CS 4375
ASSIGNMENT 1

Names of students in your group: Natalie Pedigo (nbp220000),
Amir Sabry (ahs210005)

Number of free late days used: 0

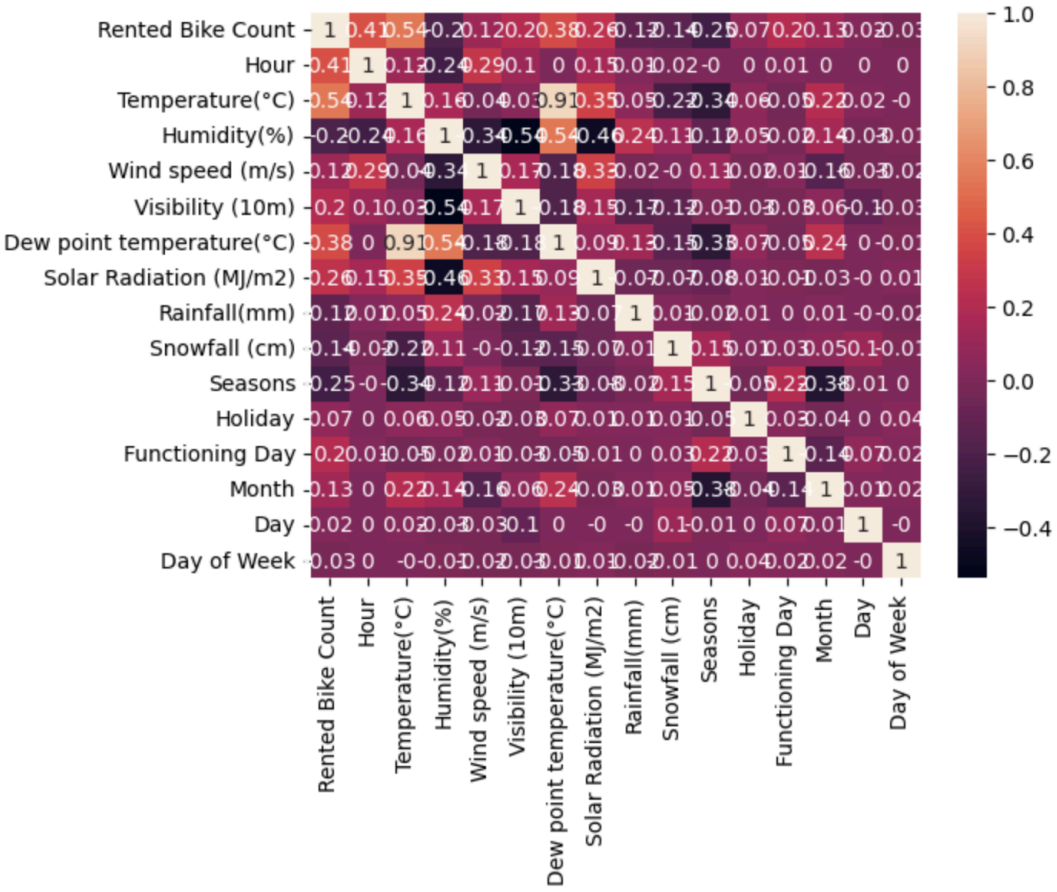
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Overview

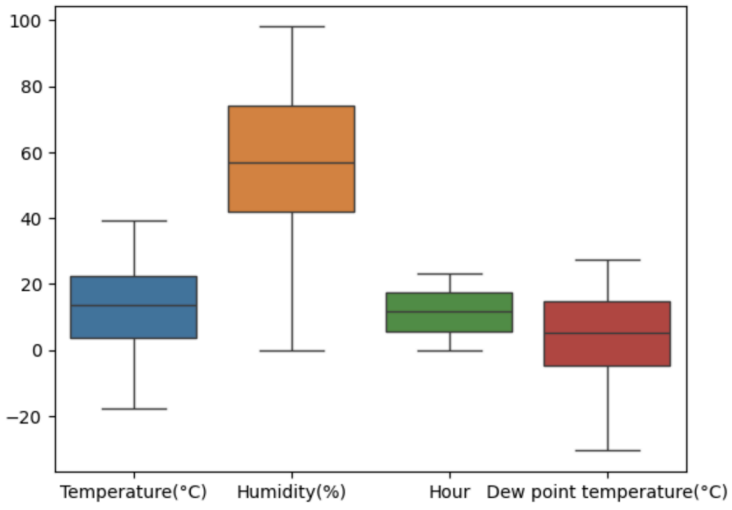
Our model examines weather and other related data to predict the number of bikes rented in an hour in Seoul, South Korea. Our model implements the SGDRegressor method.

Plots

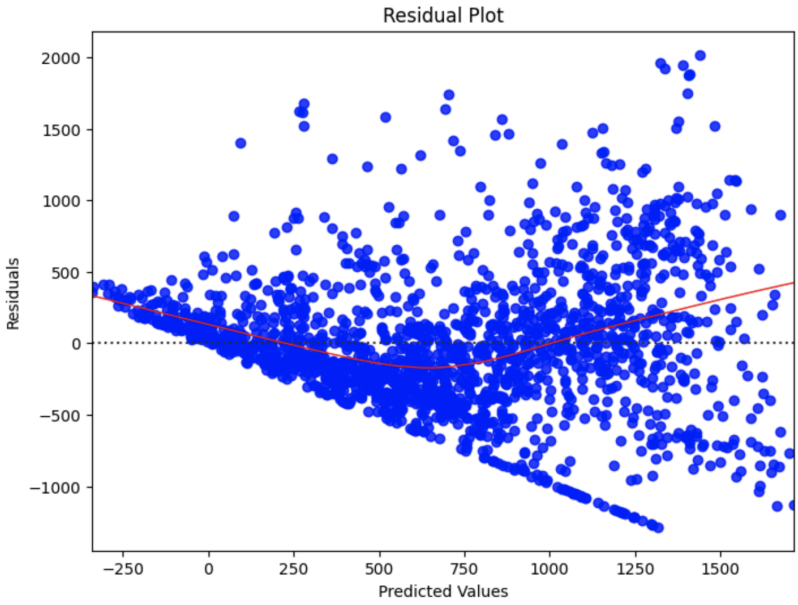
1. Correlation Matrix



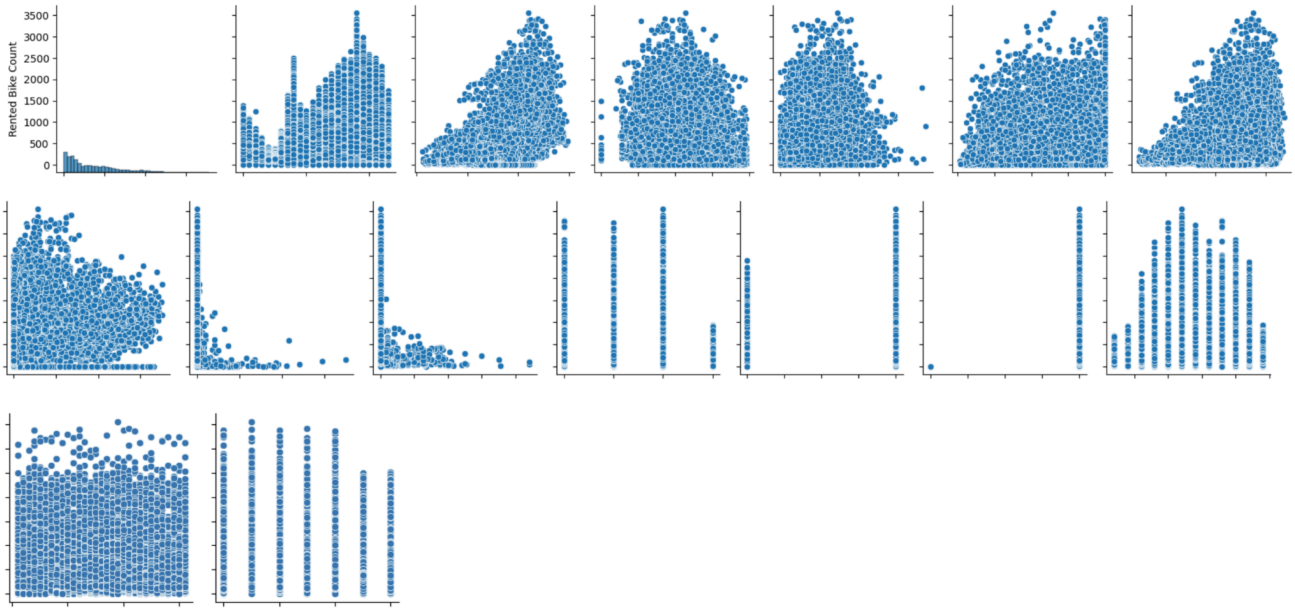
2. Box and Whisker Plot



3. Residual Plot



4. Pairplots (Rented Bike Count vs Hour, Temperature, etc…)



Hyper-parameter Experiment Log

trial	max_iter	tolerance	early stopping	learning rate	alpha	MSE	R-sqaured
1	1000	0.001	TRUE	invscaling	0.0001	231389.3767	0.443596815
2	1000	0.001	FALSE	invscaling	0.0001	231242.0237	0.443951142
3	100000	0.001	TRUE	invscaling	0.001	231372.0366	0.443638512
4	100000	0.001	TRUE	constant	0.001	242072.7268	0.417907434
5	10000	0.0001	TRUE	invscaling	0.01	231212.2285	0.444022789
6	100000	0.001	FALSE	invscaling	0.001	231213.6729	0.444019316

Model Results
MSE Train: 226813.50325400545
MAE Train: 351.4206786135883
EV Train: 0.45477124678870373
R2 Train: 0.45477072582253586
Coefficients: [214.7319617 -204.61904835 202.55152297 160.83217773]

MSE Test: 231212.22847996265
MAE Test: 358.05682165122766
EV Test: 0.4440481614887323
R2 Test: 0.4440227892340881
Coefficients: [214.7319617 -204.61904835 202.55152297 160.83217773]

Result Analysis

We are not satisfied that we found the best solution. The relatively low R-squared value indicates that the model failed to capture the complexity of the data. Through 6 trials of hyperparameter tuning, the highest R-squared value achieved is 0.444 (Trial 5). This means that the model is only explaining 44% of the variance in the target variable. The residual plot shows a curved pattern indicating the relationship between the predictors and the target variable in non-linear. Given that the relationship is non-linear, it is clear that to derive an optimal solution a linear model is not sufficient.