**Machine Learning**

Week Five Solution



**Week Five**

**Task:** Aura customer FloridaBikeRentals.com is unable to predict peaks and troughs in demand for their high-end bikes. They have approached Aura to customize a marketing tool to predict bike-sharing demand. To stabilize the demand, devise marketing strategies using the bike-sharing dataset.

Based on rented bike count, the hour of the day, the day's temperature, humidity, wind speed, rainfall, holidays, and many other factors, build a model to predict the bike count required at each hour for the stable supply of rental bikes.

1. Load the dataset
2. Check for null values in any columns and handle the missing values
3. Convert Date columns to Date format and extract day, month, day of week and weekdays/weekend from date column
4. Check correlation of features using a heatmap
5. Plot the distribution plot of Rented Bike Count
6. Plot the histogram of all numerical features
7. Plot the box plot of Rented Bike Count against all the categorical features (Hint: Categorical features on X-axis and Rented Bike Count on Y-axis)
8. Plot the Seaborn catplot of Rented Bike Count against features like Hour, Holiday, Rainfall(mm), Snowfall (cm), weekdays\_weekend and give your inferences
9. Encode the categorical features into numerical features.

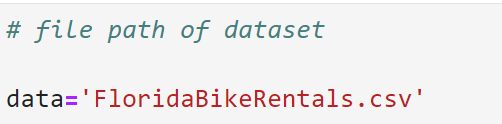
(Hint: use get\_dummies())

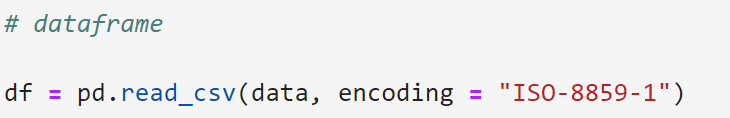
1. Identify the target variable and split the dataset into train and test with a ratio of 75:25 and random state 1
2. Perform Standard Scaling of the train dataset.
3. Perform Linear Regression, Lasso Regression and Ridge Regression for predicting the bike count required at each hour and compare the results.

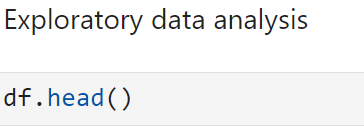
**Steps**:

1. Load the dataset

**Output:**







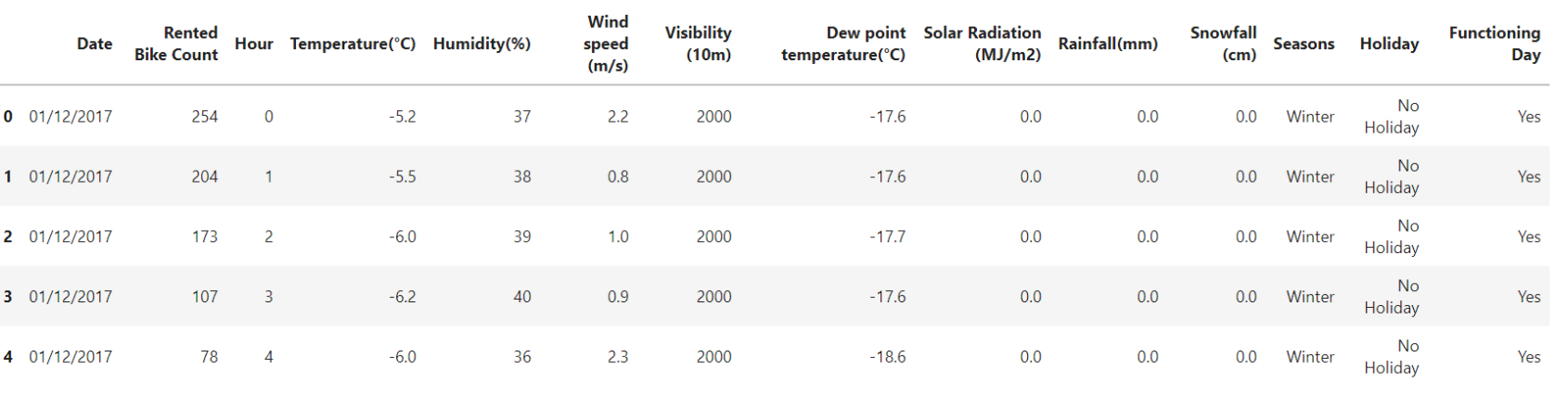


Table Part 1:

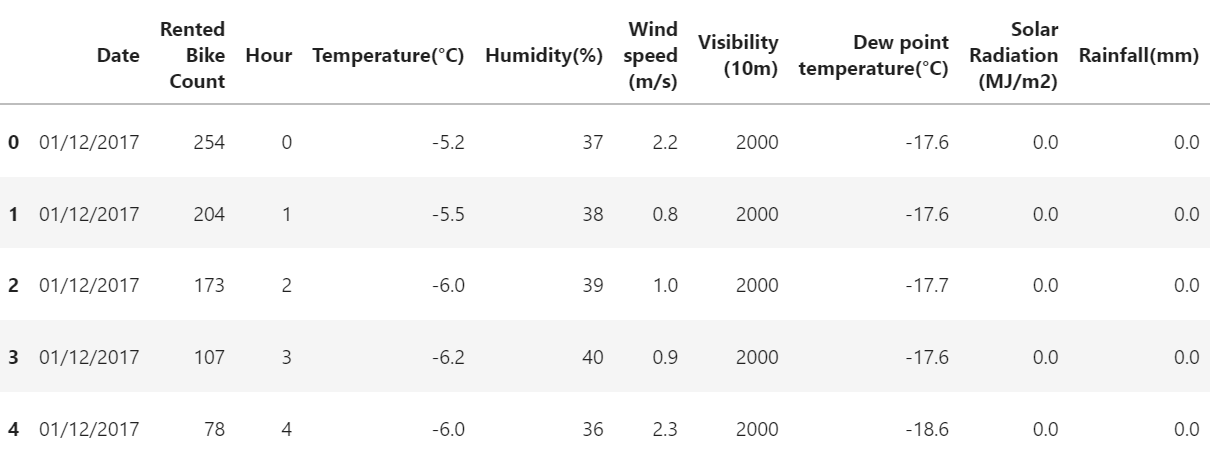
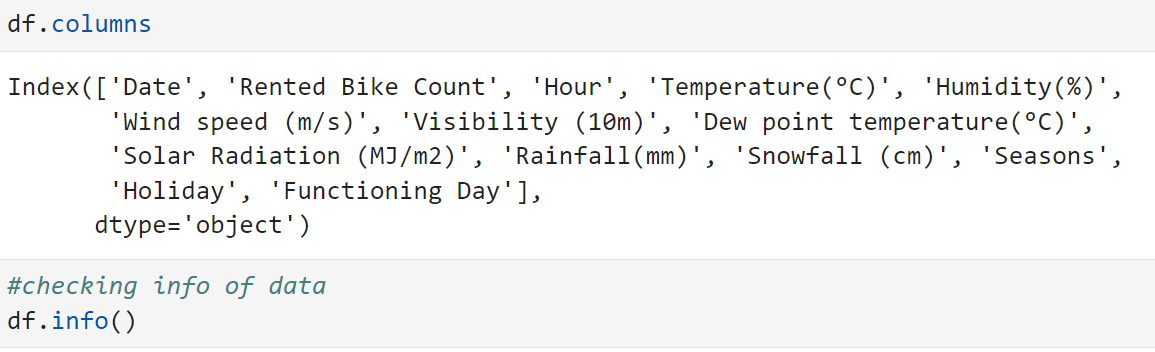
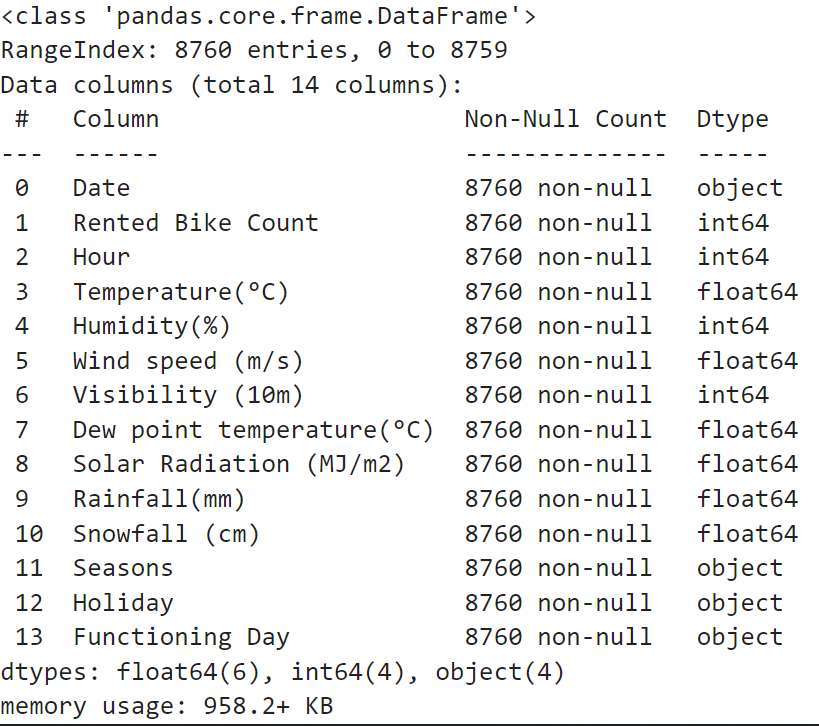


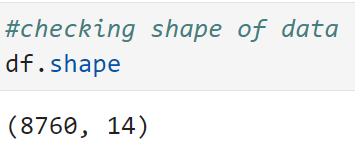
Table Part 2:





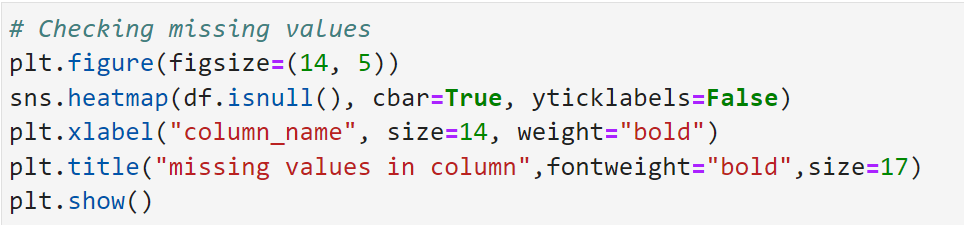


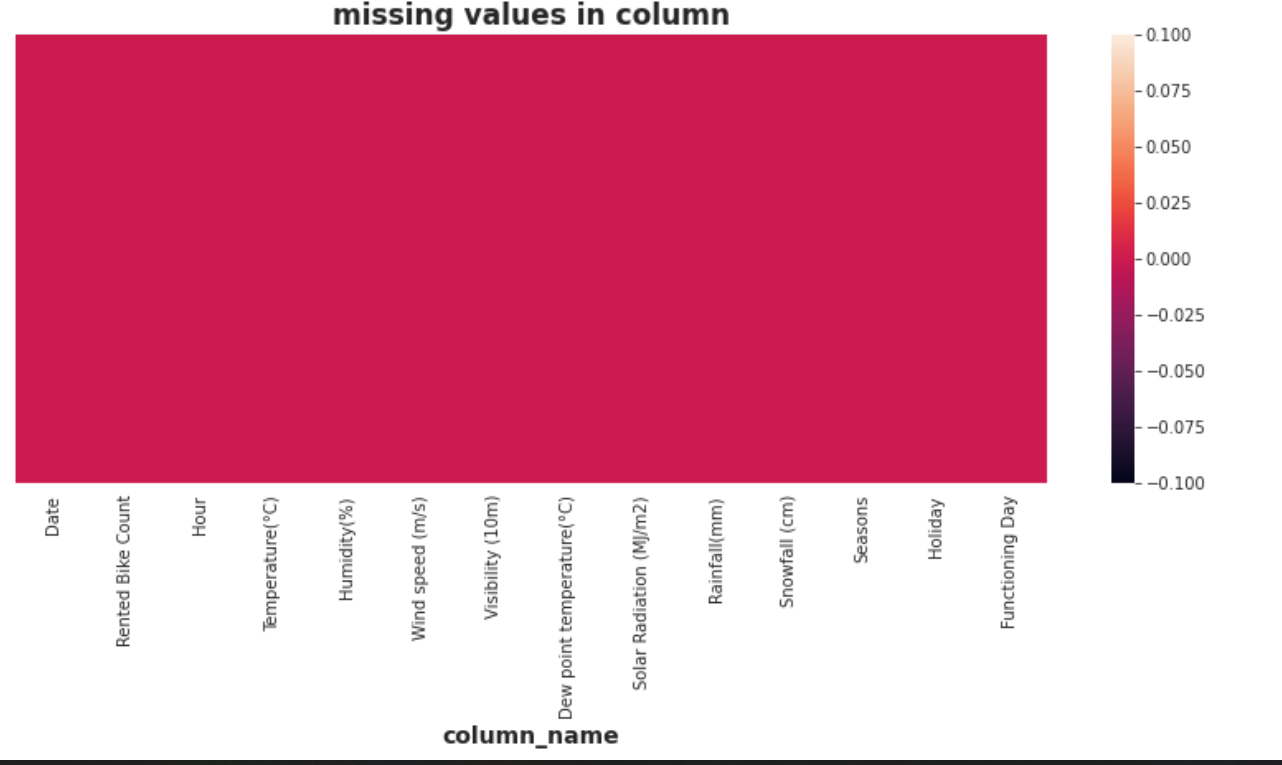




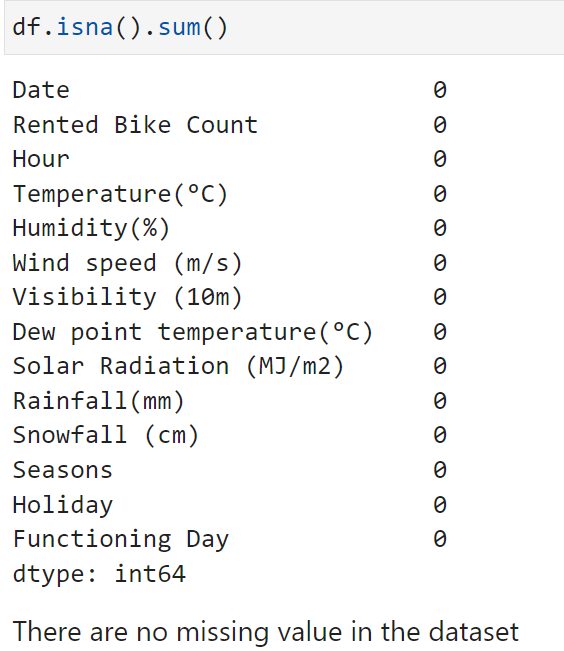
1. Check for null values in any columns and handle the missing values

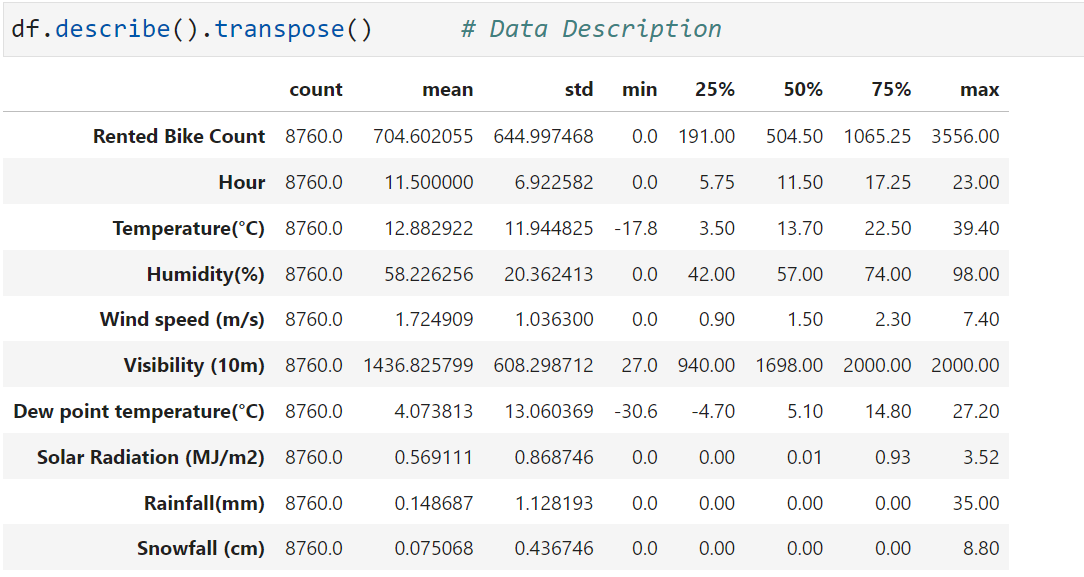
**Output:**

****



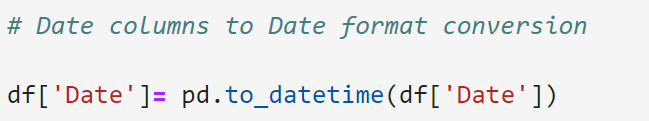
***Observation:***

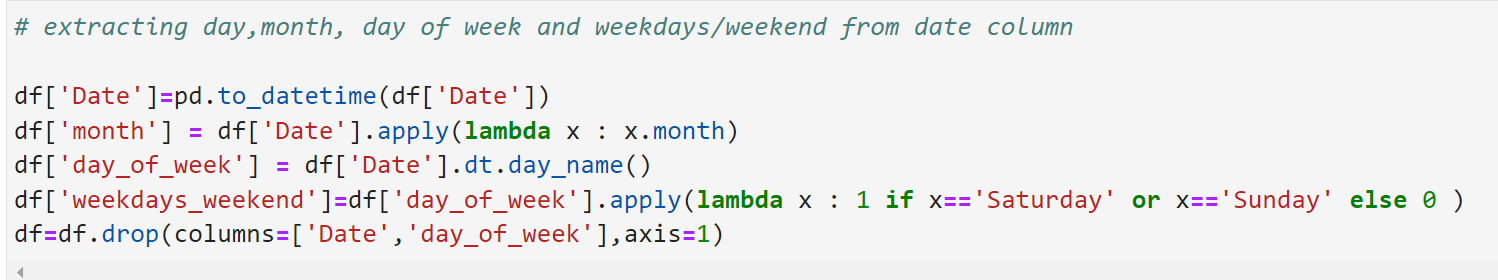
******



1. Convert Date columns to Date format and extract day, month, day of week and weekdays/weekend from date column

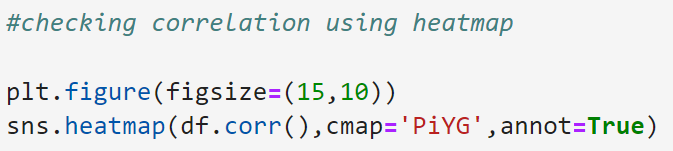
**Output:**



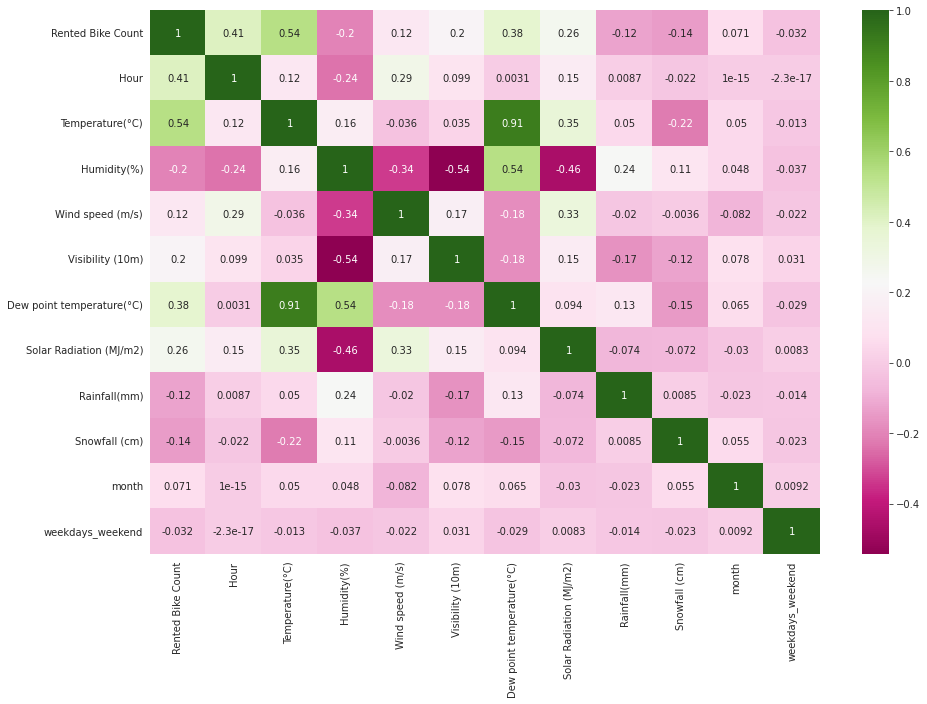


1. Check correlation of features using a heatmap

**Output:**





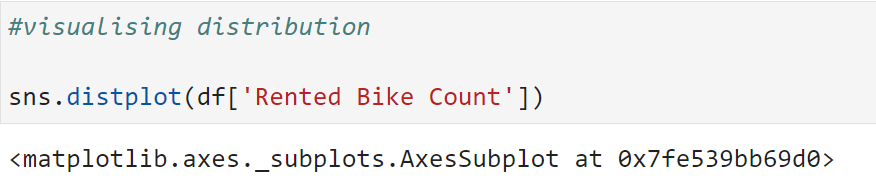


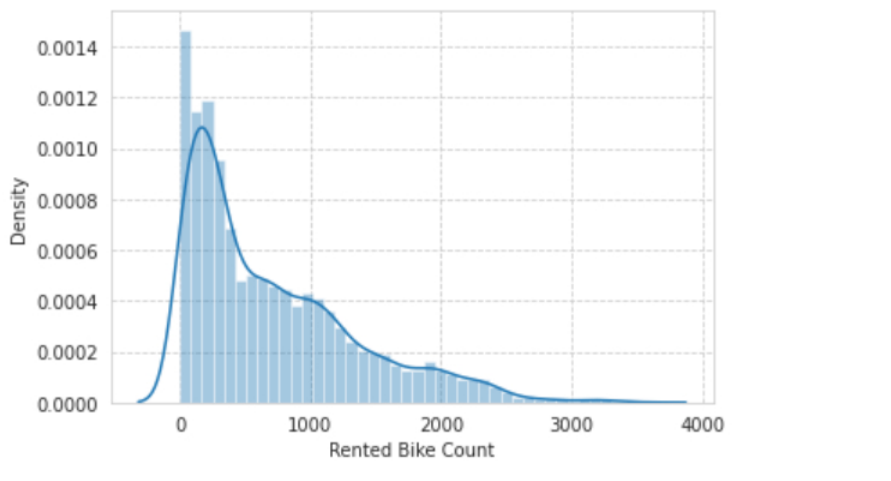
***Observation:***

The Correlation heatmap shows the range of correlation between each variables. The more darker the colour the more they are correlated.

1. Plot the distribution plot of Rented Bike Count

**Output:**

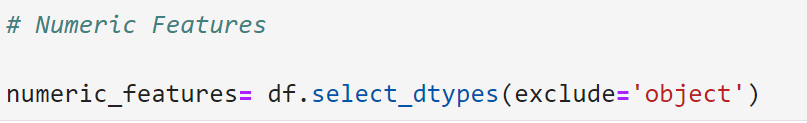




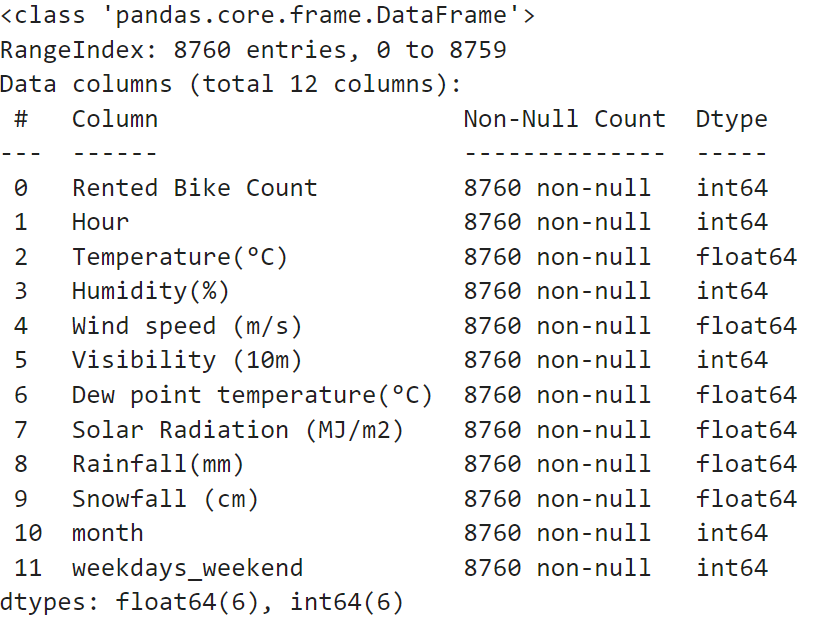
**Observation:** The distribution plot shows that the density of the Rented Bike.Count is at the peak around 350

1. Plot the histogram of all numerical features

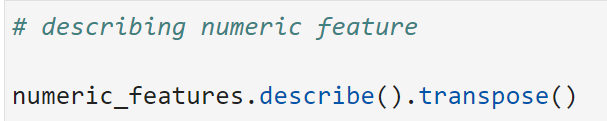
**Output:**

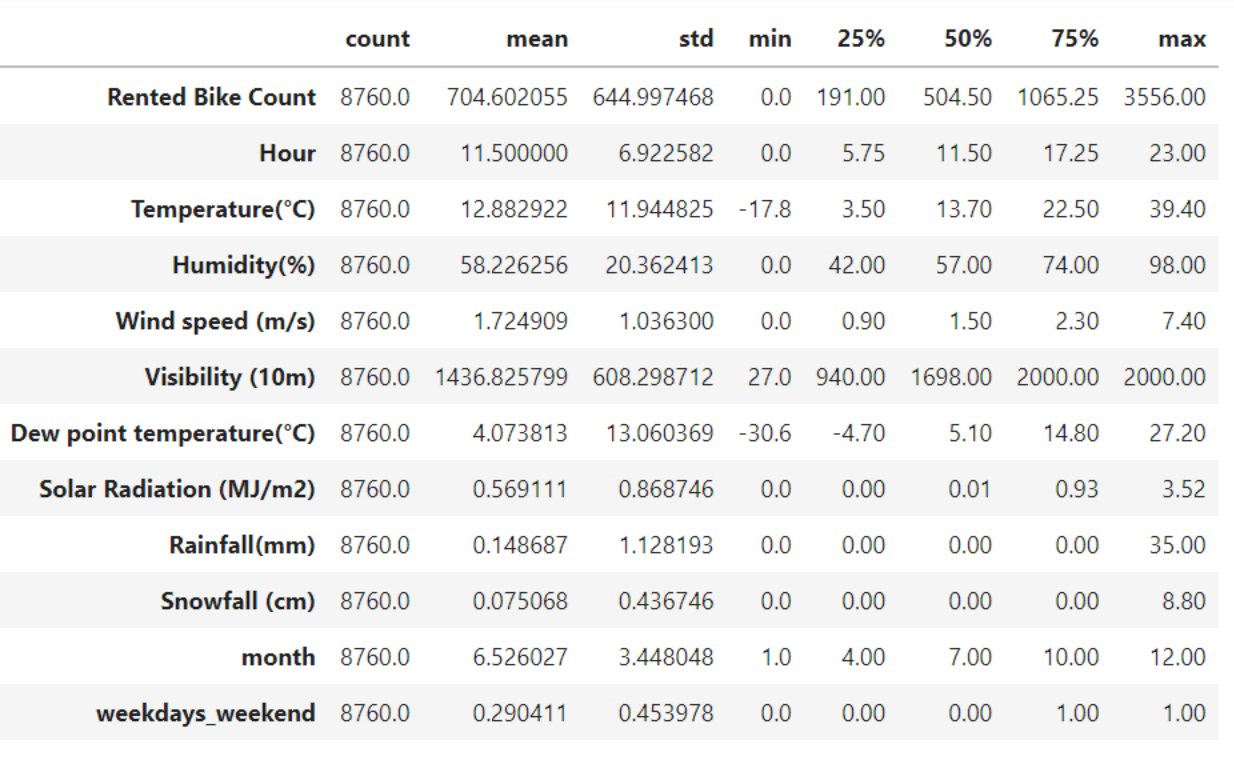


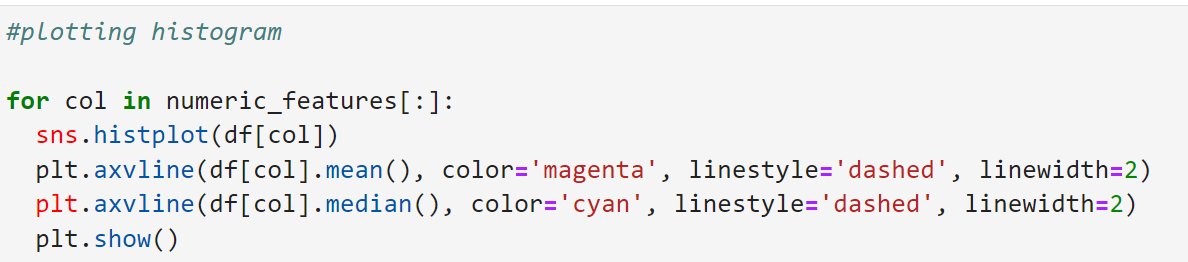


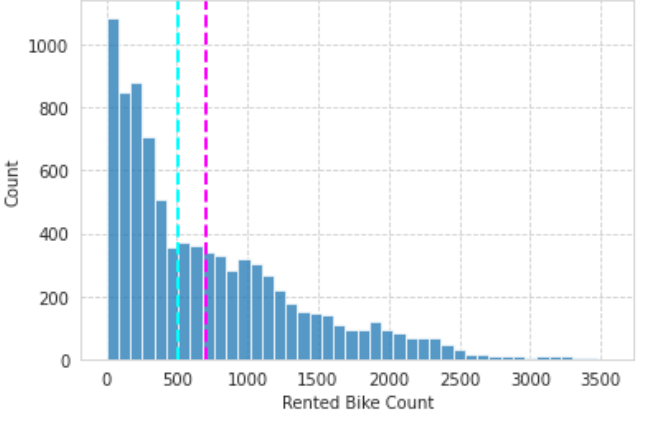




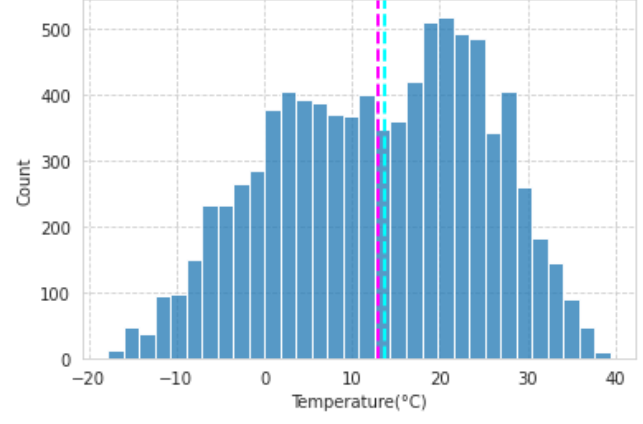


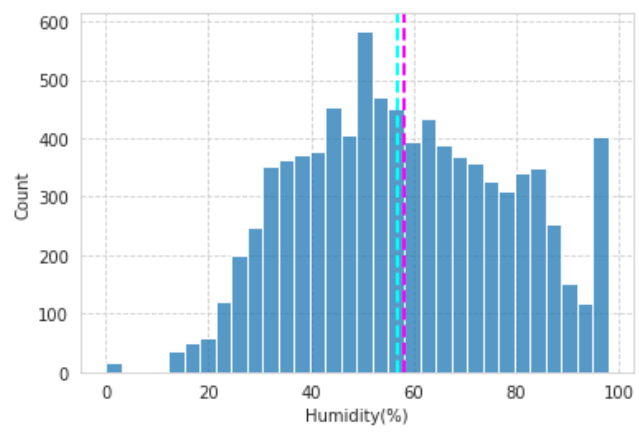


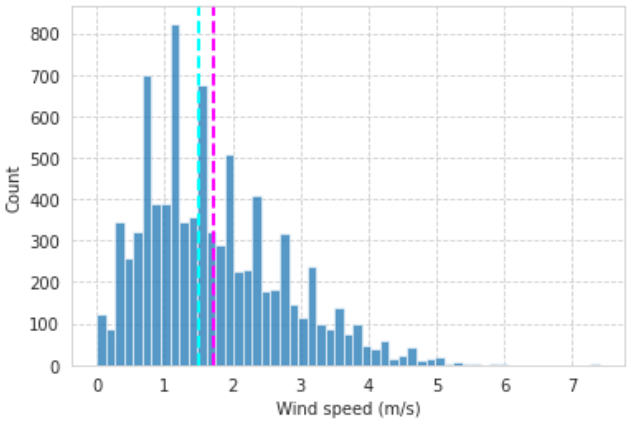


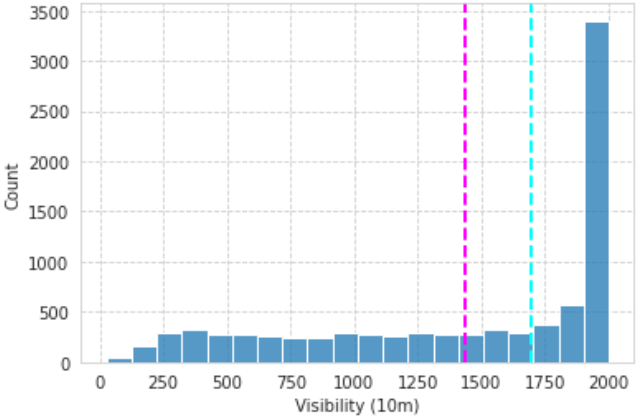


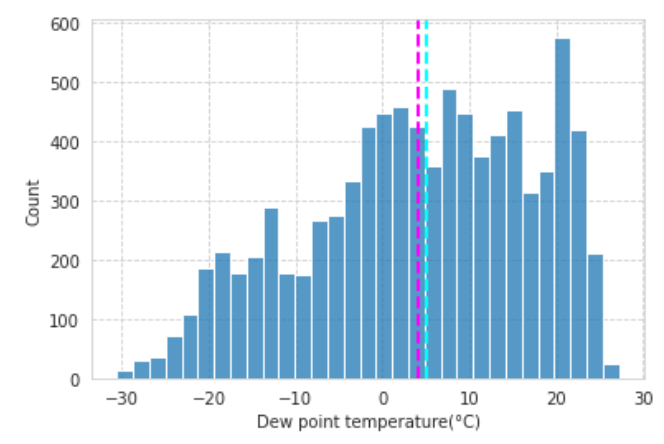


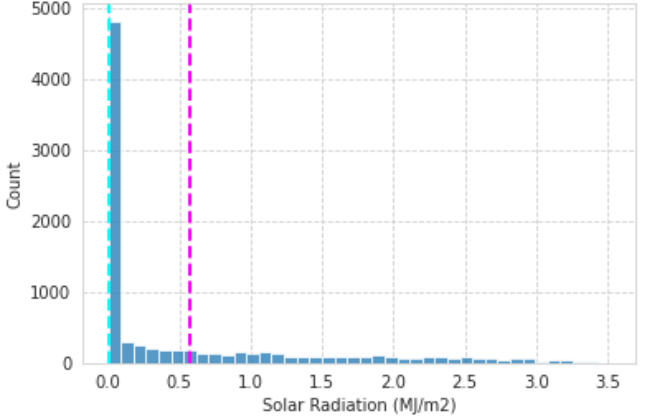


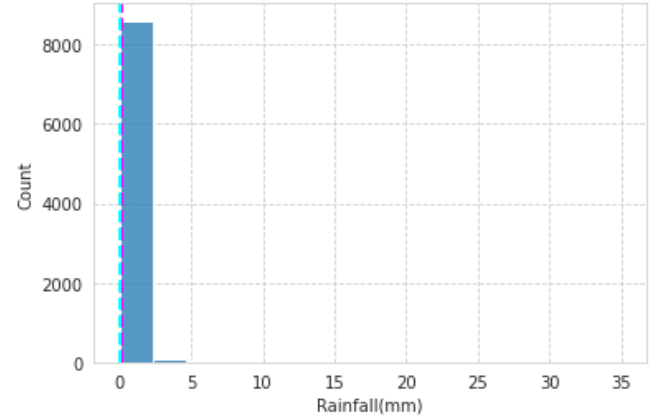


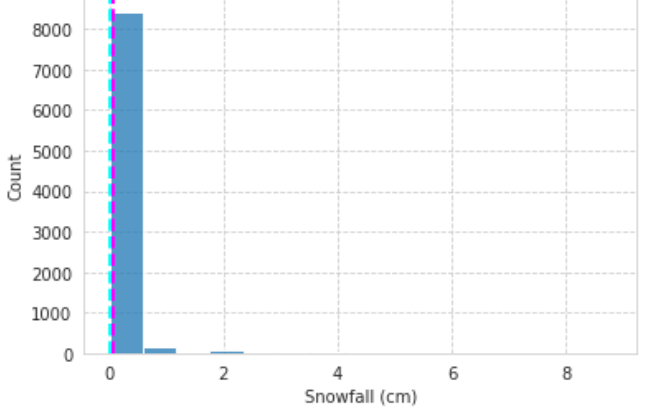


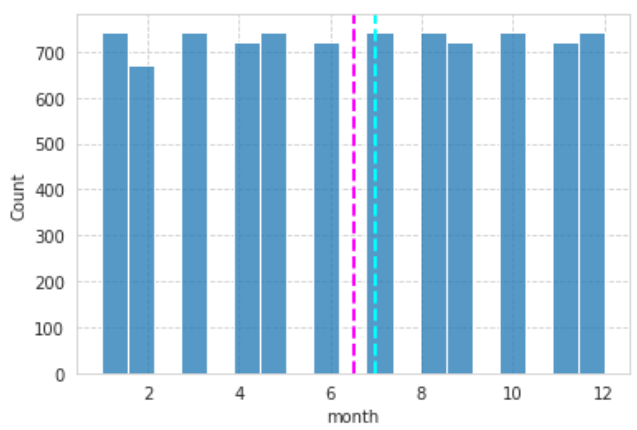


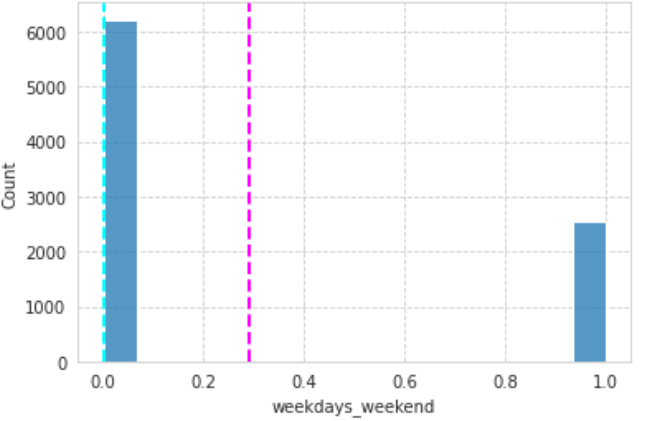










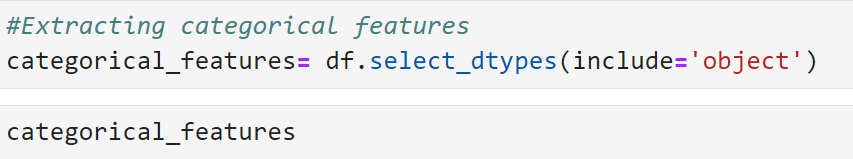


**Observation:**

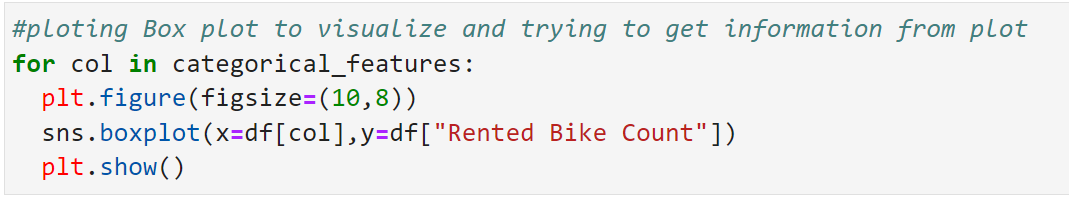
* The histogram plots shows the distribution of each numerical variables across its bins
* Each graphs shows when does their count reached its peak

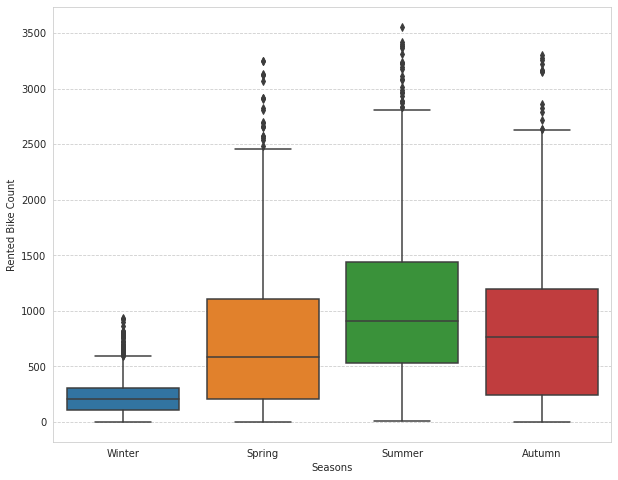
1. Plot the box plot of Rented Bike Count against all the categorical features (Hint: Categorical features on X-axis and Rented Bike Count on Y-axis)

**Output:**

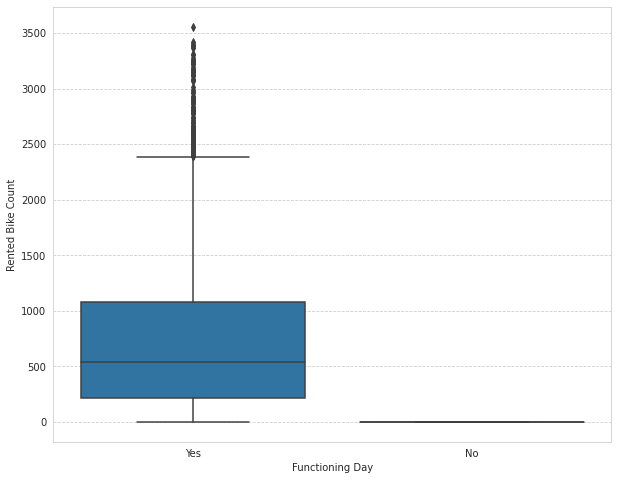










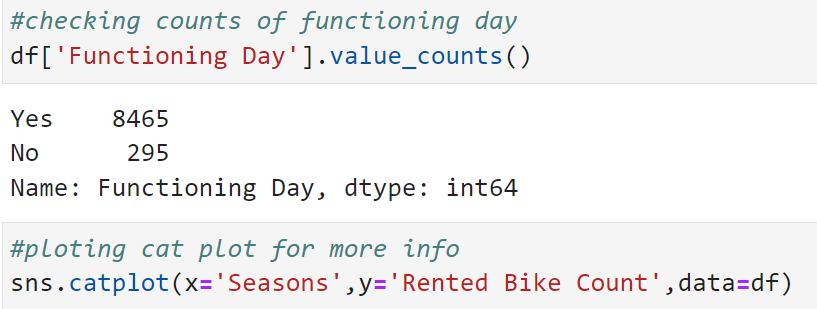


**Observation:**

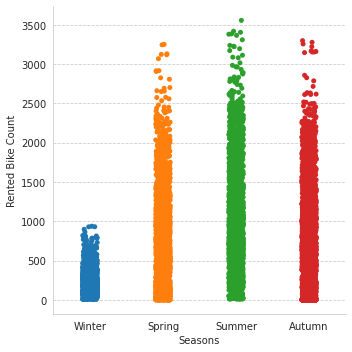
*- Less demand on winter seasons*

*- Sligthly Higher demand during Non holidays*

*- Almost no demand on non functioning day*





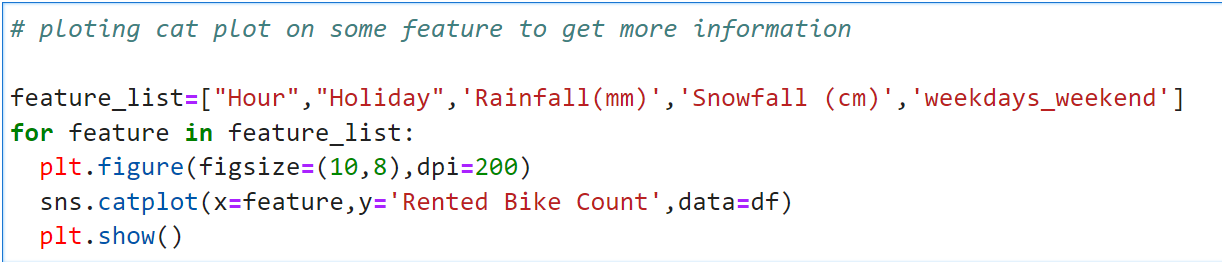


***Observation:***

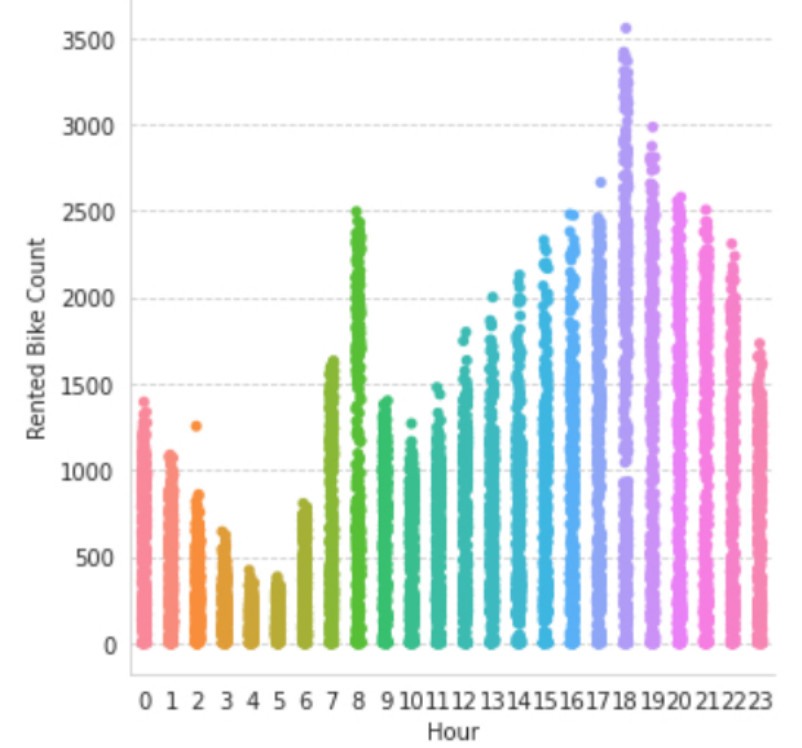
We can clearly see that there is less demand for rented bike in winter season.

1. Plot the Seaborn catplot of Rented Bike Count against features like Hour, Holiday, Rainfall(mm), Snowfall (cm), weekdays\_weekend and give your inferences

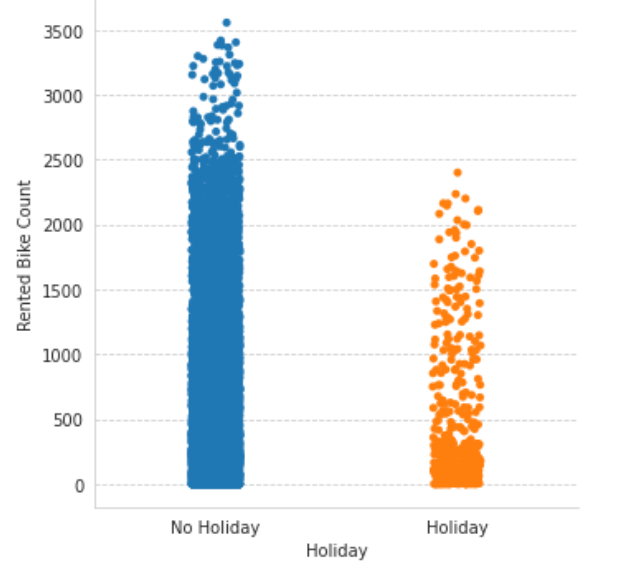
**Output:**

****

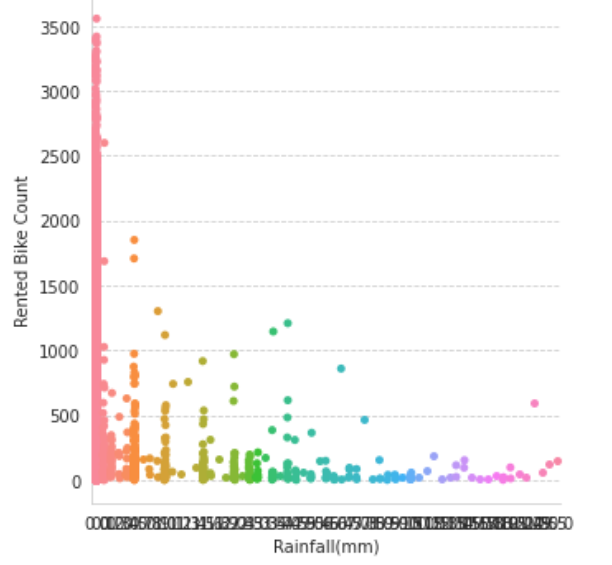
<Figure size 2000x1600 with 0 Axes>

****

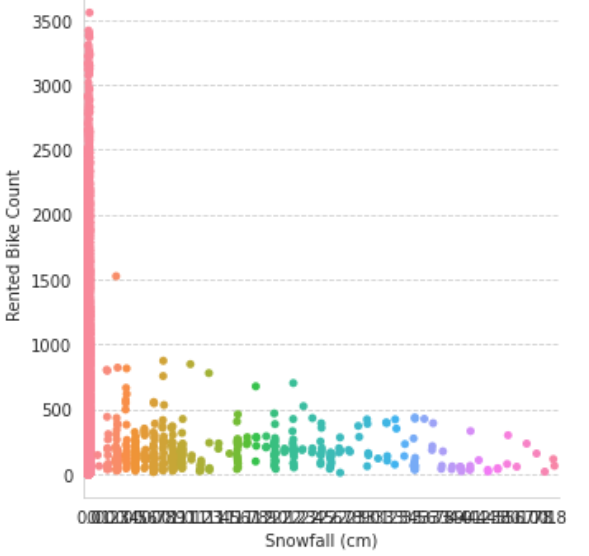
<Figure size 2000x1600 with 0 Axes>

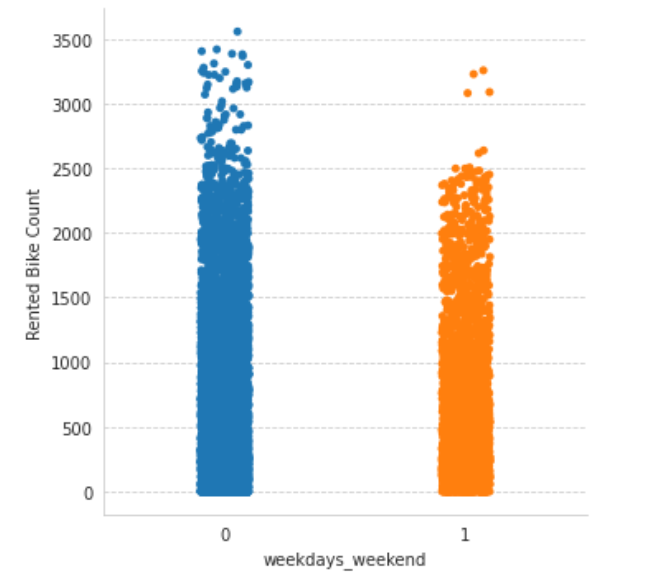
****

<Figure size 2000x1600 with 0 Axes>

****

<Figure size 2000x1600 with 0 Axes>

****

****

**Observation:**

*\*\*From hour v/s rented bike\*\**

* we can clearly see there is high demand of Rented bike between the office hours.

*\*\*From working-nonworking v/s rented bike\*\**

* As cleared from 2nd plot working days has comparatively high demand of rented bike as compared to non working day

*\*\*From Rainfall v/s rented bike\*\**

* we can see that if Rainfall increase demand of Rented Bike Decreases

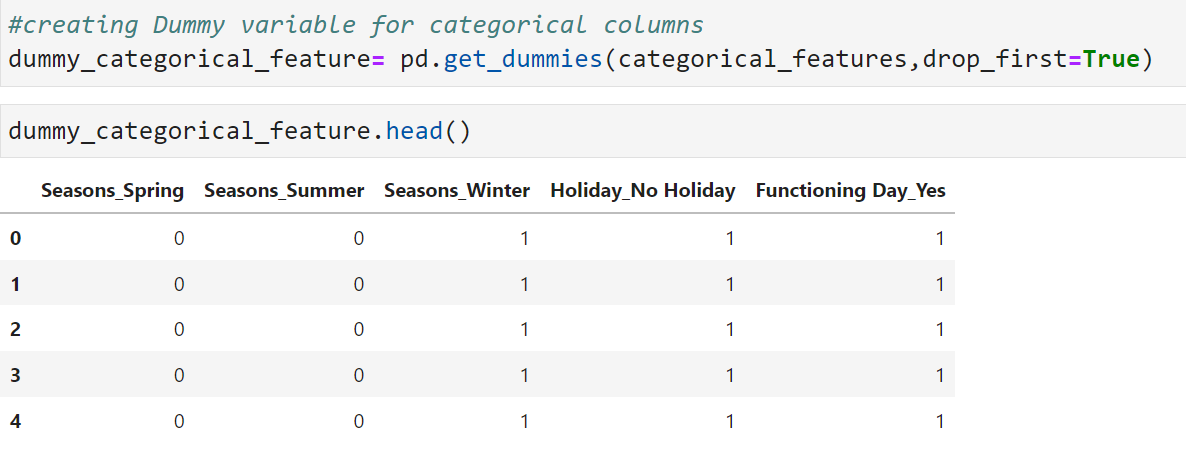
*\*\*From Snowfall v/s rented bike\*\**

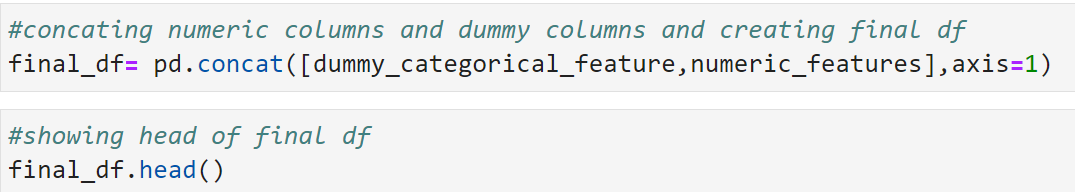
* we can see that if Snowfall increase demand of Rented Bike Decreases

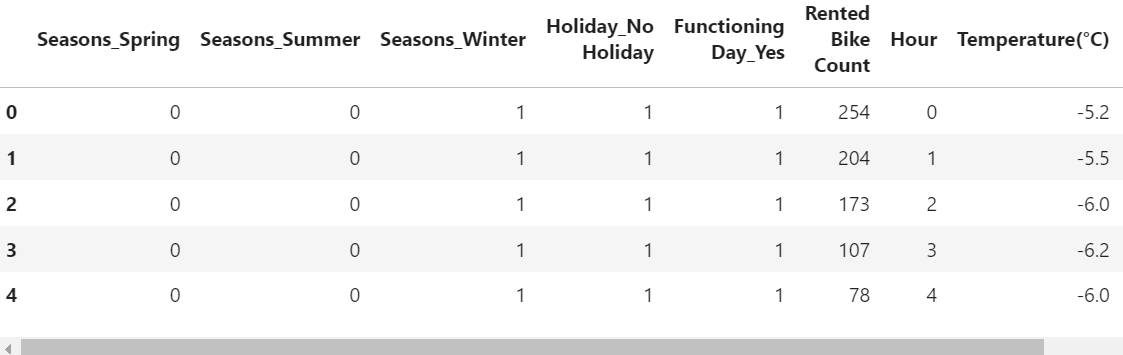
1. Encode the categorical features into numerical features.

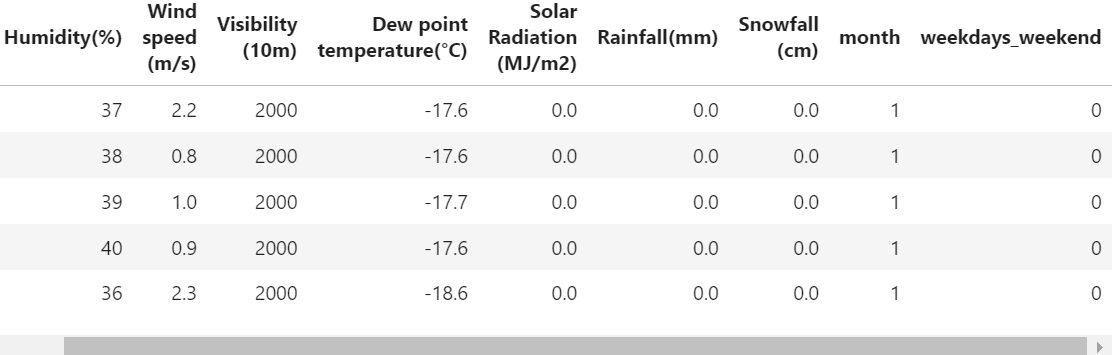
(Hint: use get\_dummies())

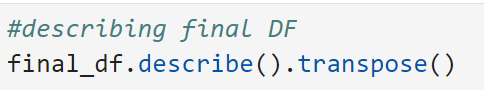
**Output:**

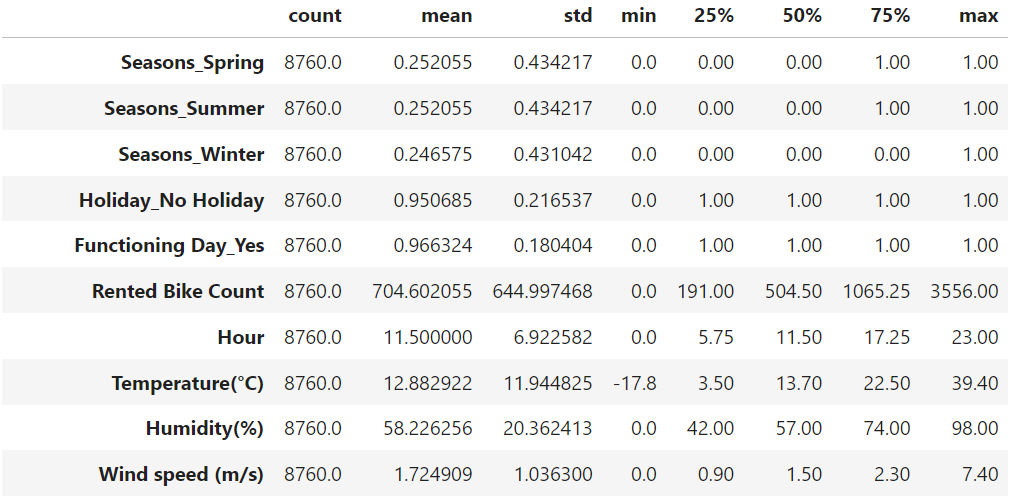








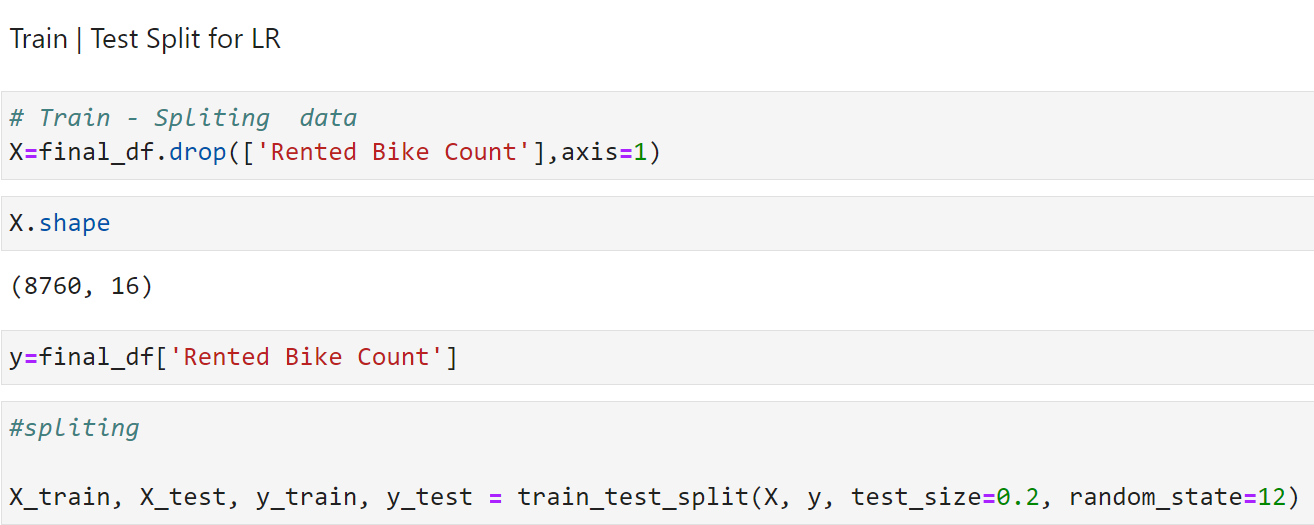






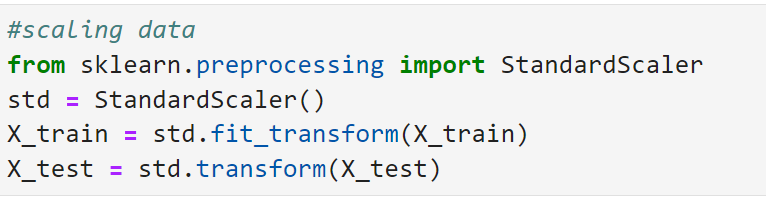
1. Identify the target variable and split the dataset into train and test with a ratio of 80:20 and random state 12

**Output:**



1. Perform Standard Scaling of the train dataset.

**Output:**

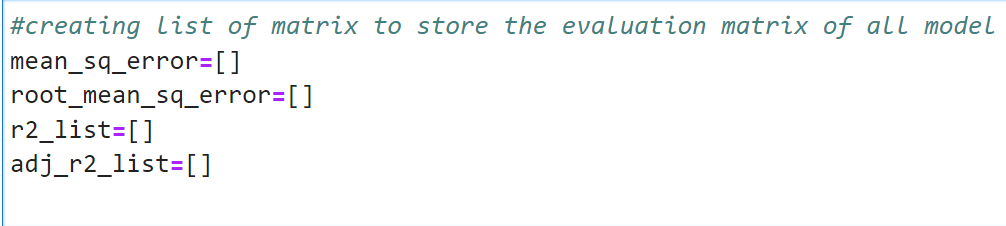


1. Perform Linear Regression, Lasso Regression and Ridge Regression for predicting the bike count required at each hour and compare the results.

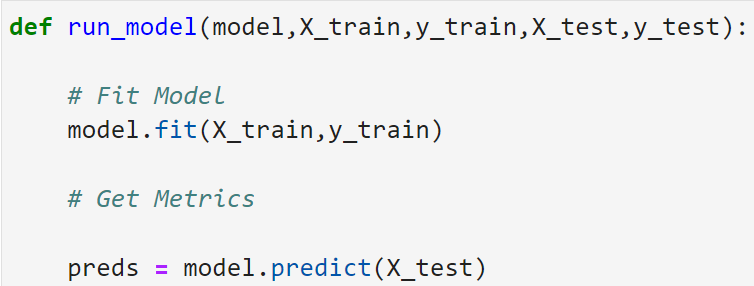
**Output:**

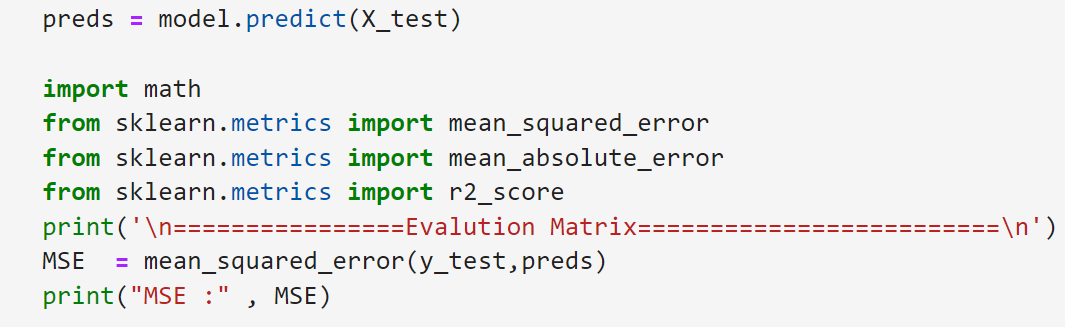
Functions

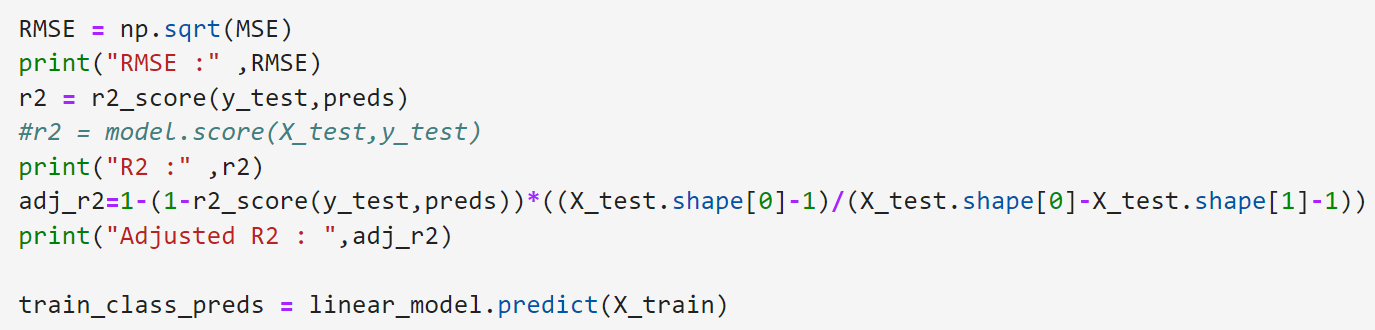
Creating Function to train linear models and calculate scores

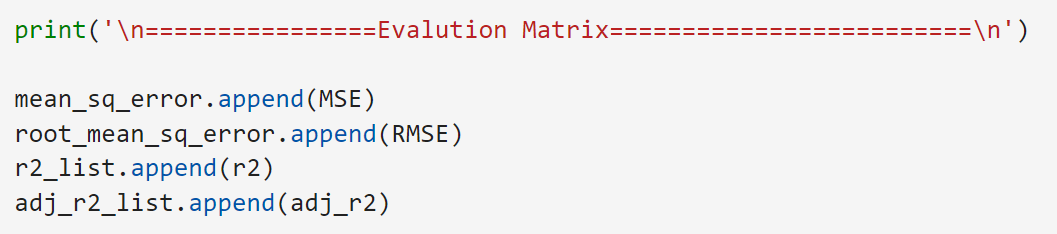


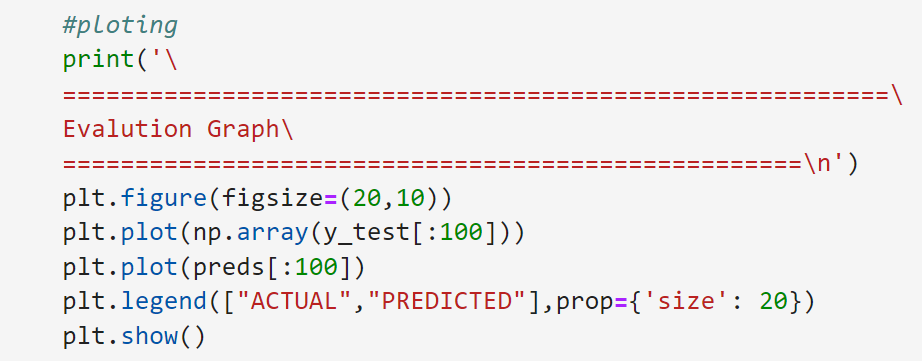
Creating functions to run different models:





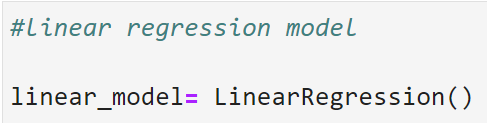


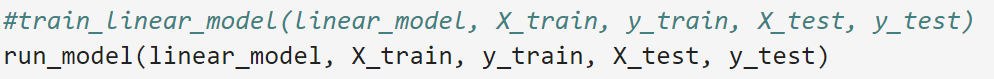


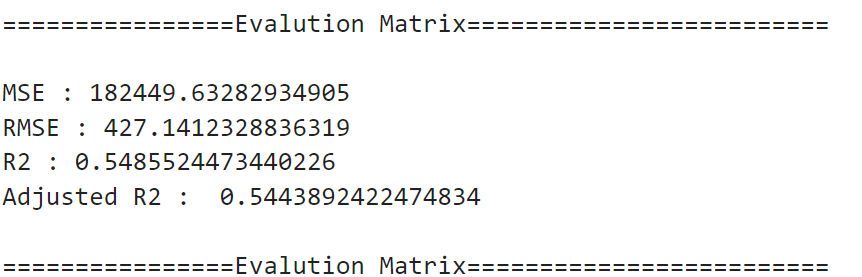


***Linear Regression:***

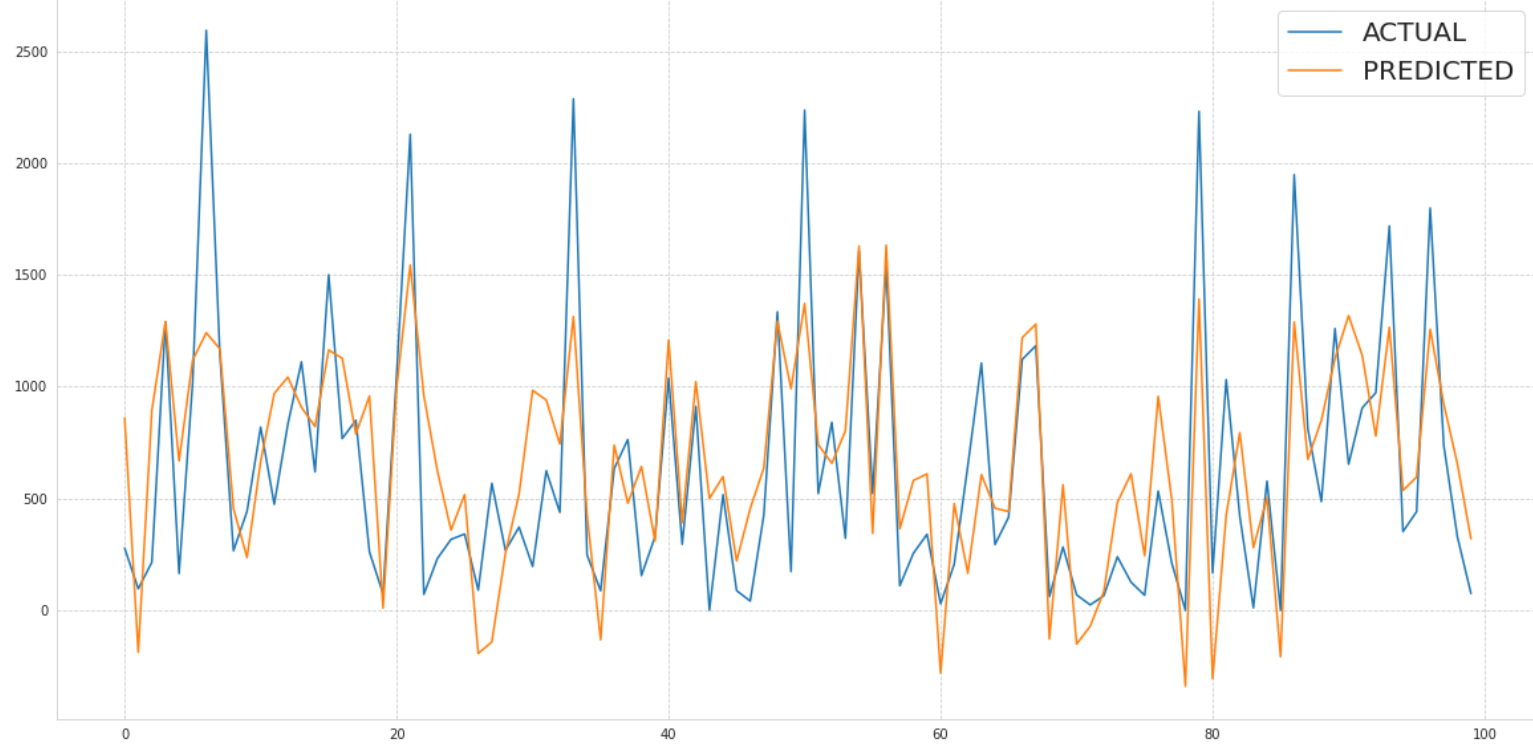
******

******

******

******

==================Evaluation Graph==========================

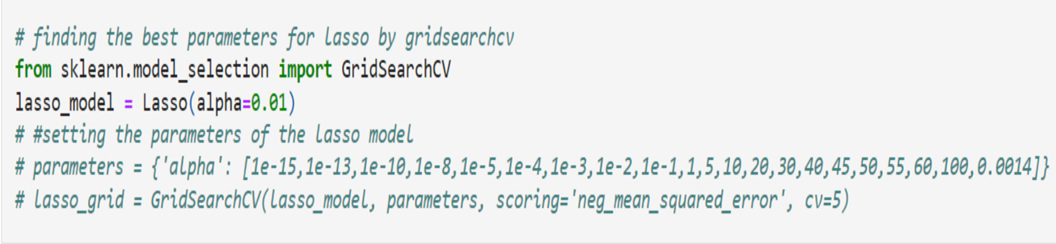


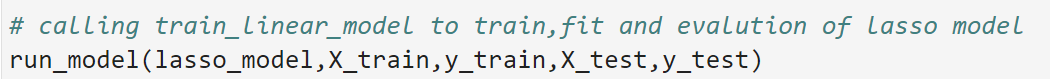
***Observation***: The above plot shows how far the predicted values are away from actual values in case of Linear Regression

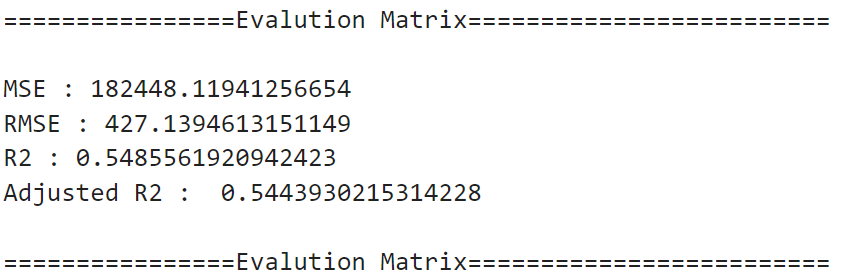
Regularisation

***Lasso Regression:***

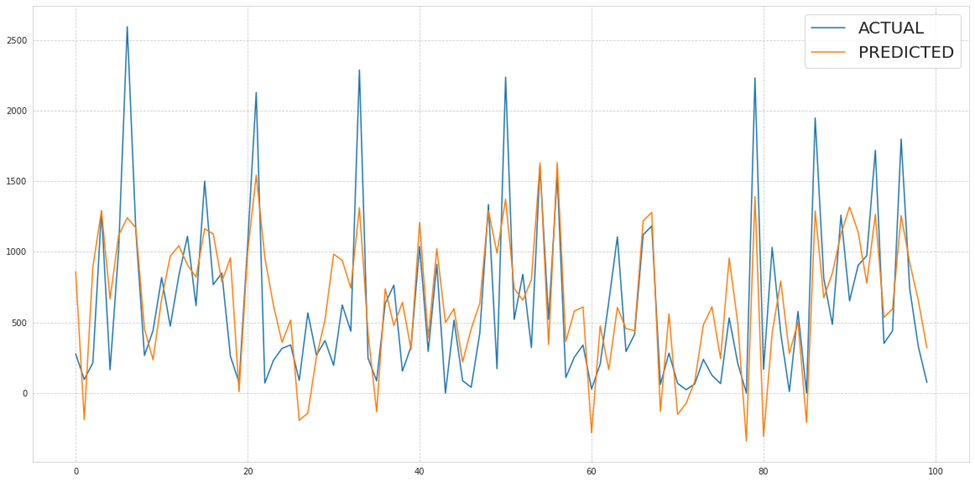
******

******

******

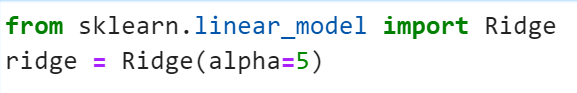
******

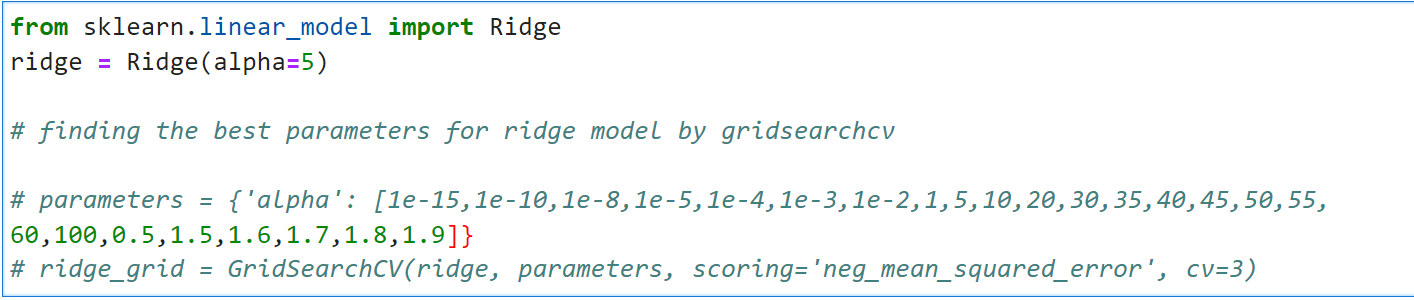
==================Evaluation Graph==========================

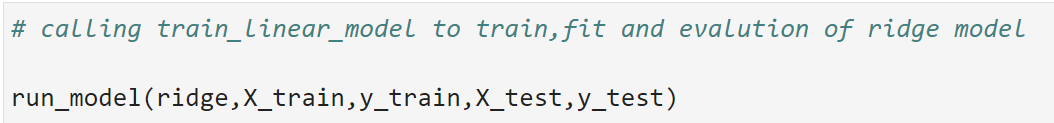


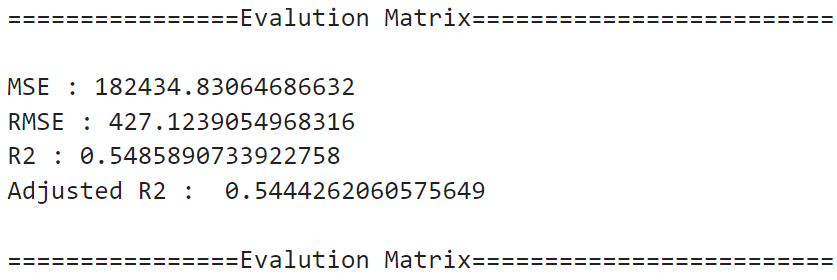
***Observation***: The above plot shows how far the predicted values are away from actual values in case of Lasso Regression

***Ridge Regression:***

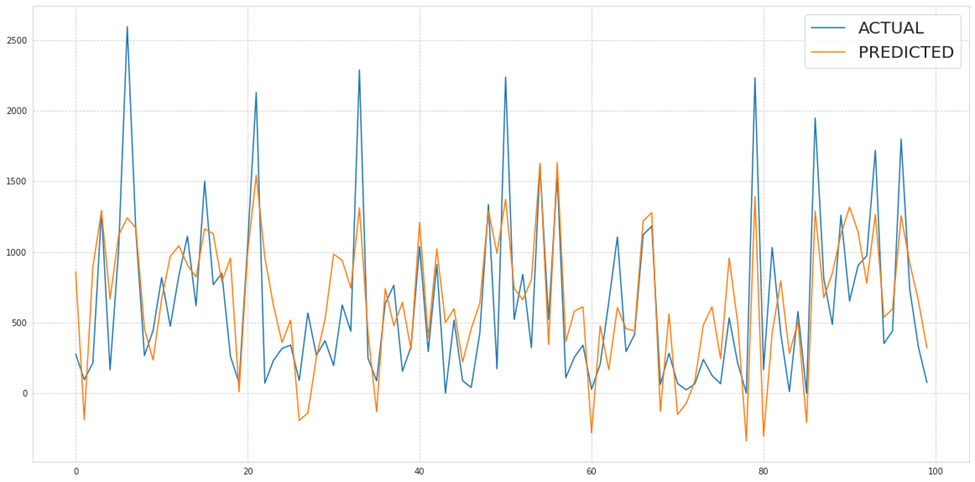
******

******

******

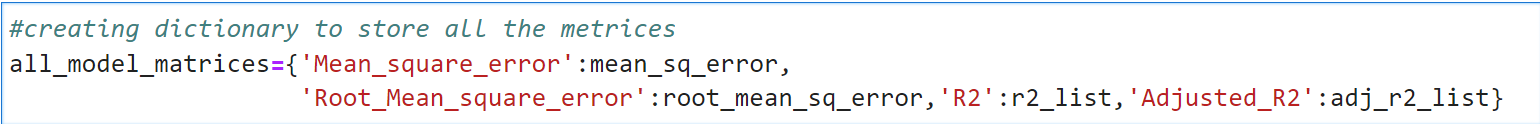


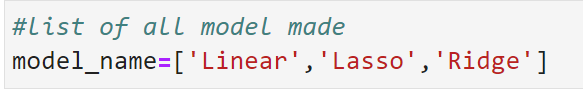
==================Evaluation Graph==========================

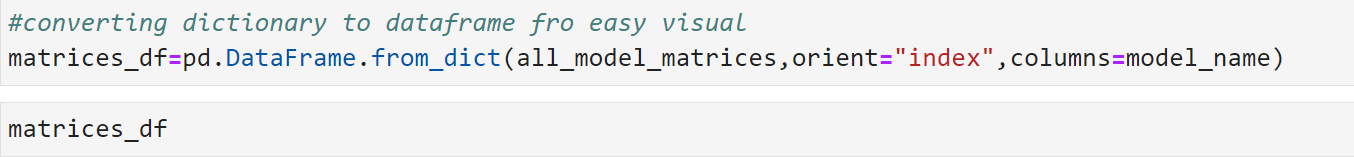


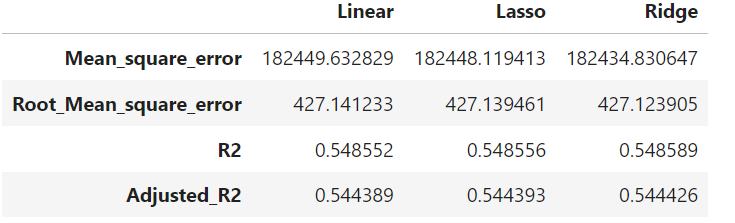
***Observation***: The above plot shows how far the predicted values are away from actual values in case of Ridge Regression

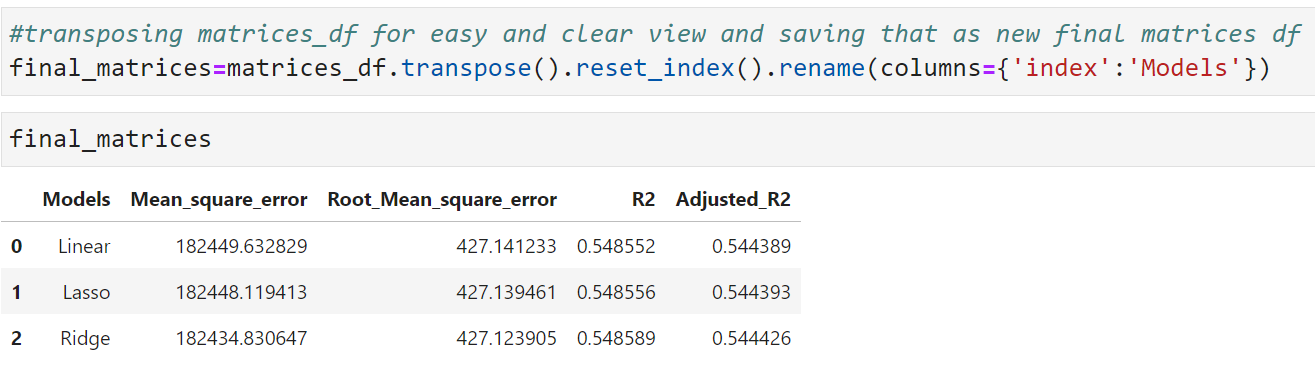
***Combined Evaluation Matrix of all models:***

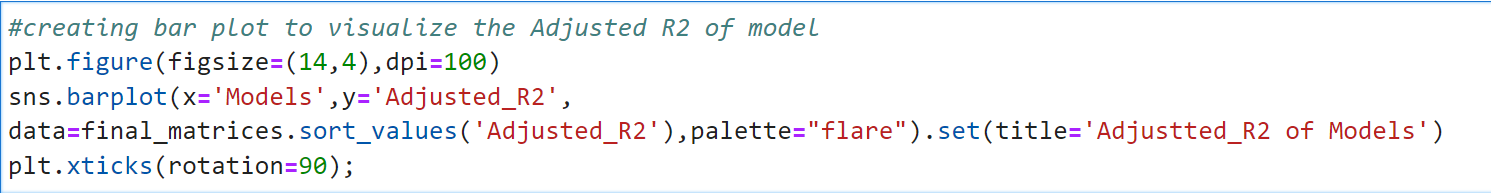
******

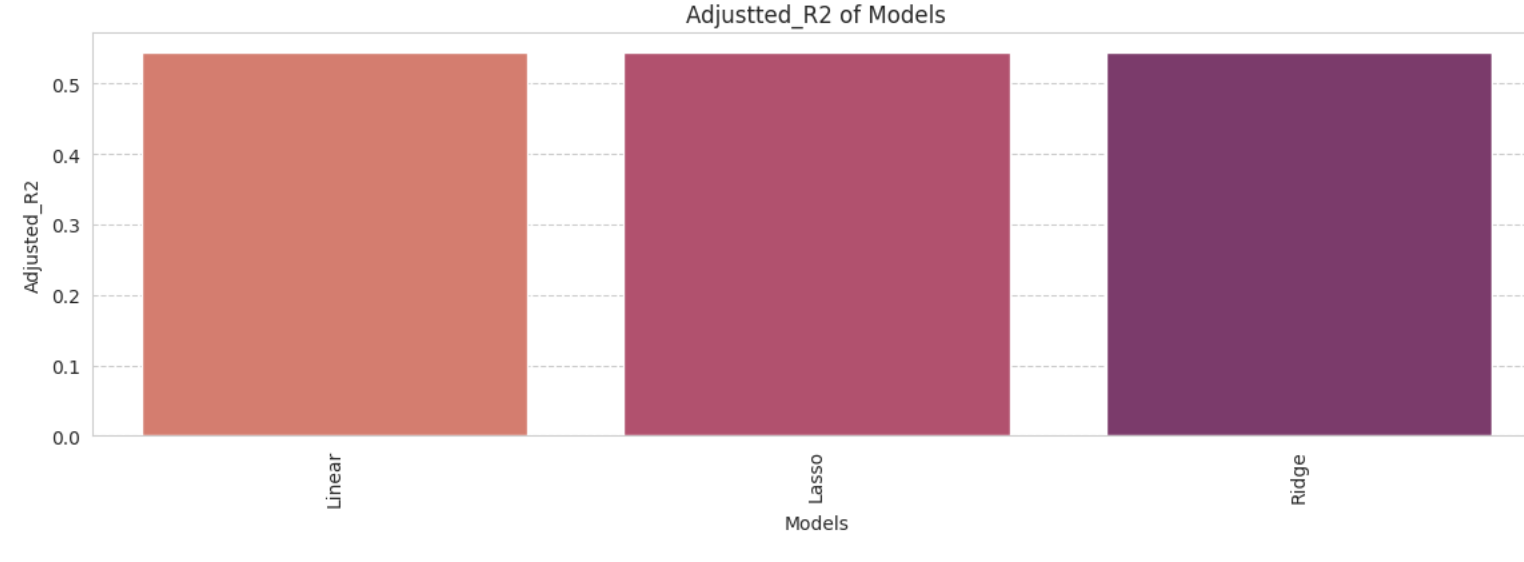
******

******







****

***Observation***: The bar plot shows adjusted R2 score of all the three models.

**Final Conclusion:**

1.In holiday or non-working days there is demands in rented bikes.

2.There is a surge of high demand in the morning 8AM and in evening 6PM

as the people might be going to their work at morning 8AM and returning from their work at the evening 6PM.

3.People prefered more rented bikes in the morning than the evening.

4.When the rainfall was less, people have booked more bikes except some few cases.

7.The Temperature, Hour & Humidity are the most important features that positively drive the total rented bikes count.

8.After performing the various models the Lasso and Ridge found to be the slightly better model that can be used for the Bike Sharing Demand Prediction since the performance metrics (mse,rmse) shows lower and (r2,adjusted\_r2) shows a higher value for the Lasso and Ridge models !

9.We can use either Lasso or Ridge model for the bike rental stations.

10. For further improvement in the performance, one need to try out a more complex models like RandomForest Regressor, GradientBoosting Regressor, LightGBM regressor, etc.