Special Offer

# BLACK FRIDAY

## SALE

@blackfridaysale

# INTRODUCTION

THE SHOPPING SECTOR HAS GREATLY EVOLVED DUE TO THE INTERNET REVOLUTION. NOWADAYS, MOST PEOPLE USE THE ONLINE SHOPPING METHOD AS IT IS EASIER THAN THE TRADITIONAL METHOD OF SHOPPING. THE BIGGEST ADVANTAGES OF ONLINE SHOPPING ARE CONVENIENCE, BETTER PRICES, MORE VARIETY, EASY PRICE COMPARISONS, NO CROWDS, ETC. THE COVID PANDEMIC HAS BOOSTED ONLINE SHOPPING.

BASICALLY, BLACK FRIDAY ORIGINATED IN THE USA. THIS DAY IS REFERRED TO AS THANKSGIVING DAY.  THE PURPOSE OF THIS DAY IS TO ORGANIZE THIS SALE TO PROMOTE CUSTOMERS TO BUY MORE PRODUCTS ONLINE SO THAT THEY CAN BOOST THE ONLINE SECTOR OF SHOPPING.

.

THE PREDICTION MODEL WHICH WE BUILT WILL PROVIDE A PREDICTION BASED ON THE AGE OF THE CUSTOMER, CITY CATEGORY, OCCUPATION, ETC. THE PREDICTION MODEL IS IMPLEMENTED BASED ON MODELS LIKE LINEAR REGRESSION, RIDGE REGRESSION, LASSO REGRESSION, DECISION TREE REGRESSOR, RANDOM FOREST REGRESSOR.

# DATASET

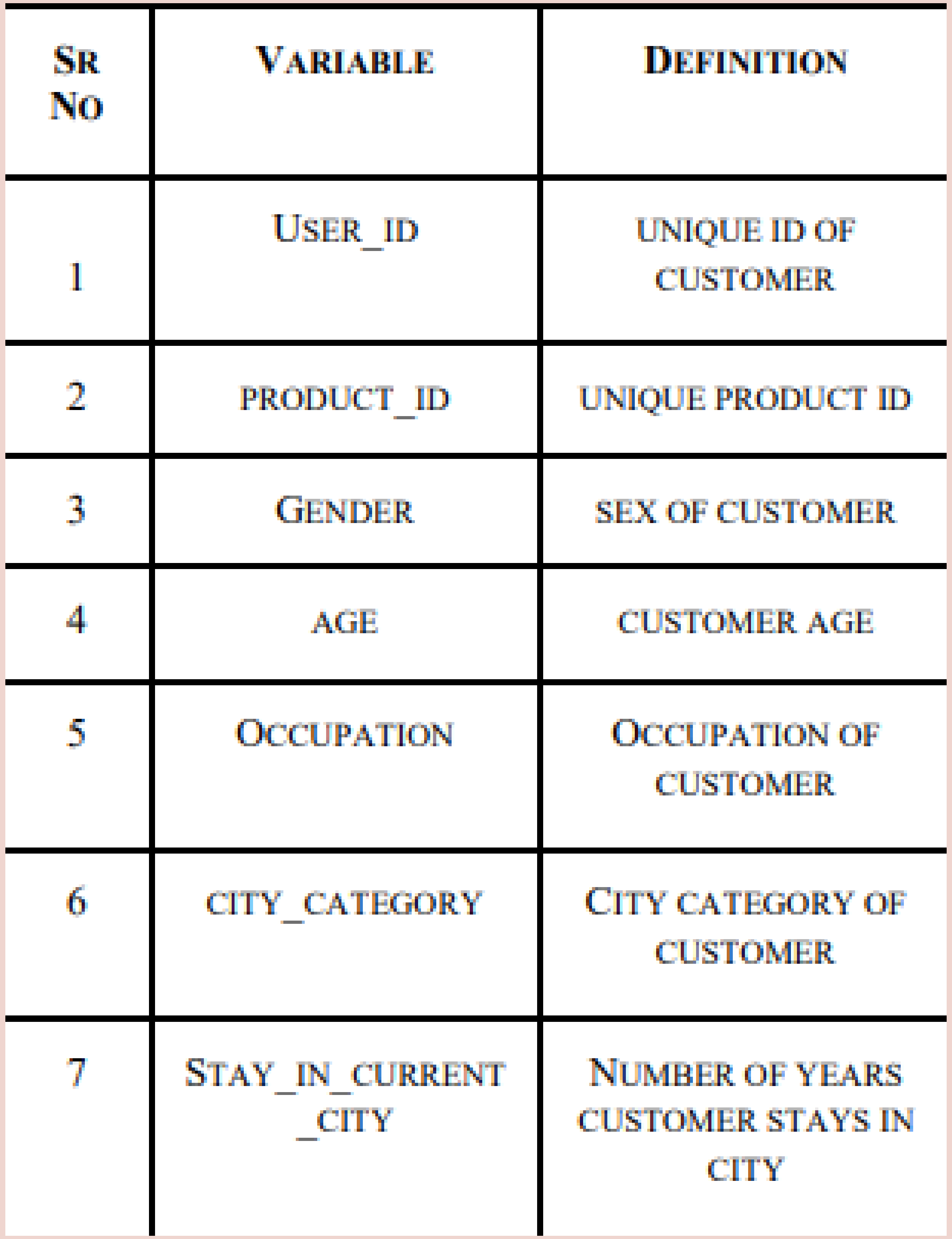
THE STUDY USES BLACK FRIDAY SALES DATASET PUBLICLY AVAILABLE ON KAGGLE. THE DATASET CONSISTS OF SALES TRANSACTION DATA. THE DATASET CONSISTS OF 5,50,069 ROWS.

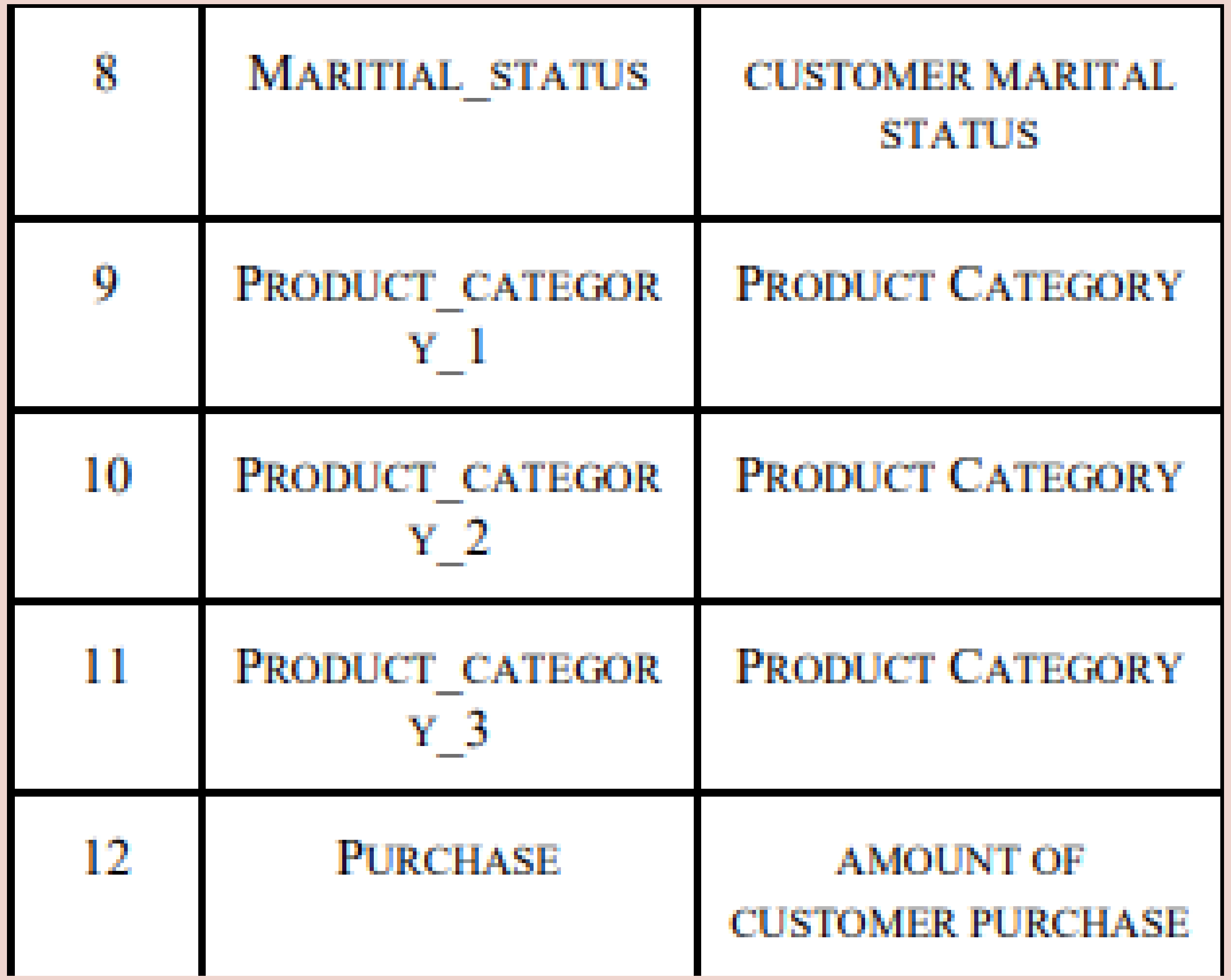
THE DATASET CONSISTS OF ATTRIBUTES SUCH AS

1. USER_ID
2. PRODUCT_ID
3. MARTIAL_STATUS
4. CITY_CATEGORY
5. OCCUPATION

   ETC.

THE DATASET DEFINITION IS MENTIONED IN TABLE IN THE NEXT SLIDE.





| Sr No | Variable | Definition |
|---|---|---|
| 1 | User_ID | unique ID of customer |
| 2 | product_id | unique product ID |
| 3 | Gender | sex of customer |
| 4 | age | customer age |
| 5 | Occupation | Occupation of customer |
| 6 | city_category | City category of customer |
| 7 | Stay_in_current_city | Number of years customer stays in city |
| 8 | Maritial_status | customer marital status |
| 9 | Product_category_1 | Product Category |
| 10 | Product_category_2 | Product Category |
| 11 | Product_category_3 | Product Category |
| 12 | Purchase | amount of customer purchase |

THE PURCHASE VARIABLE WILL BE THE PREDICTOR VARIABLE.

THE PURCHASE VARIABLE WILL PREDICT THE AMOUNT OF PURCHASES MADE BY A CUSTOMER ON BLACK FRIDAY SALES.

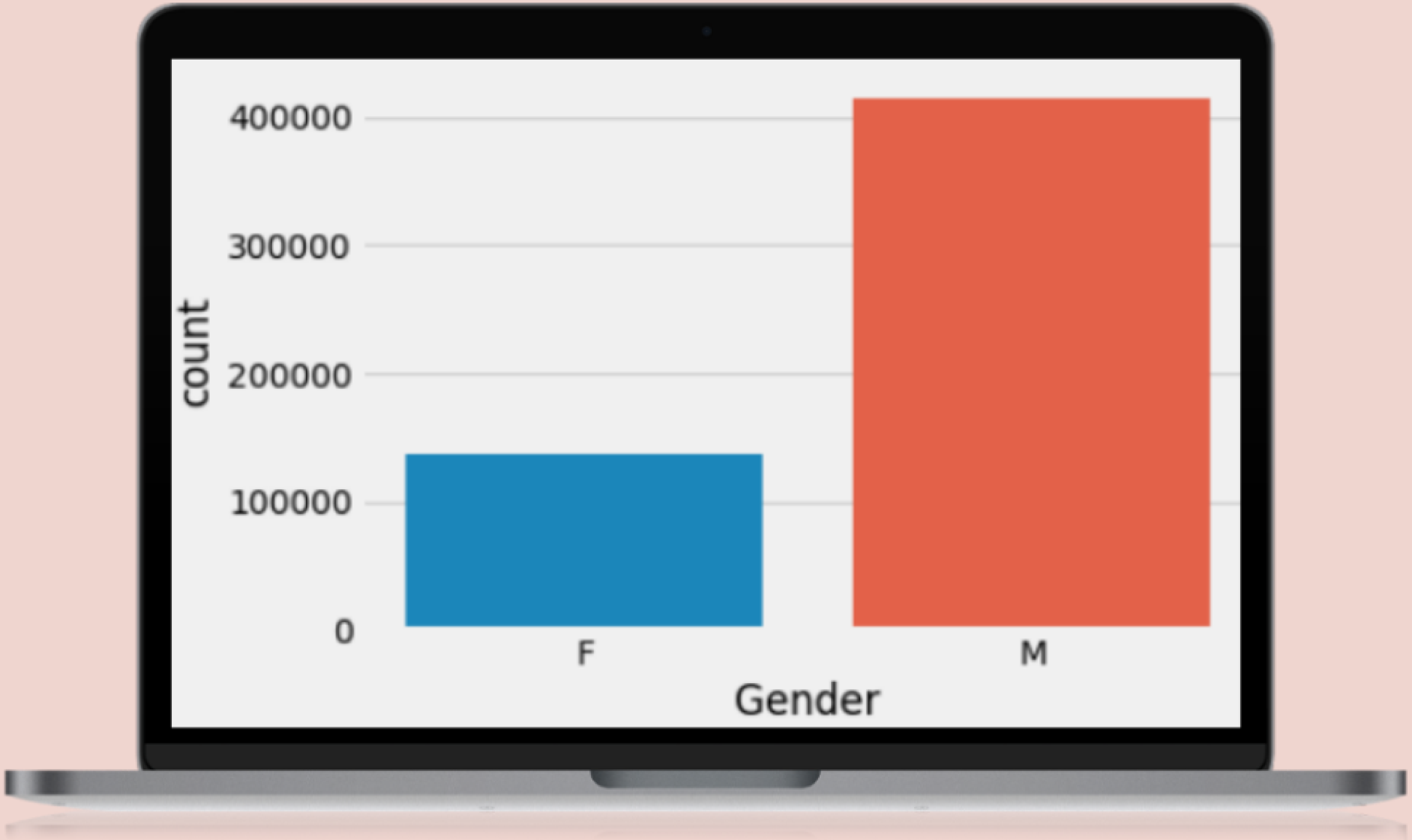
# DATA VISUALIZATION

DATA VISUALIZATION IS THE REPRESENTATION OF DATA THROUGH USE OF COMMON GRAPHICS, SUCH AS CHARTS, PLOTS, INFOGRAPHICS, AND EVEN ANIMATIONS. THESE VISUAL DISPLAYS OF INFORMATION COMMUNICATE COMPLEX DATA RELATIONSHIPS AND DATA-DRIVEN INSIGHTS IN A WAY THAT IS EASY TO UNDERSTAND.

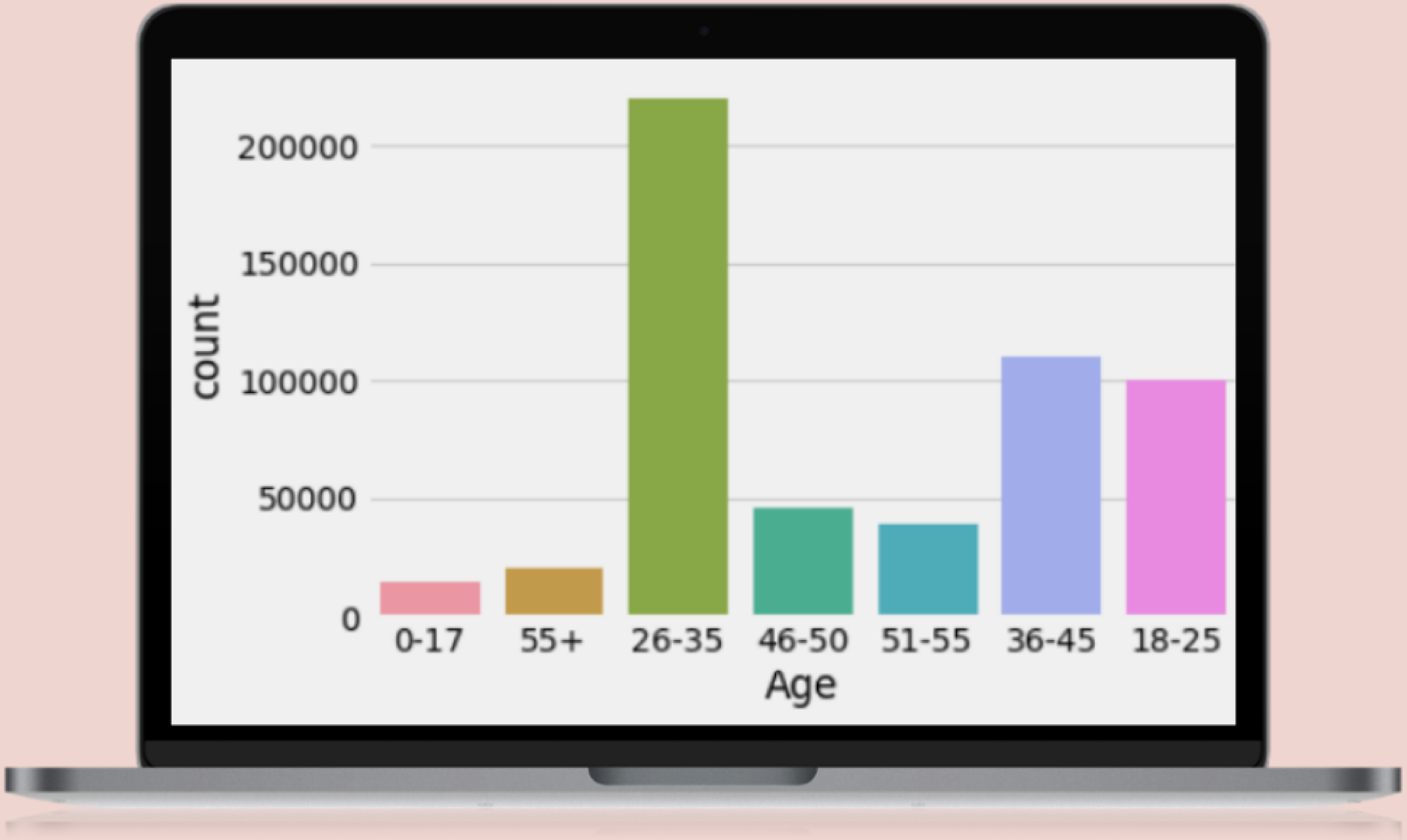
BASICALLY THE THREE MAIN GOALS OF DATA VISUALIZTION ARE:

- TO UNDERSTAND THE DATASET
- WORK WITHIN A CLEAR FRAMEWORK
- TELL A GOOD STORY

The count plots for different attributes are visualized as different figures given below.
The count plot for gender attributes is given below.

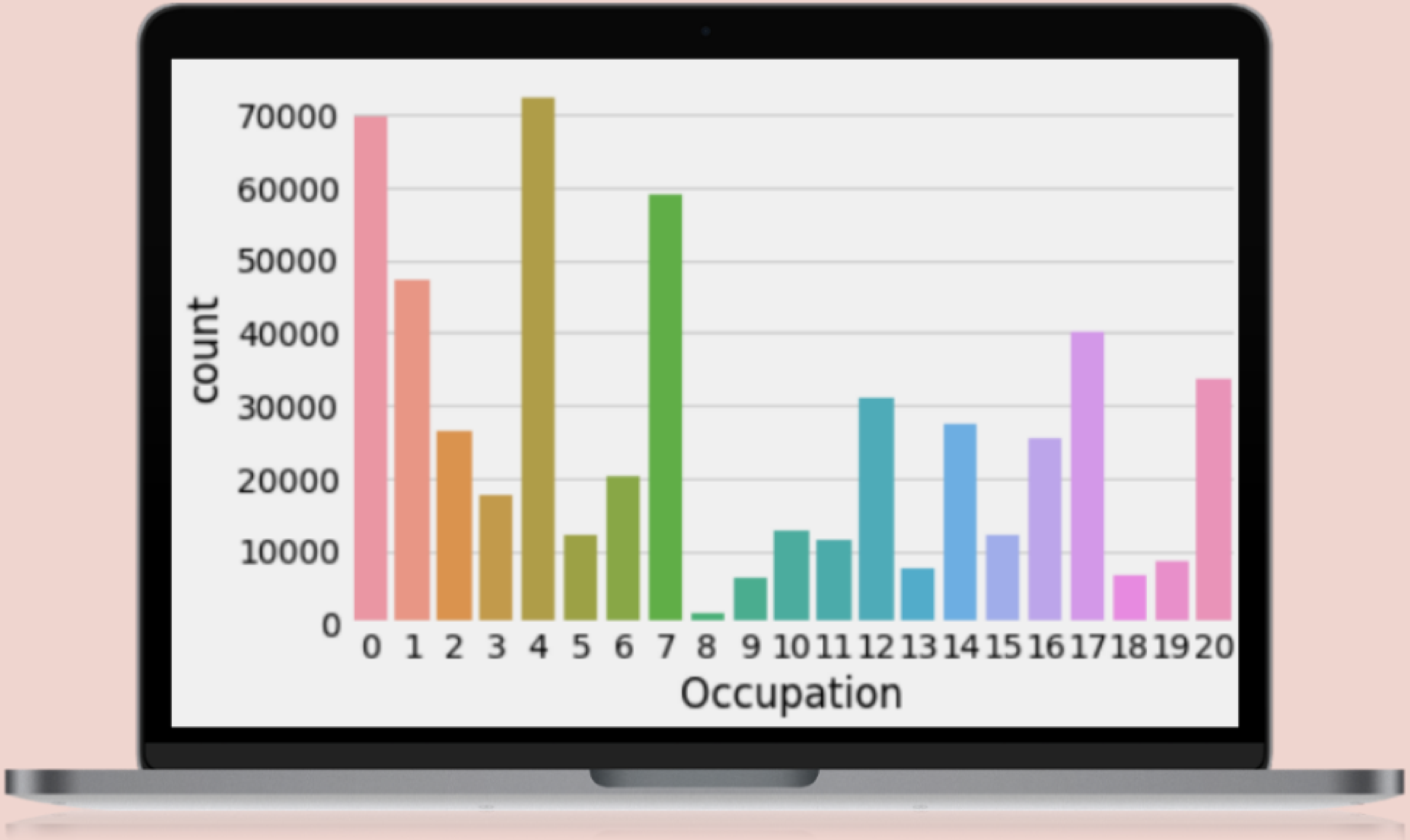


The count plot for the age attribute is given below.

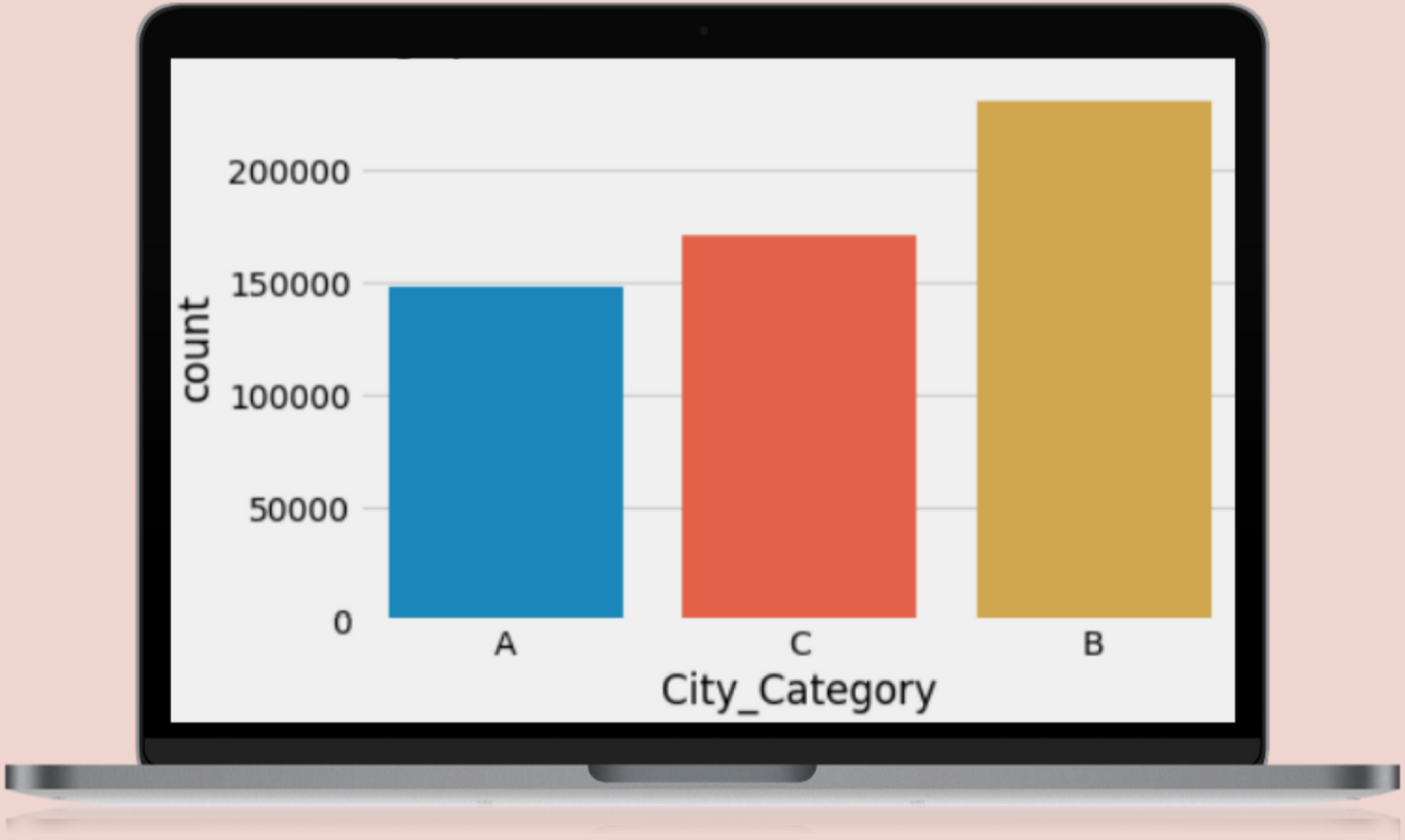


@blackfridaysale

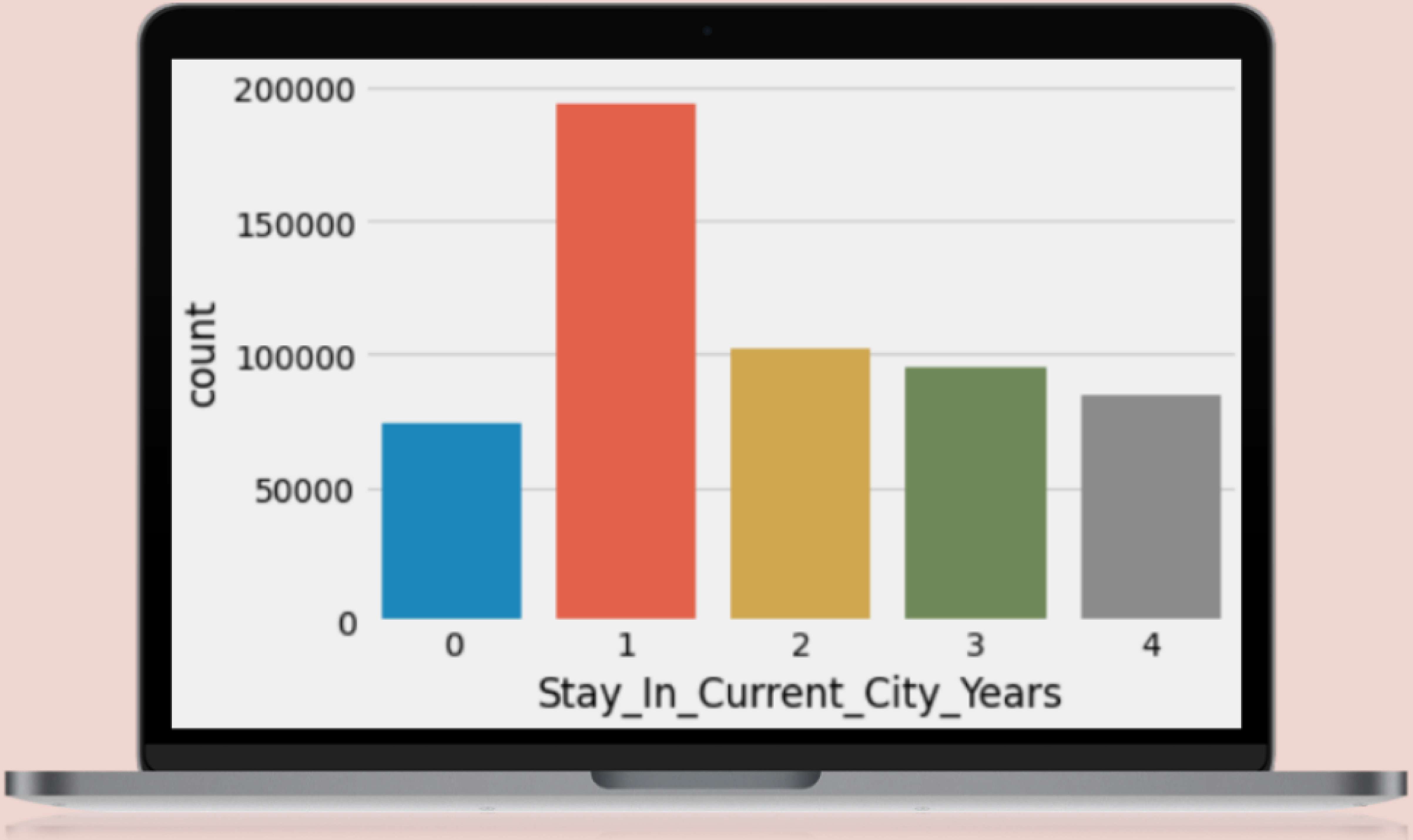
## THE COUNT PLOT FOR THE OCCUPATION ATTRIBUTE IS AS BELOW.



## The count plot for city_category is as below.

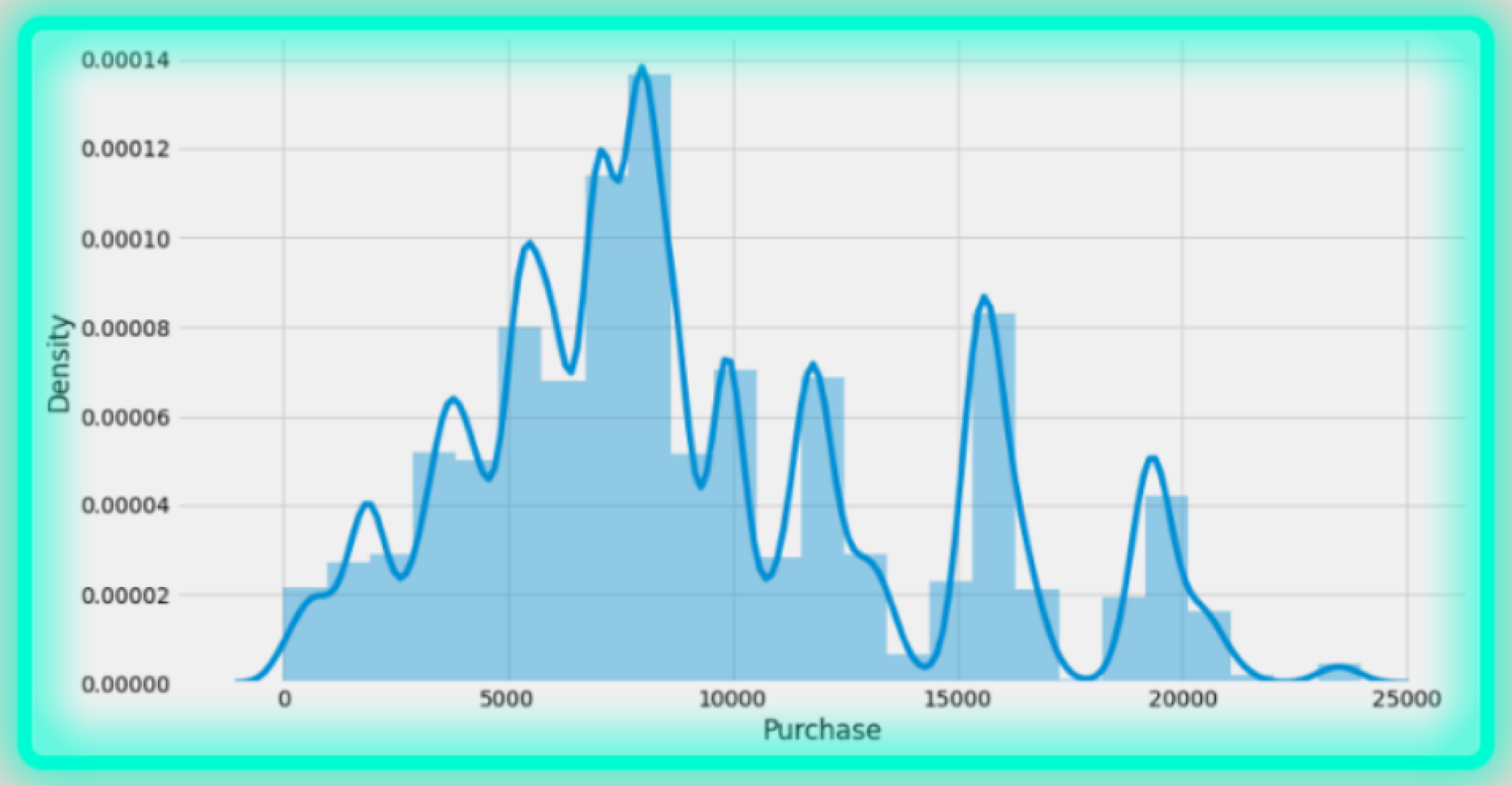


The count plot for Stay_In_Current_City is as below.



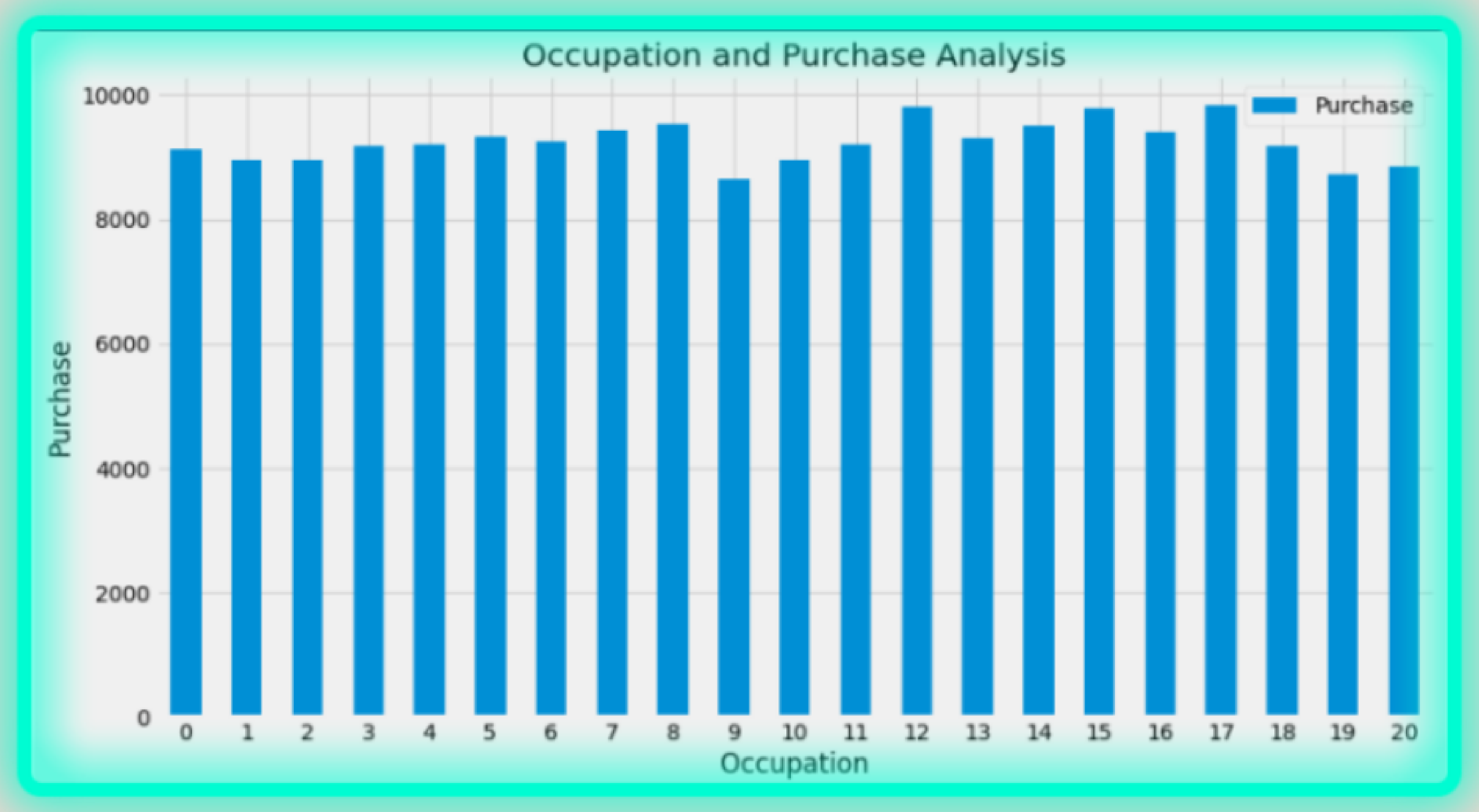
@blackfridaysale

## The density of purchase is as given below.



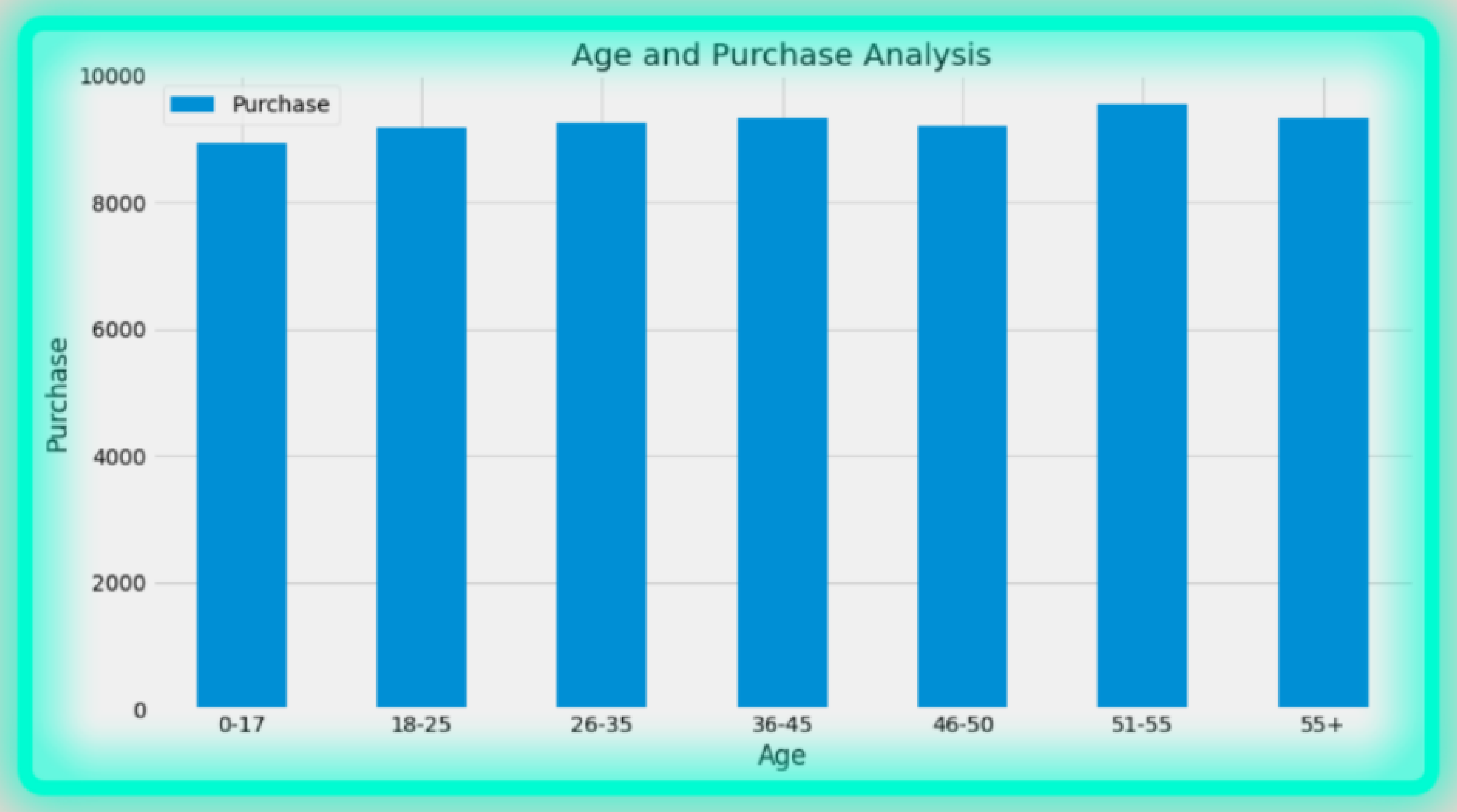
@blackfridaysale

## Bivariate analysis between occupation and purchase is shown below.



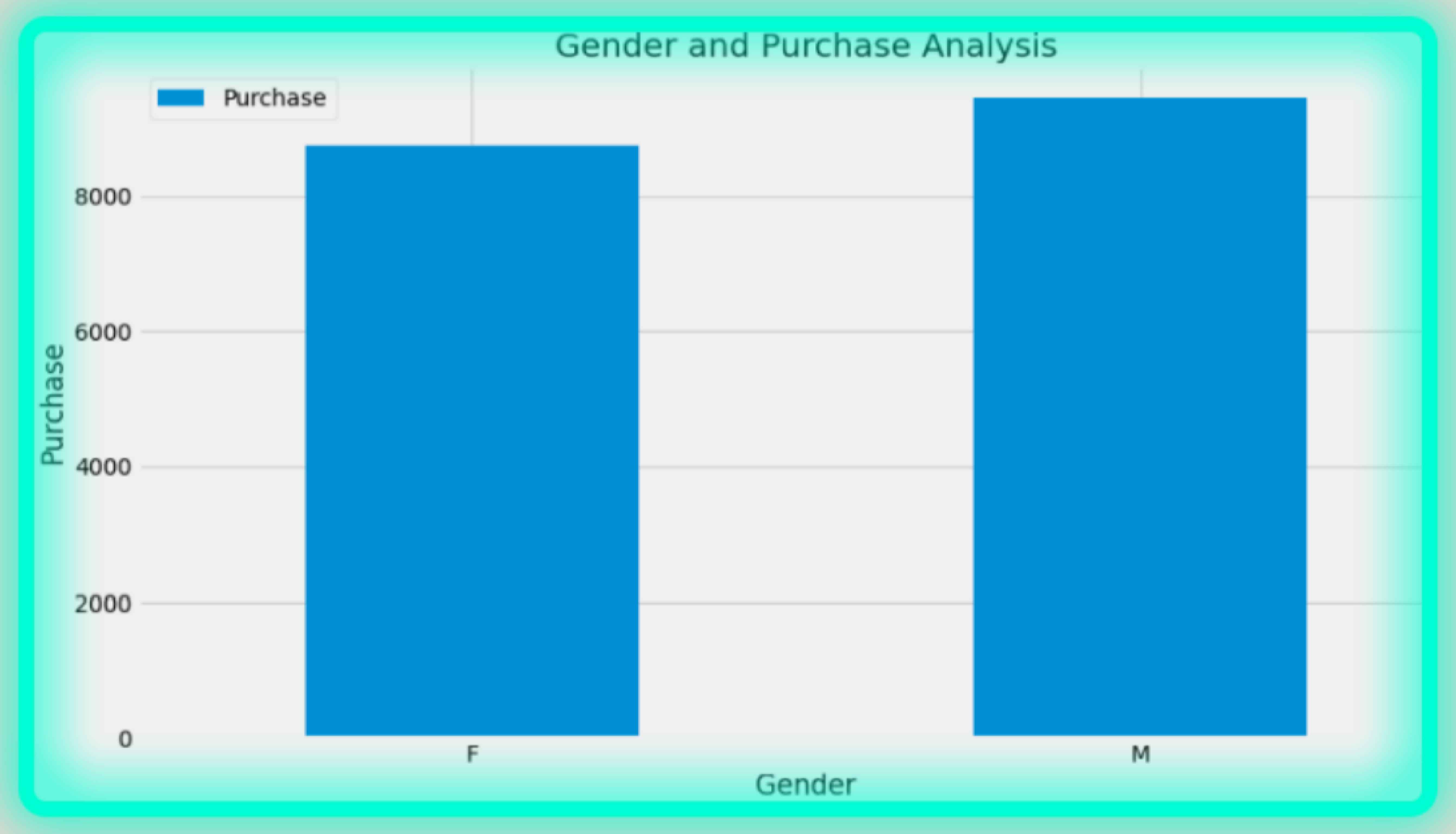
@blackfridaysale

## Bivariate analysis between age and purchase is shown below



@blackfridaysale

## Bivariate analysis between gender and purchase is shown  below.



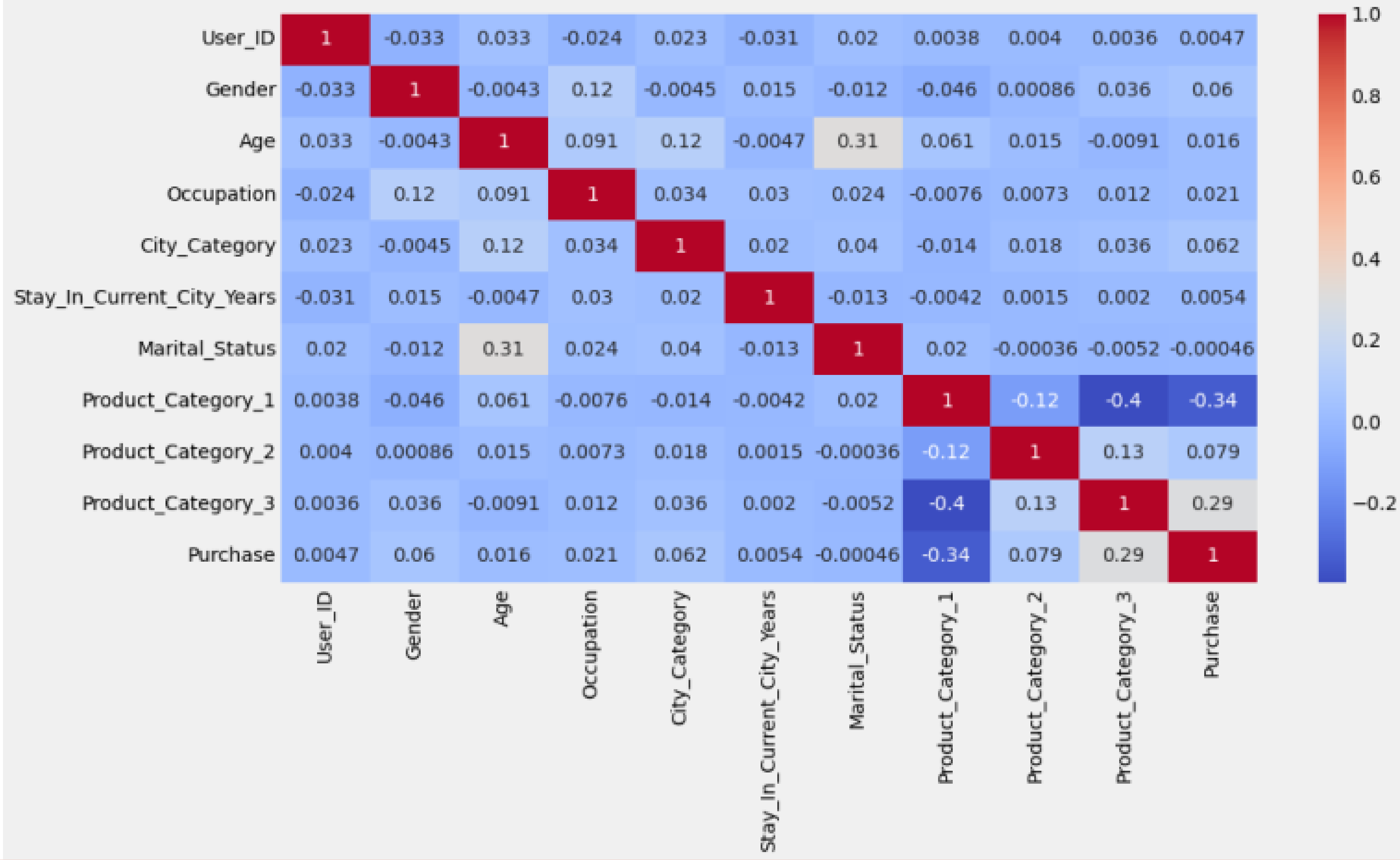
@blackfridaysale

# HEATMAP

Heatmap is used for determining the correlation between dataset attributes. The data of a given dataset can be easily represented graphically by using a Heatmap. It uses a color system to represent the correlation among different attributes.

Based on the Black Friday Sales Dataset, the heatmap obtained gives output which we will see in next slide. The observation based on the heatmap is the attributes age and martial_status, product_category_3 and purchase have a correlation.

From correlation matrix we take age, gender, and occupation as they have high correlation to purchase

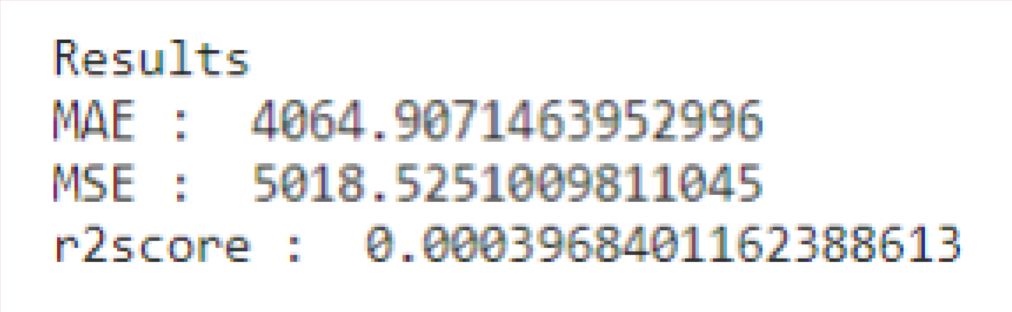


| | User_ID | Gender | Age | Occupation | City_Category | Stay_In_Current_City_Years | Marital_Status | Product_Category_1 | Product_Category_2 | Product_Category_3 | Purchase |
|---|---|---|---|---|---|---|---|---|---|---|---|
| User_ID | 1 | -0.033 | 0.033 | -0.024 | 0.023 | -0.031 | 0.02 | 0.0038 | 0.004 | 0.0036 | 0.0047 |
| Gender | -0.033 | 1 | -0.0043 | 0.12 | -0.0045 | 0.015 | -0.012 | -0.046 | 0.00086 | 0.036 | 0.06 |
| Age | 0.033 | -0.0043 | 1 | 0.091 | 0.12 | -0.0047 | 0.31 | 0.061 | 0.015 | -0.0091 | 0.016 |
| Occupation | -0.024 | 0.12 | 0.091 | 1 | 0.034 | 0.03 | 0.024 | -0.0076 | 0.0073 | 0.012 | 0.021 |
| City_Category | 0.023 | -0.0045 | 0.12 | 0.034 | 1 | 0.02 | 0.04 | -0.014 | 0.018 | 0.036 | 0.062 |
| Stay_In_Current_City_Years | -0.031 | 0.015 | -0.0047 | 0.03 | 0.02 | 1 | -0.013 | -0.0042 | 0.0015 | 0.002 | 0.0054 |
| Marital_Status | 0.02 | -0.012 | 0.31 | 0.024 | 0.04 | -0.013 | 1 | 0.02 | -0.00036 | -0.0052 | -0.00046 |
| Product_Category_1 | 0.0038 | -0.046 | 0.061 | -0.0076 | -0.014 | -0.0042 | 0.02 | 1 | -0.12 | -0.4 | -0.34 |
| Product_Category_2 | 0.004 | 0.00086 | 0.015 | 0.0073 | 0.018 | 0.0015 | -0.00036 | -0.12 | 1 | 0.13 | 0.079 |
| Product_Category_3 | 0.0036 | 0.036 | -0.0091 | 0.012 | 0.036 | 0.002 | -0.0052 | -0.4 | 0.13 | 1 | 0.29 |
| Purchase | 0.0047 | 0.06 | 0.016 | 0.021 | 0.062 | 0.0054 | -0.00046 | -0.34 | 0.079 | 0.29 | 1 |

# METHODOLOGY

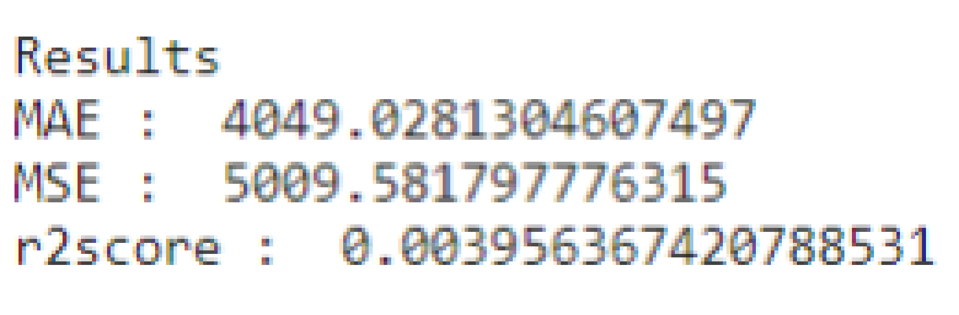
1) Polynomial regression: Polynomial features are those features created by raising existing features to an exponent. Here we are using Polynomial Features with a degree as 3.

a) Age:



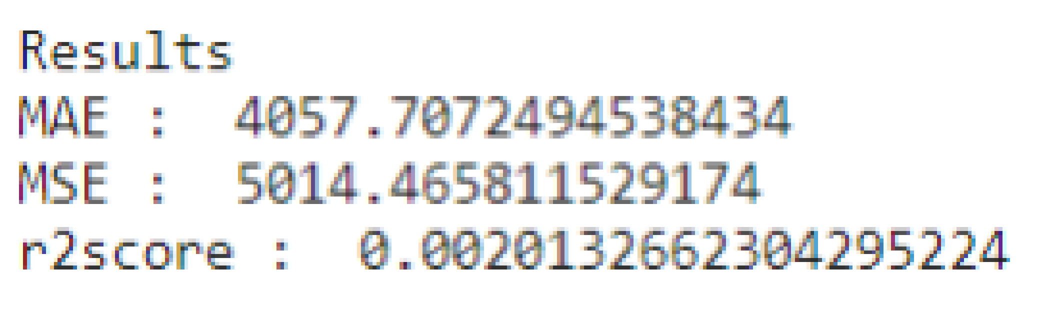
```
Results
MAE :  4064.9071463952996
MSE :  5018.5251009811045
r2score :  0.0003968401162388613
```

b) Gender:



```
Results
MAE :  4049.0281304607497
MSE :  5009.581797776315
r2score :  0.003956367420788531
```

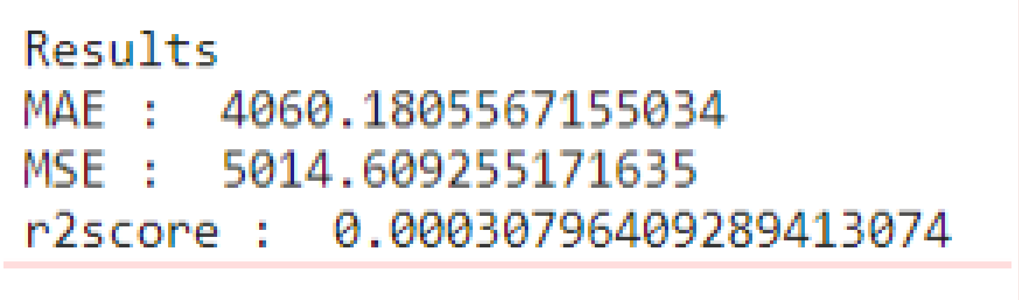
c) Occupation:



```
Results
MAE :  4057.7072494538434
MSE :  5014.465811529174
r2score :  0.002013266230429522 4
```

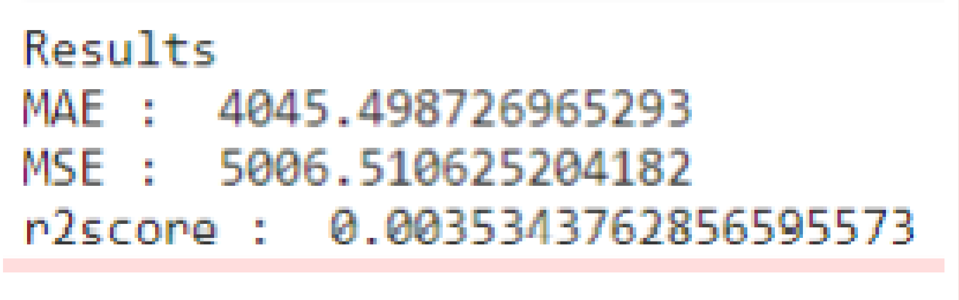
2) **Linear Regression: Linear Regression is one of the supervised machine learning algorithms. A regression problem can be stated as a case when the output variable is continuous. Linear regression predicts a dependent variable (y) based on a given independent variable (x).**

### a) Age:



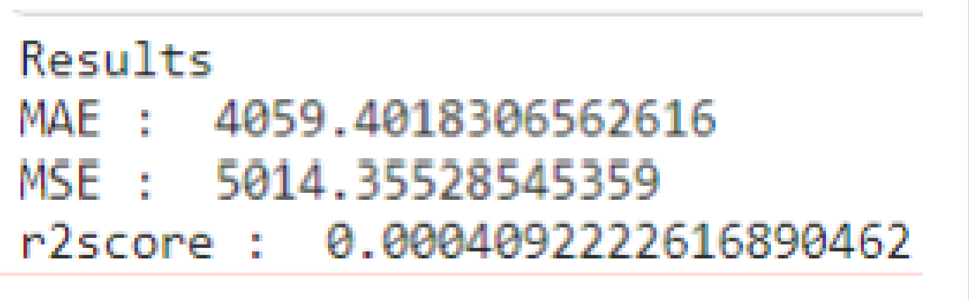
```
Results
MAE :  4060.1805567155034
MSE :  5014.609255171635
r2score :  0.00030796409289413074
```

### b) Gender:



```
Results
MAE :  4045.498726965293
MSE :  5006.510625204182
r2score :  0.0035343762856595573
```

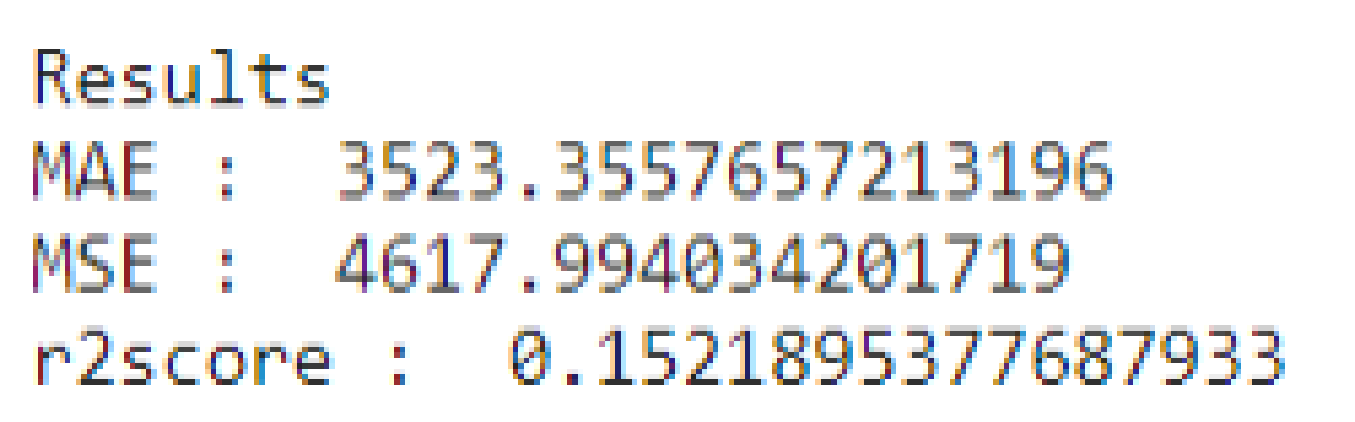
### c) Occupation:



```
Results
MAE :  4059.4018306562616
MSE :  5014.35528545359
r2score :  0.0004092222616890462
```

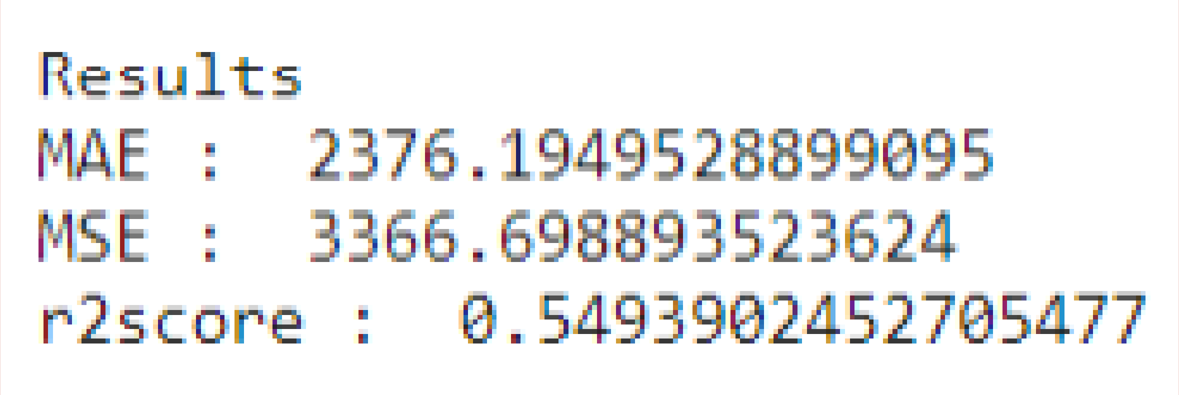
3) Linear regression on multiple parameters: also known simply as multiple regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

(User_ID', 'Product_ID', 'Purchase') the given parameters are removed.



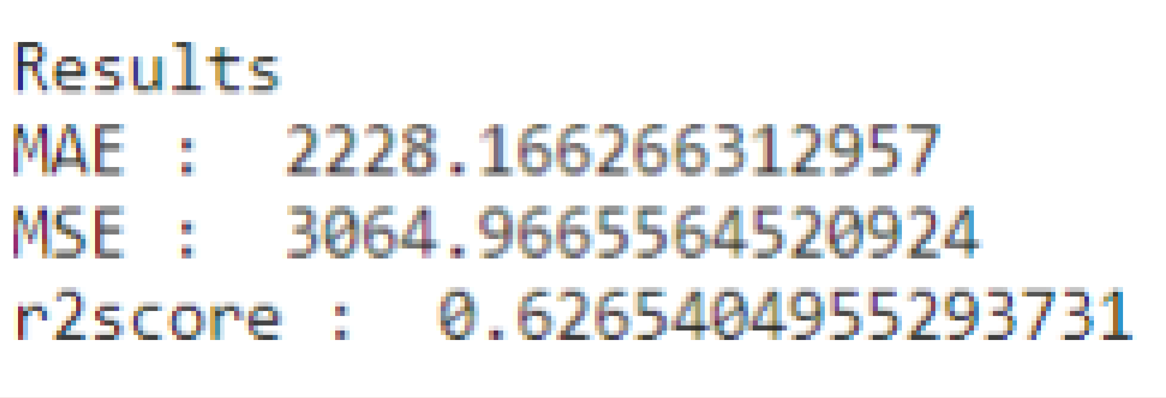
```
Results
MAE :  3523.3557657213196
MSE :  4617.994034201719
r2score :  0.1521895377687933
```

4) Decision Tree Regressor: The Decision Tree model builds a tree-like structure for regression or classification models. The dataset is simply broken down into smaller subsets. In a DT the control statements or values are a basis for branching, and the splitting node contains data points on either side, depending on the value of a specific attribute.



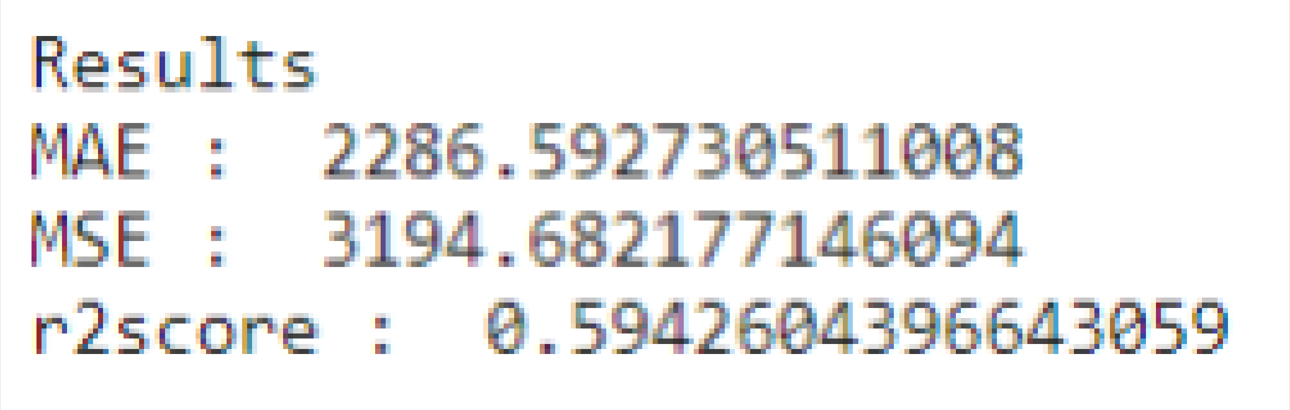
```
Results
MAE :  2376.1949528899095
MSE :  3366.698893523624
r2score :  0.5493902452705477
```

5) Random Forest Regressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The subsample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.



```
Results
MAE :  2228.166266312957
MSE :  3064.9665564520924
r2score :  0.6265404955293731
```

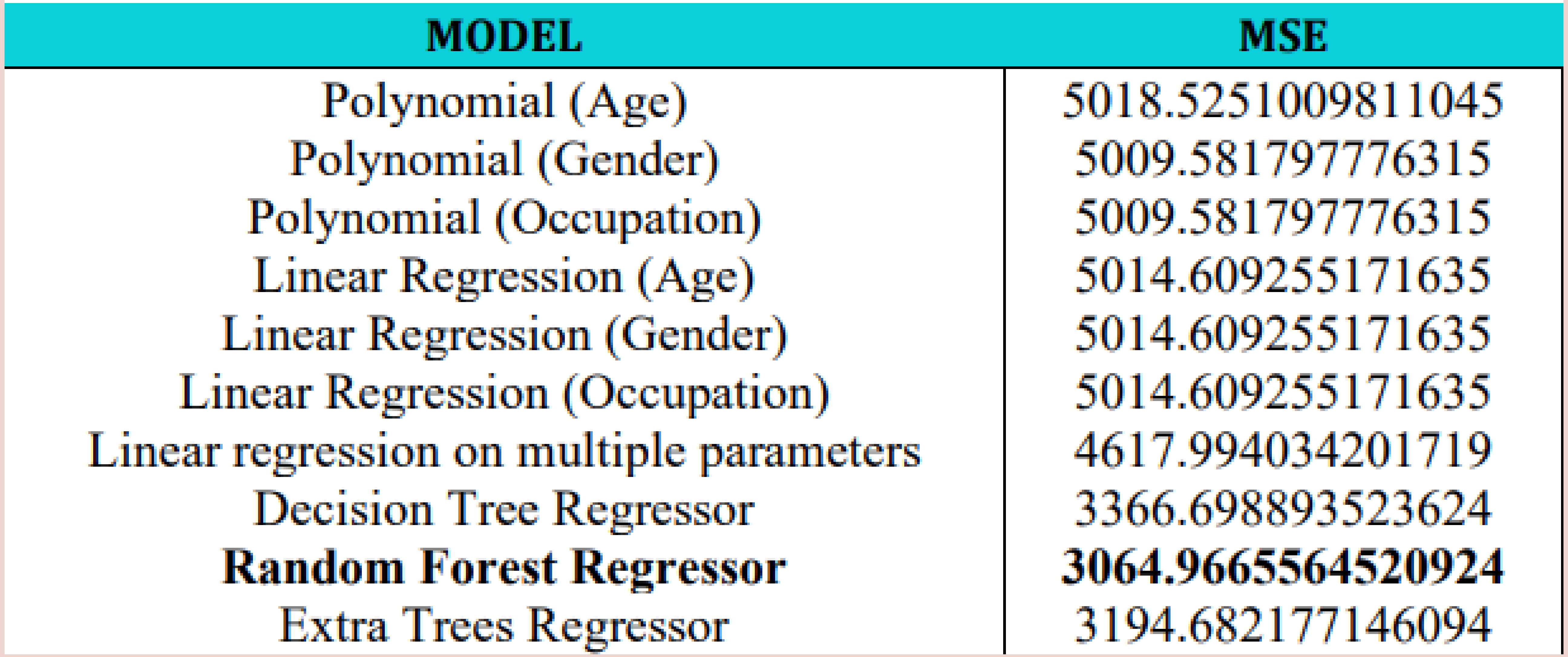
**6) Extra tree regressor model: This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting.**



```
Results
MAE :  2286.592730511008
MSE :  3194.682177146094
r2score :  0.5942604396643059
```

# RESULT

**The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points.**



| MODEL | MSE |
|---|---|
| Polynomial (Age) | 5018.5251009811045 |
| Polynomial (Gender) | 5009.581797776315 |
| Polynomial (Occupation) | 5009.581797776315 |
| Linear Regression (Age) | 5014.609255171635 |
| Linear Regression (Gender) | 5014.609255171635 |
| Linear Regression (Occupation) | 5014.609255171635 |
| Linear regression on multiple parameters | 4617.994034201719 |
| Decision Tree Regressor | 3366.698893523624 |
| **Random Forest Regressor** | **3064.9665564520924** |
| Extra Trees Regressor | 3194.682177146094 |

@blackfridaysale

# CONCLUSION

With traditional methods not being of much help to business growth in terms of revenue, the use of Machine learning approaches proves to be an important point for the shaping of the business plan taking into consideration the shopping pattern of consumers.  Thus, the dataset is used for the experimentation, Black Friday Sales Dataset from Kaggle.

The models used are Linear Regression, Polynomial Regression, Linear regression on multiple parameters, Decision Tree Regressor, Extra Trees Regressor and Random Forest Regressor. The evaluation measure used is Mean Squared Error (MSE).

As future research, we can perform hyperparameter tuning and apply different machine learning algorithms

# THANK YOU

NAMAHA SHIVAYA

@blackfridaysale