Assignement-3: Linear Regression

Name- Aritra Ray

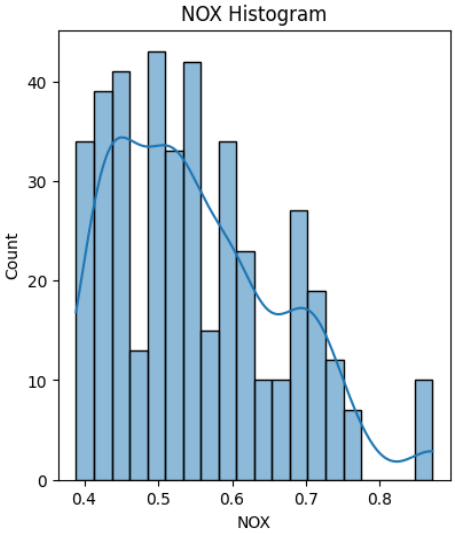
Roll No.- 21IM10008

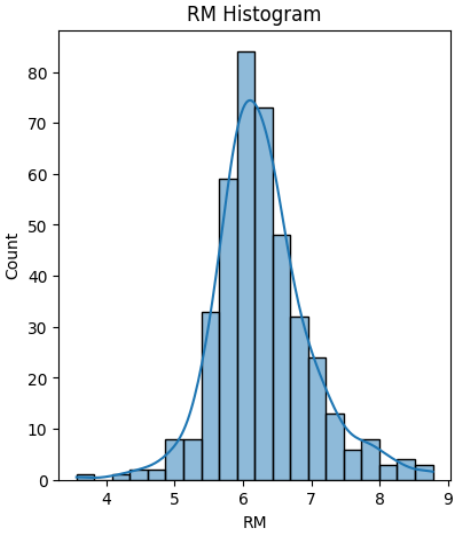
#Experiment 1

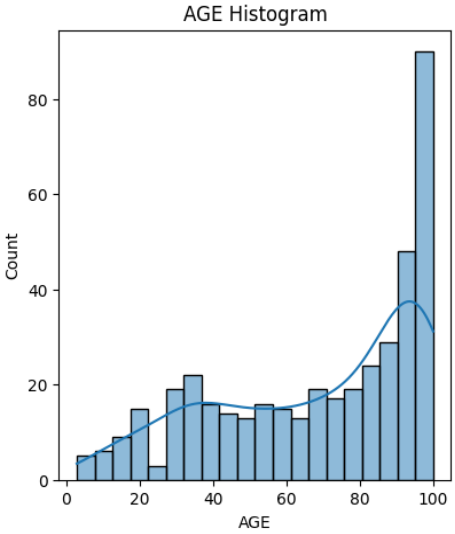
- First 10 rows of dataset_altered:

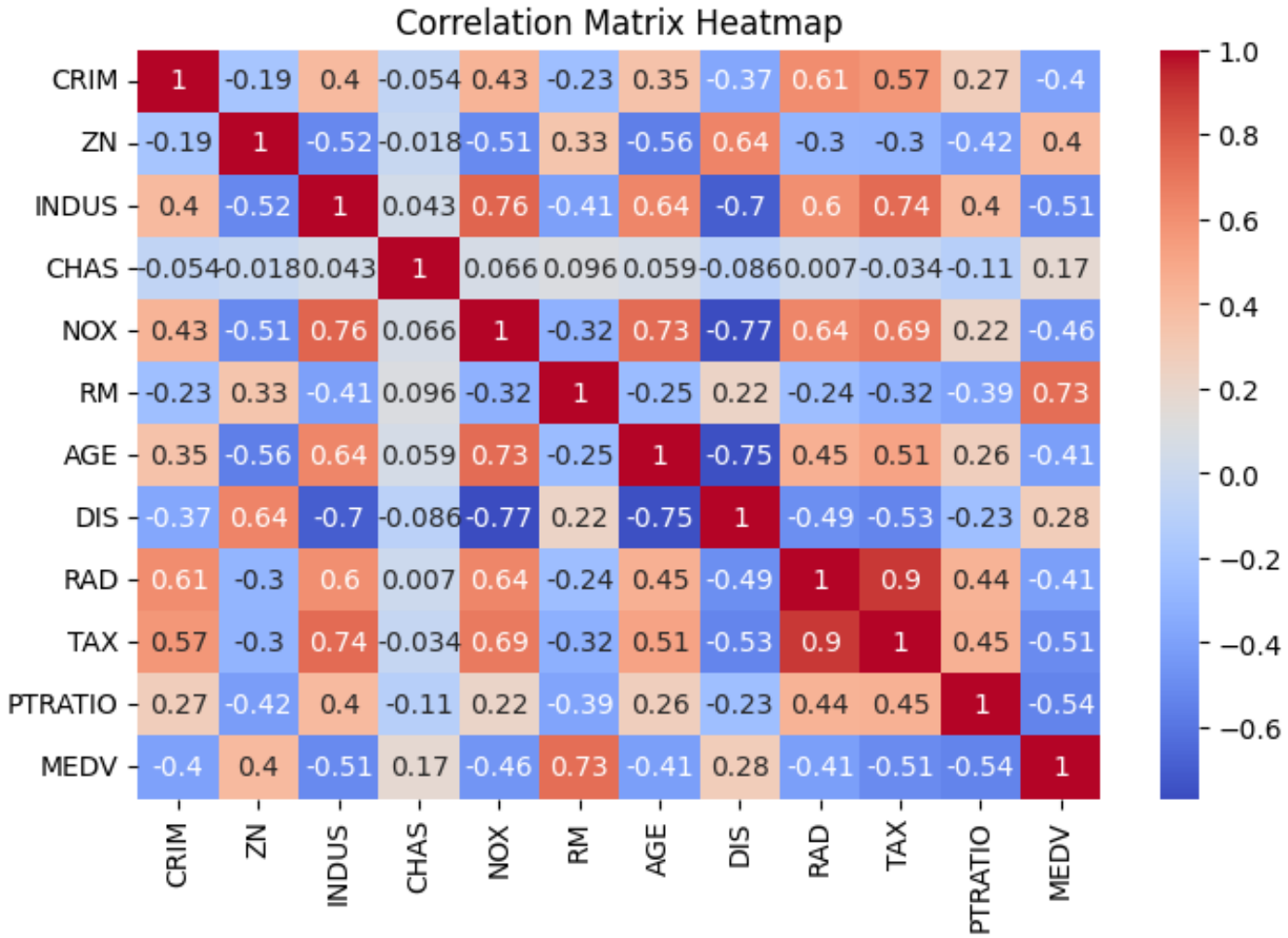
CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	24
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	36.2
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	28.7
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	27.1
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	16.5
0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	15
0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	18.9

#Experiment 2









- **Positive Correlations**: Features like average number of rooms (RM) positively correlate with home values (MEDV), suggesting that larger houses tend to have higher values.
- **Negative Correlations**: Factors like TAX negatively correlate with home values, indicating that areas with increasing TAX rates tend to have lower home values.
- **Multicollinearity**: High correlations between certain features suggest interdependence. For example, high property taxes (TAX) might coincide with better highway accessibility (RAD), leading to multicollinearity.
- **Feature Selection**: Features strongly correlated with MEDV, like RM, may be crucial predictors for estimating home values.
- **Correlation with MEDV**: Analyzing feature correlations with the target variable (MEDV) helps prioritize important predictors for home value estimation.
  Understanding these correlations guides feature selection and model building, enhancing predictive accuracy for real estate valuation.

#Experiment 3

Shapes of training and testing subsets:

dataset_altered_features_train: (370, 11)

dataset_altered_features_test: (42, 11)

dataset_altered_target_train: (370,)

dataset_altered_target_test: (42,)

NOTE-> For (value1, value2) value1 denotes number of rows and value 2 denotes number of columns. If value2 is empty it means number of columns=1.

#Experiment 4

Intercept: 23.654679658166707

Coefficients of CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO are respectively: [-1.50465492e-01, 4.21681638e-02, -9.28484407e-03, 3.05838078e+00, -2.03866768e+01, 6.41859192e+00, -4.21672641e-02, -1.43761981e+00, 2.90381743e-01, -1.78134026e-02, -9.46614092e-01]

RMSE for Closed Form Linear Regression: 3.0214987736656296

#Experiment 5

Optimal Learning Rate: 0.001

Intercept of Optimal Learning Rate 0.0

Coefficients of Optimal Learning Rate of CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO are respectively: [-0.46632116, 0.0284623, 0.67973974, 0.94777539, -0.80404544, 4.12528662, -0.23621331, -3.91685118, 2.14380601, -0.4706321, -1.32172245]

RMSE with Optimal Learning Rate: 284.6034629170736

Additional Information: I experimented with increasing the number of iterations but observed that with the learning rates (a) 0.001, (b) 0.01, and (c) 0.1, the coefficients were diverging. Therefore, to increase the accuracy of the gradient descent algorithm, it's necessary to decrease the learning rate.

Conclusion -

This report summarizes the experiments conducted on the Boston Housing Dataset, including tabular data, correlation analysis, model coefficients, performance metrics, and additional insights on the gradient descent algorithm's optimization.