Capstone Project-2

Bike Sharing Demand Prediction

By-Bharath Kumar A

Github link- https://github.com/bharath967/Seoul-Bike-Sharing-Demand-Prediction

Agenda

- 1) Introduction
- 2) Problem Statement
- 3) Overview of the dataset
- 4) EDA and Feature engineering
- 5) Implementation of model
- 6) Model Performance
- 7) Hyperparameters tuning
- 8) Conclusions

Indroduction

Bike sharing systems are a type of bicycle rental service in which the company provides renting a bike, and returning the bike when we use it by traveling throughout our desired location around a city.

People can rent a bike from one location and return it to a different location on an as needed basis using these systems.

The purpose of this study is to estimate bike rental demand by combining past bike usage trends with meteorological data, the data set consists of two years worth of hourly rental data

Problem Statement

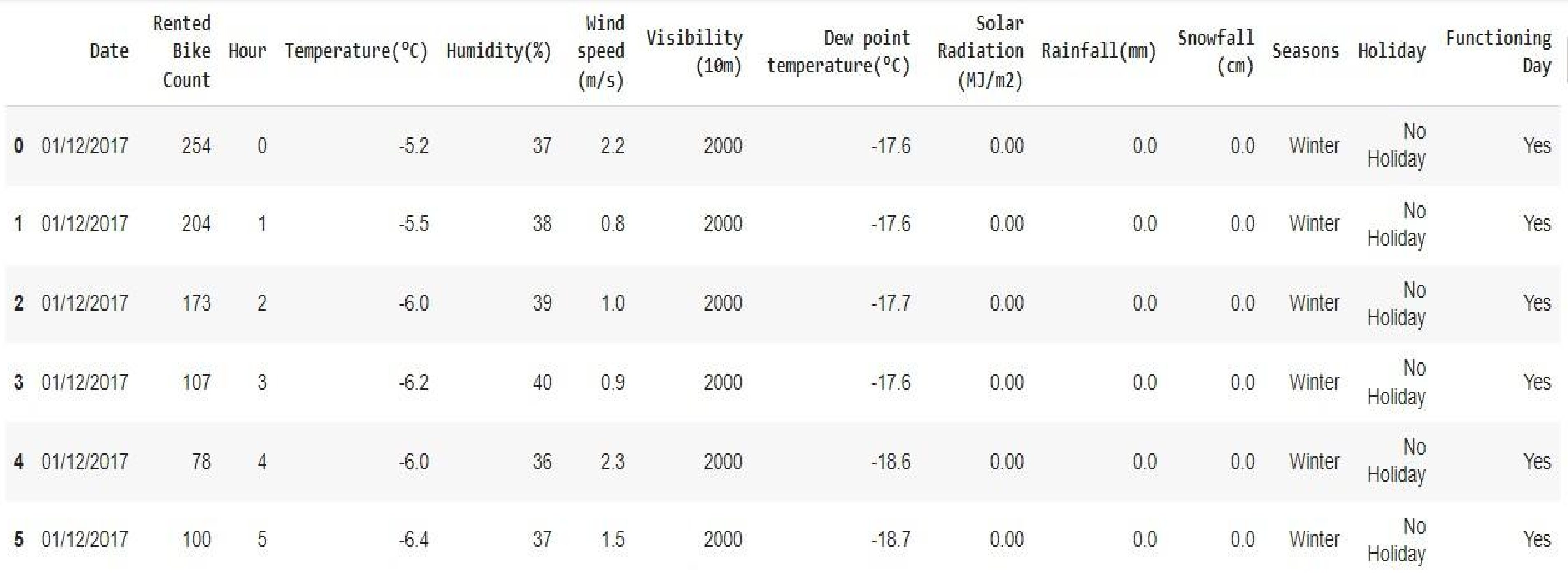
- Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it reduces the waiting time.
- Eventually, providing the city with a stable supply of rental bikes becomes a
- major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Overview of the dataset

- This dataset contains weather information about the number of bikes rented per hour and date information.
- Attribute Information:
- ● Date : in (Year-Month-Day)
- ● Rented Bike count - Count of bikes rented at each hour
- ● Hour - Hour of the day
- ● Temperature-Temperature in Celsius
- ● Humidity - %
- ● Wind Speed - m/s
- ● Visibility
- ● Dew point temperature - Celsius
- ● Solar radiation
- ● Rainfall - mm
- ● Snowfall - cm
- ● Seasons - Winter, Spring, Summer, Autumn
- ● Holiday - Holiday/No holiday
- ● Functional Day

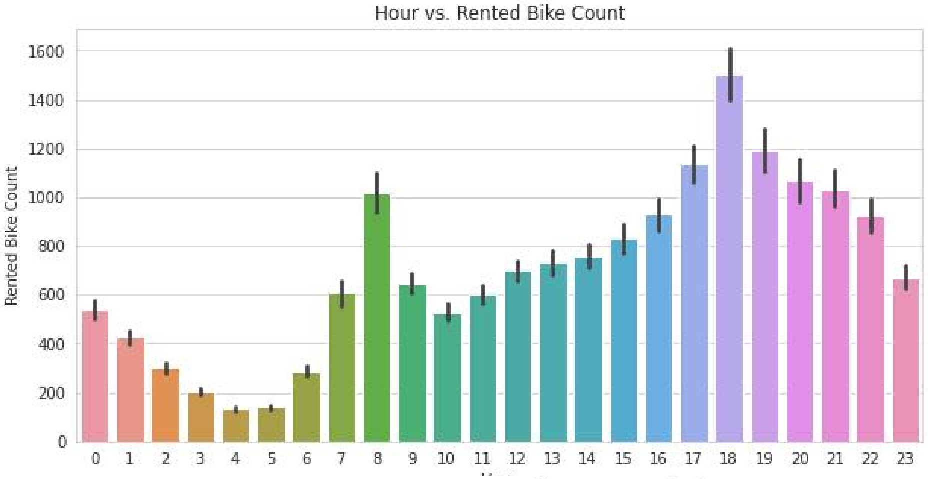
About Dataset

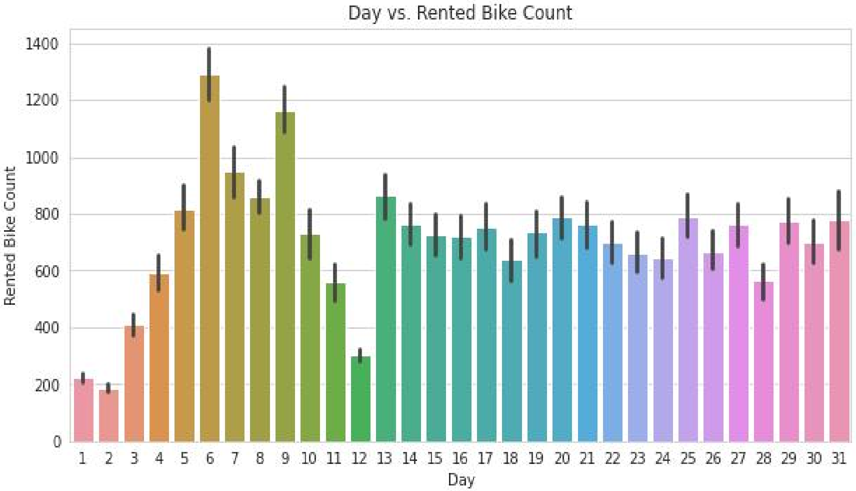
- This dataset contains 14 columns and 8760 rows

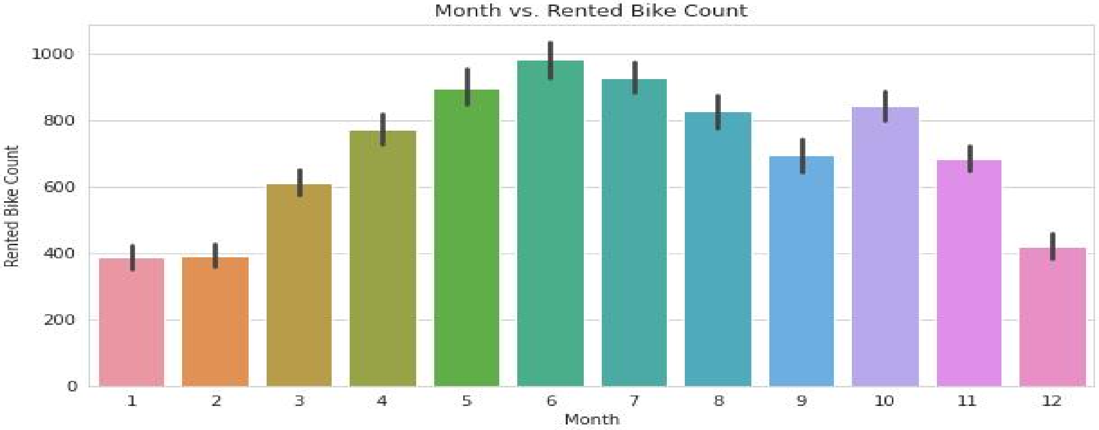


	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.00	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.00	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.00	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.00	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.00	0.0	0.0	Winter	No Holiday	Yes
5	01/12/2017	100	5	-6.4	37	1.5	2000	-18.7	0.00	0.0	0.0	Winter	No Holiday	Yes

EDA and Feature engineering


Hour vs. Rented Bike Count


Day vs. Rented Bike Count

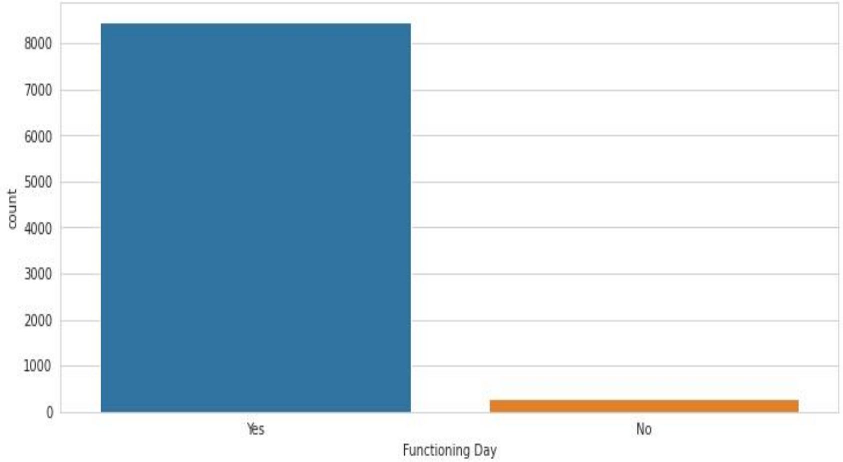

Month vs. Rented Bike Count

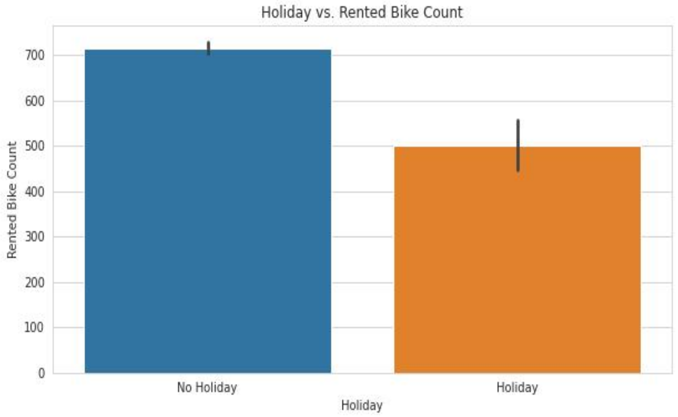
6 pm has maximum bike demand

first week of month has maximim bike demand.

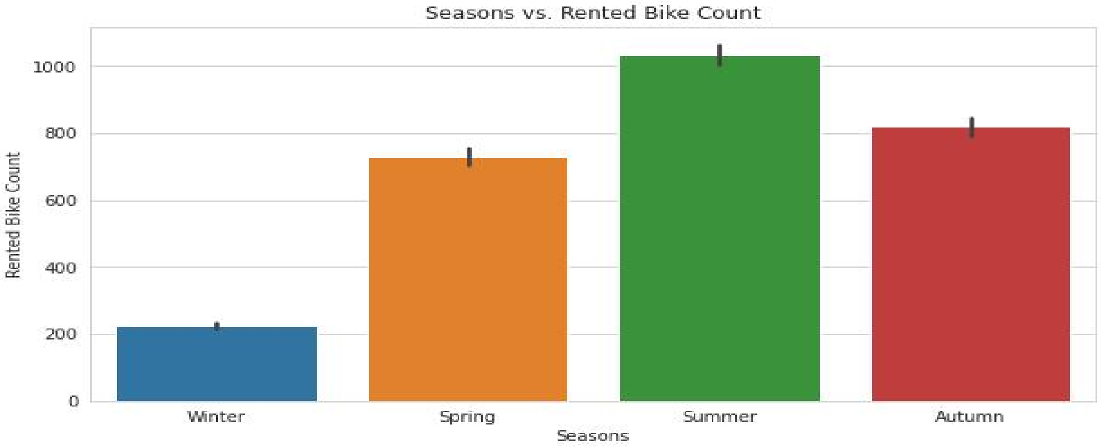
June has the highest bike rented in a year

EDA cont.

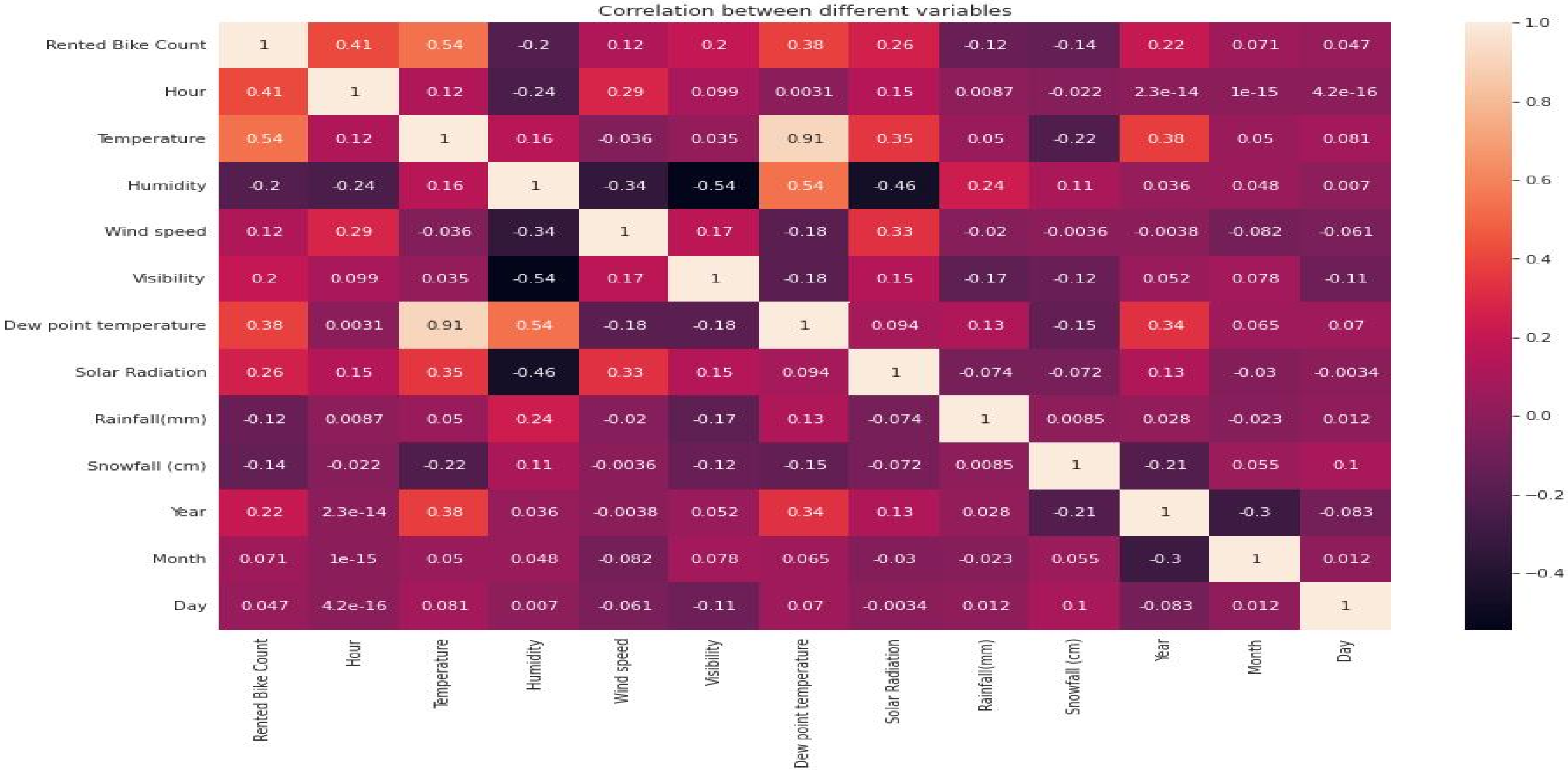




* functioning day has the highest
bike count when compared to
non functioning day
* when its not a holiday bike rent demand is highter
then compared to holiday
* off all the seasons summar has the highest bike
rented and winter has the lowest bike rent count



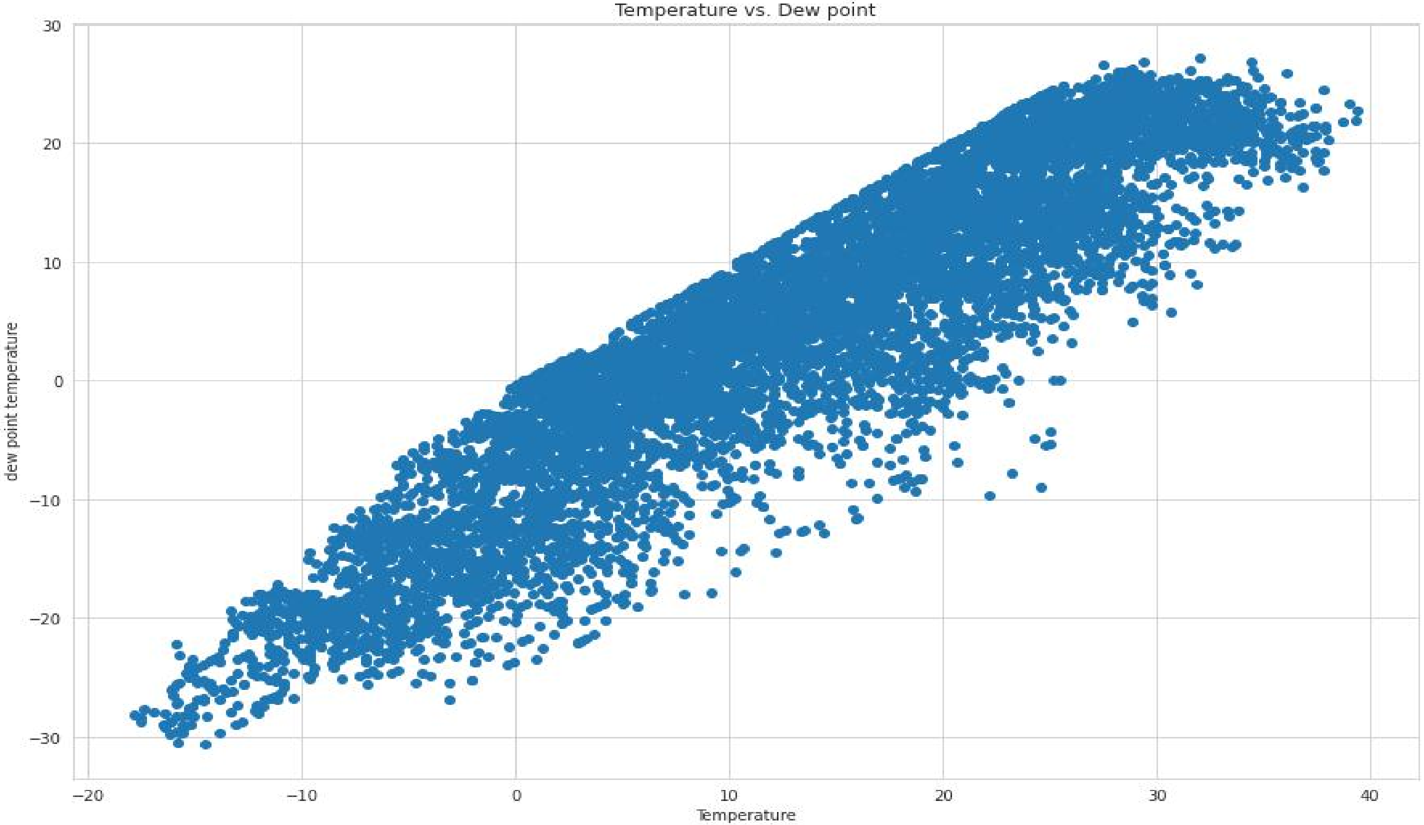
Correlation between different variables



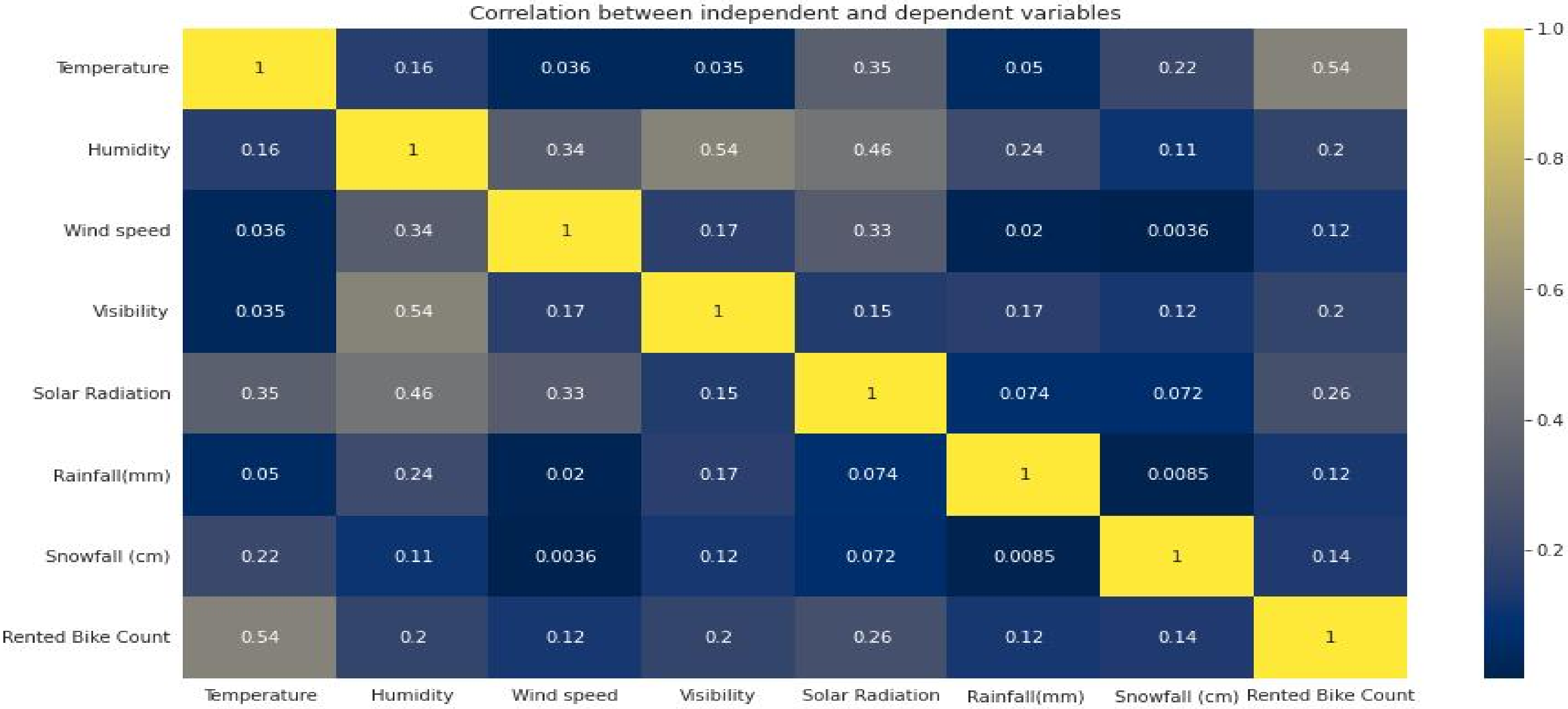
Correlation between different variables

	Rented Bike Count	Hour	Temperature	Humidity	Wind speed	Visibility	Dew point temperature	Solar Radiation	Rainfall(mm)	Snowfall (cm)	Year	Month	Day
Rented Bike Count	1	0.41	0.54	-0.2	0.12	0.2	0.38	0.26	-0.12	-0.14	0.22	0.071	0.047
Hour	0.41	1	0.12	-0.24	0.29	0.099	0.0031	0.15	0.0087	-0.022	2.3e-14	1e-15	4.2e-16
Temperature	0.54	0.12	1	0.16	-0.036	0.035	0.91	0.35	0.05	-0.22	0.38	0.05	0.081
Humidity	-0.2	-0.24	0.16	1	-0.34	-0.54	0.54	-0.46	0.24	0.11	0.036	0.048	0.007
Wind speed	0.12	0.29	-0.036	-0.34	1	0.17	-0.18	0.33	-0.02	-0.0036	-0.0038	-0.082	-0.061
Visibility	0.2	0.099	0.035	-0.54	0.17	1	-0.18	0.15	-0.17	-0.12	0.052	0.078	-0.11
Dew point temperature	0.38	0.0031	0.91	0.54	-0.18	-0.18	1	0.094	0.13	-0.15	0.34	0.065	0.07
Solar Radiation	0.26	0.15	0.35	-0.46	0.33	0.15	0.094	1	-0.074	-0.072	0.13	-0.03	-0.0034
Rainfall(mm)	-0.12	0.0087	0.05	0.24	-0.02	-0.17	0.13	-0.074	1	0.0085	0.028	-0.023	0.012
Snowfall (cm)	-0.14	-0.022	-0.22	0.11	-0.0036	-0.12	-0.15	-0.072	0.0085	1	-0.21	0.055	0.1
Year	0.22	2.3e-14	0.38	0.036	-0.0038	0.052	0.34	0.13	0.028	-0.21	1	-0.3	-0.083
Month	0.071	1e-15	0.05	0.048	-0.082	0.078	0.065	-0.03	-0.023	0.055	-0.3	1	0.012
Day	0.047	4.2e-16	0.081	0.007	-0.061	-0.11	0.07	-0.0034	0.012	0.1	-0.083	0.012	1

- From the above heatmap we can see that temperature and dew point temperature are highly correlated with each other
- We can also observe that as the temperature increases,the dew point temperature also increases



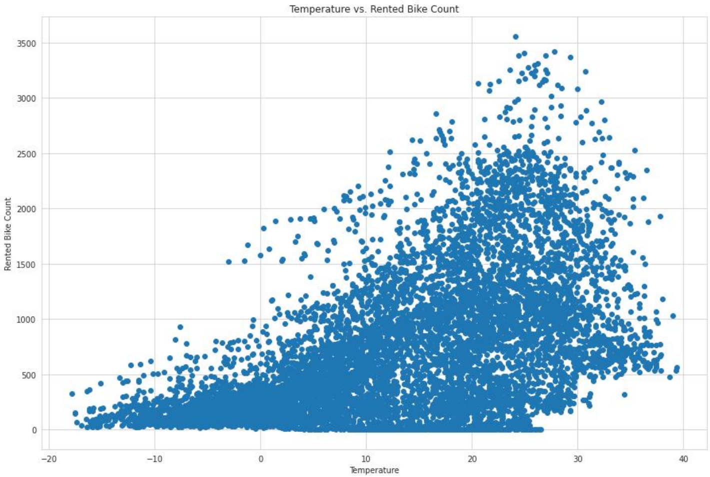
Correlation between only independent variables and dependent variable

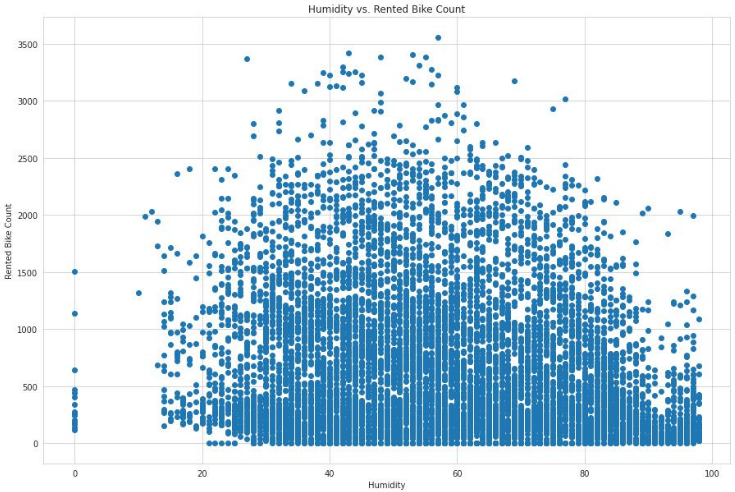


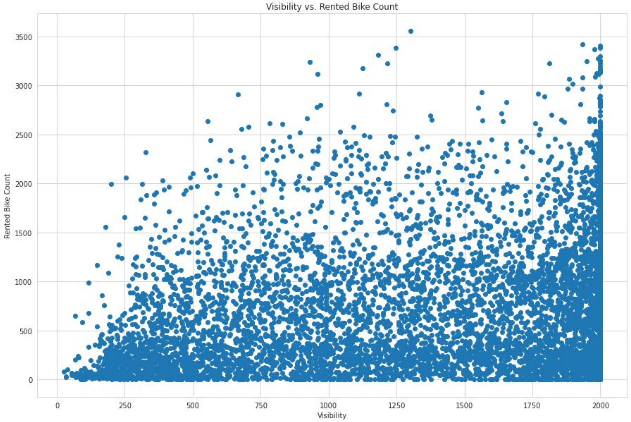
Correlation between independent and dependent variables

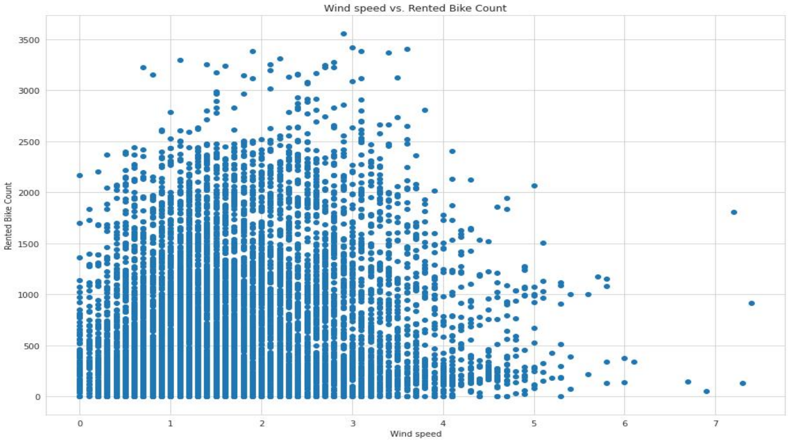
	Temperature	Humidity	Wind speed	Visibility	Solar Radiation	Rainfall(mm)	Snowfall (cm)	Rented Bike Count
Temperature	1	0.16	0.036	0.035	0.35	0.05	0.22	0.54
Humidity	0.16	1	0.34	0.54	0.46	0.24	0.11	0.2
Wind speed	0.036	0.34	1	0.17	0.33	0.02	0.0036	0.12
Visibility	0.035	0.54	0.17	1	0.15	0.17	0.12	0.2
Solar Radiation	0.35	0.46	0.33	0.15	1	0.074	0.072	0.26
Rainfall(mm)	0.05	0.24	0.02	0.17	0.074	1	0.0085	0.12
Snowfall (cm)	0.22	0.11	0.0036	0.12	0.072	0.0085	1	0.14
Rented Bike Count	0.54	0.2	0.12	0.2	0.26	0.12	0.14	1

Dependent variable vs Independent variables


Temperature vs. Rented Bike Count


Humidity vs. Rented Bike Count

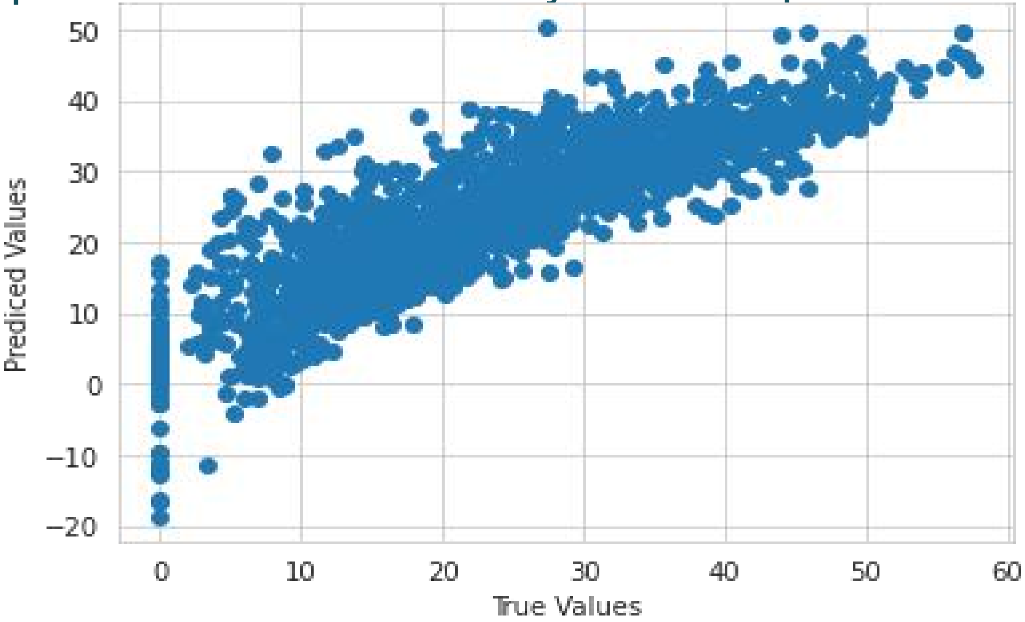

Visibility vs. Rented Bike Count


Wind speed vs. Rented Bike Count

Model Implementation

1. Linear Regression

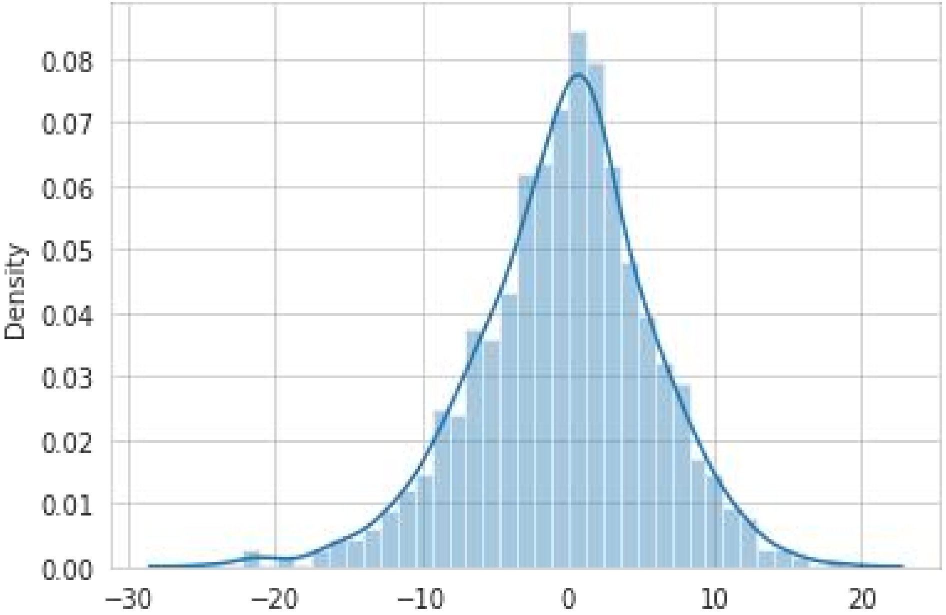
- Used linear regression model on the data and then we have got
- prediction model accuracy in train equals to 77.66% and for test data it is 77 also.



Linear Regression

The data was scaled to be used in the model implementation and fit the trained dataset to the model.

The regressor score got from the model is 0.536337.

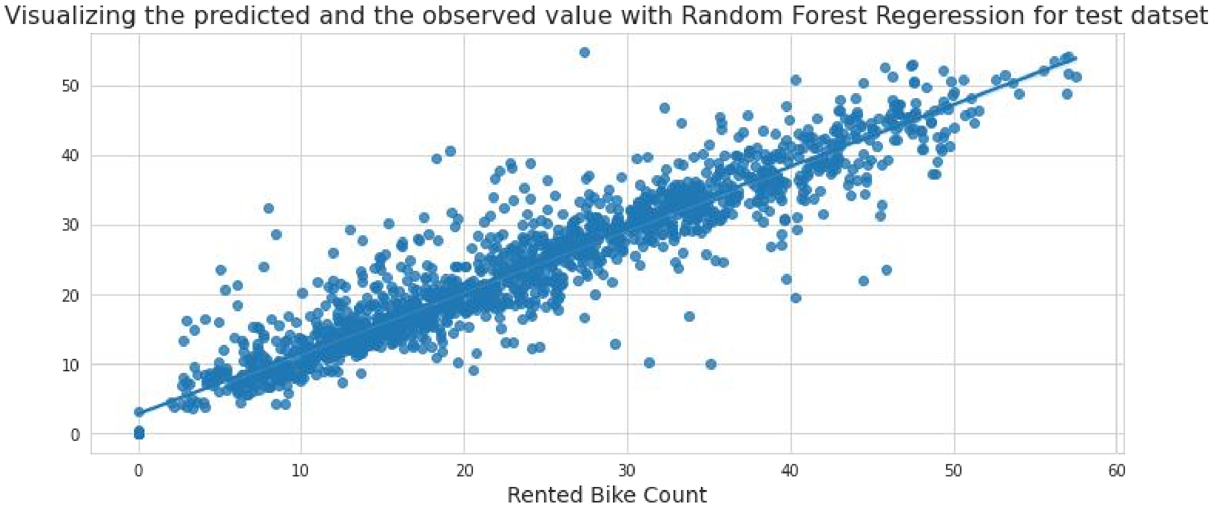


Residual = actual y value(test) − predicted y value(test)

2. Random Forest Regressor

After implementing the linear regression model with 77% accuracy,it was still not able to give satisfactory result on the training data and testing data, so we used the random forest algorithm to improve our performance.

After implementing the random forest algorithm, we are able to improve the model prediction now the accuracy after implementing random forest regressor is 98.5 % on training data and 88.66 % on testing data.



Hyperparameter tuning using Grid Search CV on Random Forest Regressor

With random forest, we were able to achieve 98.5 % accuracy on our training data and 88.66 % accuracy on our testing data, resulting in a gap of 9.84% between our training and testing accuracy

As a result of these numbers, we can say our model is overfitting

The algorithm may overfit if we use this model on unknown data to predict the number of rental bikes required at each hour to maintain a stable supply

As a result, we must overcome this challenge, and we are doing so by implementing Hyperparameter tuning using Grid Search CV

We got the same accuracy on our training data and testing data after implementing Hyperparameter tuning using Grid Search CV on Random Forest Regressor.

Conclusions

* The rented bike demand is good when the weather is clear and good in all the seasons.
* And in the seasons the demand is somewhat high during summer.
* More demand at 8 in the morning and 6 in the evening, which seems to be the working hours of office.
*After 6 in the evening there is slightly more demand in the bike rentals.
* As per the model evaluation it is better to implement the Random Forest Regression rather that going for Linear Regression.
* When it comes to the accuracy the Random Forest Regression is performing well on the test dataset with the accuracy of 98.5 %

Thank You