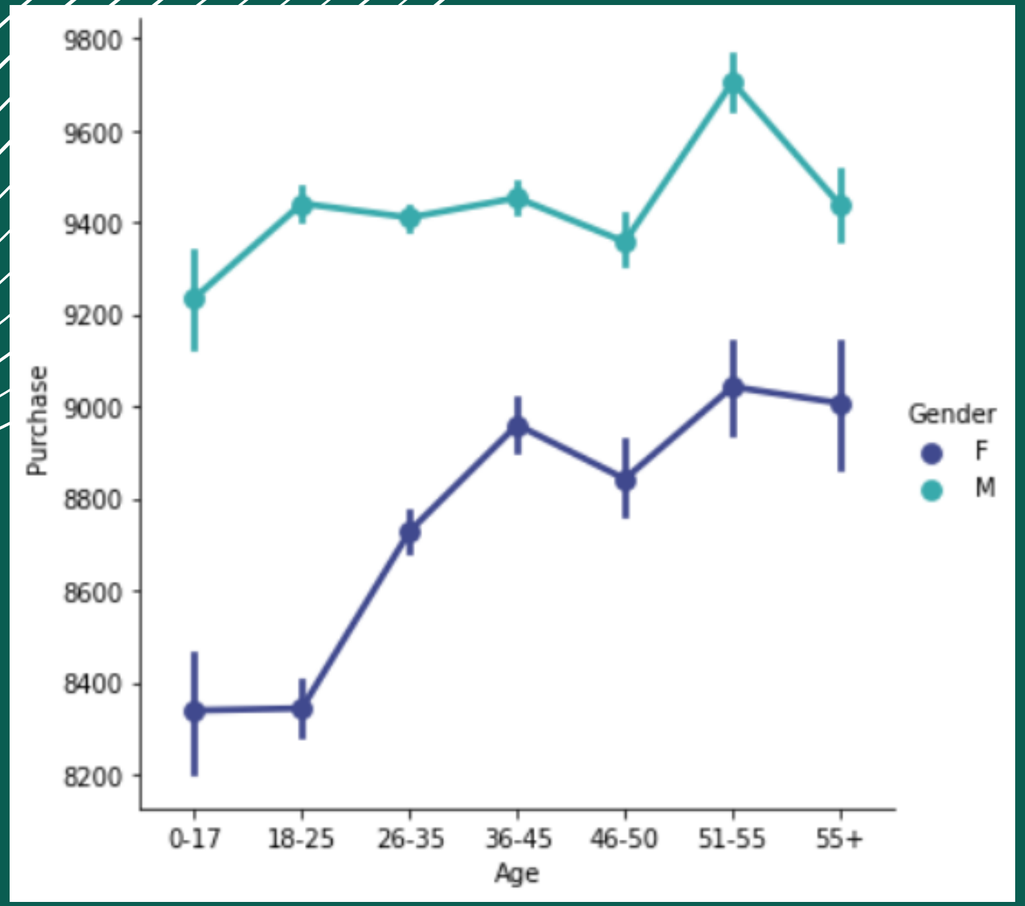
- The figure depicts the customer distribution across different age groups wrt to gender.

Total Purchase Amount

- The figure depicts the total purchase amount across different age groups wrt to gender.

Products Purchase

- Total products purchased distribution according to the age groups and genders within each group.



Average Purchase

- Graph depicts average purchase across the different age groups w.r.t gender.

Model Building

- Objective
- Regression Technique
- Variables
- Model Building
- Result Interpretation

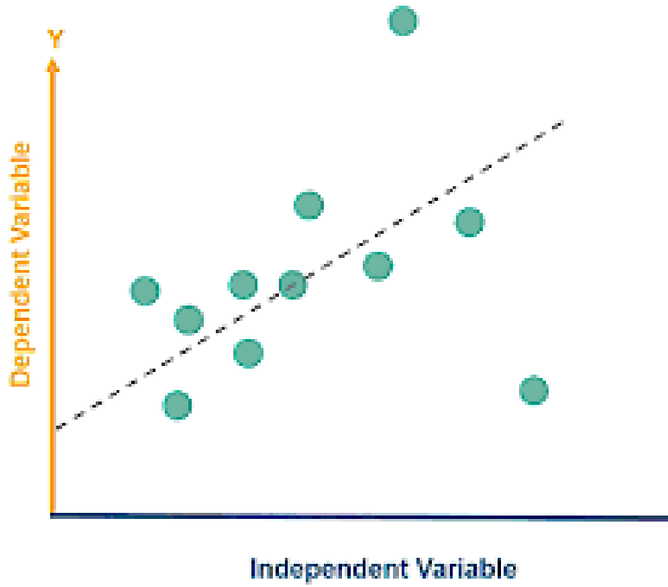
Objective

To build a model to predict the purchase amount of customers against various products which will help them to create personalized offers for customers against different products.

REGRESSION TECHNIQUE

Multiple Linear Regression

- It is an extension of Simple linear regression(SLR) where we use one dependent variable with one independent variable but in multiple linear regression(MLR) we can multiple independent variables.
- MLR is more useful in comparison with the SLR because even in real-world scenarios, predicting about one thing is based upon multiple factors.



VARIABLES

Dependent Variable:

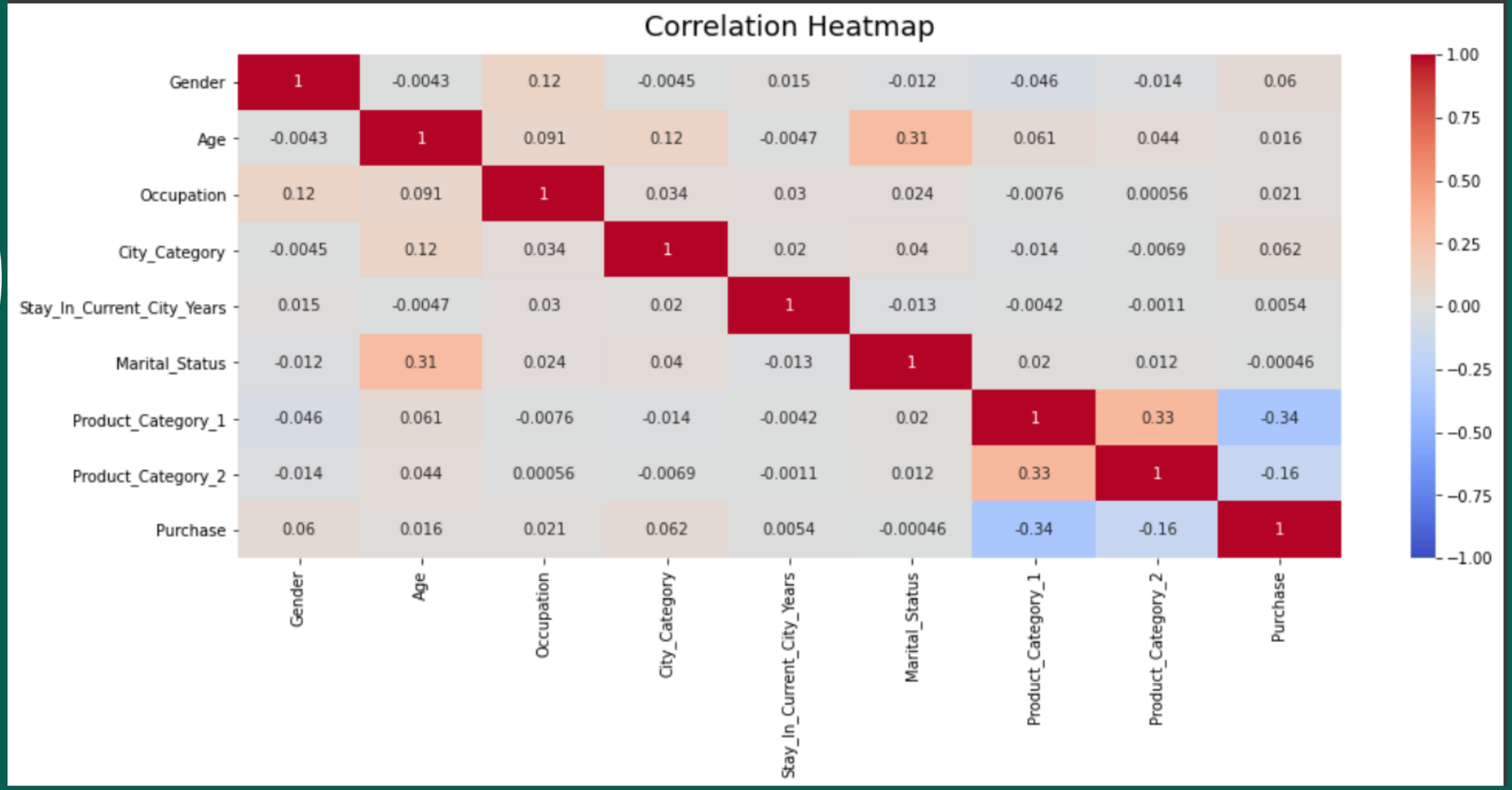
- Purchase - Predicting the purchase amount of customers against various products. It is a continuous variable.

Independent Variables:

- Gender: indicates the gender of the person making the transaction.
- Age: indicates the age group of the person making the transaction.
- Occupation: shows the occupation of the user, already labeled with numbers 0 to 20.
- City_Category: User's living city category. Cities are categorized into 3 different categories 'A', 'B', and 'C'.
- Stay_In_Current_City_Years: Indicates how long the users have lived in this city.
- Marital_Status: is 0 if the user is not married and 1 otherwise.
- Product_Category_1 to _3: Category of the product. All 3 are already labeled with numbers.

Model Building

- We first identified the missing values in our dataset and found that product_category_2 and product_category_3 had missing values.
- We replaced the missing values in product_category_2 with the median value of it and dropped product_category_3 as it had around 70% of missing values.
- Then we drop the variables such as user_id and product_id which will not add much value to our objective of predicting the sales purchase amount.
- We then encoded the following variables: age, gender, city_category, and stay_in_current_city_years.
- We split our dataset into train and test data in a 75:25 ratio.
- And lastly, we normalized our independent variables before applying Linear Regression.



Correlation Heatmap

- There is no significant correlation between the variables
- -0.34 between purchase and Product_Category_1

Model Building

- For our model building, we have taken purchase as our dependent variable and gender, age, occupation, city category, stay in current city years, marital status, product category 1, and product category 2 as our independent variables.

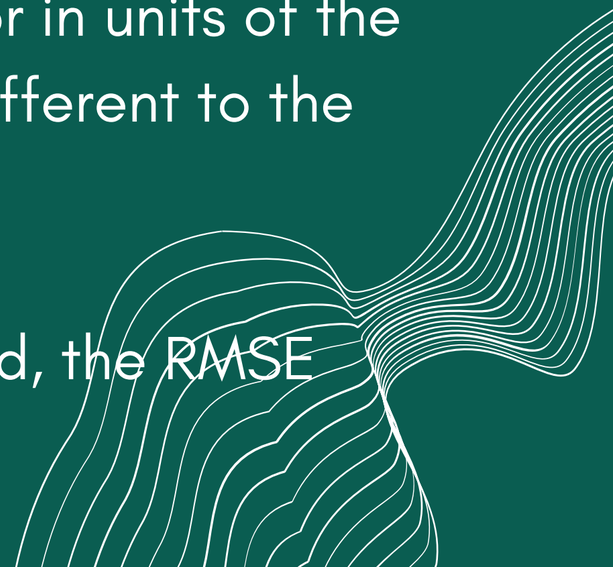
- **Model Equation:**
  Purchase = 19630.603 + 4551.56 * Gender - 711.303 * Age - 5219.498 * Occupation + 3266.152 * City_Category - 1171.31 * Stay_In_Current_City_Years -804.09 * Marital Status - 11912.08 * Product_Category_1 - 6542.90 * Product_Category_2.

Calculating Accuracy of Model

- Mean Absolute Error (MAE) and Root mean squared error (RMSE) are two of the most common metrics used to measure accuracy for continuous variables.
- MAE measures the average magnitude of the errors in a set of predictions, without considering their direction.
- RMSE is the square root of the average of squared differences between prediction and actual observation

Similarities - MAE and RMSE express average model prediction error in units of the variable of interest.Both metrics can range from 0 to ∞ and are indifferent to the direction of errors.

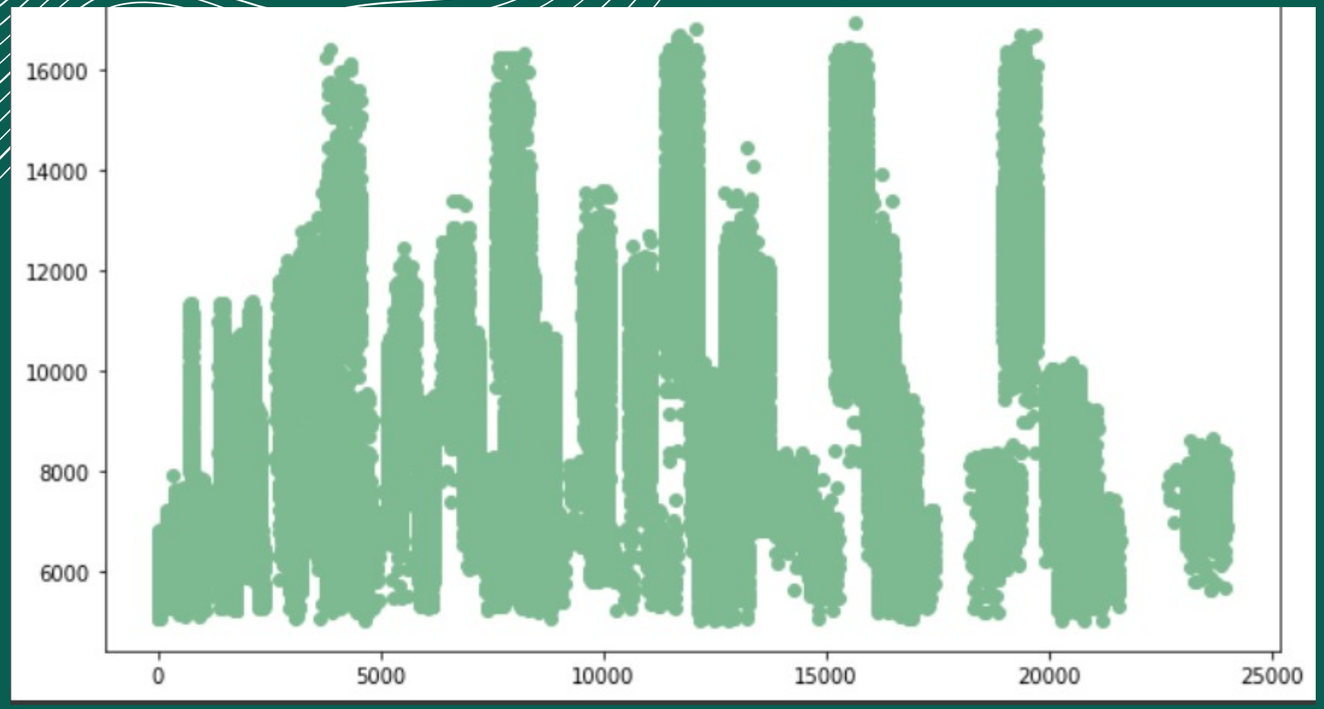
Differences - RMSE the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors.



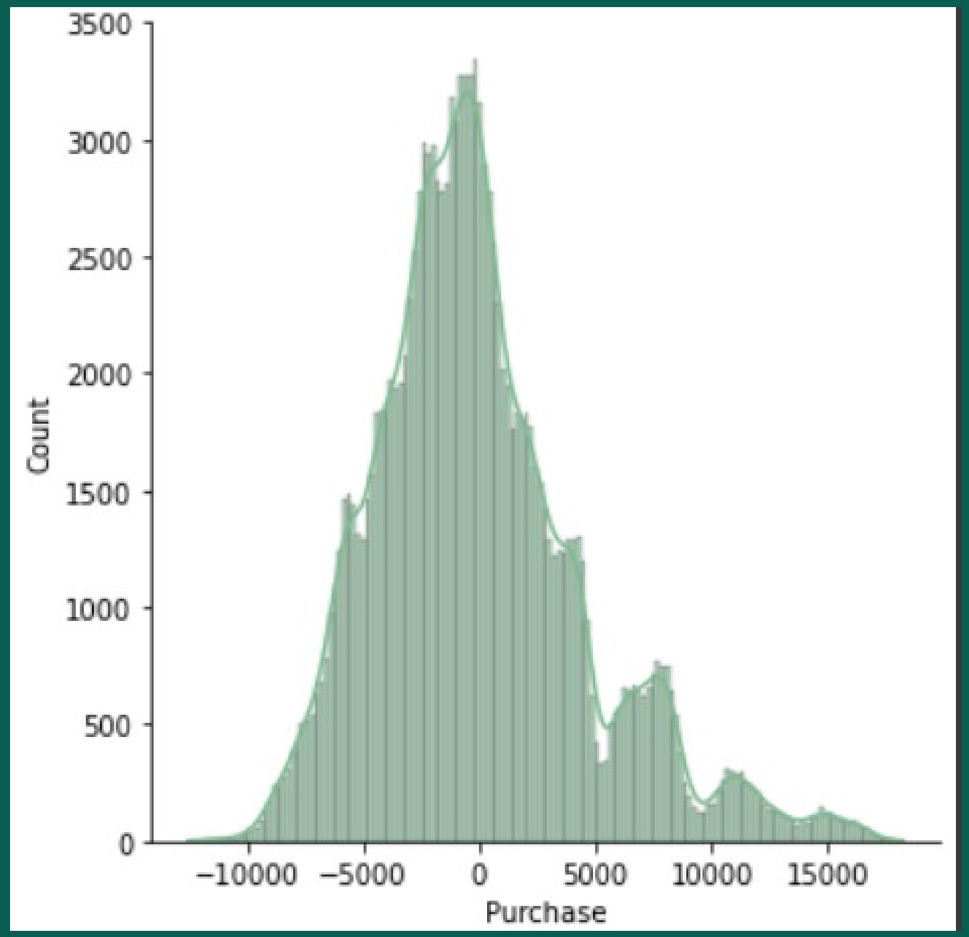
Result Interpretation

- The accuracy of our model is 17.93%
- The Mean Absolute Error is 3454.505.
- The Mean Squared Error is 20705643.072.
- The Root Mean Squared Error is 4550.345.
- The lesser the value of RMSE the better is the model.

Result Interpretation



- From this graph, we can see that there no linear relationship.
- This plot contains plots of the relationship between the observed residuals against those expected under the condition of normality. The closer the observed residuals fall in relation to the regression line, the more evidence of normality.



- From this graph, we can see that the residuals are not normally distibuted.

CONCLUSION

- We have developed a model to predict the purchase amount of customers against various products which will help them to create personalized offers for customers against different products.
- The prediction accuracy of our model is 17.93%.
- We can obtain more accuracy by using other models of Machine Learning.

Thank You