CAPSTONE PROJECT - 2

Seoul Bike Sharing Demand Prediction

Created By : Sakshi Dhyani

Project Details

In many urban cities, rental bikes are introduced to ensure mobility comfort. It is required to make the rental bike available for public at the correct time. In this project, major concern is to predict the rented bikes required at each hour for the stable supply of bikes.

Steps performed

- Data cleaning
- Data visualizations
- Data preprocessing
- Model Implementation
- Evaluation metrics

Data Summary

Independent Features

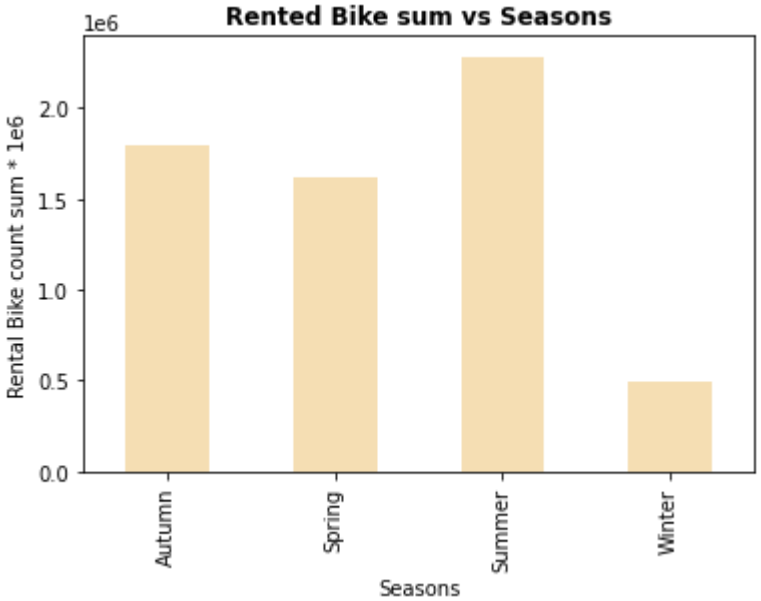
Date: year-month-day
Hour: Hour of the day
Temperature: (in Celsius)
Humidity: (in %)
Windspeed: m/s
Visibility: 10m
Dew Point Temperature (in celsius)
Solar Radiation: MJ/m2
Rainfall: mm
Snowfall: cm
Seasons : Winter, Spring, Summer, Autumn
Holiday - No Holiday/ Holiday
Functional Day - Yes/No

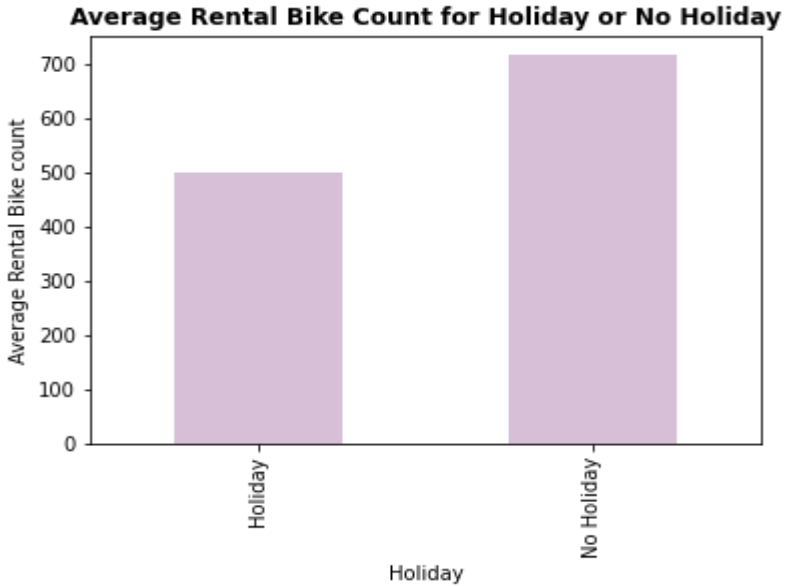
Dependent Feature

Rented Bike Count - Count of bikes rented at each hour

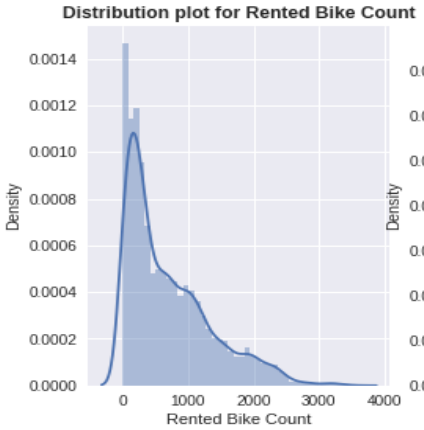
Exploratory Data Analysis

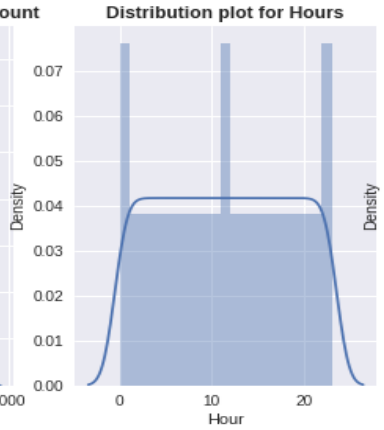
Bar Plots for categorical values

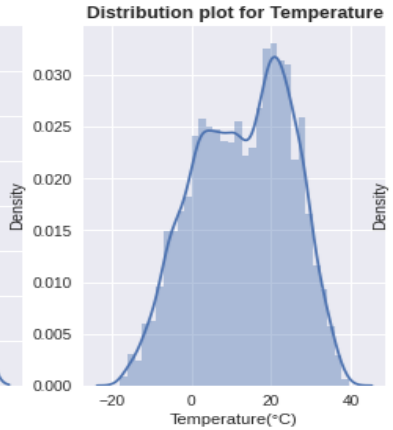


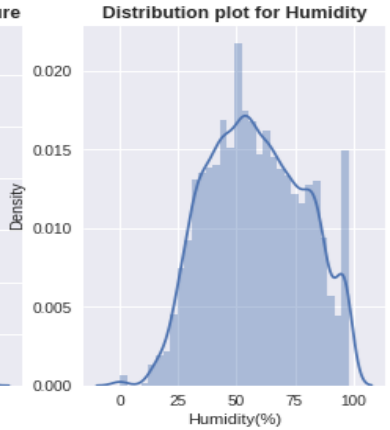


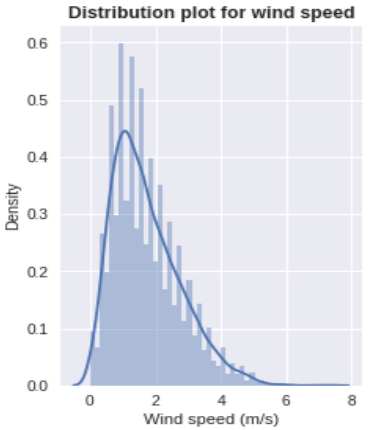
Distribution plots for numerical features (Independent and Dependent features)

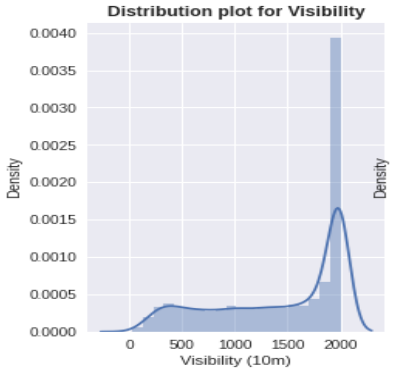


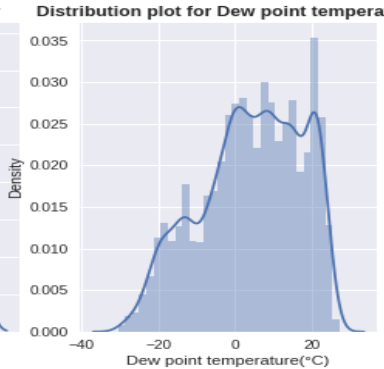


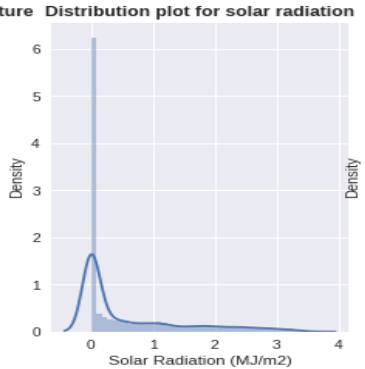


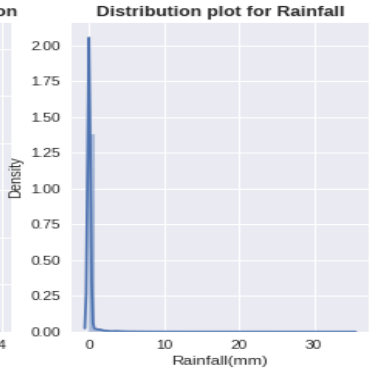


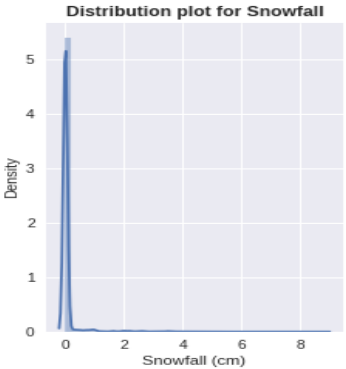




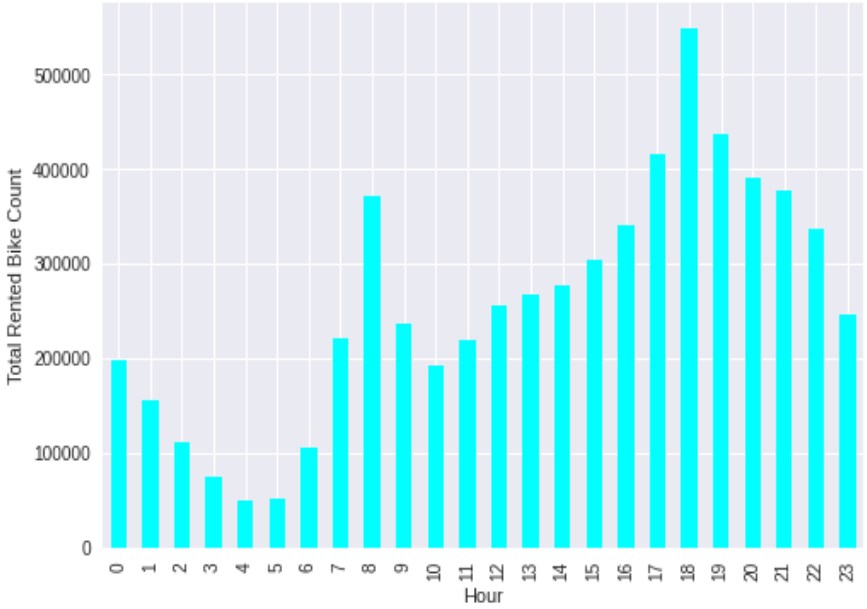


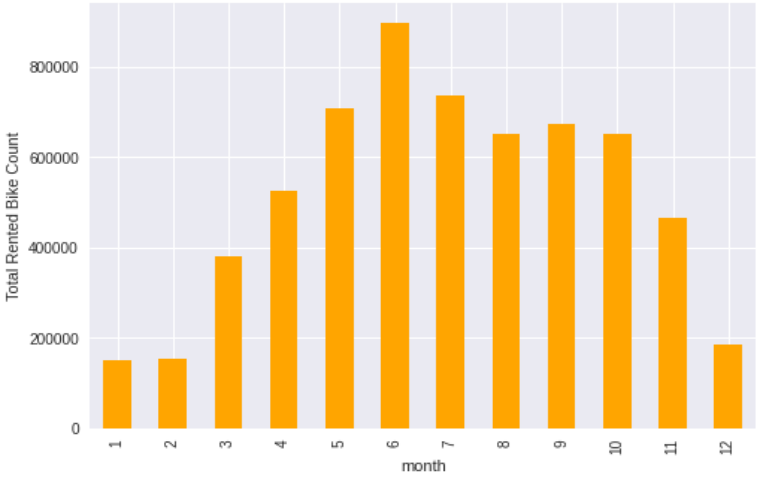


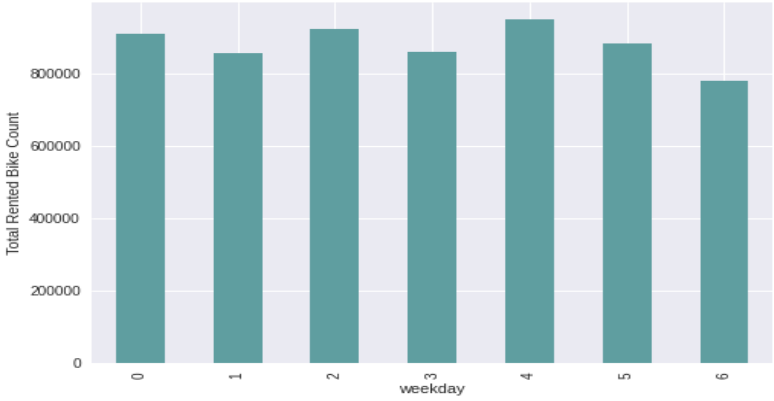




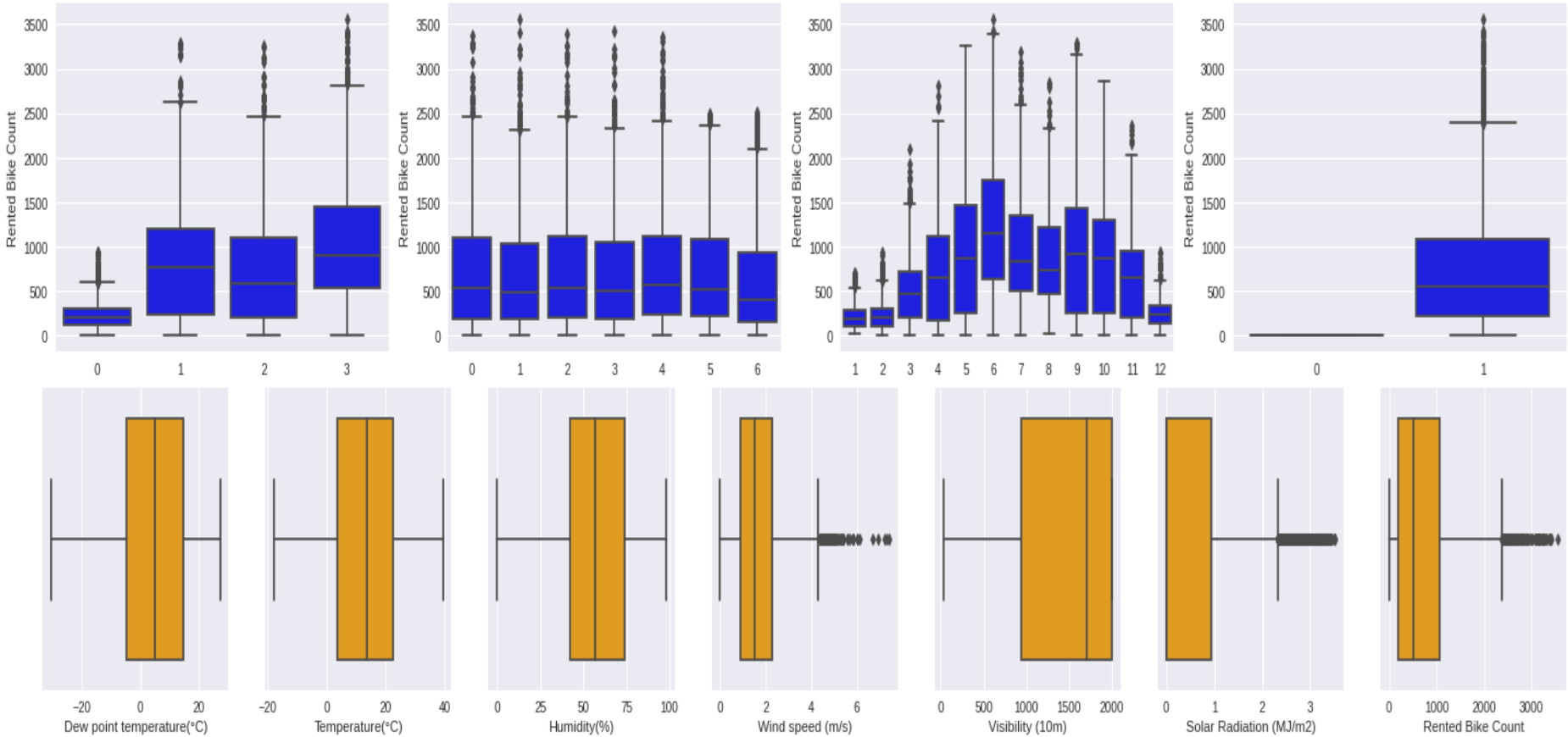
Monthly, hourly, week-wise Rental Bike Count



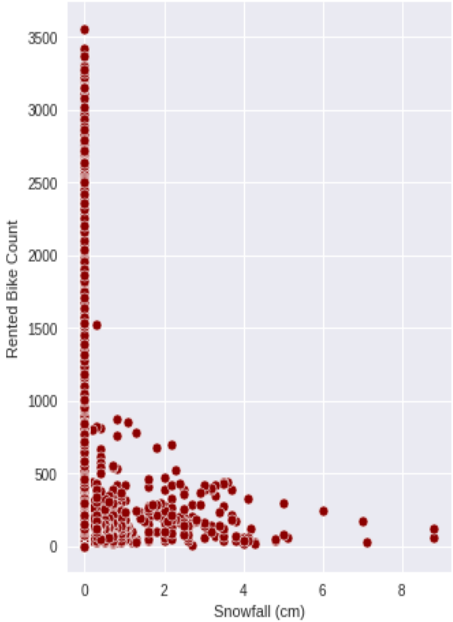


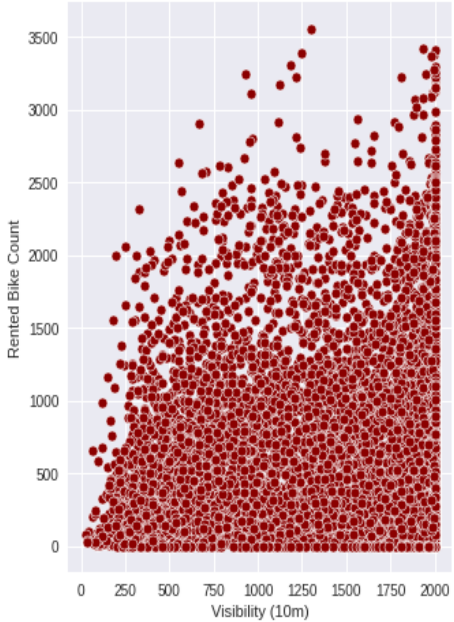


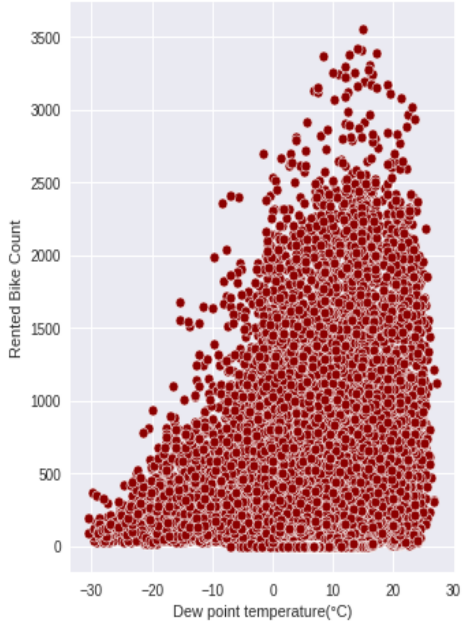
Outlier Detection using Box Plot

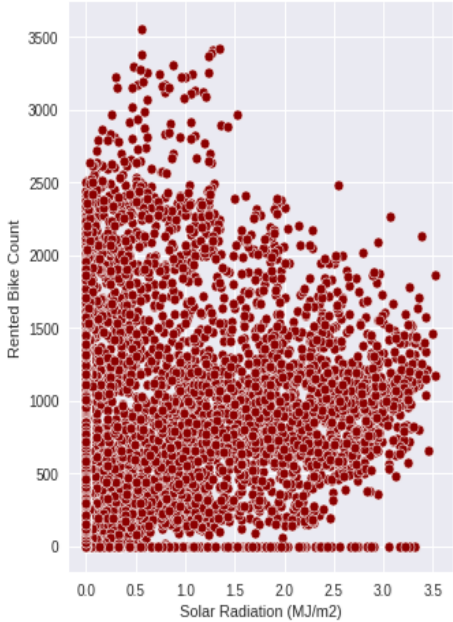


Scatter Plots for features

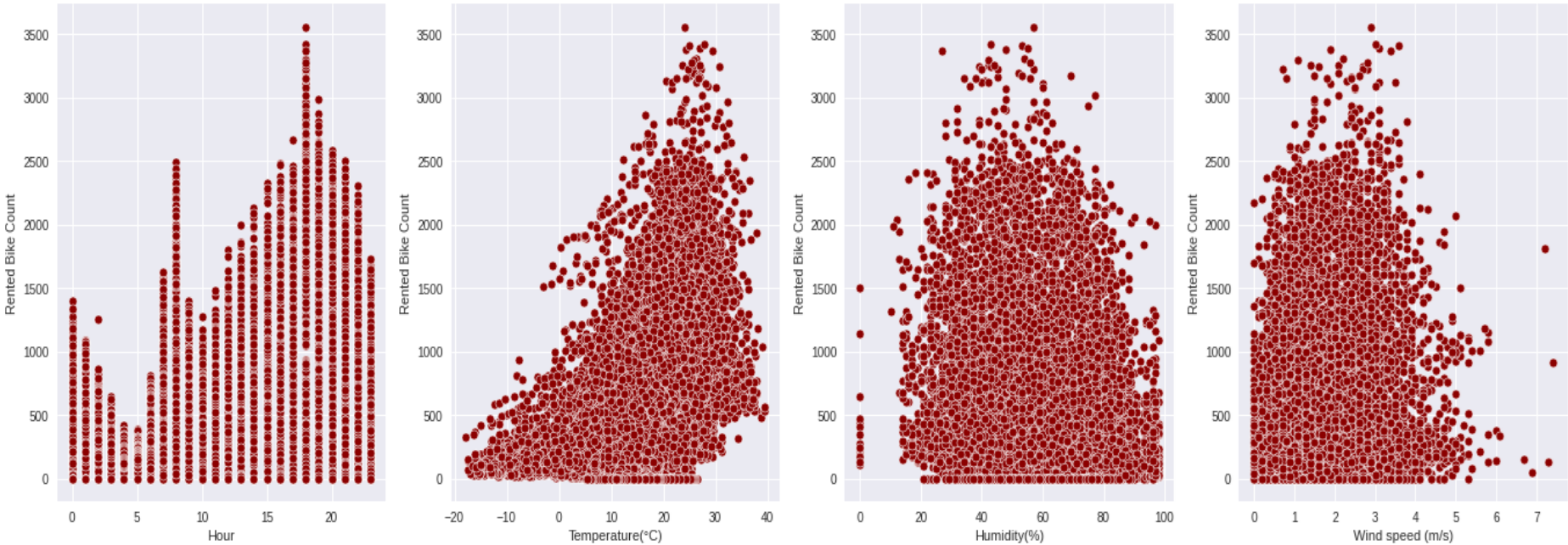




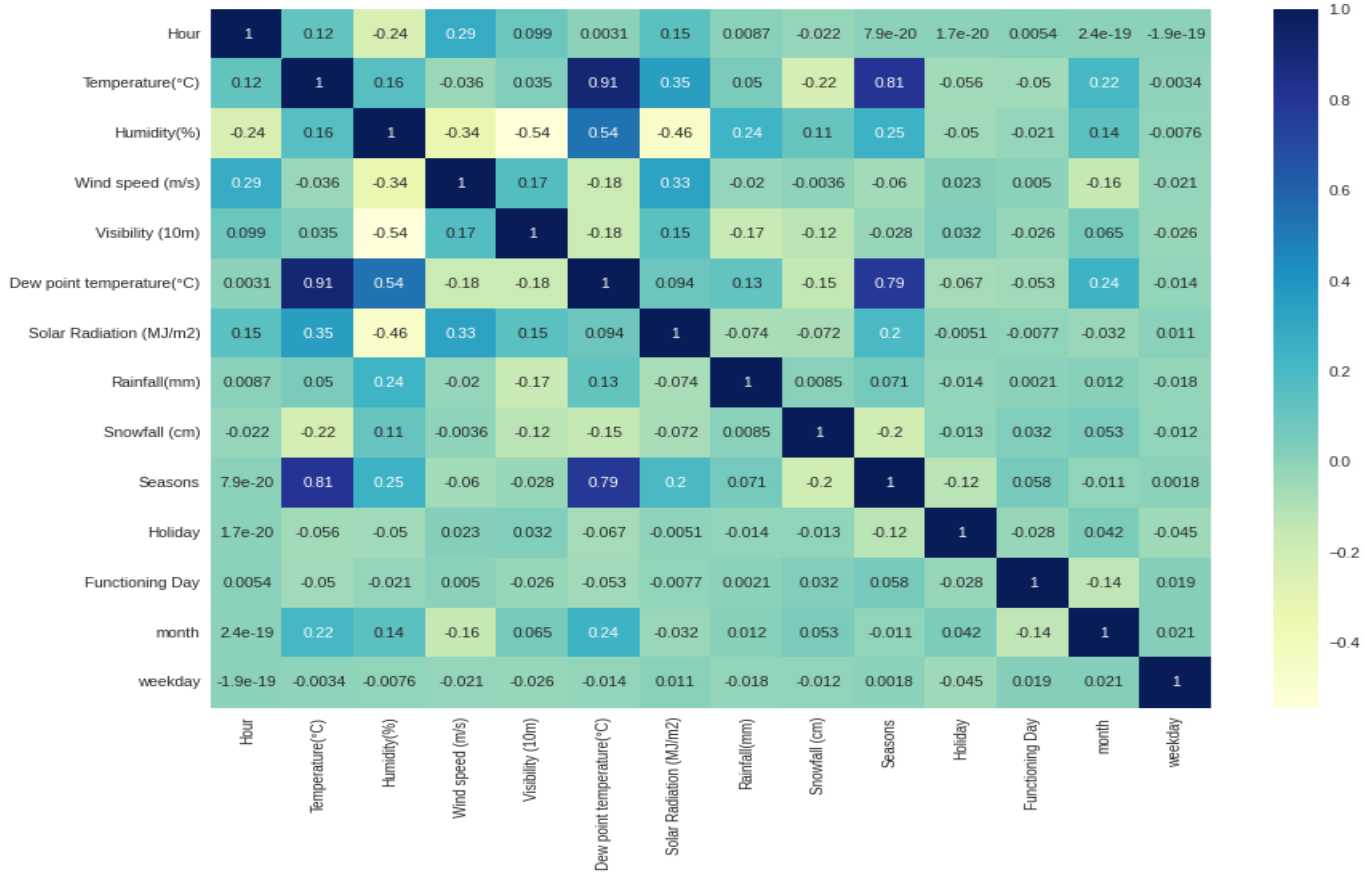




Scatter Plots for features



Heatmap



Heatmap helps to find the correlation between the features. While implementing Linear Regression, features having high collinearity will be removed.

Data preparation

Initial Data Shape - 8760 Rows and 14 Columns (1 column is dependent)

After Train-Test Split :

Training Data - 7008 Rows and 14 Columns (1 column dependent)

Test Data - 1752 Rows and 14 Columns (1 column dependent)

Dependent column will be predicted as that is the target variable named “Rented Bike Count”.

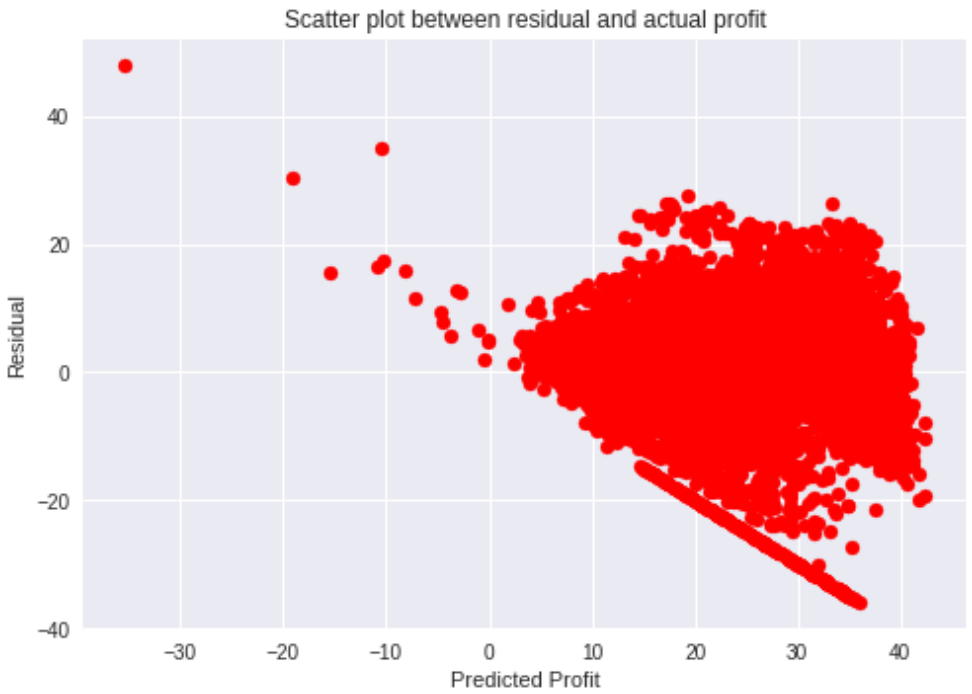
Linear Regression Model

Evaluation Metrics for training data

```
MSE is -> 82.54262440299173
RMSE is -> 9.085297155459019
R2 square -> 0.4651658504428886
MAE -> 6.714158886558394
Adjusted R2 score-> 0.464401474067932
```

Evaluation Metrics for test data

```
MSE -> 89.75365590130693
RMSE -> 9.473840609874484
R2 square -> 0.43008578964332533
MAE -> 6.967849529253253
Adjusted R2 score-> 0.42681230193306297
```



Polynomial Regression Model

Polynomial Regression is a special case of linear regression where the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial.

Polynomial regression of degree 2 was implemented for this case.

```
Evaluation Metric for training data

MSE is 69.15939833595111
RMSE is 8.316212980434731
R2 square is 0.5518823364241219
MAE is 5.9847987628546795
Adjusted R2 score is 0.5512418938579137
```

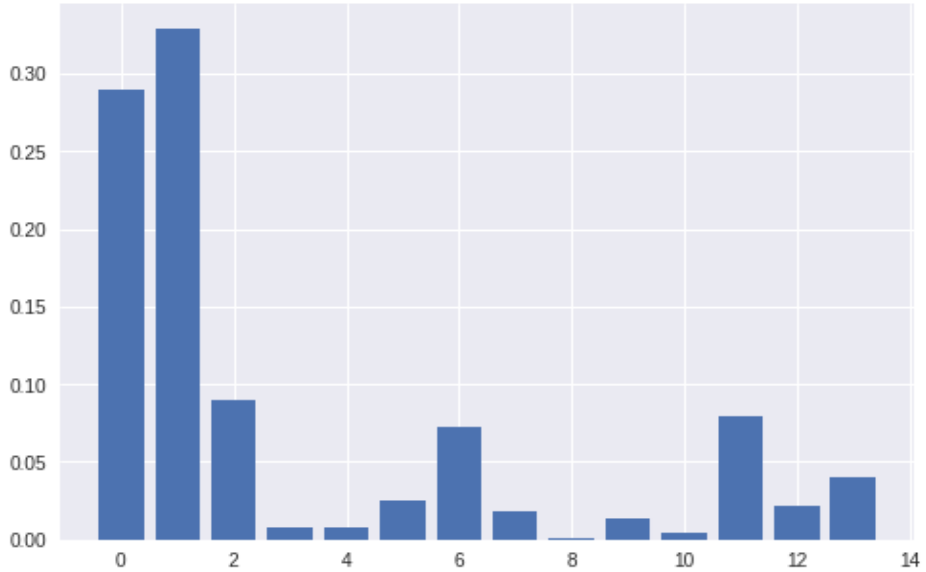
```
Evaluation Metric for test data

MSE is 78.4882252924261
RMSE is 8.85935806322479
R2 square is 0.501618574857646
MAE is 6.281442419541594
Adjusted R2 score is
0.49875595897515124
```

Decision Tree Model

Feature Importances

```
Feature: 0, Score: 0.28233
Feature: 1, Score: 0.34606
Feature: 2, Score: 0.09697
Feature: 3, Score: 0.00321
Feature: 4, Score: 0.00415
Feature: 5, Score: 0.02088
Feature: 6, Score: 0.07751
Feature: 7, Score: 0.01524
Feature: 8, Score: 0.00079
Feature: 9, Score: 0.01571
Feature: 10, Score: 0.00244
Feature: 11, Score: 0.08176
Feature: 12, Score: 0.01909
Feature: 13, Score: 0.03387
```



Feature 0 :'Hour', Feature 1: 'Temperature(°C)',Feature 2: 'Humidity(%)',Feature 3:'Wind speed (m/s)',Feature 4: 'Visibility (10m)', Feature 5: 'Dew point temperature(°C)',
Feature 6:'Solar Radiation (MJ/m2)',Feature 7:'Rainfall(mm)',Feature 8: 'Snowfall (cm)',
Feature 9:'Seasons',Feature 10:'Holiday',Feature 11: 'Functioning Day', Feature 12:'month',
Feature 13:'weekday'

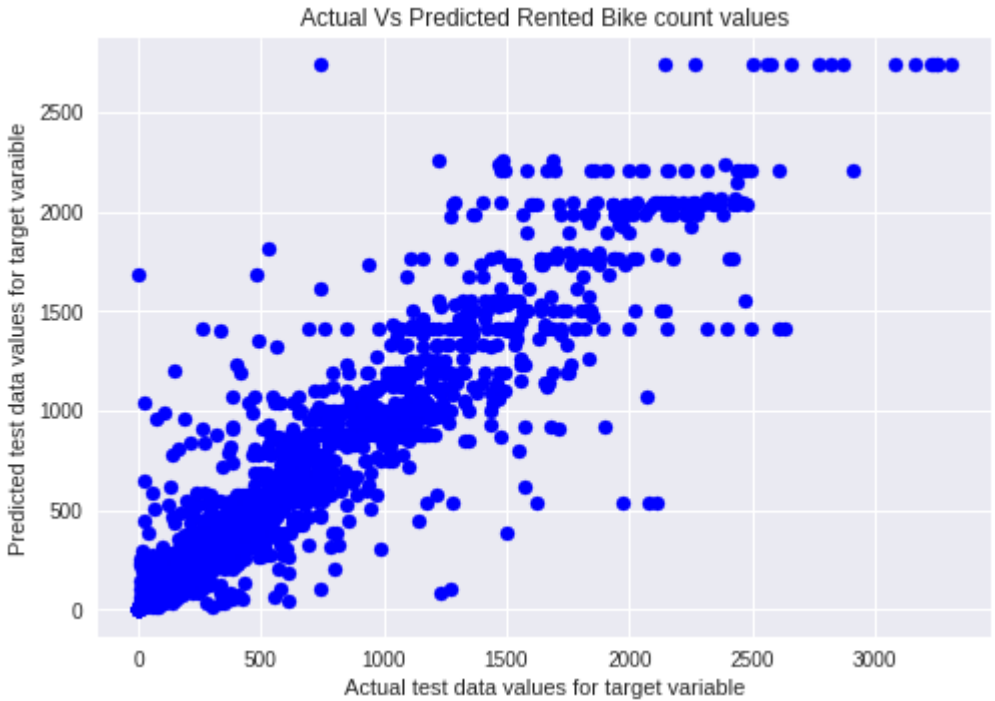
Decision Tree Model

Evaluation metric for training data

MSE is 5667.104023972603
RMSE is 75.28017019091152
R2 square is 0.9863523196242828
MAE is 38.87678367579909
Adjusted R2 score is 0.986324996940848

Evaluation metric for testing data

MSE is 52755.21314133054
RMSE is 229.68503029438062
R2 square is 0.8739489466120322
MAE is 131.53125
Adjusted R2 score is 0.8729329910867405



KNN Regressor Model

```
Evaluation metric for training data:

MSE is 0.0
RMSE is 0.0
R2 square is 1.0
MAE is 0.0
Adjusted R2 score is 1.0
```

```
Evaluation metric for test data:

MSE is 73199.23857056774
RMSE is 270.55357800363265
R2 square is 0.8251008653059814
MAE is 163.76097089242106
Adjusted R2 score 0.8236912004322242
```

Using grid search CV, k parameter value chosen : 5

Random Forest Regressor

Evaluation metric for training data:

MSE is 22022.431171319764
RMSE is 148.39956594046953
R2 square is 0.9469649576836751
MAE is 96.05389103308234
Adjusted R2 score is 0.9468587814227816

Evaluation metric for test data:

MSE is 48032.2201217954
RMSE is 219.16254269786933
R2 square is 0.8852338644392338
MAE is 142.79673689986572
Adjusted R2 score is 0.8843088639223364

XGBoost Regressor

Best parameters selected using grid search CV:

```
{'colsample_bytree': 0.7, 'learning_rate': 0.07,
'max_depth': 5, 'n_estimators': 500, 'nthread': 4,
'objective': 'reg:linear', 'subsample': 0.9}
```

```
Evaluation metric for training data:

MSE is 8028.853311451494
RMSE is 89.60386884198412
R2 square is 0.9806646881167718
MAE is 59.030774724563436
Adjusted R2 score is 0.9806259787836722
```

```
Evaluation metric for test data:

MSE is 25350.051165979992
RMSE is 159.21699396100905
R2 square is 0.9394296703085964
MAE is 97.85649094004269
Adjusted R2 score is 0.938941481122828
```

Conclusion

All metrics were evaluated for each model, MSE(Mean Squared Error), MAE (Mean Absolute Error), RMSE(Root Mean squared Error),R2 Score, Adjusted R2 Score

At the end, comparison of models stated that some models showed improvement or were able to handle the overfitting issues when hyperparameter tuning was performed.

Adjusted R2 score was used to compare models as it is a special form of R2 score. Adjusted R2 indicates how well terms fit a curve or line, and also adjusts for the number of terms in a model.

If only adjusted R2 score is considered, then XG Boost regressor performed much better in test data as compared to other models

THANK YOU