Capstone Project

Seoul Bike Sharing Demand Prediction

Shubham Kumar

Points of Discussion:

- Defining Problem Statement
- Introduction
- Data Summary
- Exploratory Data Analysis
- Model Building
- Evaluation
- Observations
- Conclusion

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 lessens 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.

Introduction and Business use of the project:

Bike sharing systems are new generation of traditional bike rental where process from membership, rentals and returns has become automatic. Users can check their travel details (distance, duration) and measure their physical activities (calories burned). Due to such smart technology and convenience along with the increased travel, the use of rental bikes is increasing day by day. Therefore, you need to able to manage the demand for rental bikes and manage the continuous and convenient service to users effectively. This study proposes a machine learning-based approach that includes different metrics to predict the rental bikes demand across the city. The ML model is used to predict the number of rental bikes required per hour. Rental bikes demand is modeled using the available independent variables. The management can use this to understand exactly how the demand vary with different features. So, they can manipulate business strategies to meet the demand levels and customer expectations. In addition, this model is a great way for management to understand the demand dynamics of a new market.

Data Summary

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature- in Celsius
- Humidity - percentage of humidity
- Wind Speed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

Data Pipeline

- **Data Processing:** In the first part, we have imported necessary libraries and data set. We then used these libraries to understand the data.
- **Data Cleaning:** After understanding the data, we got to know that there are no null values or duplicate values in our data set.
- **Data Preparation:** For the EDA, We refactored the datetime feature. We can't analyse non numerical values. We transformed it ("yyyy/mm/dd") into date, hours, day of year and year.
- **Exploratory Data Analysis:** After preparing the data set, we did some exploratory data analysis using tables and graphs to derive the observations from the data and to better understand the problem statement, and make ways to the solution to the problem statement.

Refactoring datetime

Datetime is a string. This is a problem because strings cannot be processed mathematically. We transformed the string into a date and then extracted the features hour, Day of the year, week day and year. Extraction of feature year, is what improved performance the most.

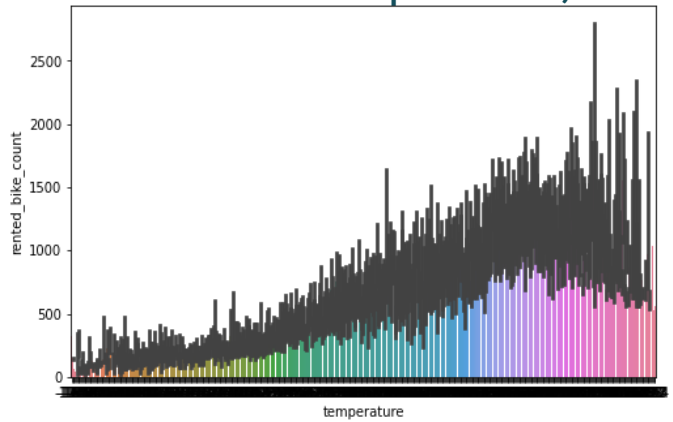
The hour feature was a particularly challenging problem. It can be considered a categorical feature on its own. There are few rentals at 1, 2, 3 am and at 10 and 11 pm. The relationship between the hour and the number of rentals is not linear.

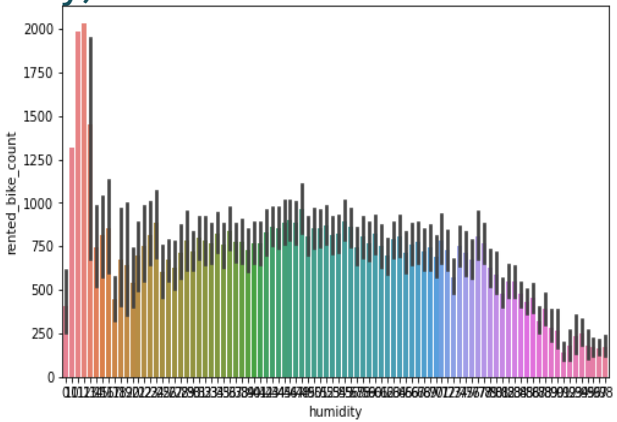
Exploratory Data Analysis

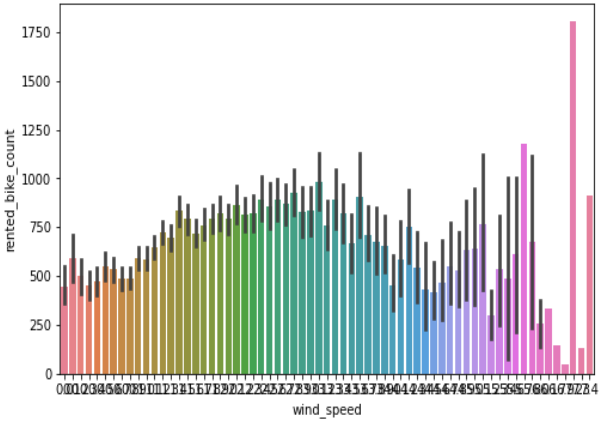
While doing the Exploratory Data analysis we tried finding the factors affecting the rental bikes demand. The factors affecting the Rental Bikes Demand are:

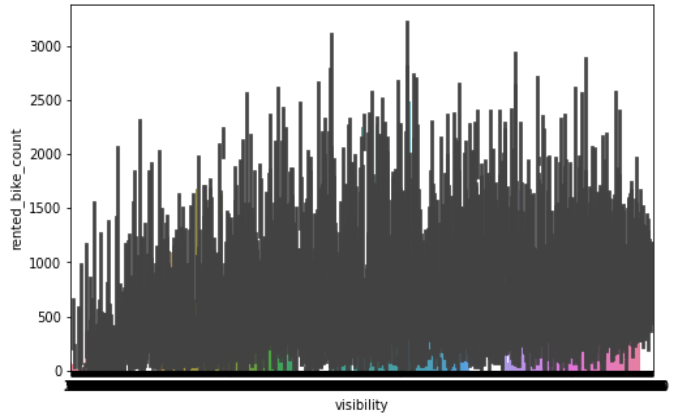
- Temperature
- Humidity
- Wind Speed
- Visibility
- Rainfall
- Snowfall

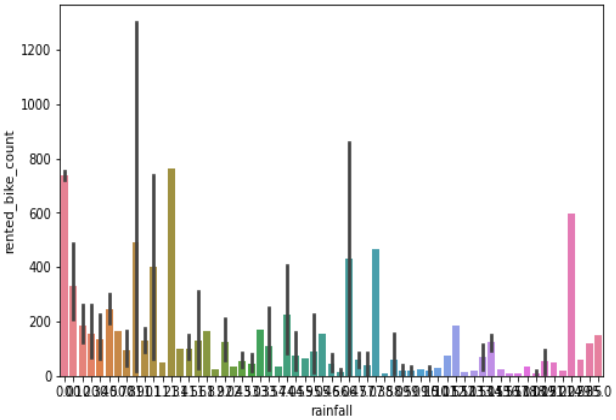
From the below graphs, it is evident that the major factor affecting the Rental bikes demand are Temperature, Visibility, Rainfall and Snowfall.

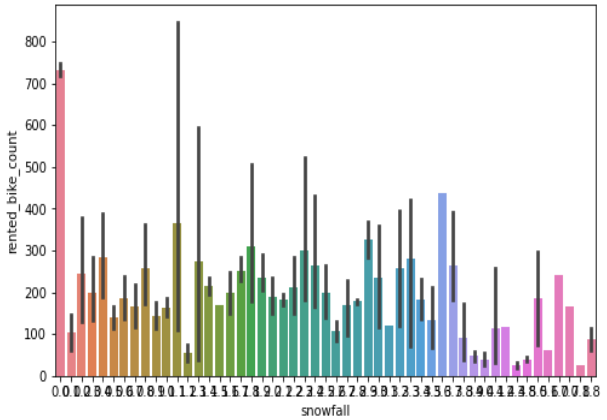








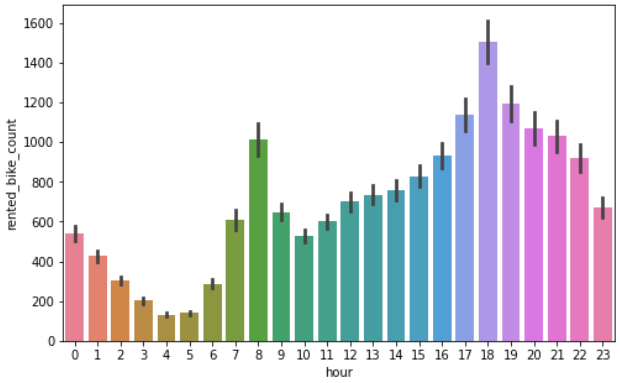


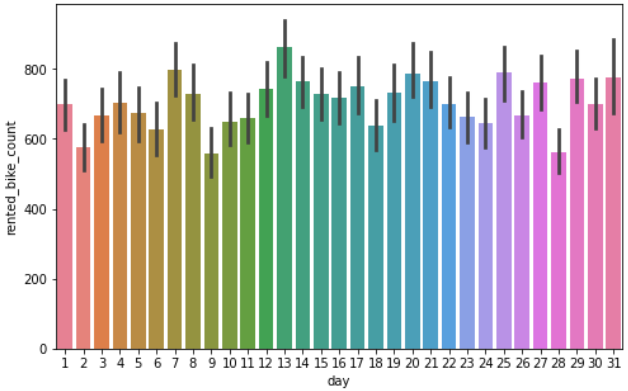


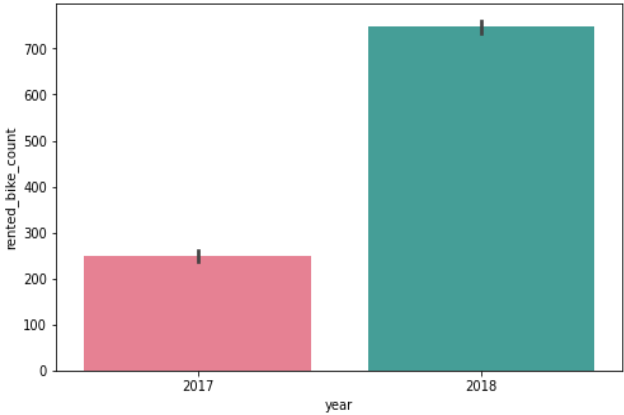
Date wise Analysis

While doing Date-wise analysis of the given rental bike demand dataset, we answered the following questions:

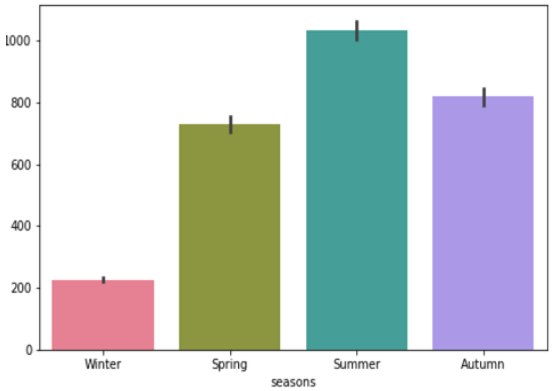
- At what time of the day the Rental bike demand is the highest?
- On which day the Rental bike demand is the highest?
- On which date the Rental bike demand is the highest?
- Which Year shows the most demand for Rental Bikes?

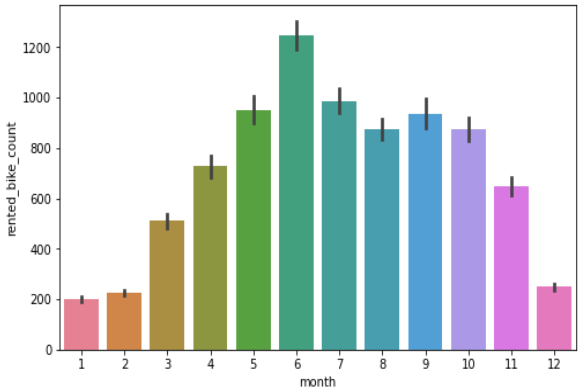






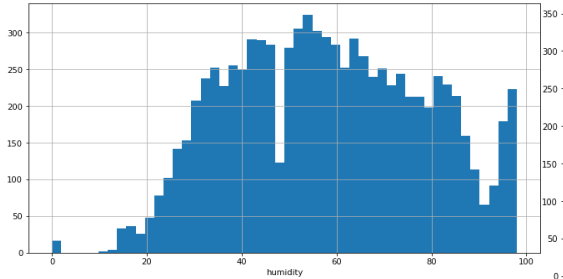
- From the above graphs, it is evident that:
- The rental bike is in most demand in the morning from 7 AM to 9 AM and in the evening from 5 PM to 8 PM.
- The Rental Booking happens the most in summers specifically from May to July.
- The Rental Bike demand is showing increasing trend.

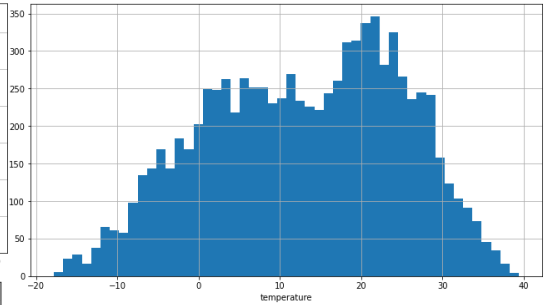


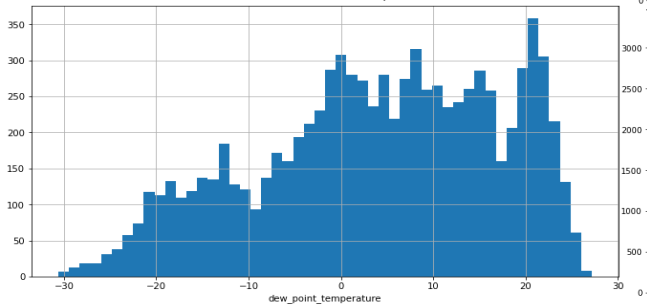


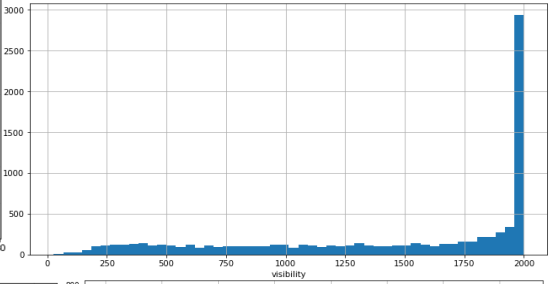
Distribution Of Features

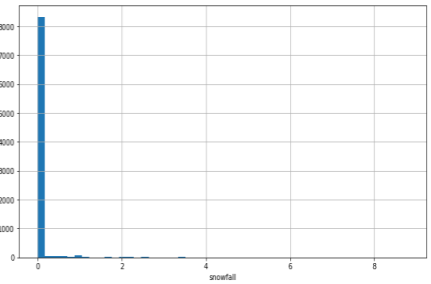
- ‘Temperature’ and ‘Humidity’ columns follows Uniform distribution.
- ‘Dew Point Temperature’ and ‘Visibility’ are Negatively skewed.
- ‘Wind Speed’ , ‘Solar Radiation’ , ‘Rainfall’ and ‘Snowfall’ are having Positively skewed distribution.

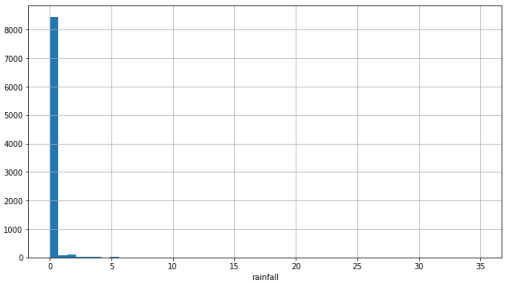


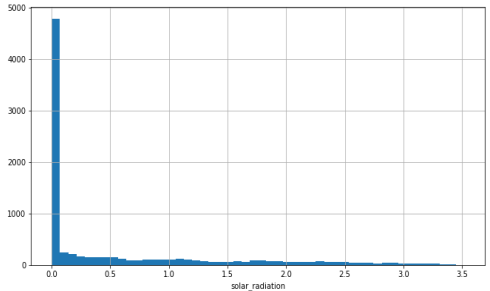


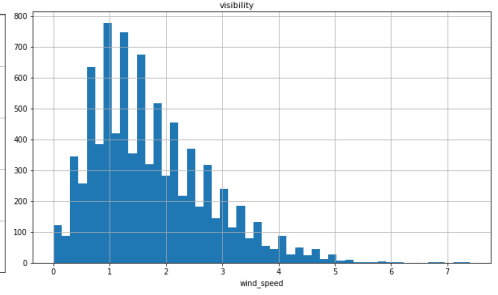


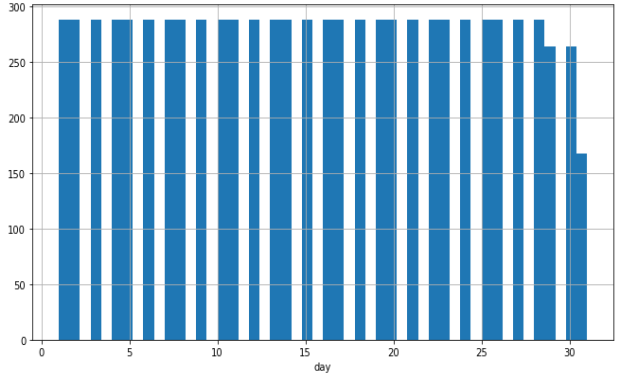


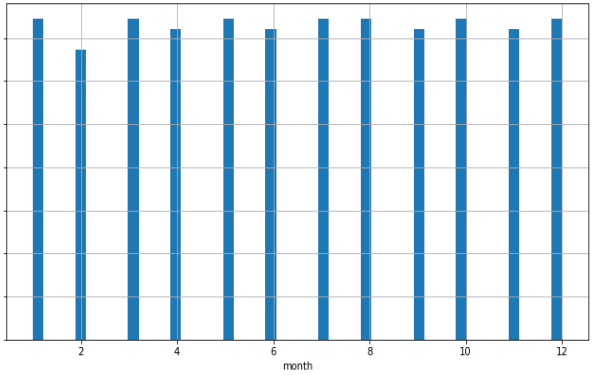


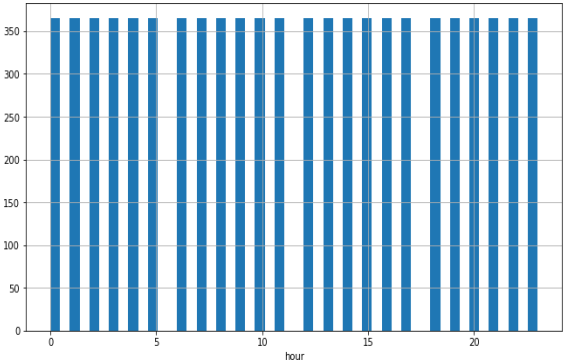


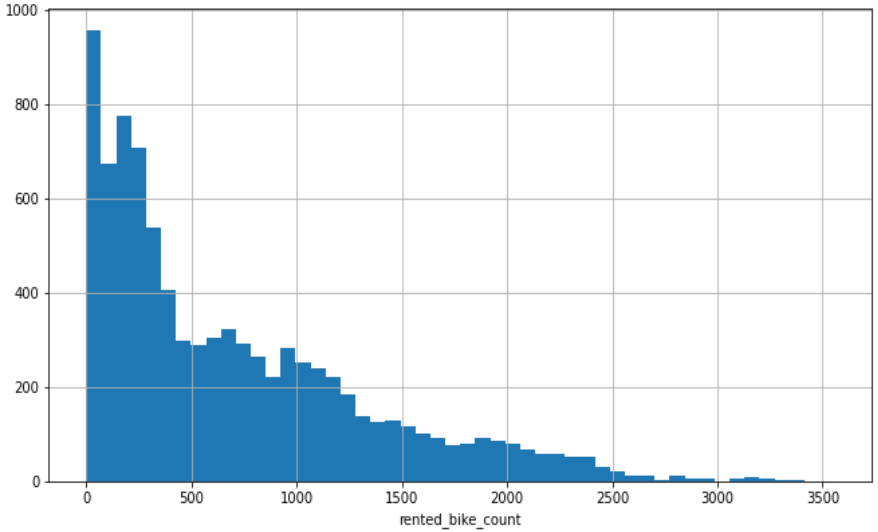





day

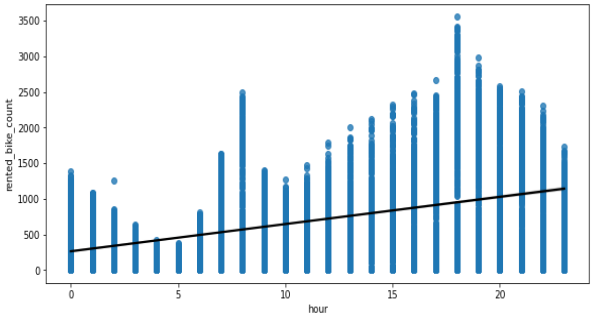

month

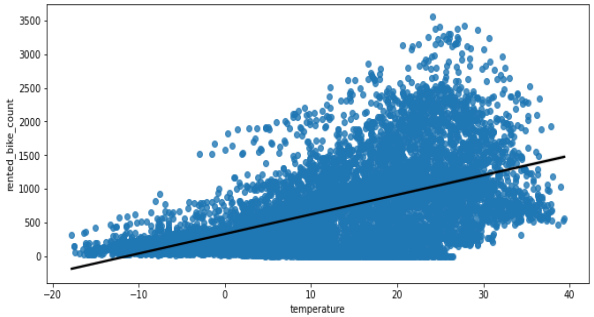

hour

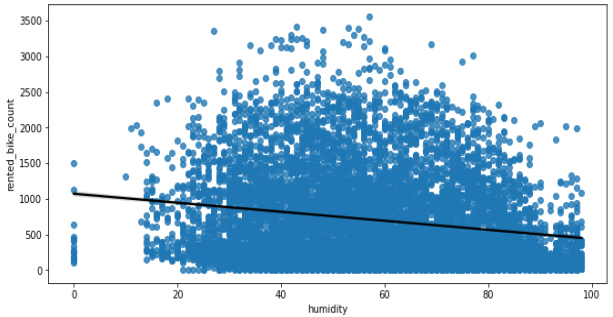

rented_bike_count

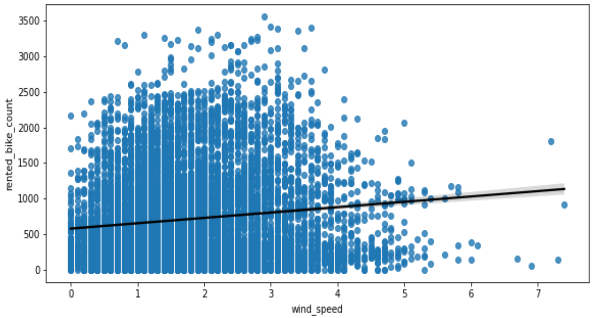

year

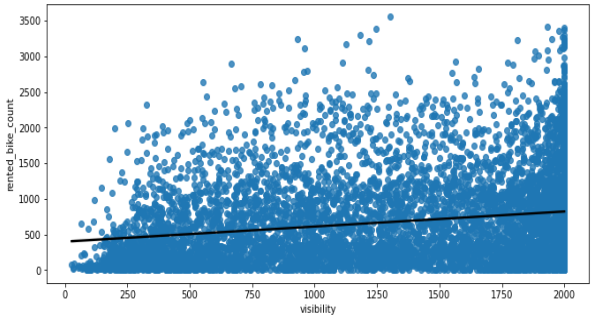
Correlation

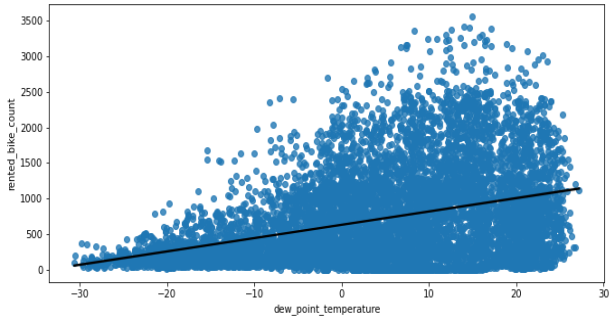




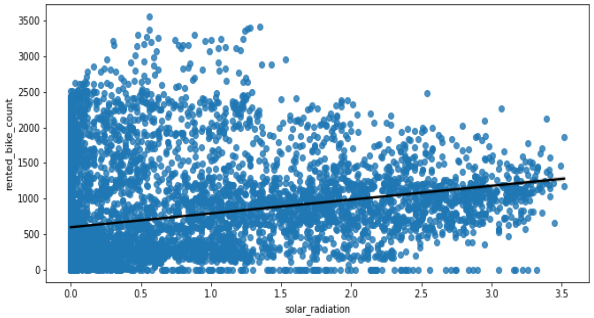


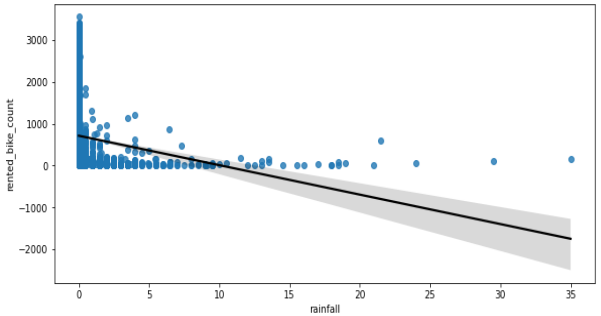


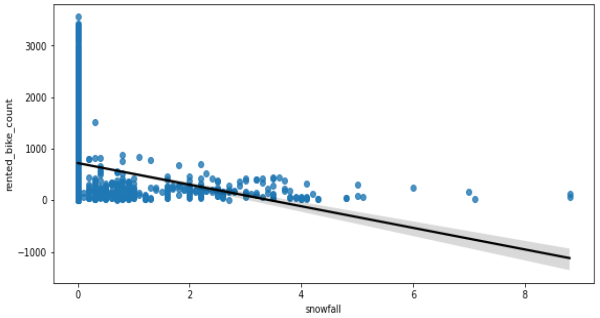


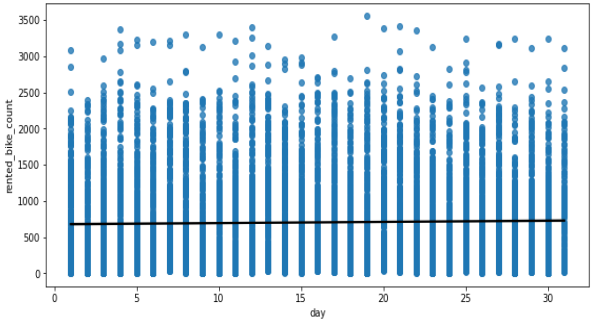


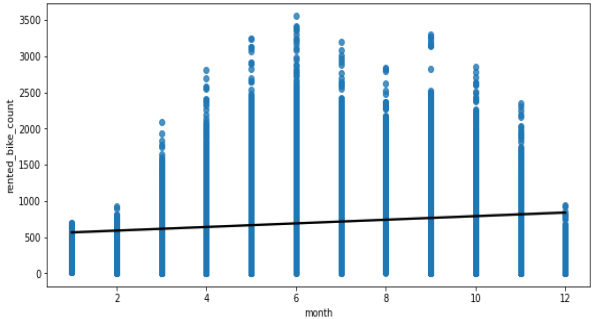
Correlation(continued)

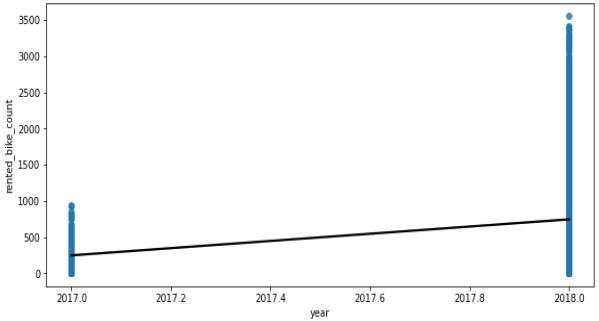




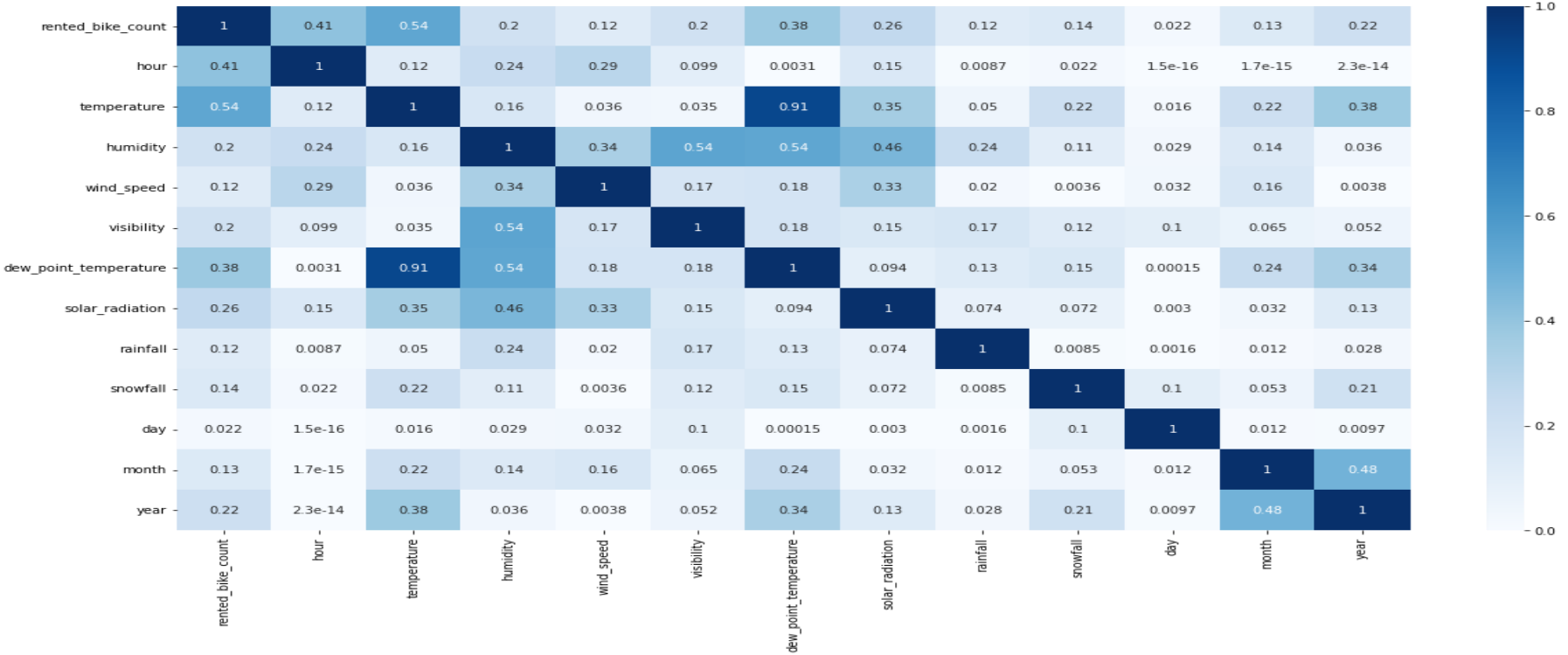








Multicollinearity

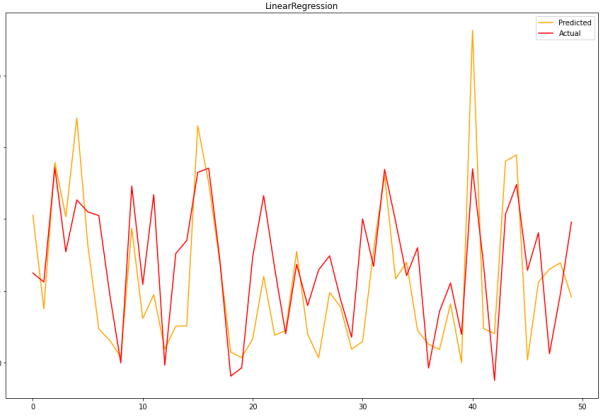


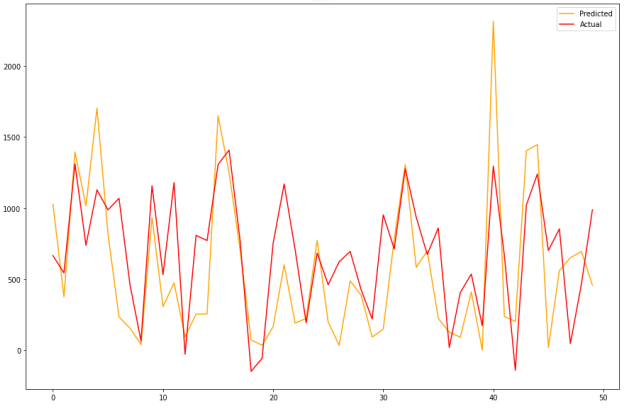
From the above graph, we can see that Temperature and Dew_point_temperature are highly correlated, keeping the factor of 0.91 . And, then we have hour in the graph which is having high correlation with our dependent variable.

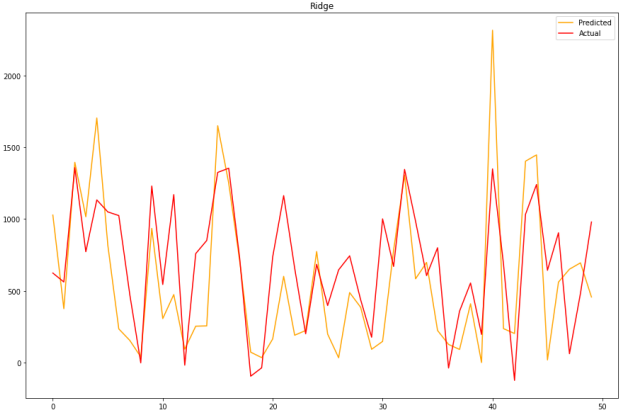
Data Preprocessing

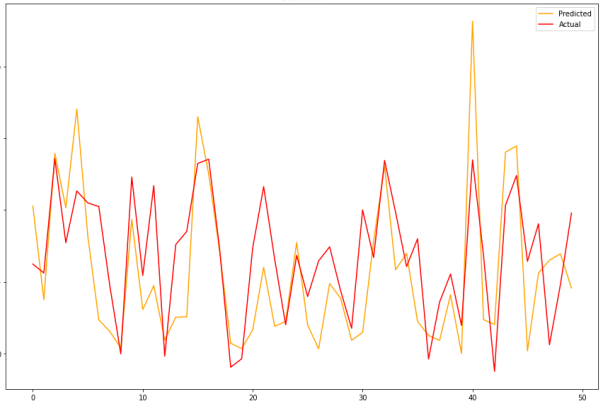
- Implemented the Principal Component Analysis for ‘Dew Point Temp’ and ‘Temperature’ as they were highly correlated.
- Removed observations where it was ‘Nonfunctional Day’ and bike rent was zero and removed the column as well.
- We have also dropped ‘Date’ column as it would not help in giving good prediction for model.

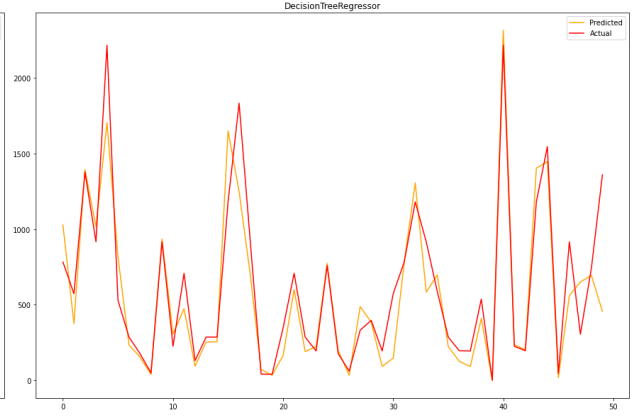
Machine Learning - Supervised Learning - Regression





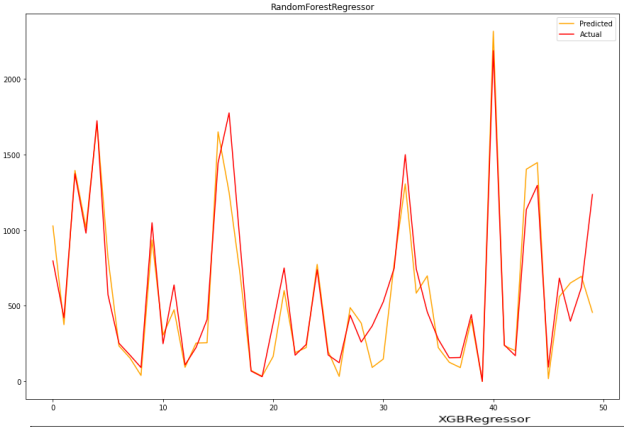


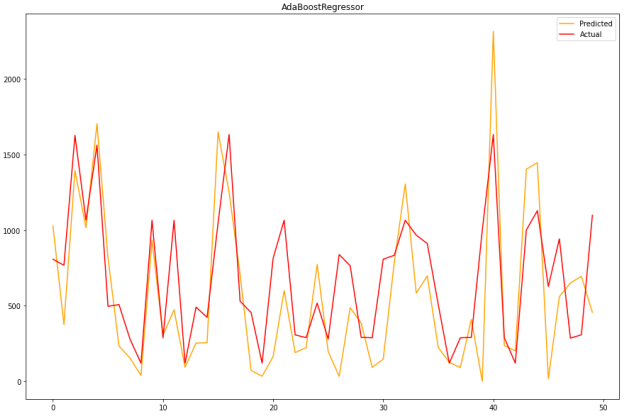


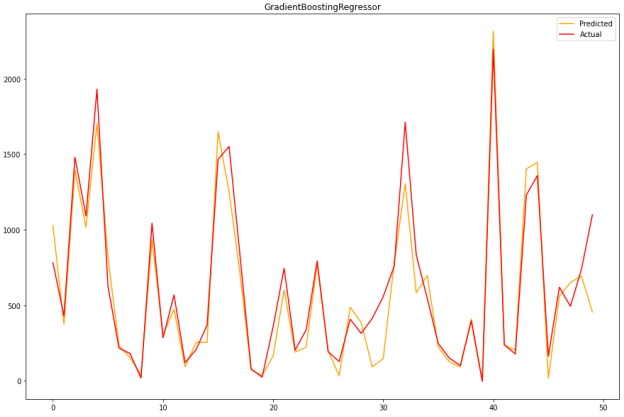


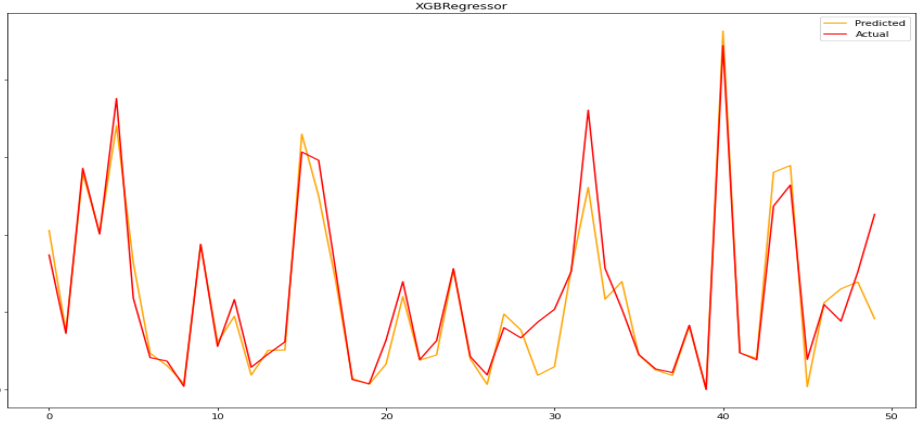
We have used different ML Models to determine the prediction of Rental Bikes needed per hours. The Linear regression model is giving an accuracy of 78%. The Lasso regression is giving the accuracy of 38%. The Ridge Model is giving the accuracy of 77%. Where as Elastic Net is giving accuracy of 62% and Decision Tree is giving accuracy of 63%.

Machine Learning - Supervised Learning - Regression(continued)









Here, the Random forest Regressor and Gradient Boosting gridsearchcv gives the highest R2 score of 99% and 95% respectively for Train Set and 92% for Test set.

Observations Summary

- The major factor affecting the Rental bikes demand are Temperature, Visibility, Rainfall and Snowfall. The rentals are independent of the wind speed and the humidity, because they are almost constant over the months.
- The rental bike is in most demand in the morning from 7 AM to 9 AM and in the evening from 5 PM to 8 PM. Hence, the Rental Bike companies need better planning for the availability of the rental bikes for this time period the most.
- The Rental Booking happens the most in summers specifically from May to July. Hence, the rental companies needs to plan the staffings required for the maintenance of bikes and smooth running of the business for this period the most.
- The Rental Bike demand is showing increasing trend. Hence, it is a booming business and there is a great scope of expansion.
- We implemented 7 machine learning algorithms those are: Linear Regression, Lasso Regression, Ridge Regression, Elastic Net, Decision Tree, Random Forest and XGBoost.
- We did hyperparameter tuning to improve our model performance.
- Random forest Regressor and Gradient Boosting gridsearchcv gives the highest R2 score of 99% and 95% respectively for Train Set and 92% for Test set.
- No overfitting is seen.
- Feature Importance value for Random Forest and Gradient Boost are different.

Conclusion

We have used many ML models such as Linear Regression, Lasso Regression, Ridge Regression, Elastic Net, Decision Tree, Random Forest and XGBoost. We have chosen Random forest Regressor and Gradient Boosting gridsearchcv as this gives the highest R2 score of 99% and 95% respectively for Train Set and 92% for Test set.

As we can see the total amount of bike rentals increases with the temperature per month. Whereas it seems that the rentals are independent of the wind speed and the humidity, because they are almost constant over the months. This also confirms on the one hand the high correlation between rentals and temperature and on the other hand that nice weather could be a good predictor. So people mainly rent bikes on nice days and nice temperature. This could be important of planning new bike rental stations.