MLFA Assignment 2 - REPORT

Name : Shiva Ganesh Reddy Lakkasani
Roll Number : 20EE10069

Characteristics of given dataset :

RangeIndex: 550068 entries, 0 to 550067
Data columns (total 12 columns):

#	Column	Non-Null Count	Dtype
0	User_ID	550068 non-null	int64
1	Product_ID	550068 non-null	object
2	Gender	550068 non-null	object
3	Age	550068 non-null	object
4	Occupation	550068 non-null	int64
5	City_Category	550068 non-null	object
6	Stay_In_Current_City_Years	550068 non-null	object
7	Marital_Status	550068 non-null	int64
8	Product_Category_1	550068 non-null	int64
9	Product_Category_2	376430 non-null	float64
10	Product_Category_3	166821 non-null	float64
11	Purchase	550068 non-null	int64

dtypes: float64(2), int64(5), object(5)

Data size after one-hot encoding :
[550068 rows x 22 columns]

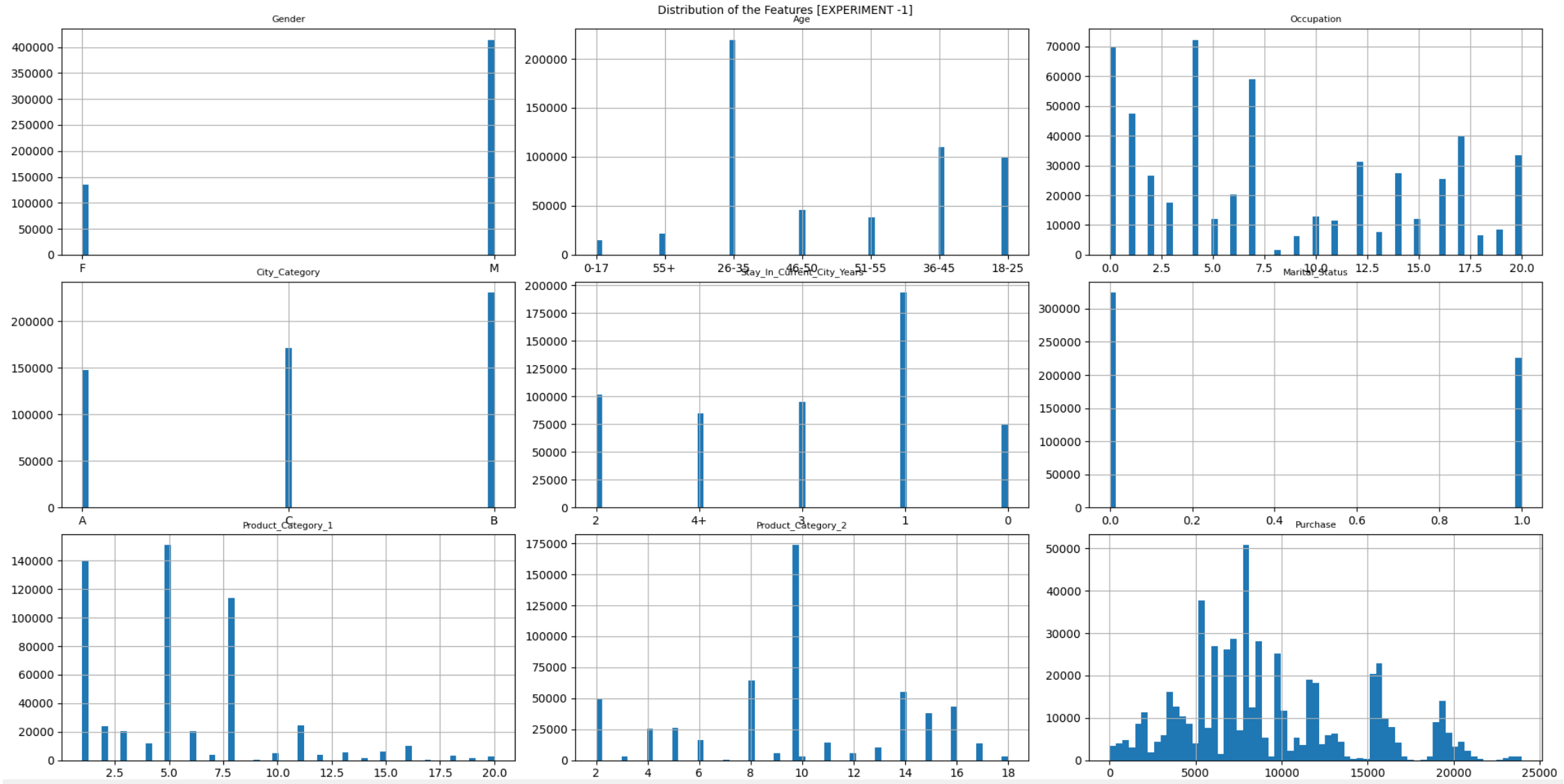
For X :
Actual set size: 550068
Training set size: 330040
Validation set size: 110014
Test set size: 110014

For y :
Actual set size: 550068
Training set size: 330040
Validation set size: 110014
Test set size: 110014

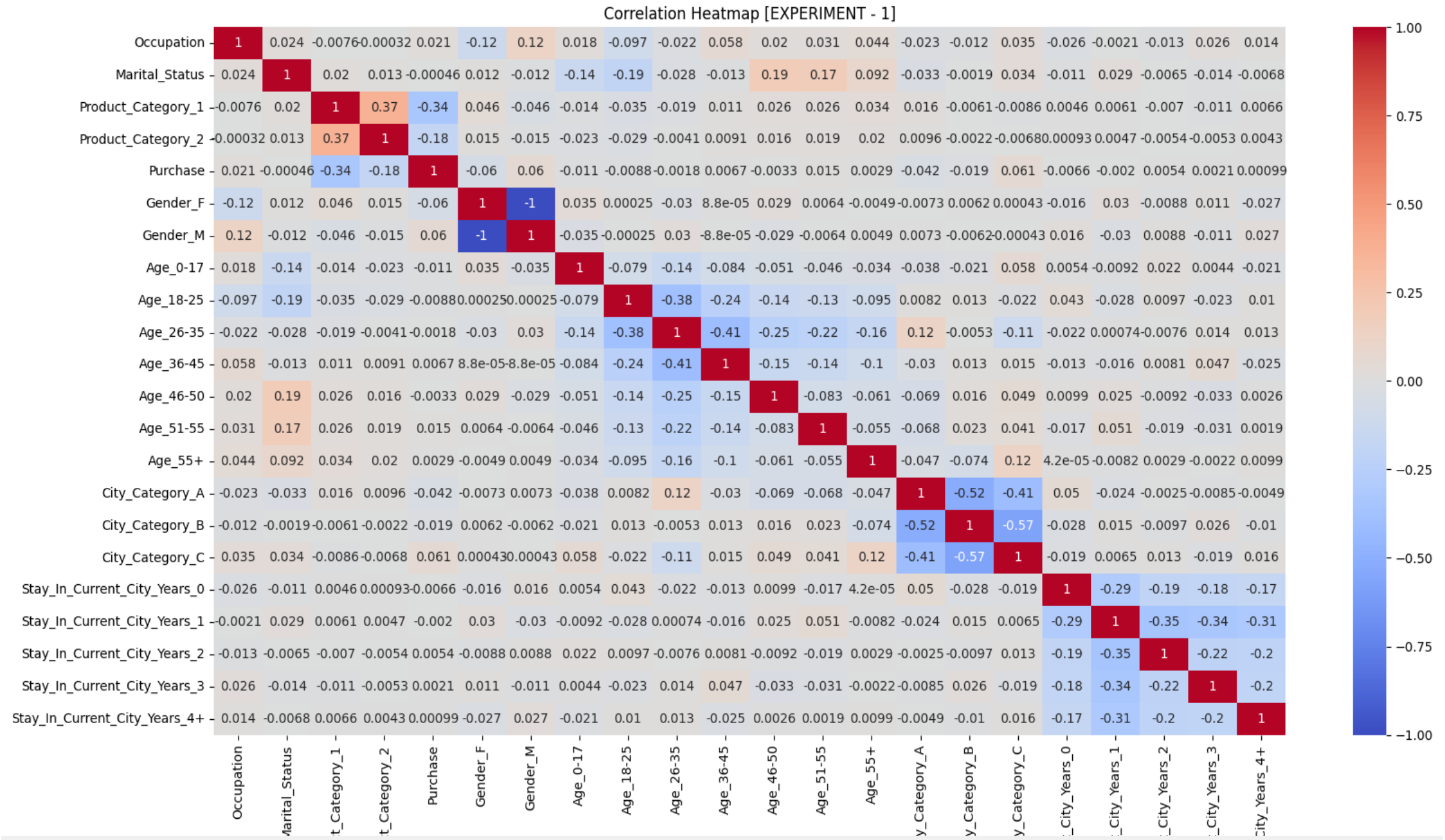
A. Experiment 1:

EDA Plots and correlation heatmap

1. Distribution of Features :



2. Correlation Heatmap [After one-hot encoding of categorical features] :



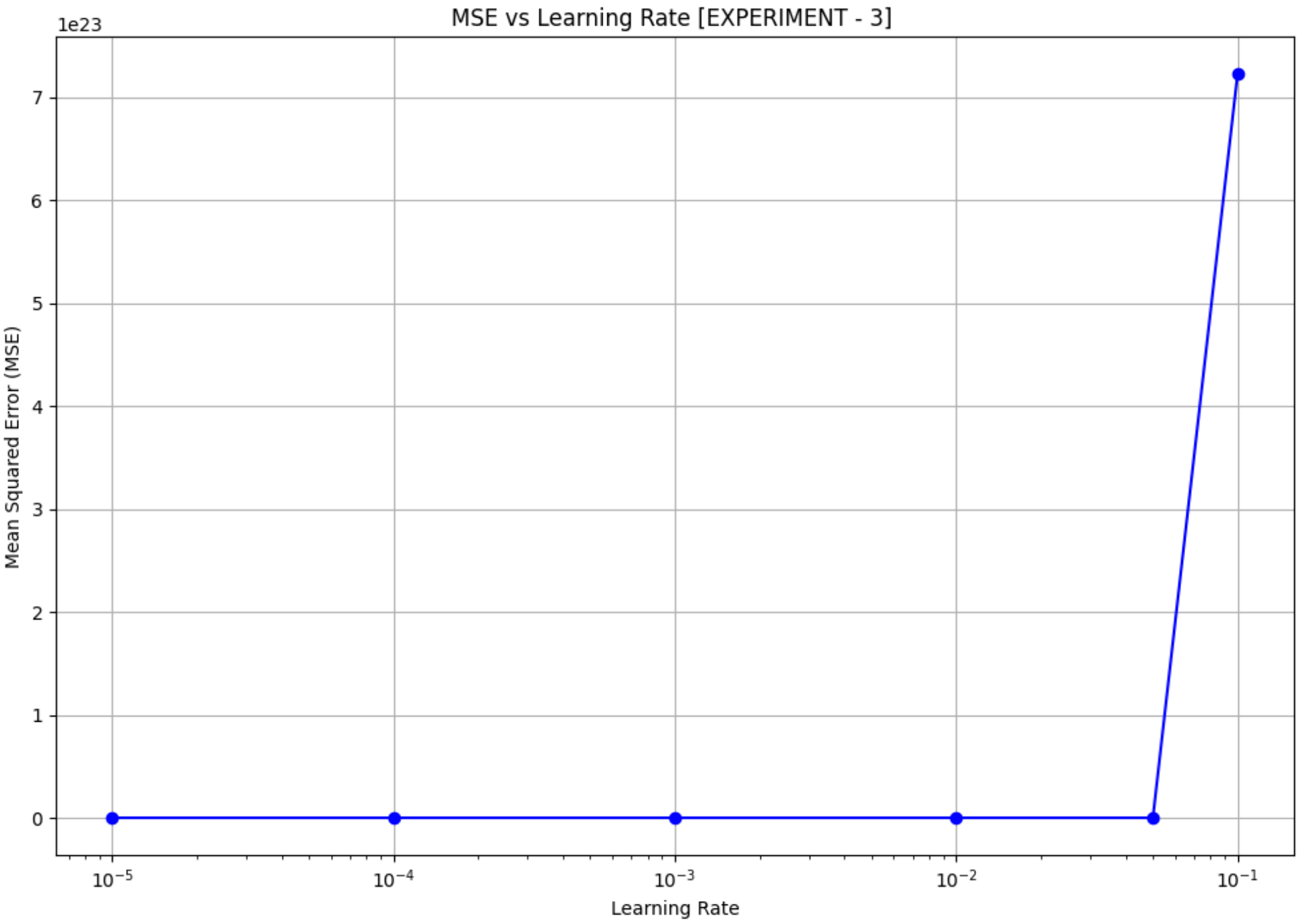
B. EXPERIMENT - 2 :

Performance (MSE) values for LIN_MODEL_CLOSED with and without feature scaling :

###	**Performance Values (MSE Values) [Mean Squared Error]**
LIN_MODEL_CLOSED without Scaling	289682250606827.25
LIN_MODEL_CLOSED with Scaling	35656142956.45556

C. EXPERIMENT - 3 :

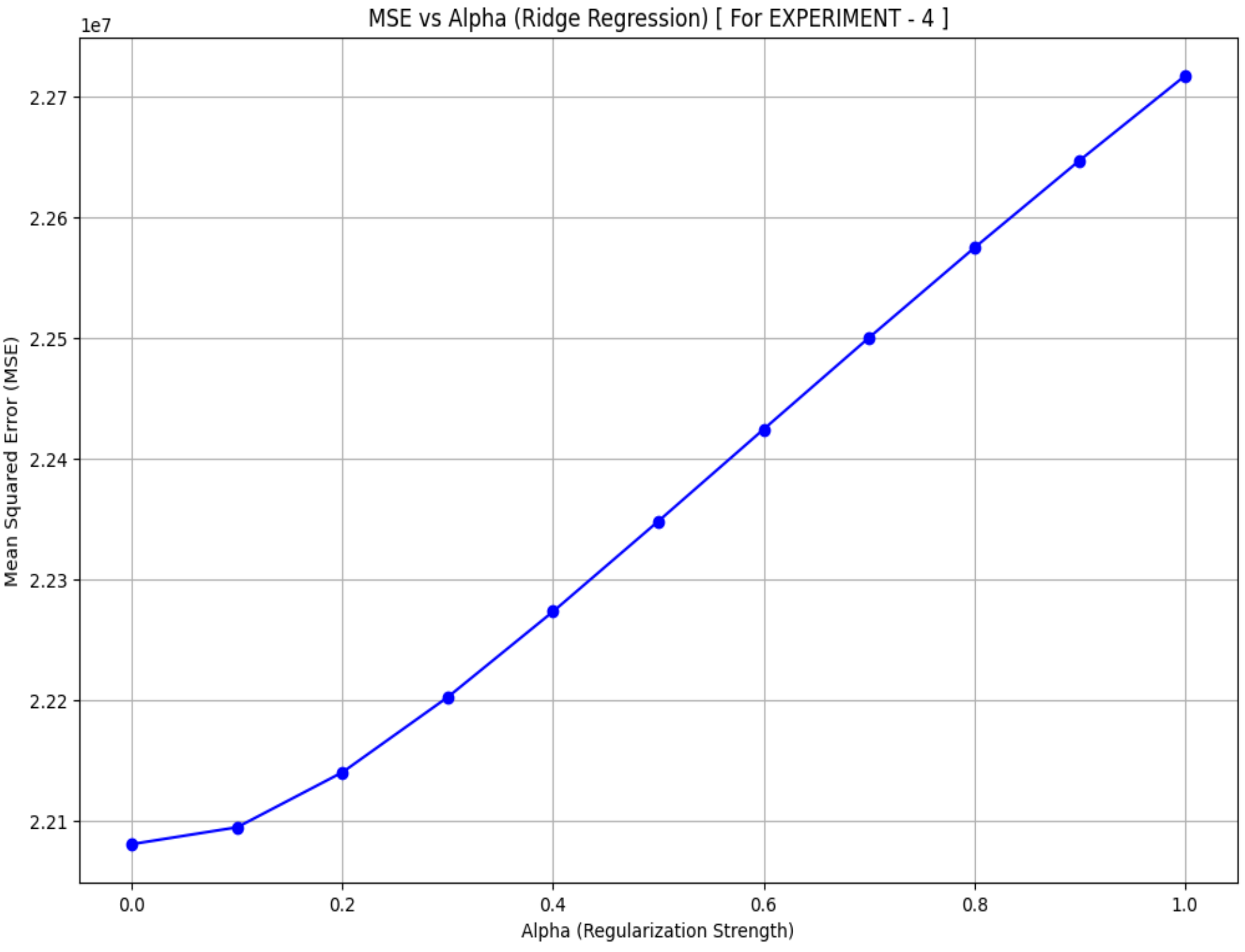
1. Mean Squared Error (MSE) Vs. Learning Rate :



2. The best choice for the learning rate is : **1e-05**

D. EXPERIMENT - 4 :

1. Mean Squared Error (MSE) Vs. Alpha (Ridge regression hyperparameter) :



2. The best choice for alpha value : **0.0**

E. EXPERIMENT - 5 :

###	**Performance Values (MSE Values) [Mean Squared Error]**
LIN_MODEL_CLOSED with Scaling	35656142956.45556
LIN_MODEL_GRAD	21868890.276889168
LIN_MODEL_RIDGE	21867454.64520962

OBSERVATIONS :

1. Closed-Form Linear Regression:

- The MSE [Mean Squared Error] for the closed-form solution is much higher than the other two methods. This suggests that simple linear regression may not be the best fit for the given data, or there may be some influential points that are disproportionately affecting the closed-form model.

2. Gradient Descent Linear Regression:

- The gradient descent approach yields a much lower MSE [Mean Squared Error] compared to the closed-form solution. This shows that the gradient descent is better in iteratively refining model parameters to minimize the cost function.
- Also here we have done feature scaling before we have implemented gradient descent approach, which could make convergence faster and produce better results, especially in the cases where features have very different scales.

3. Ridge Regression [Gradient Descent Linear Regression with regularization] :

- The Ridge regression (L2 regularization) has an MSE slightly lower than the gradient descent approach without regularization. This indicates that introducing a penalty for large coefficients (regularization) can enhance the performance.
- We can prevent overfitting by adding a penalty term in Ridge regression, which forces the magnitude of the coefficients to be smaller. Given that its MSE is slightly lower than the gradient descent approach without regularization, this suggests that overfitting might have been a minor issue, and Ridge regression effectively addressed this issue.
- Also, the Ridge regression model used feature-scaled data, which can also help in the efficiency and performance of the model.