CAPSTONE PROJECT. 2

SUPERVISED ML - REGRESSION

SEOUL BIKE SHARING DEMAND PREDICTION

Presented By

SANJU KHANRA

Data Science Trainee, Alma Better

WHY ANALYZE THE BIKE SHARING DEMAND PREDICTION ?

- **The global bicycle market is projected to grow from $82.50 billion in 2022 to $127.83 billion by 2029 at a CAGR of 6.5% in forecast period, 2022-2029**
- **Their central concept is to provide free or affordable access to bicycles for short-distance trips in an urban area as an alternative to private vehicles, thereby reducing congestion, noise, and air pollution.**
- **bike sharing system in Shanghai saved 8,358 tones of petrol and decreased Carbon dioxide and NOx emissions by 25,240 and 64 tones, respectively.**

CONTENTS

1. INTRODUCTION

- A **bicycle-sharing system**, **bike share program**, **public bicycle scheme**, or **public bike share (PBS) scheme**, is a shared transport service where bicycles are available for shared use by individuals at low cost.
- The program themselves include both docking and dock less systems, where docking systems allow users to rent a bike from a dock, i.e., a technology-enabled bicycle rack and return at another node or dock within the system — and dock less systems, which offer a node-free system relying on smart technology
- systems may incorporate smartphone web mapping to locate available bikes and docks. In July 2020, Google Maps began including bike share systems in its route recommendations.
- With its antecedents in grassroots mid-1960s efforts; by 2022, approximately 3,000 cities worldwide offer bike-sharing systems available.

2A. Problem Statement

- ***EDA Part*** :

1. Relation between 'Month' and 'Rented-bike -count'
2. Count of Rented bikes according to Weekdays-Weekend and Time
3. Relation between Rented Bikes count and Hours
4. Count of Rented bikes according to *Seasons*
5. Analyze of Numerical variables
6. Numerical of Rented-Bike-Count
7. percentage distribution of the value counts of the categorical features
8. Correlation Heat-map
9. Pair Plot

- ***Feature Engineering & Data Pre-processing Part :***

1 . Regression plot
2 . Handling Missing Values
3. Checking of Correlation between variables
4 . Handling Outliers.
5 . Data Scaling
6 . Categorical Encoding
7 . Data Splitting

2B. Problem Statement

- ***ML Model Implementation Part :***
  - ML Model-1. LINEAR REGRESSION
  - *ML Model-2. LASSO REGRESSION*
  - ML Model - 3. RIDGE REGRESSION
  - ML Model - 4. RANDOM FOREST REGRESSION
  - ML Model-5.Gtadient Boosting

- **Hyper-parameter tuning :**
  - Gradient Boosting with Grid-Search-CV

3.Data Summary

- *Bike sharing has been gaining importance over the last few decades. More and more people are turning to healthier and more live able cities where activities like bike sharing are easily available. there are many benefits from bike sharing, such as environmental benefits. It was a green way to travel*

- *The dataset contains weather information (Temperature, Humidity, Wind-speed, Visibility, Dew-point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.*

- *This dataset contains the hourly and daily count of rental bikes between years 2017 and 2018 in Capital bike share system with the corresponding weather and seasonal information. The dataset contains 8760 rows (every hour of each day for 2017 and 2018) and 14 columns (the features which are under consideration).*

4. INITIAL PREPARATION:

- In this section I've loaded in the dependencies, like Numpy , Pandas ,Matplotlib ,seaborn, and many more from the scikit learn library.
- The next step was to mount the drive where the data was stored.
- After mounting the drive I used the pandas.read_csv() function to read the data given to us in csv format.

5. DATA DESCRIPTION

- **Dependent variable:**

• Rented Bike count - Count of bikes rented at each hour

- **Independent variables :**
  - • Date : year-month-day
  - • Hour - Hour of he day
  - • Temperature-Temperature in Celsius
  - • Humidity - %
  - • Wind speed - m/s
  - • Visibility - 10 m
  - • Dew point temperature - Celsius
  - • Solar radiation - MJ/m2
  - • Rainfall - mm
  - • Snowfall - cm
  - • Seasons - Winter, Spring, Summer, Autumn
  - • Holiday - Holiday/No holiday
  - • Functional Day – No Function (Non Functional Hours), Fun(Functional hours)

6. CLEANING THE DATA

- **Handling Null values: In this project the first step in Data cleaning is Handling the null values. Null values can affect the accuracy and quality of our ML models, therefore it is a good practice to handle null values. In our case, fortunately there are no null values in our dataset, so we are good to go.**

- **Handling duplicate values: Duplicate values can have adverse effects on our ML models, therefore we have to try and remove it. Luckily we don't have any duplicate values either so we can move on to the next step in data cleaning.**

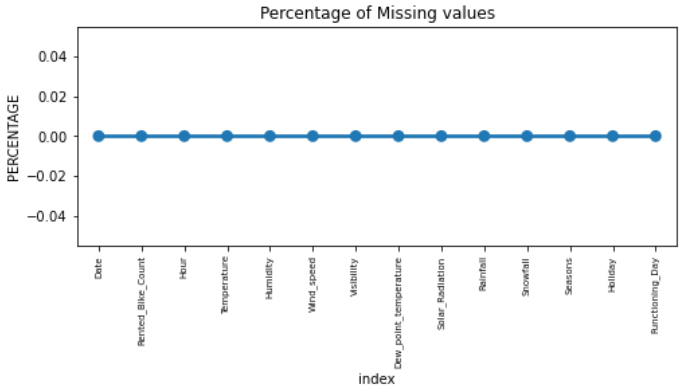
7. CLEAN UP(CONTD):

- ❖ *Removing Outliers: First of all we find the variables that may contain outliers, to detect this I've used the box plot offered by sea born library.*

As we can see here that out of the possible columns the columns that contain outliers are Rainfall, Snowfall, Solar Radiation and Wind speed. Now we will operate on these columns and try to remove the outliers from them using the IQR method.

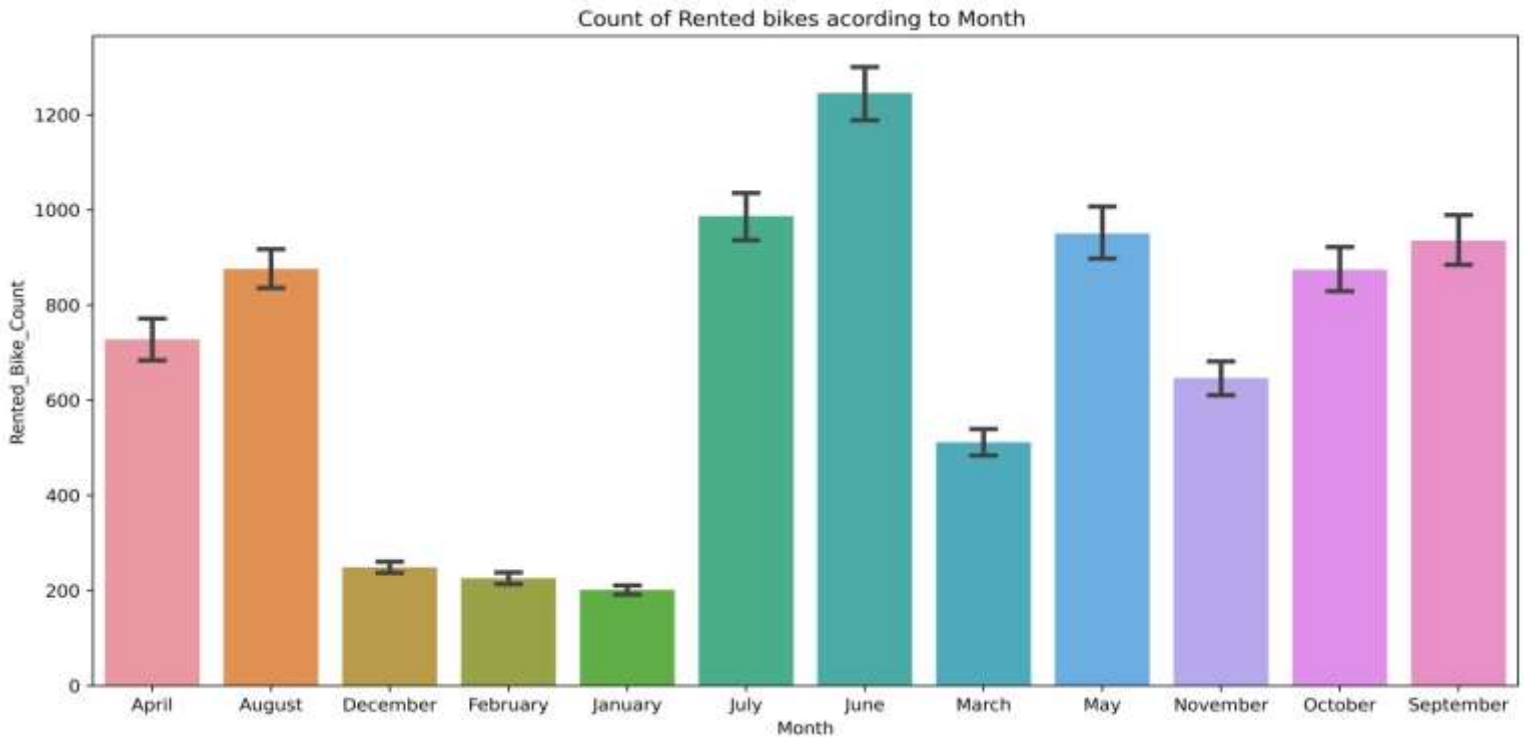
- Percentage of Missing values :

Using point plot we can see that 0-(zero) Missing value present



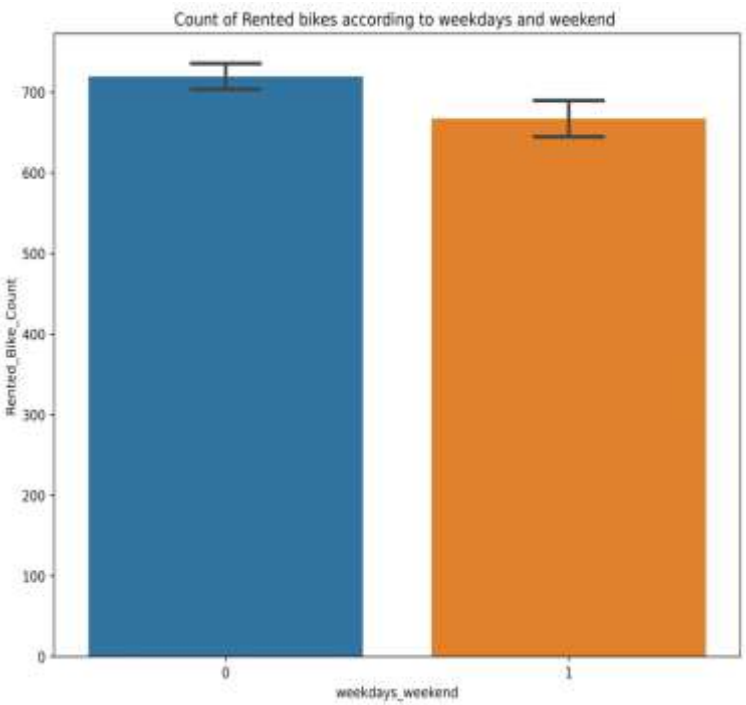
1. Relation between 'Month' and 'Rented-bike -count'(EDA)

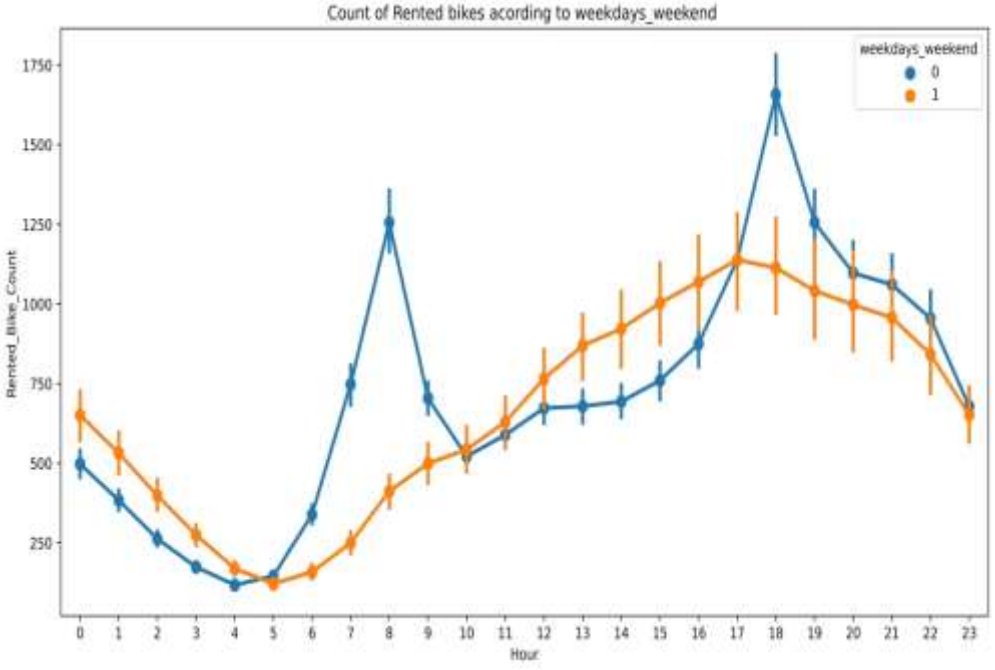
- From the above bar plot we can clearly say that from the month June, JULY, May to October or September the demand of the rented bike is high as compare to other months.
- In the summer season the rented bike business will go high as compare to winter season.



2. *Count of* Rented bikes according to Weekdays-Weekend and Time*(EDA)*

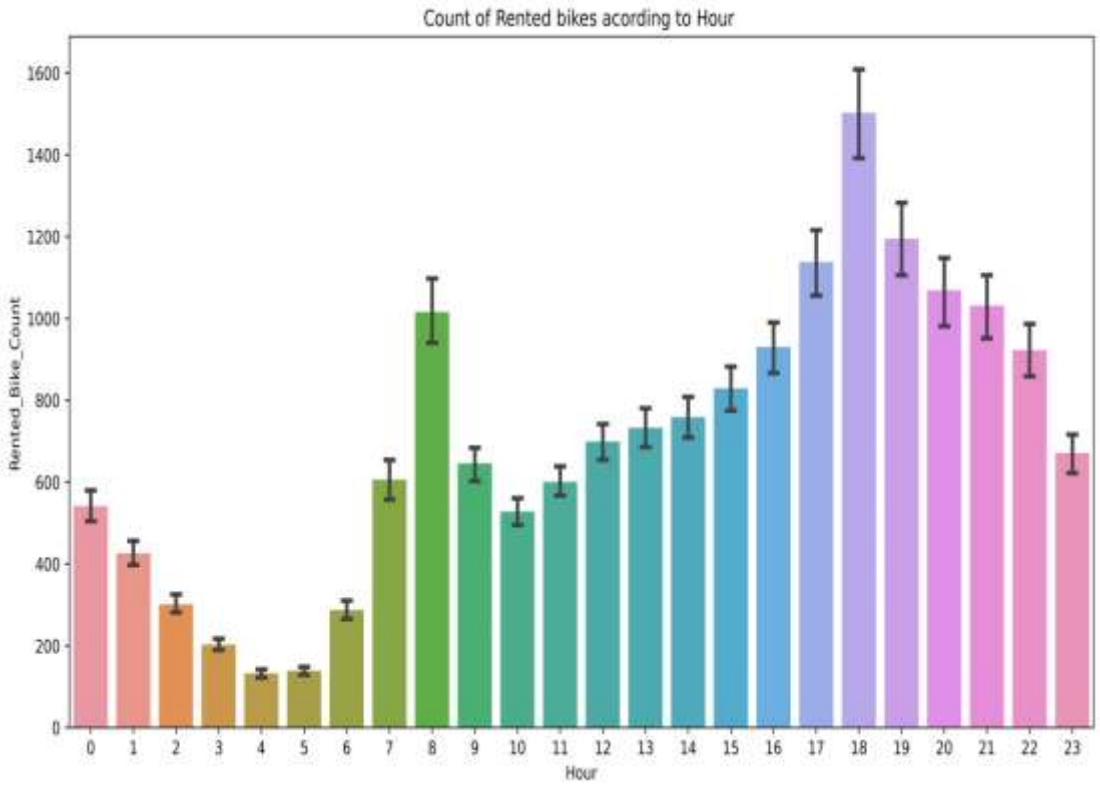
- From the above bar plot we can clearly say that the mean distribution of rented bike between Weekdays and weekend is almost same. But in weekdays its slightly higher due to office and weekend its slightly lower.
- From the above point plot and bar plot we can say that in the week days which represent in blue color show that the demand of the bike higher because of the office. Peak Time are 7 am to 9 am and 5 pm to 7 pm
- The orange color represent the weekend days, and it show that the demand of rented bikes are very low specially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases.





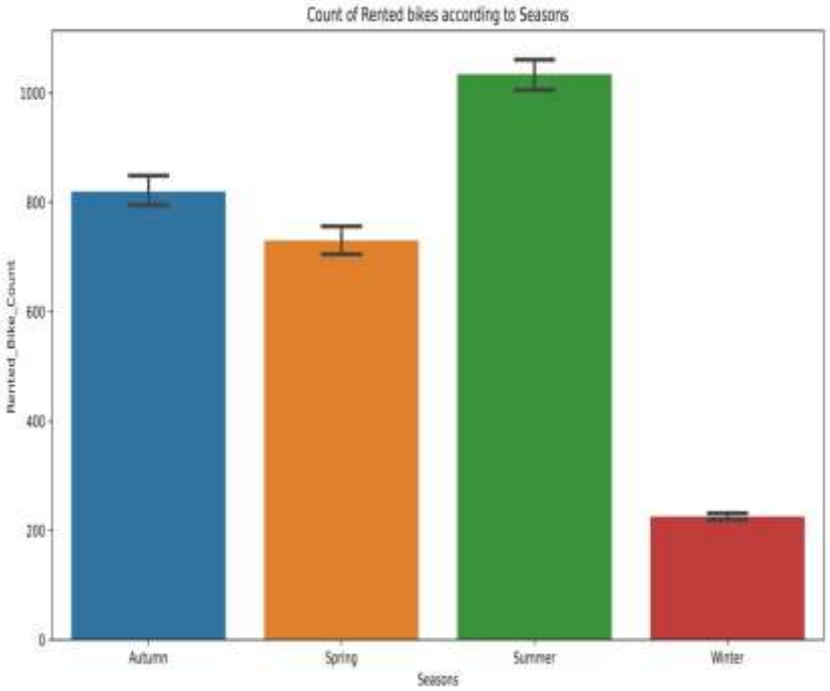
3. Relation between Rented Bikes count and Hours(EDA)

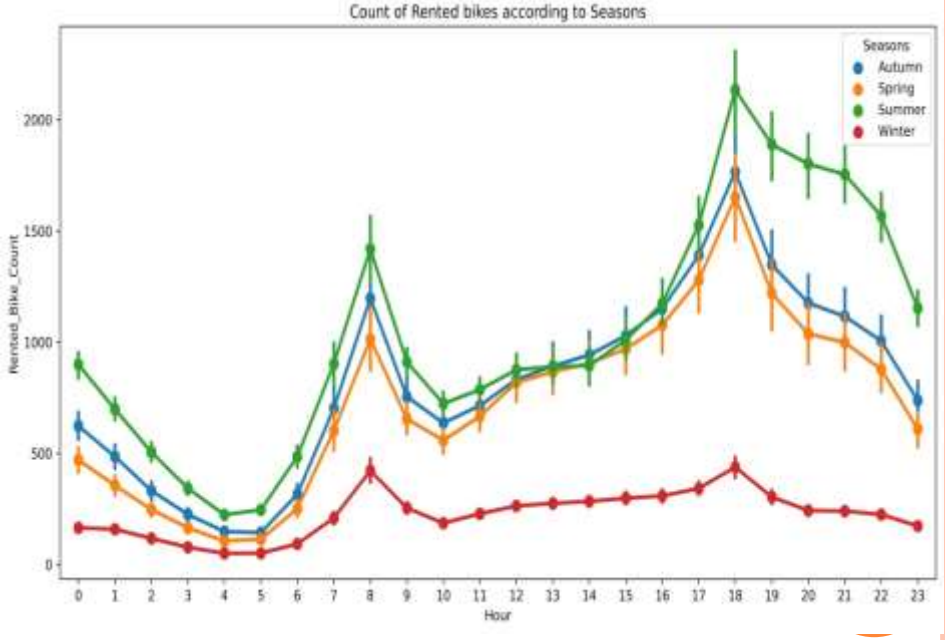
- From the above graph which show that maximum demand of rented bikes comes at the time working hour from 7am to 9am and 5pm to 7pm and minimum demand of rented bikes comes in the morning .



4. *Count of Rented bikes according to Seasons(EDA)*

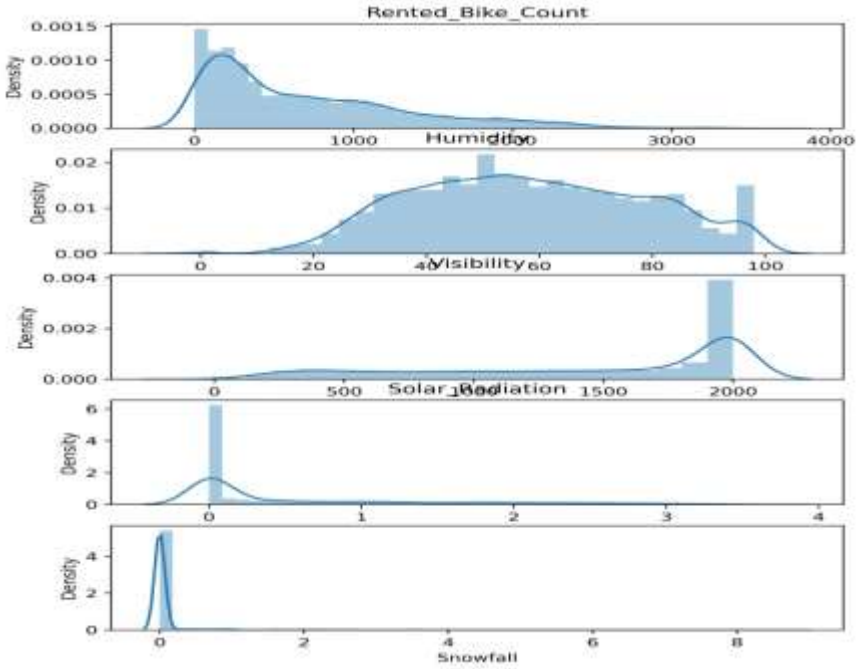
- In the above bar plot and point plot which shows the use of rented bike in in four different seasons, and it clearly shows that,
- In summer season the use of rented bike is high and peak time is 7am-9am and 7pm-5pm.
- In winter season the use of rented bike is very low because of snowfall.

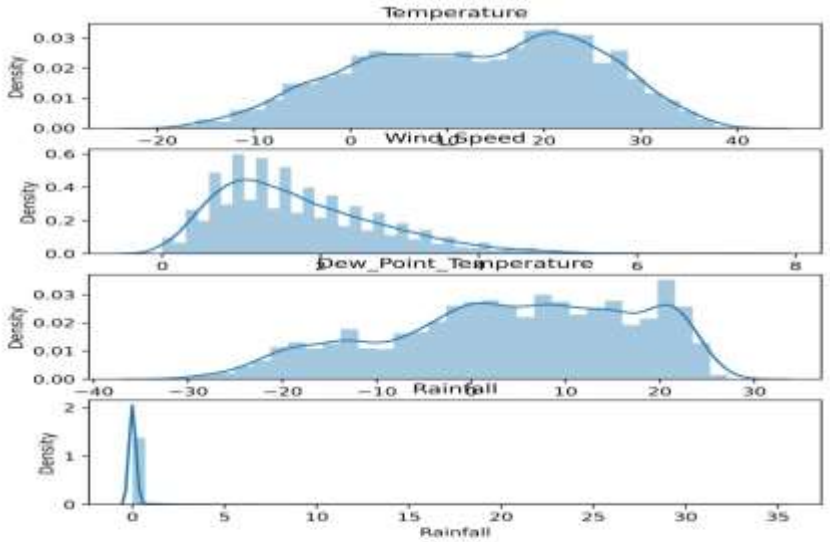




5. Analyze of Numerical variables(EDA)

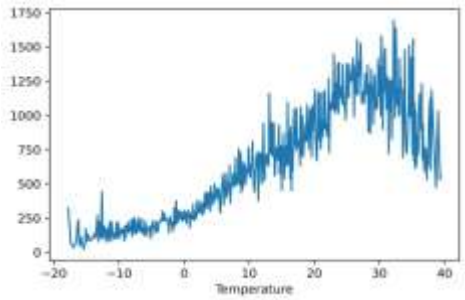
- From the above plot which shows that mean ranted bike count was 650 and ranted bike count maximum goes to above 3000 in a day and this is right skew distribution.
- From the above plot which shows that mean temperture was 14 degree and the maximum distribution lies between 0 to 30 degree celcius.
- From the above plot which shows that mean Humidity was 58.
- From the above plot which shows that mean wind speed in the year was 1.7 m/s and its normal.
- From the above plot which shows that the maximum days visibility was good and the mean visibility in the year was 1700.
- From the above plot which shows that the mean Dew point tempration was 5DegC.and the maximum distribution lies between -5 to +25 deg
- From the above plot which shows that mean solar radiation lies about 0.6 and maximum days solar radiation lies close to zero.
- From the above plot which shows that in a year maximum days were dry.
- From the above plot which shows that in a year maximum days sky was clear and did not have snowfall.

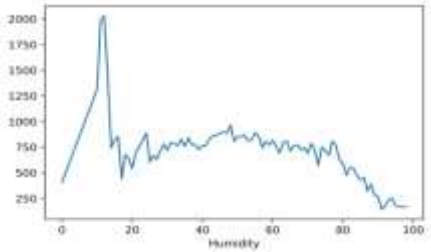


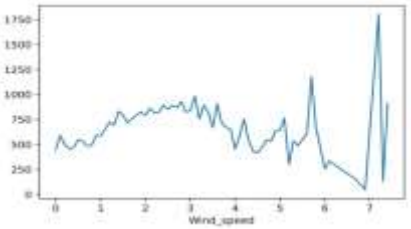


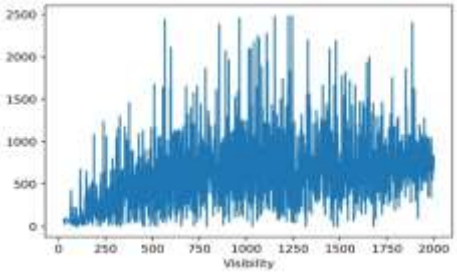
6. Numerical & Rented-Bike-Count(EDA)

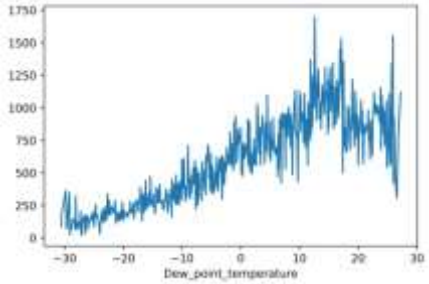
- From the above plot we see that people like to ride bikes when it is pretty hot around 25°C in average
- We can see from the above plot that the demand of rented bike is uniformly distribute from 20% to 80% Humidity but when the Humidity was above 80% then the demand of rented bike decrease and below 20% Humidity the demand of rented bike was increased.
- We can see from the above plot that the demand of rented bike is uniformly distribute despite of wind speed but when the speed of wind was 7 m/s then the demand of rented bike also increase that clearly means peoples love to ride bikes when its little windy.
- We can see from the above plot that the demand of rented bike is uniformly distribute above 500 visibility but below 500 visibility the demand of rented bike slightly less.
- From the above plot of "Dew-point-temperature' is almost same as the 'temperature' there is some similarity present we can check it in our next step.
- from the above plot we see that, the amount of rented bikes is huge, when there is solar radiation, the counter of rents is around 1000
- e can see from the above plot that even if it rains a lot the demand of rent bikes is not decreasing, here for example even if we have 20 mm of rain there is a big peak of rented bikes
- We can see from the plot that, on the y-axis, the amount of rented bike is very low When we have more than 4 cm of snow, the bike rents is much lower

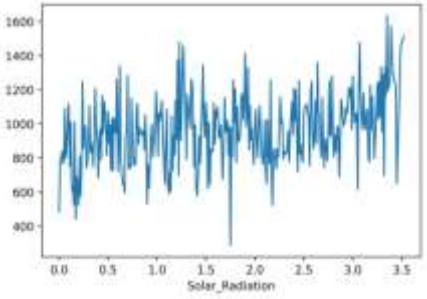


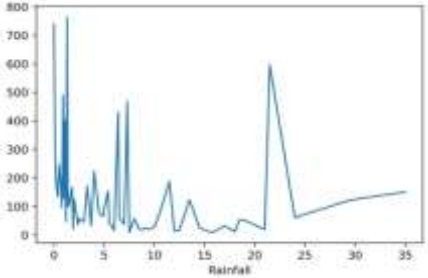


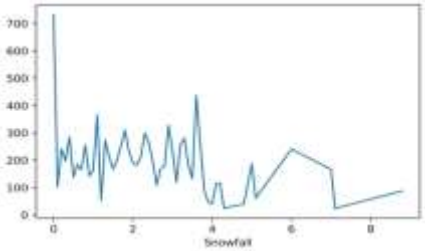








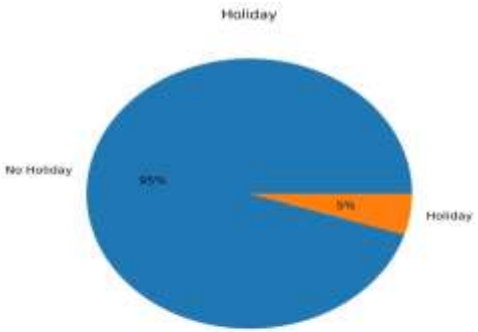


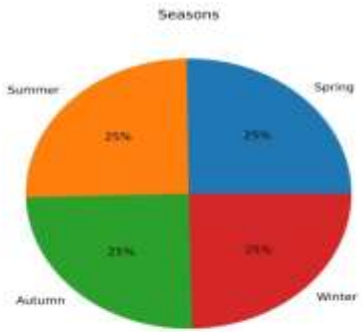


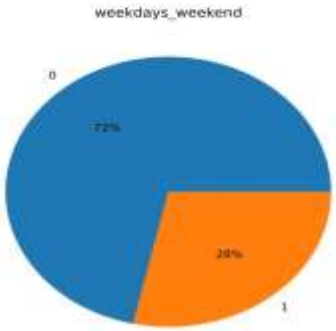
7. PERCENTAGE DISTRIBUTION OF THE VALUE COUNTS OF THE CATEGORICAL FEATURES(EDA)

- Month, Holiday, Seasons, Hour,

'weekdays-weekend‘ all the value counts of the categorical features

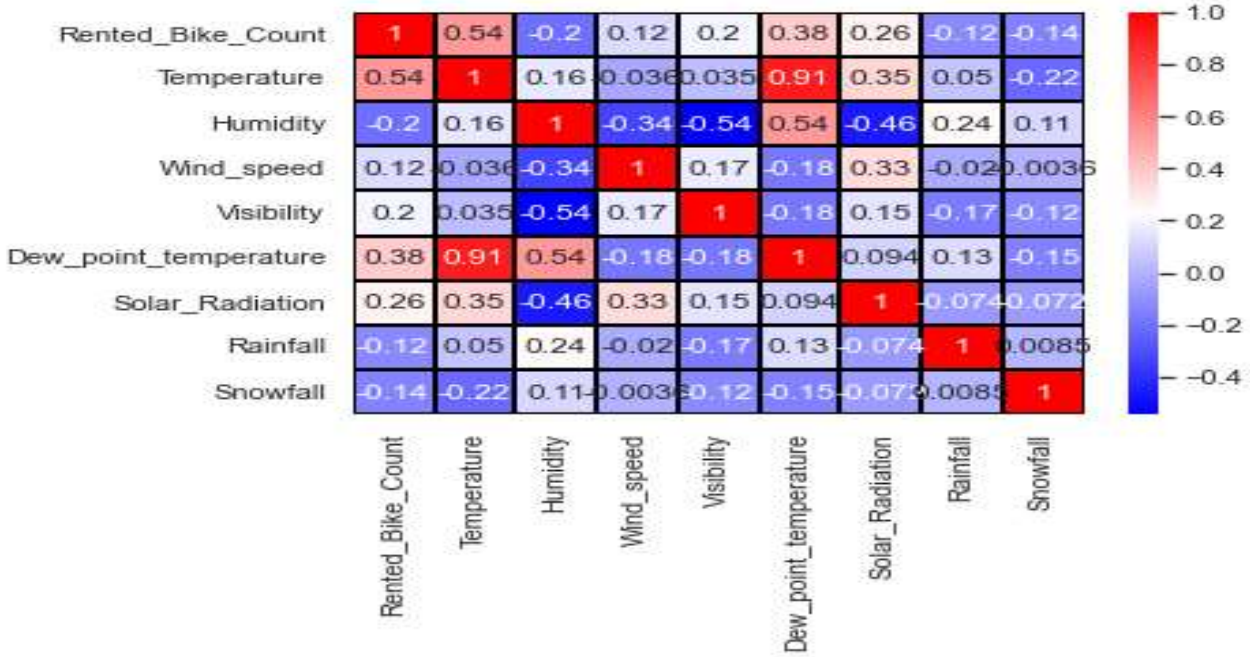






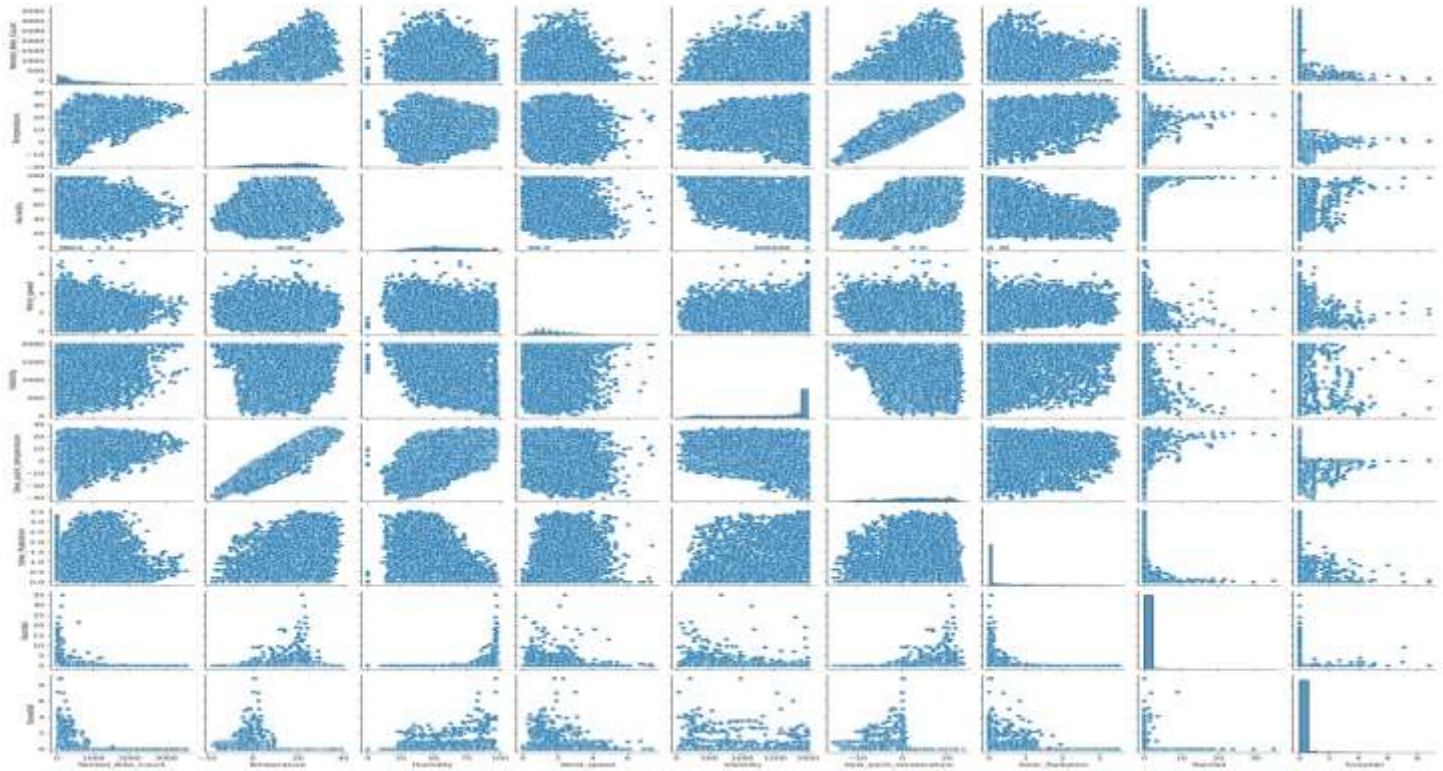
8. CORRELATION HEAT –MAP(EDA)

- We can infer the following from the above correlation heat map -
- Temperature and Dew point temperature are highly correlated to each other.
- We see a positive correlation between Rented bike count and temperature. Note that this is only true for the range of temperatures provided.
- We see a negative correlation among rented bike count with humidity, Rainfall and Snowfall. The more the humidity, Rainfall and Snowfall the less people prefer to bike.
- visibility has a weak dependence on Humidity.



9. Pair Plot(EDA)

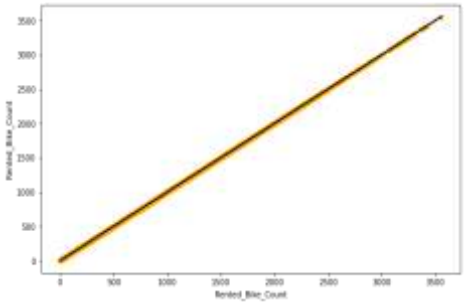
- We can infer the following from the above Pair Plot -
- When the snowfall and Rainfall increased Rented bike count decreased.
- With the increased of Temperature Rented bike count also increased.
- There is a positive relation between count and visibility.
- Visibility has a weak dependence on Humidity
- When the snowfall and Rainfall are decreased solar radiation increased.

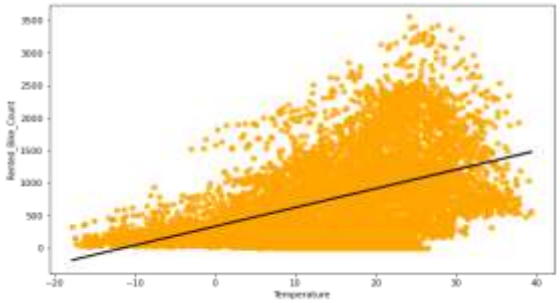


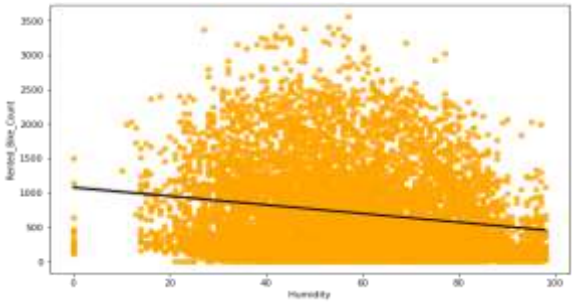
Feature Engineering & Data Pre-processing

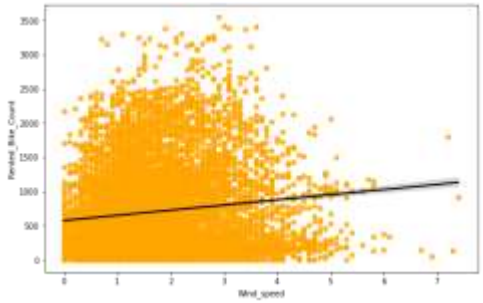
1. Regression plot

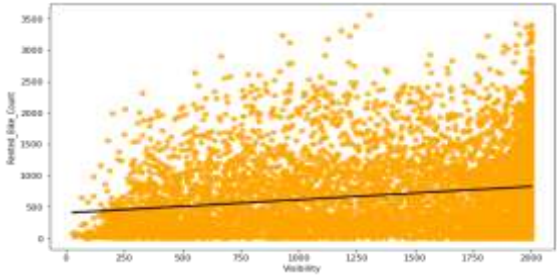
- which means the rented bike count increases with increase of these features.
- 'Rainfall‘ 'Snowfall‘ 'Humidity' these features are negatively related with the target variable which means the rented bike count decreases when these features increase.

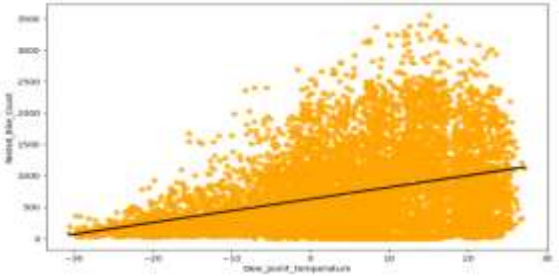


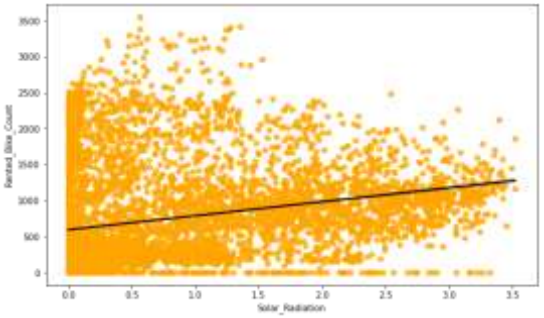


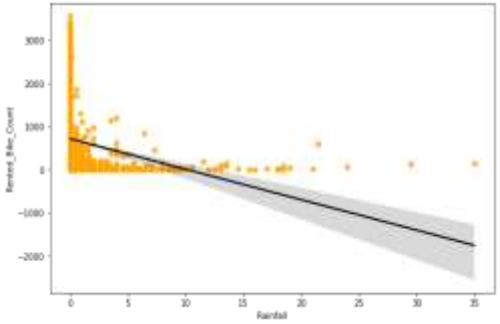


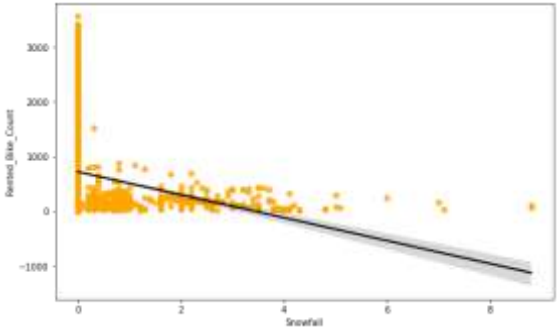






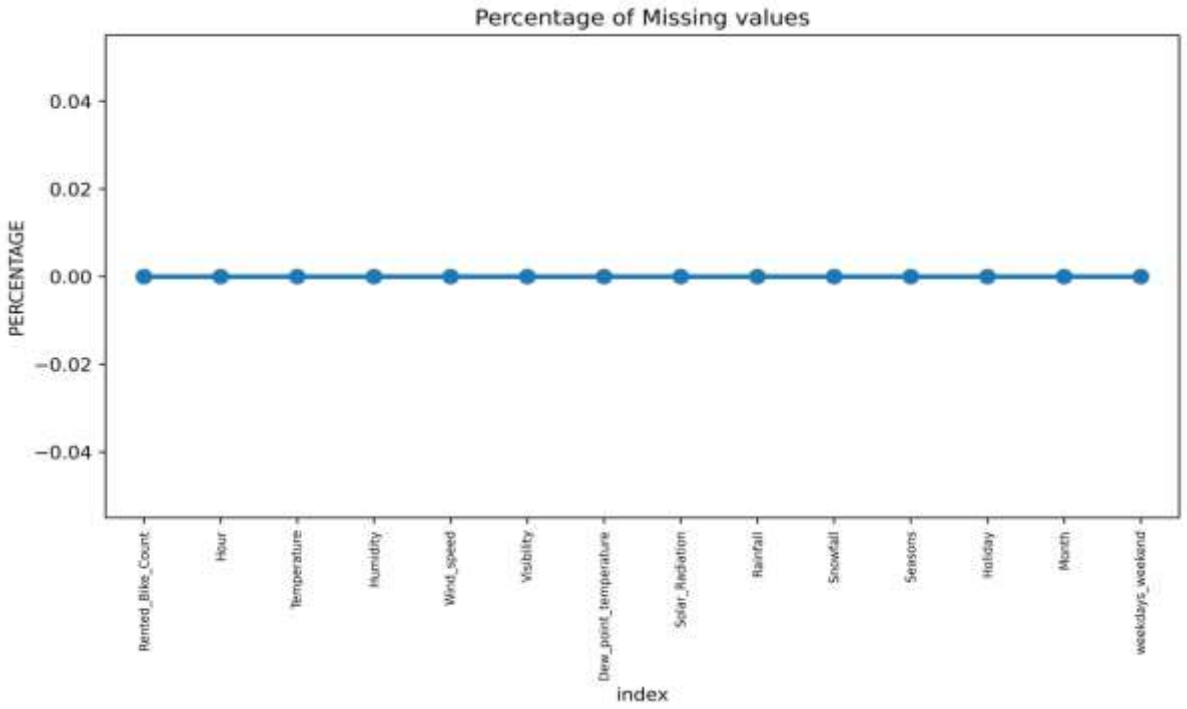






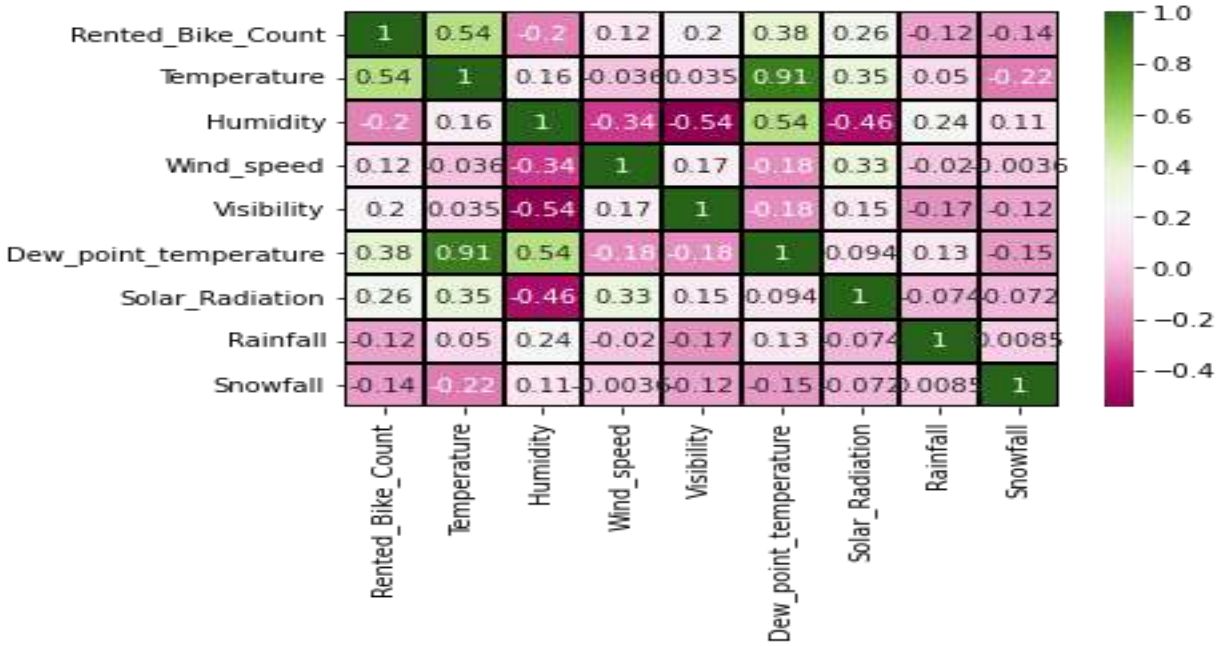
2. Handling Missing Values

- Temperature and Dew point temperature are almost 0.91 correlated, So it's generate multi-col-linearity issue.



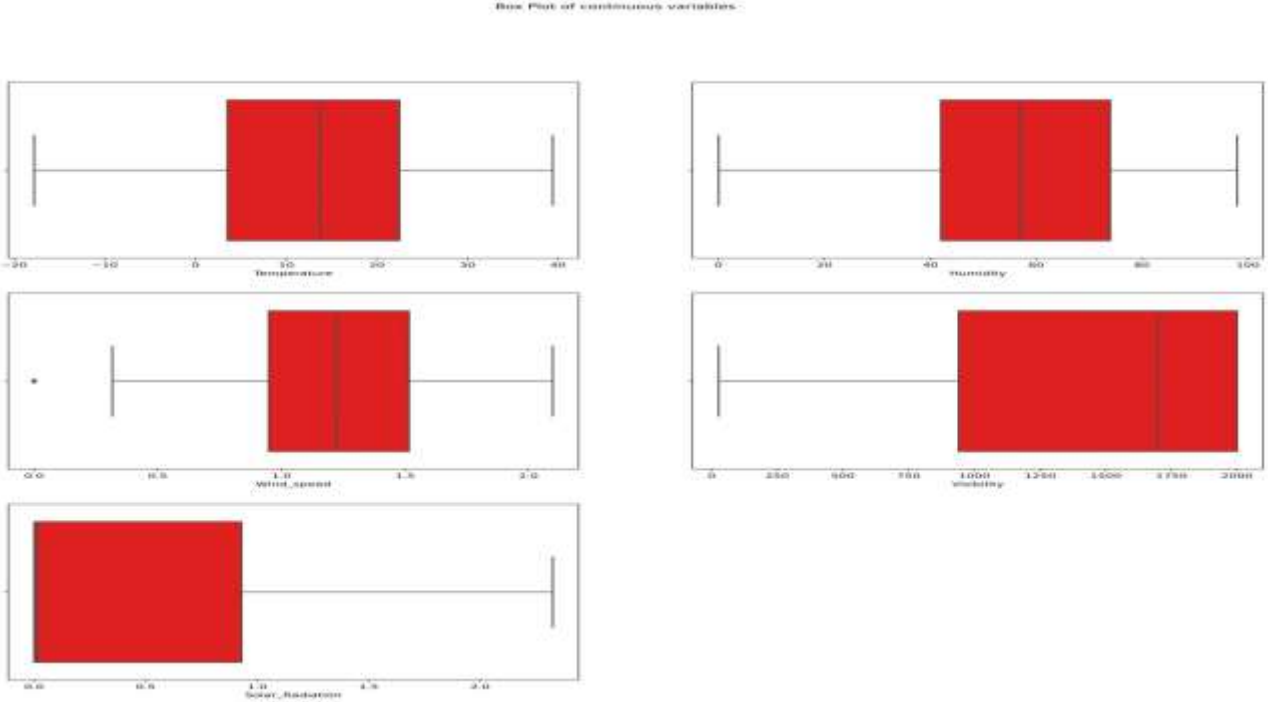
3. Checking of Correlation between variables

- **We can observe on the heat-map that on the target variable line the most positively correlated variables to the rent are :**
  - The temperature
  - The dew point temperature
  - The solar radiation
- **And most negatively correlated variables are:**
  - Humidity
  - Rainfall
  - From the above correlation heat-map, We see that there is a positive
  - correlation between columns 'Temperature' and 'Dew point temperature' 0.91 so the outcome of our analysis.



4. HANDLING OUTLIERS

- Median: In the box plot, the median is displayed rather than the mean.
- Q1: The first quartile (25%) position.
- Q3: The third quartile (75%) position.
- Interquartile range (IQR): a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles. It represents how 50% of the points were dispersed.
- Lower and upper 1.5*IQR whiskers: These represent the limits and boundaries for the outliers.
- Outliers: Defined as observations that fall below Q1 − 1.5 IQR or above Q3 + 1.5 IQR. Outliers are displayed as dots or circles.

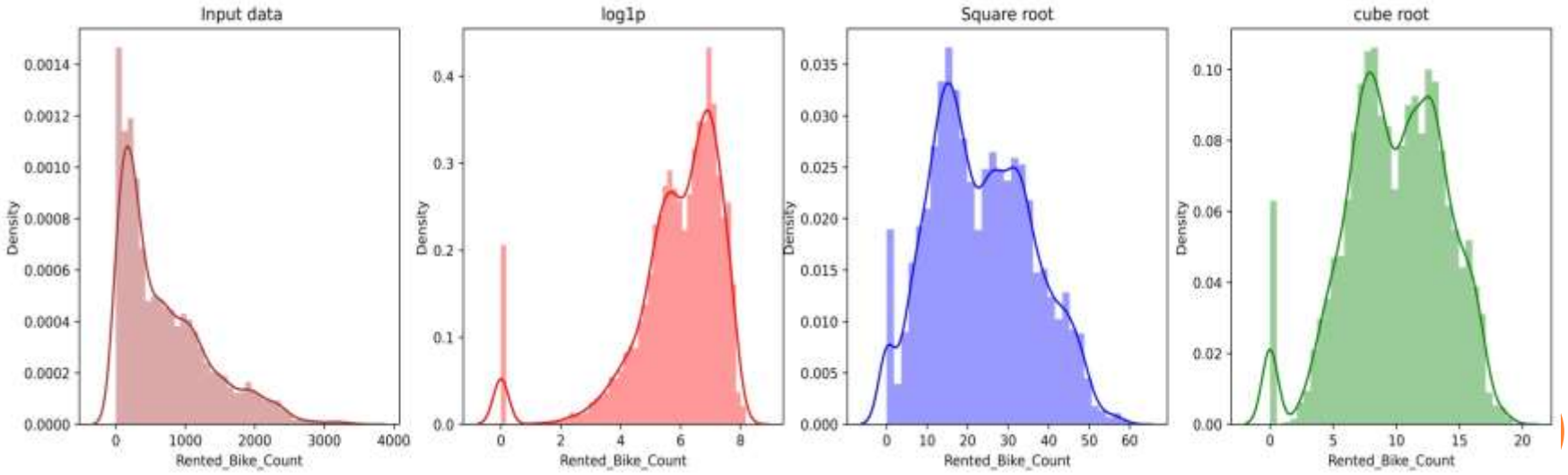


5. DATA SCALING

- **What all scaling techniques have you used & why did you use those techniques?**

- **In here I used square root ,Cube root and log transformation**

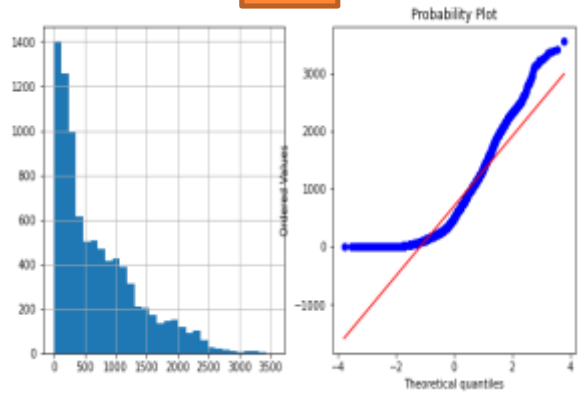
- Square Root The square root method is typically used when your data is moderately skewed. Now using the square root (e.g., sqrt(x)) is a transformation that has a moderate effect on distribution shape. It is generally used to reduce right skewed data. Finally, the square root can be applied on zero values and is most commonly used on counted data. Square Root Transformation: Transform the values from y to √y.
- Log Transformation The logarithmic is a strong transformation that has a major effect on distribution shape. This technique is, as the square root method, often used for reducing right skew ness. Worth noting, however, is that it can not be applied to zero or negative values. Log Transformation: Transform the values from y to log(y).
- Cube root transformation involves converting x to x^ (1/3). This is a fairly strong transformation with a substantial effect on distribution shape: but is weaker than the logarithm. It can be applied to negative and zero values too. Negatively skewed data Cube Root Transformation: Transform the values from y to y^(1/3)



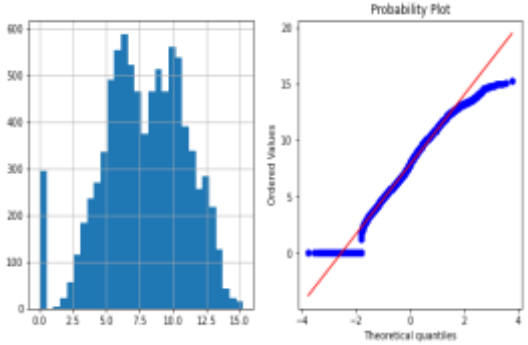
5. Data Scaling

- Rented-bike-count plot variable

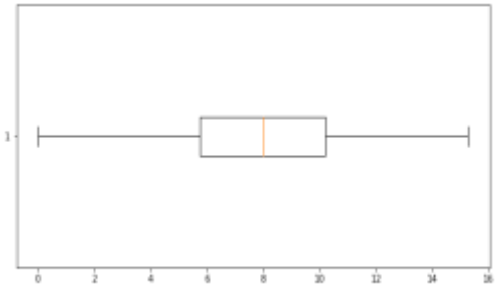
1



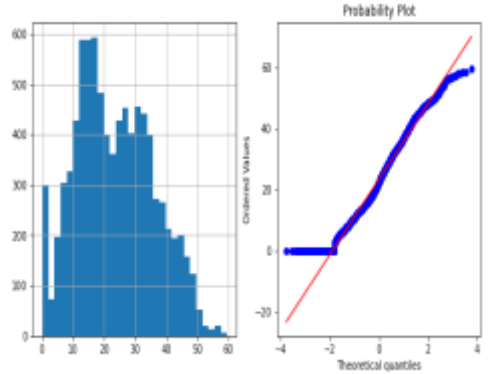
2



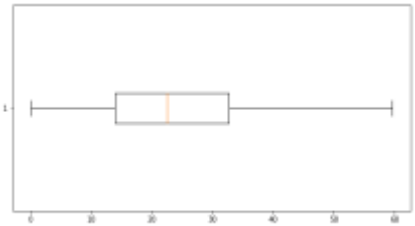
3



our Rented Bike Count target this not normally distributed ,so we need to make some transformations before supply to the model

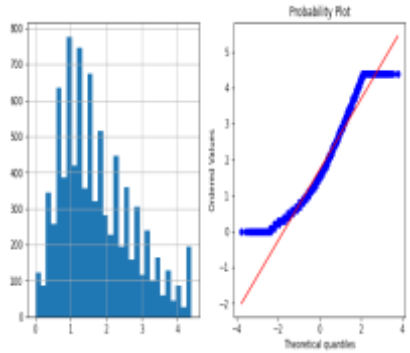


4



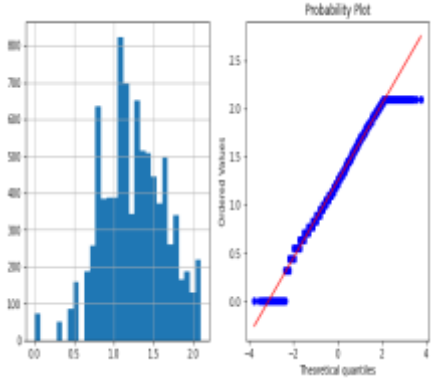
Now, Its Look Like Normal Distribution

5



not look like normal and right screwed distribution so need to apply transformation

6



look like a normal distribution.

7

6. Categorical Encoding

- **What all categorical encoding techniques have you used & why did you use those techniques?**
  - **In here I used both Ordinal Encoder on 'Seasons' feature and One Hot Encoder on Hour, Holiday, Month, weekdays-weekend, features.**
  - Ordinal Encoder is used when the variables in the data are ordinal, ordinal encoding converts each label into integer values and the encoded data represents the sequence of labels.
  - In One-Hot Encoding, each category of any categorical variable gets a new variable. It maps each category with binary numbers (0 or 1). This type of encoding is used when the data is nominal. Newly created binary features can be considered dummy variables. After one hot encoding, the number of dummy variables depends on the number of categories presented in the data.

7. Data Splitting

- What data splitting ratio have you used and why?
  - The foregoing data splitting methods can be implemented once we specify a splitting ratio. A commonly used ratio is 80:20, which means 80% of the data is for training and 20% for testing which I did in here. Other ratios such as 70:30, 60:40, and even 50:50 are also used in practice. There does not seem to be clear guidance on what ratio is best or optimal for a given dataset. The 80:20 split draws its justification from the well-known Pareto principle, but that is again just a thumb rule used by practitioners.

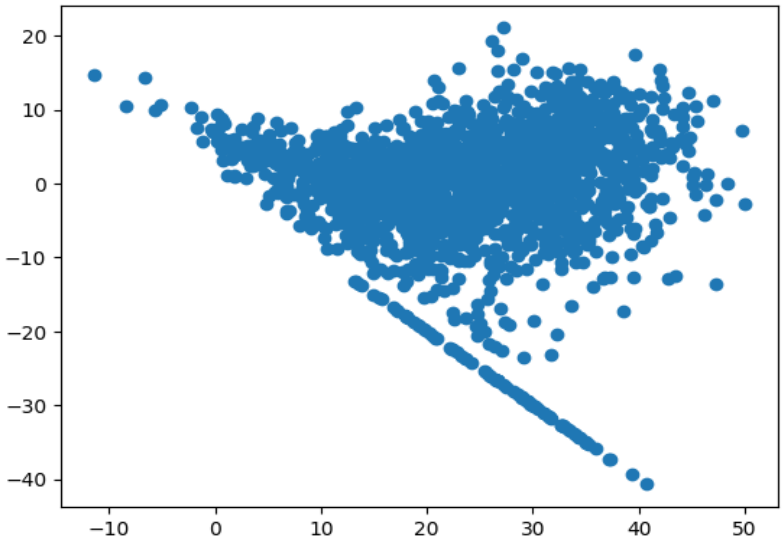
ML Model Implementation

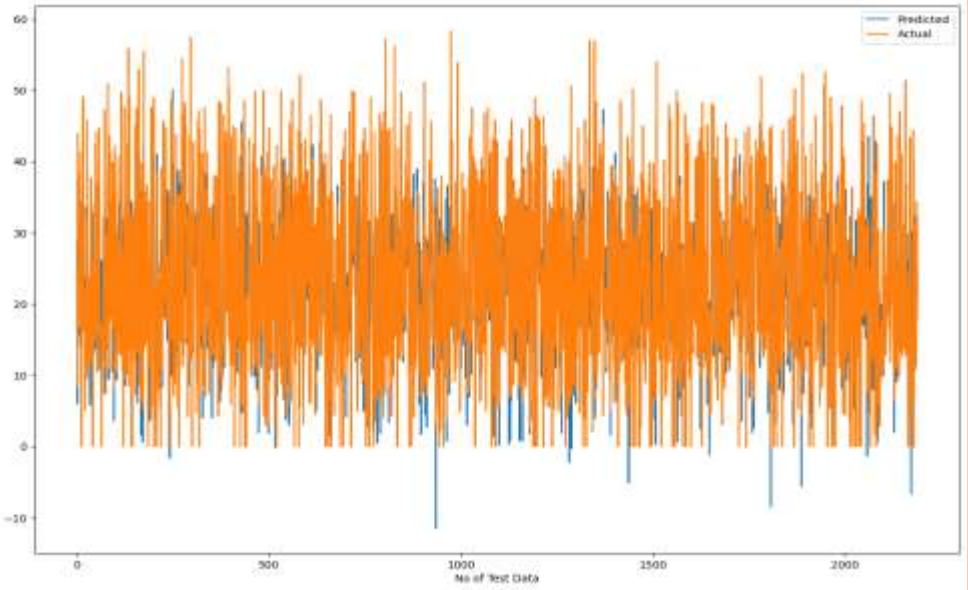
ML Model-1. LINEAR REGRESSION

- . Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable)
- Explain the ML Model used and it's performance using Evaluation metric Score Chart.
- **Looks like our r2 score value is 0.63 that means our model is able to capture most of the data variance. Lets save it in a data frame for later comparisons.**
- **The r2_score for the test set is 0.60. This means our linear model is performing well on the data. Let us try to visualize our residuals and see if there is Hetero-scedasticity (unequal variance or scatter).**

Evaluation Metrics :

MSE :	62.212
RMSE :	7.8874
MAE :	5.4147
R2 :	0.5996
Adjusted R2 :	0.5910





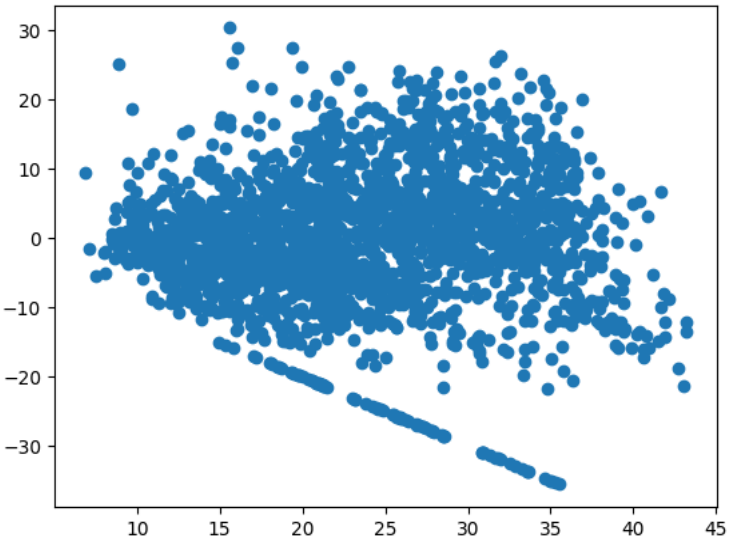
ML MODEL-2. LASSO REGRESSION

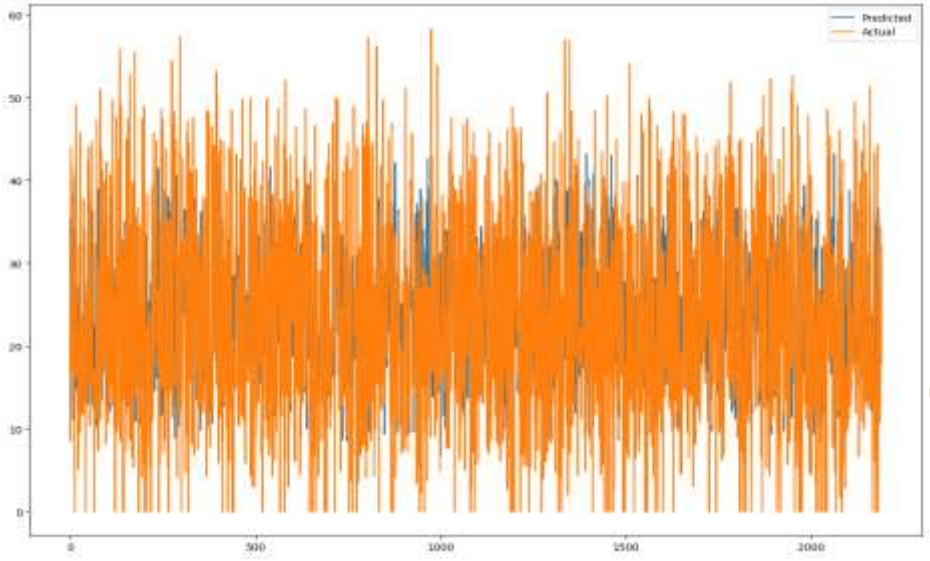
- Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multi-co-linearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.
- Lasso Regression uses L1 regularization technique (will be discussed later in this article). It is used when we have more features because it automatically performs feature selection.

Explain the ML Model used and it's performance using Evaluation metric

- **Looks like our r2 score value is 0.40 that means our model is able to capture most of the data variance. Lets save it in a data frame for later comparisons.**

Evaluation Metrics :	
MSE :	92.956
RMSE :	9.6414
MAE :	7.2726
R2 :	0.4018
Adjusted R2 :	0.3889





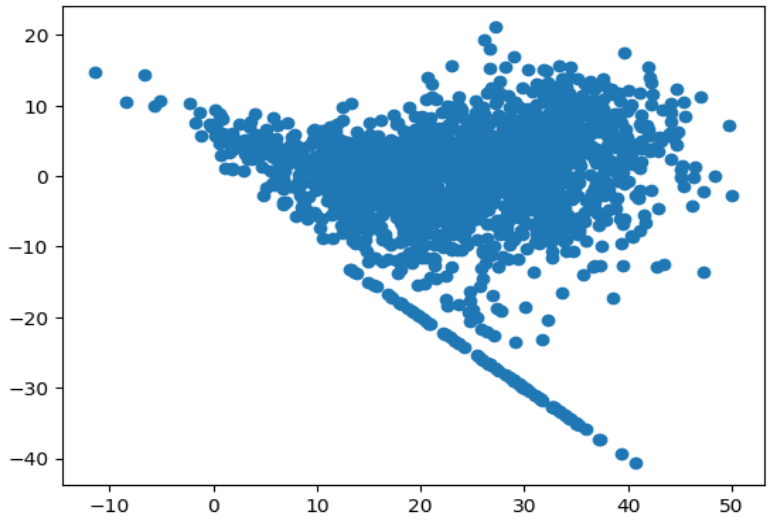
ML Model - 3. RIDGE REGRESSION

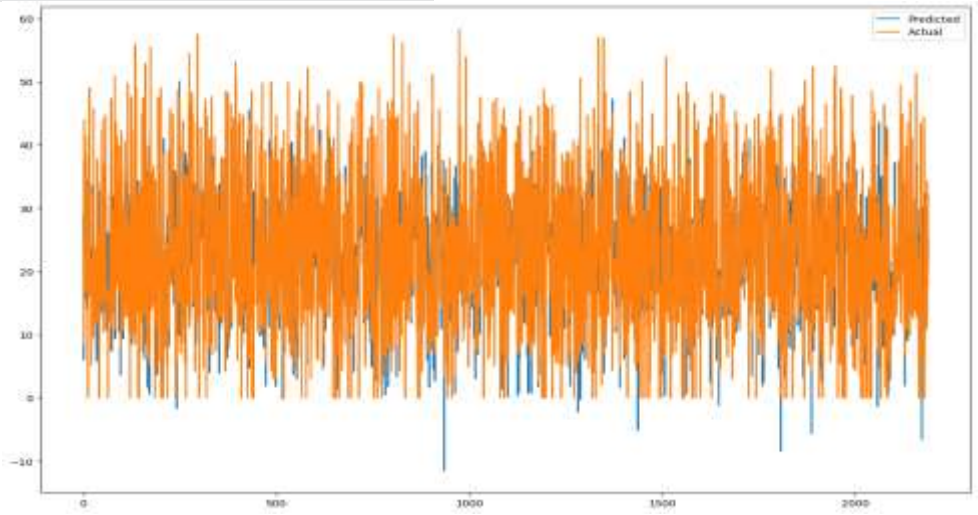
- Ridge regression is a specialized technique used to analyze multiple regression data that is multi-co-linearity in nature. It is a fundamental regularization technique, but it is not used very widely because of the complex science behind it. However, it is fairly easy to explore the science behind ridge regression in r if you have an overall idea of the concept of multiple regression. Regression stays the same, but in regularization, the way the model coefficients are determined is different, The main idea of ridge regression focuses on fitting a new line that does not fit.

Explain the ML Model used and it's performance using Evaluation metric

- **Looks like our r2 score value is 0.60 that means our model is able to capture most of the data variance. Lets save it in a data frame for later comparisons.**
- **Looks like our r2 score value is 0.63 that means our model is able to capture most of the data variance. Lets save it in a data frame for later comparisons.**

Evaluation Metrics :	
MSE :	62.211
RMSE :	7.8874
MAE :	5.4147
R2 :	0.5996
Adjusted R2 :	0.5910

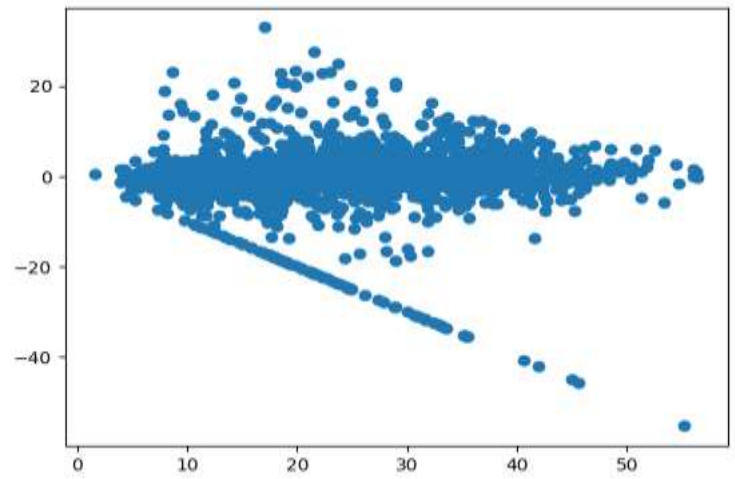


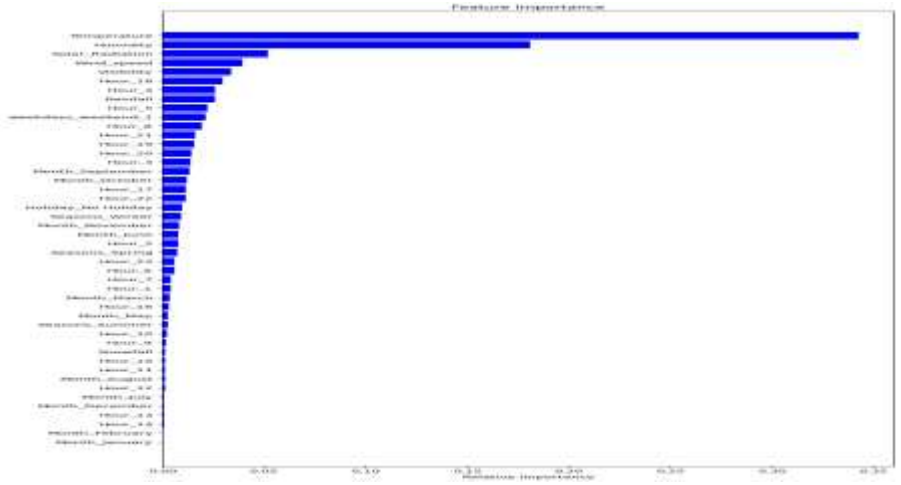


ML Model-4.RANDOM FOREST

- Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.
  - **Looks like our r2 score value is 0.97 that means our model is able to capture most of the data variance. Lets save it in a data frame for later comparisons.**
  - **Looks like our r2 score value is 0.97 that means our model is able to capture most of the data variance. Lets save it in a data frame for later comparisons.**
  - **The r2_score for the test set is 0.74. This means our linear model is performing well on the data. Let us try to visualize our residuals and see if there is heteroscedasticity(unequal variance or scatter).**

Evaluation Metrics :	
MSE :	42.360
RMSE :	6.5084
MAE :	3.6135
R2 :	0.7274
Adjusted R2 :	0.7215

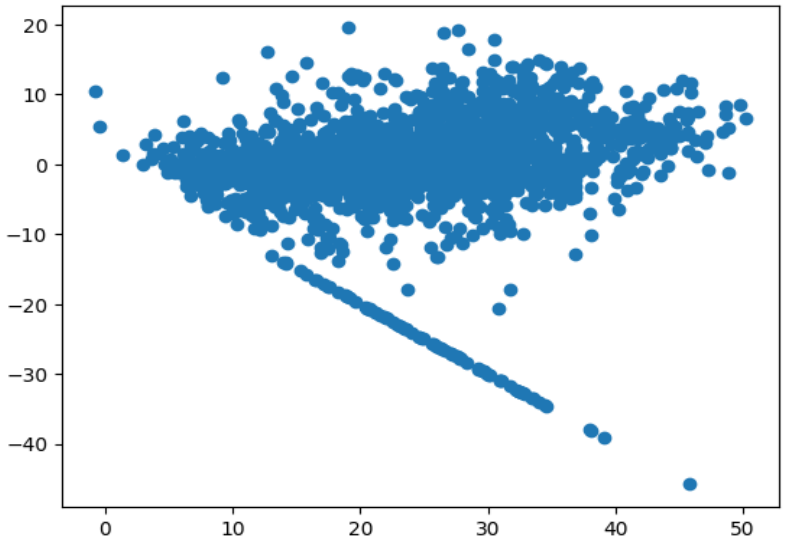


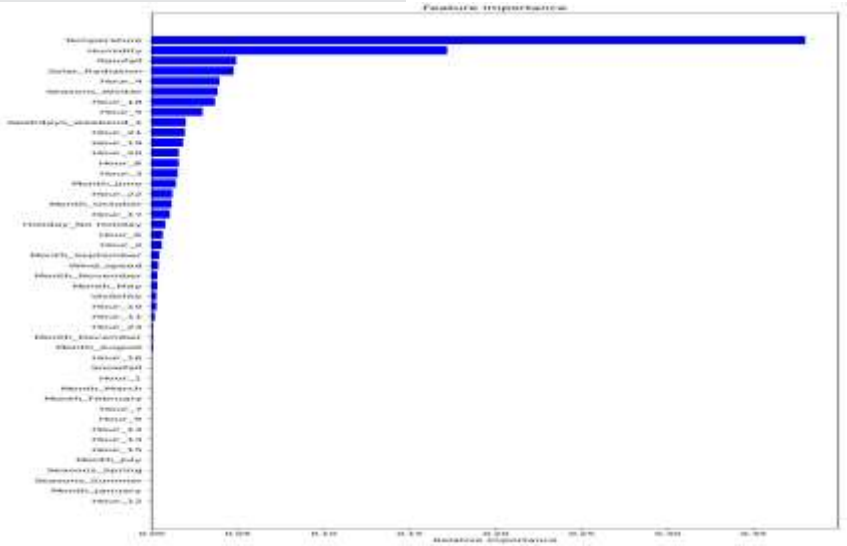


ML Model-5.Gtadient Boosting

- Gradient boosting Regression calculates the difference between the current prediction and the known correct target value. This difference is called residual. After that Gradient boosting Regression trains a weak model that maps features to that residual.
- **Looks like our r2 score value is 0.74 that means our model is able to capture most of the data variance. Lets save it in a data frame for later comparisons.**
- **The r2_score for the test set is 0.69. This means our linear model is performing well on the data. Let us try to visualize our residuals and see if there is hetero-scedasticity(unequal variance or scatter).**

Evaluation Metrics :	
MSE :	49.339
RMSE :	7.0242
MAE :	4.4155
R2 :	0.6825
Adjusted R2 :	0.6756





Gradient Boosting with Grid-Search-CV

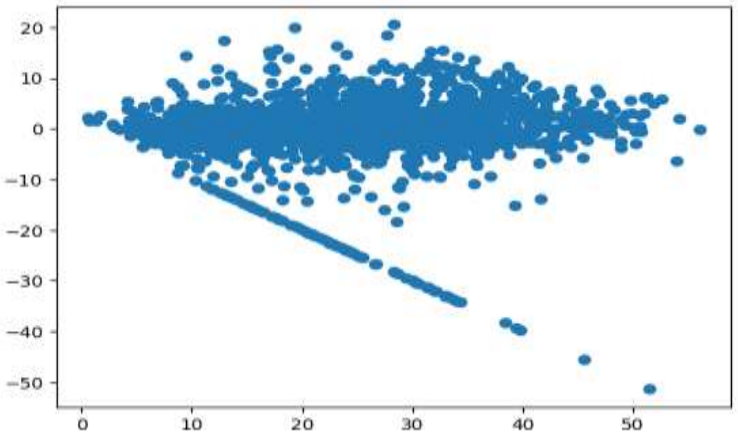
Hyper-parameter tuning

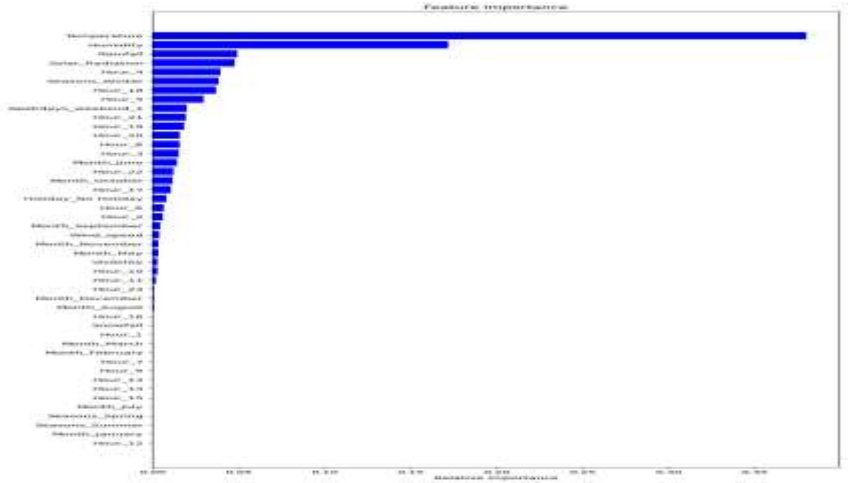
- Before pro-ceding to try next models, let us try to tune some hyper-parameters and see if the performance of our model improves.
- Hyper-parameter tuning is the process of choosing a set of optimal hyper-parameters for a learning algorithm. A hyper-parameter is a model argument whose value is set before the learning process begins. The key to machine learning algorithms is hyper-parameter tuning.

Using Grid-Search CV

- Grid-Search CV helps to loop through predefined hyper parameters and fit the model on the training set. So, in the end, we can select the best parameters from the listed hyper parameters.

Evaluation Metrics :	
MSE :	39.643
RMSE :	6.2963
MAE :	3.6012
R2 :	0.7448
Adjusted R2 :	0.7394





CONCLUSION

- During the time of our analysis, we initially did EDA on all the features of our dataset. We first analyzed our dependent variable,
- 'Rented Bike Count' or The month June, JULY, May to October or September the demand of the rented bike is high as compare to other months.
- **Rented bikes according to Weekdays-Weekend and Time** the demand of the bike higher because of the office. Peak Time are 7 am to 9 am and 5 pm to 7 pm that the demand of rented bikes are very low specially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases.
- In summer season the use of rented bike is high and peak time is 7am-9am and 7pm-5pm. In winter season the use of rented bike is very low because of snowfall.
- From the above plot we see that people like to ride bikes when it is pretty hot around 25°C in average
- We can see from the above plot that the demand of rented bike is uniformly distribute from 20% to 80% Humidity but when the Humidity was above 80% then the demand of rented bike decrease and below 20% Humidity the demand of rented bike was increased
- Next we analyzed categorical variable and dropped the variable who had majority of one class, we also analyzed numerical variable, found out the correlation, distribution and their relationship with the dependent variable.
- We also removed some numerical features who had mostly 0 values and hot encoded the categorical variables.
- Next we implemented 5 machine learning algorithms Linear Regression lasso , ridge, Random Forest and XG-Boost. We did hyper parameter tuning to improve our model performance.

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	5.35	58.45	7.64	0.62	0.61
	1	Lasso regression	7.33	93.04	9.65	0.40	0.39
	2	Ridge regression	5.42	62.21	7.89	0.60	0.59
	3	Random forest regression	1.28	5.11	2.26	0.97	0.97
	4	Gradient boosting regression	4.13	41.72	6.46	0.73	0.72
	5	Gradient Boosting gridsearchcv	2.97	26.19	5.12	0.83	0.83
Test set	0	Linear regression	5.42	62.21	7.89	0.60	0.59
	1	Lasso regression	7.27	92.96	9.64	0.40	0.39
	2	Ridge regression	5.35	58.45	7.64	0.62	0.61
	3	Random forest regression	3.62	42.67	6.53	0.72	0.72
	4	Gradient boosting regression	4.41	49.30	7.02	0.68	0.68
	5	Gradient Boosting gridsearchcv	3.60	39.64	6.30	0.74	0.74

- • No over fitting is seen.
- • Random forest and Gradient Boosting grid-search-cv gives the highest R2 score of 97% and 84% respectively for Train Set and 75% for Test set.
- • Feature Importance value for Random Forest and Gradient Boost are different.
- • We can deploy this model

I'M VERY GLAD THAT YOU HAVE TAGGED ALONG UNTIL THE END. I HOPE YOU ENJOYED IT AND IF YOU HAVE ANY SUGGESTIONS ABOUT MY WORK, PLEASE LET ME KNOW :)