Capstone Project

Bike Sharing Demand Prediction

By-Aman kumar

Table of Contents:

- Introduction.
- Dataset description.
- Percentage of Rented Bike on Holiday and No Holiday.
- Distribution of the Rented bike count.
- Which Season has the most number of bike rented?
- Analysing for different temperatures.
- Count of Rented Bike in different hour.
- Correlation matrix.

Table of Contents:

- Fitting all features with Rented Bike Count.
- Linear regression.
- Ridge Regression.
- Lasso Regression
- Random Forest Regression.
- Conclusion.

Introduction:

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place an there need. Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern.

Dataset Description:

- The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

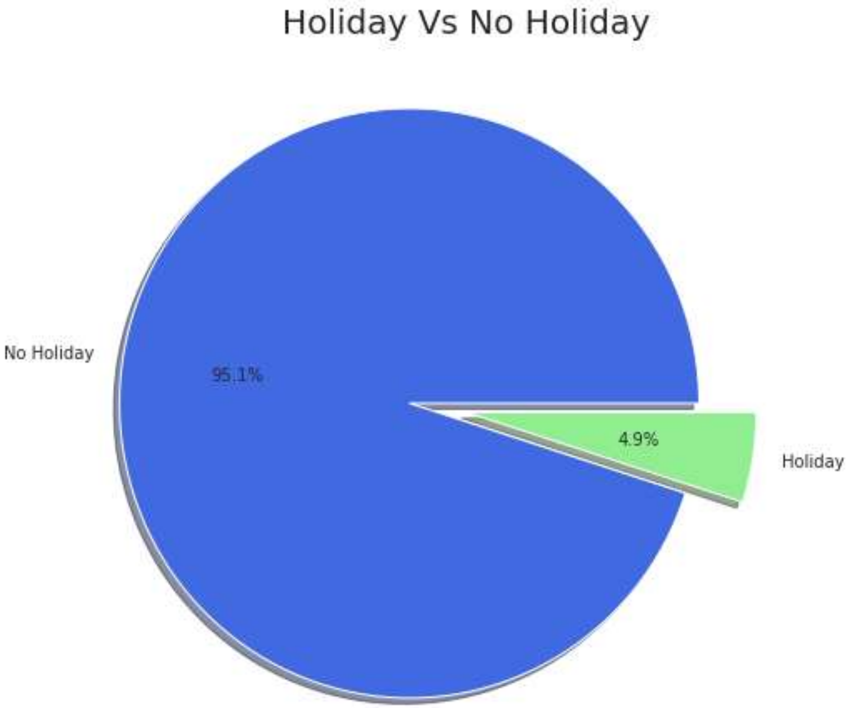
- **Attribute Information:**
- 1) Date : year-month-day
- 2) Rented Bike count - Count of bikes rented at each hour
- 3) Hour - Hour of day
- 4) Temperature-Temperature in Celsius
- 5) Humidity - %
- 6) Windspeed - m/s
- 7) Visibility - 10m
- Hours), Fun(Functional hours)

Dataset Description:

8) Dew point temperature - Celsius
9) Solar radiation - MJ/m2
10) Rainfall - mm
11) Snowfall - cm
12) Seasons - Winter, Spring, Summer, Autumn
13) Holiday - Holiday/No holiday
14) Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours).

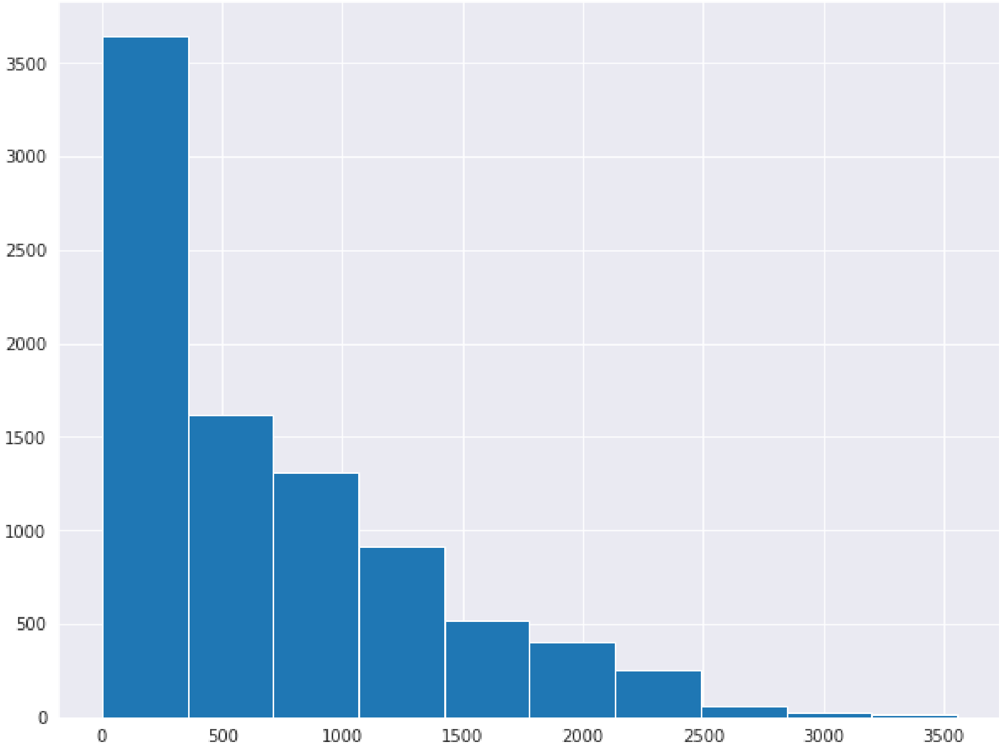
Percentage of Rented Bike on Holiday and No Holiday :

From this plot, we can see that the majority of the bikes rented are on days which are considered as No Holiday.



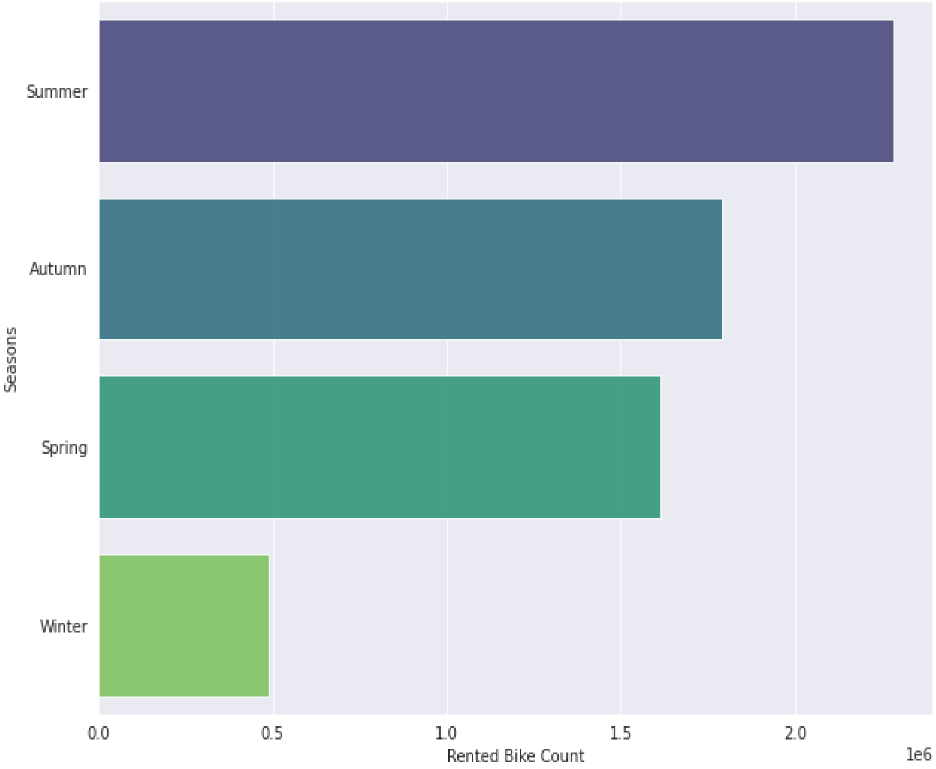
Distribution of the Rented bike count:

From this plot, we can come to the conclusion that most of the bike rented between 0 to 400.



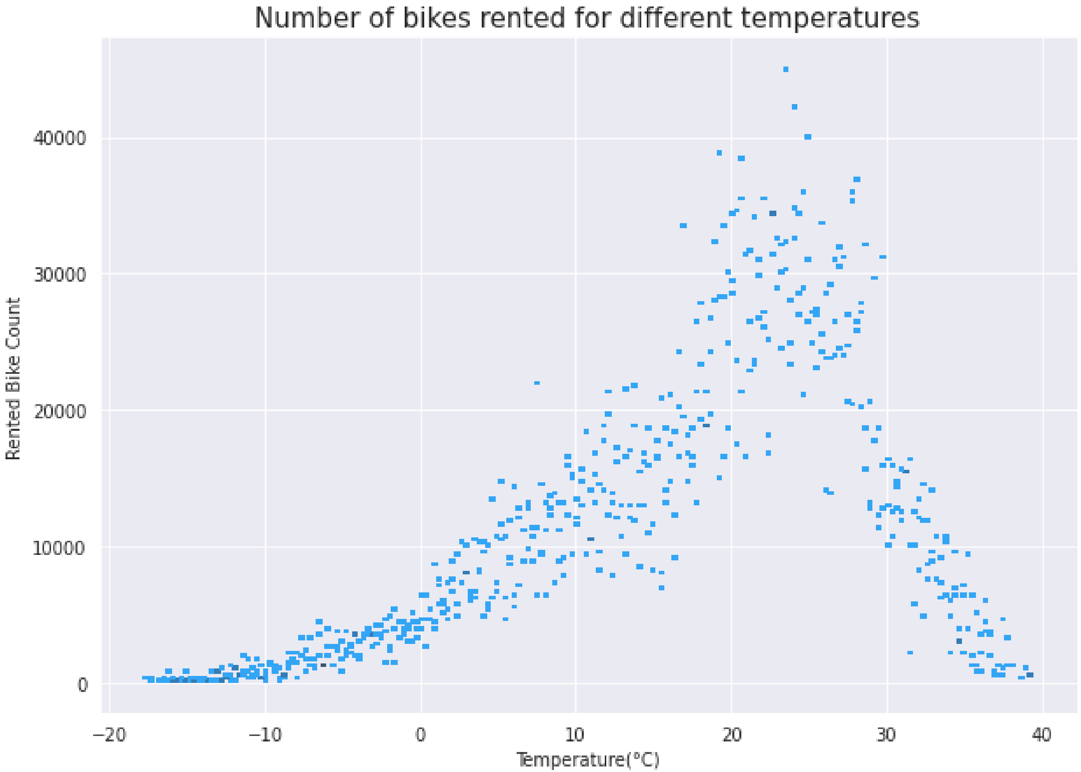
Which Season has the most number of bike rented?

- As we can see that the
- majority of the bikes
- rented in the Summer season and the lowest in
- the winter season.



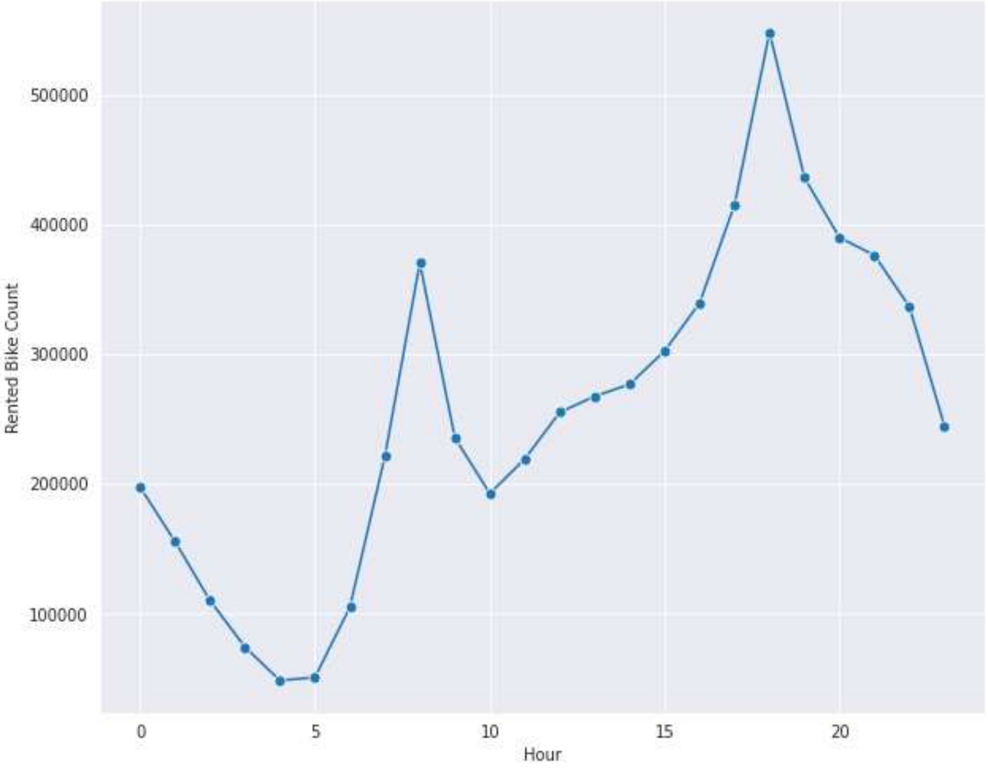
Analysing for different temperatures:

From this visualization, we can see that the Most number of bikes are rented in the temperature range of 20 degrees to 30 degrees.



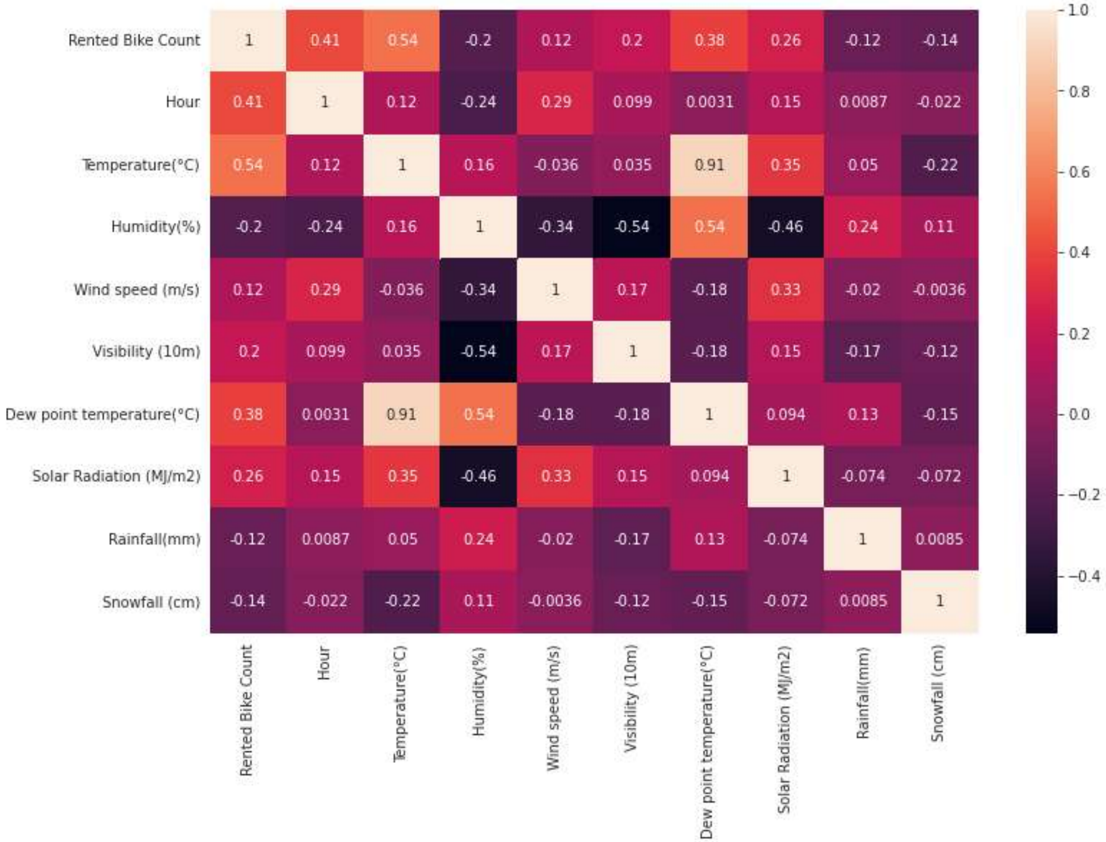
Count of Rented Bike in different hour:

From this graph, we can come to the conclusion that the highest number of bike rentals have been the 18th hour i.e 6pm, and lowest in the 4th hour i.e 4am.



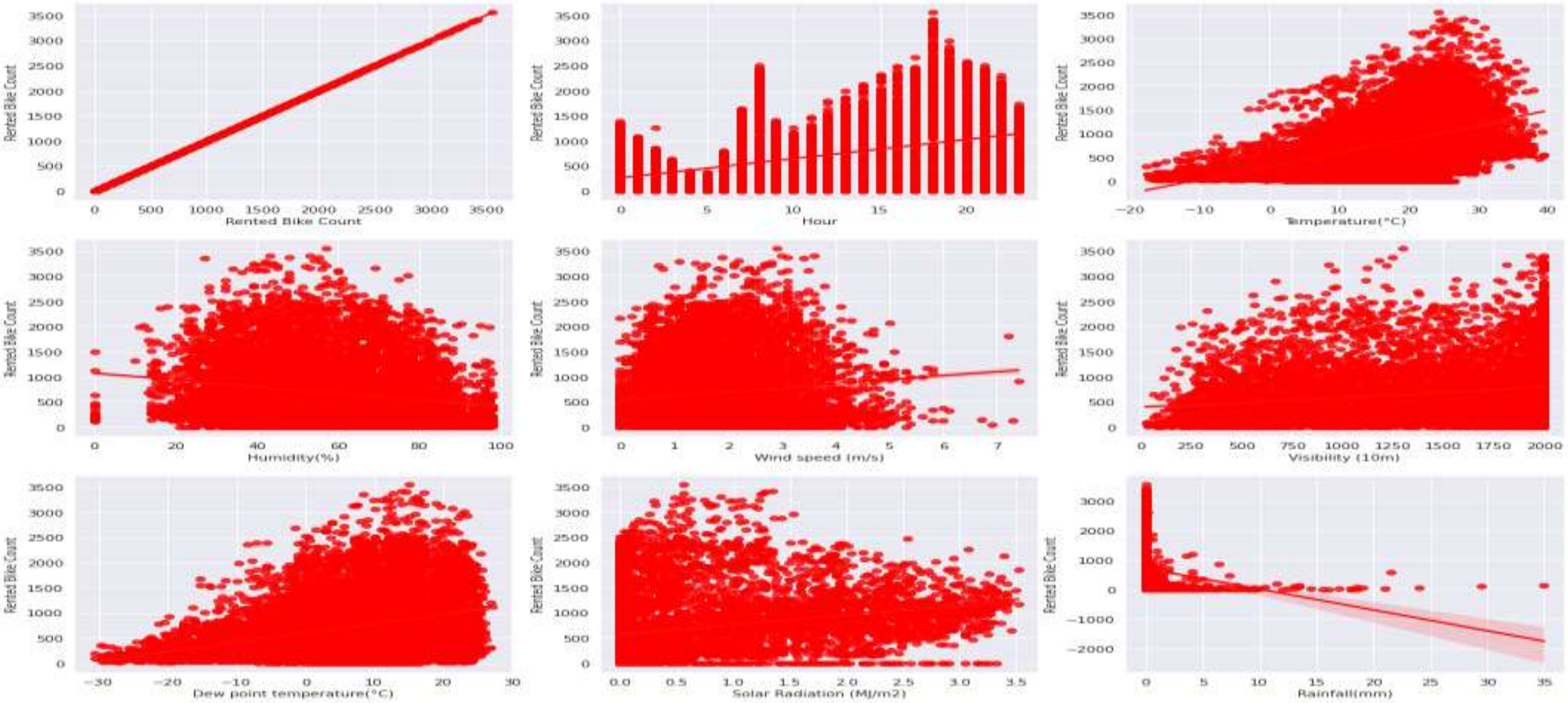
Correlation matrix:

- From this graph, we can come to the Conclusion that 'Rented Bike Count' has high positive correlation with Temperature, Hour, and solar Radiation and negative correlation Rainfall, snowfall and Humidity.



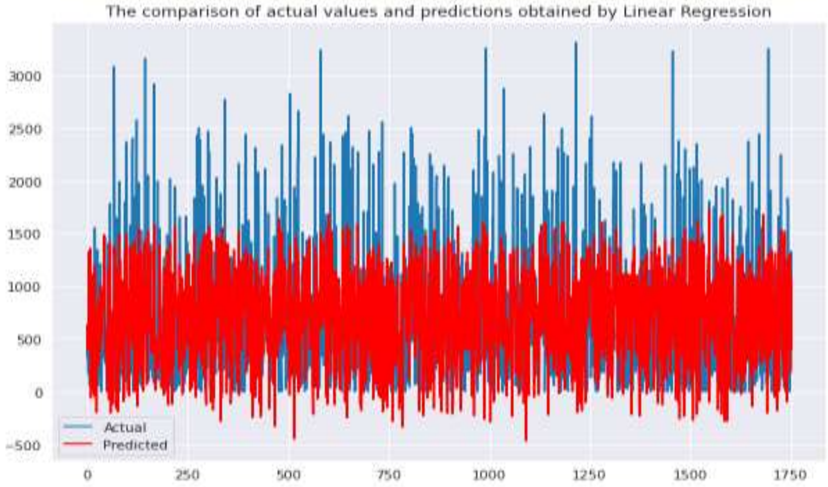
	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)
Rented Bike Count	1	0.41	0.54	-0.2	0.12	0.2	0.38	0.26	-0.12	-0.14
Hour	0.41	1	0.12	-0.24	0.29	0.099	0.0031	0.15	0.0087	-0.022
Temperature(°C)	0.54	0.12	1	0.16	-0.036	0.035	0.91	0.35	0.05	-0.22
Humidity(%)	-0.2	-0.24	0.16	1	-0.34	-0.54	0.54	-0.46	0.24	0.11
Wind speed (m/s)	0.12	0.29	-0.036	-0.34	1	0.17	-0.18	0.33	-0.02	-0.0036
Visibility (10m)	0.2	0.099	0.035	-0.54	0.17	1	-0.18	0.15	-0.17	-0.12
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	-0.18	-0.18	1	0.094	0.13	-0.15
Solar Radiation (MJ/m2)	0.26	0.15	0.35	-0.46	0.33	0.15	0.094	1	-0.074	-0.072
Rainfall(mm)	-0.12	0.0087	0.05	0.24	-0.02	-0.17	0.13	-0.074	1	0.0085
Snowfall (cm)	-0.14	-0.022	-0.22	0.11	-0.0036	-0.12	-0.15	-0.072	0.0085	1

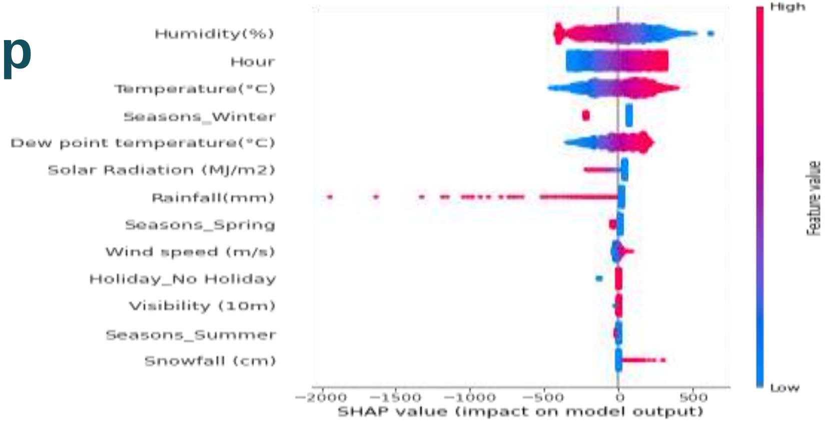
Fitting all features with Rented Bike Count:



Linear regression:

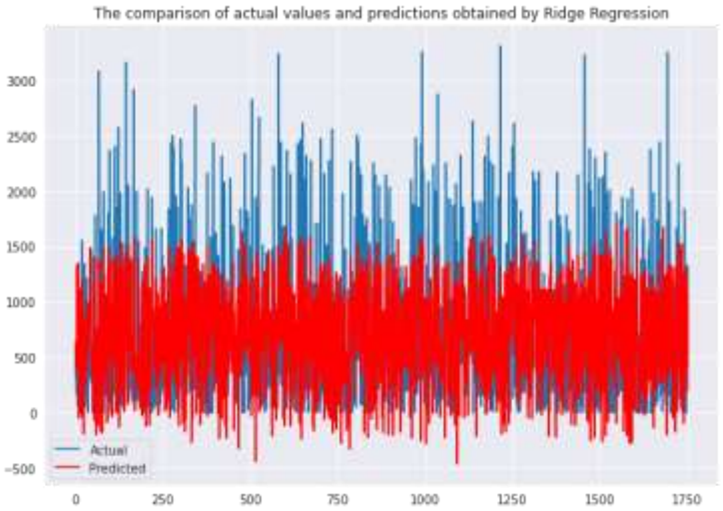
Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable value (x). So, this regression technique finds out a linear relationship between x (input) and y(output) with r2 score = 0.4719.

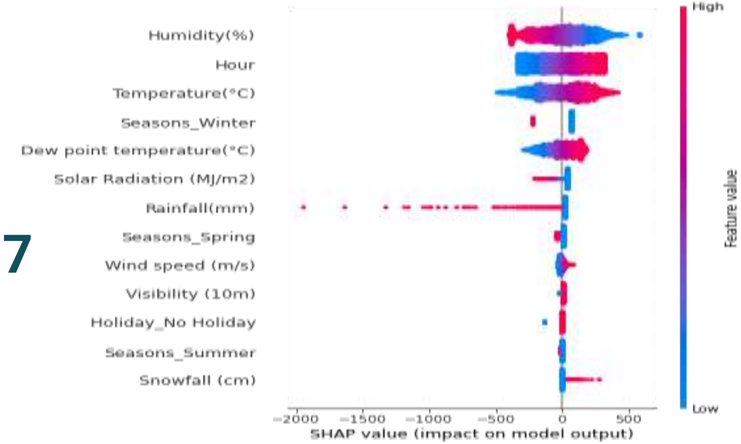




Ridge Regression:

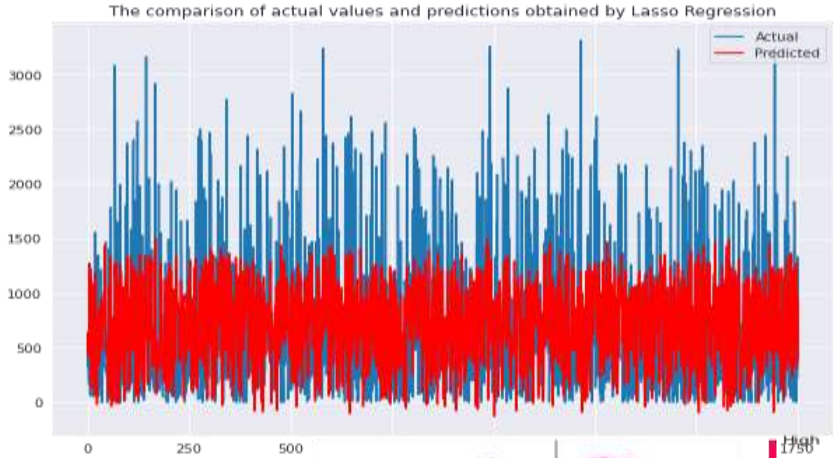
Ridge regression is a method of estimating the coefficients of regression models in scenarios where the independent variables are highly correlated. It uses the linear regression model with the L2 regularization method. As we can see there is only a very slight difference between the results achieved through Linear regression and Ridge regression with r2 score = 0.4717

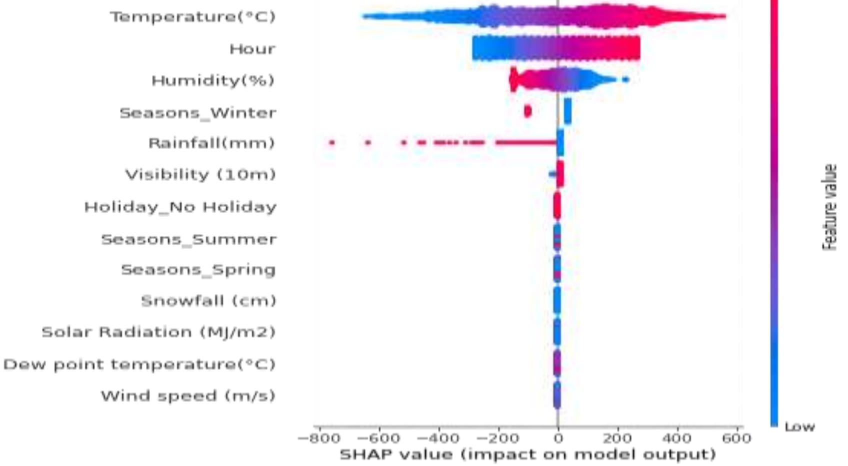




Lasso Regression:

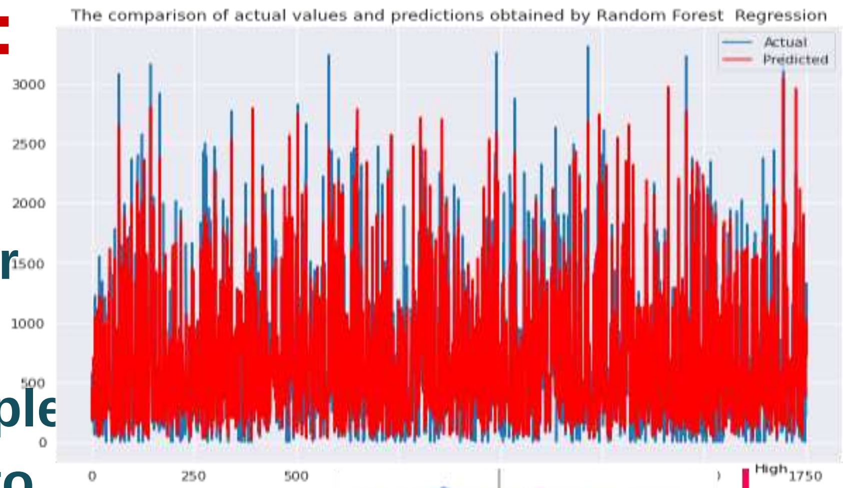
Lasso regression analysis is a shrinkage and variable selection Method for linear regression models. The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. Lasso Regression(L1 regularization) is the worst performing model with an r2 score of 0.4359.

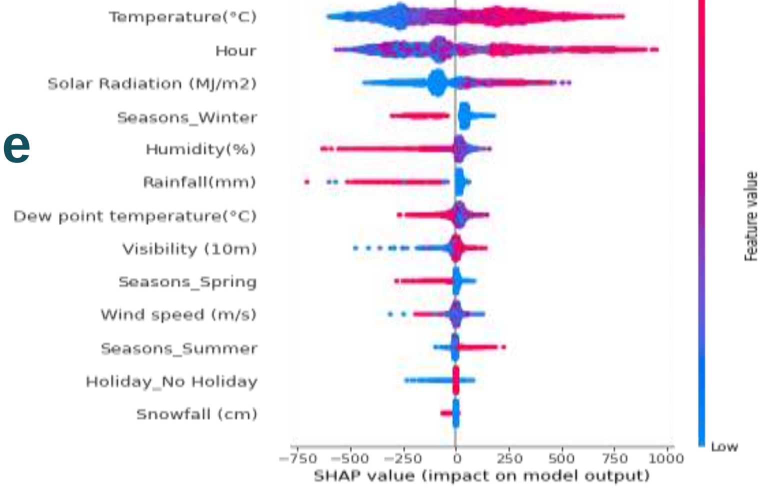




Random Forest Regression:

- Fg
- A random forest is a meta estimator that fits a number of classifying decision trees on various sub-sample of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. As we can see the quality of model prediction has drastically improved with an r2 score of 0.7347





Conclusion:

That's it! We reached the end of our exercise.

- Starting with loading the data so far we have done EDA, Most number of bikes are rented in the Summer season and the lowest in the winter season. 95.1% of the bikes are rented on days that are considered as No Holiday. Most number of bikes are rented in the temperature range of 20 degrees to 30 degrees and Majority of the bikes are rented for a humidity percentage range of 30 to 70. The highest number of bike rentals have been done in the 18th hour and lowest in the 4th hour. Most of the bike rentals have been made when there is high visibility.

Conclusion:

- Results from ML models:
- Random Forest Regression is the best performing model with an r2 score of 0.7347. Lasso Regression(L1 regularization) is the worst performing model with an r2 score of 0.4359. All 4 models have been explained with the help of SHAP library. Temperature and Hour are the two most important factors according to all the models.

THANK YOU!