Capstone project

Seoul Bike Sharing Demand Prediction

Team members

Ajinkya Dakhale

Harshjyot Singh

Suvir Kapse

CONTENT

- Problem Statement
- Data summary
- Feature summary
- Data Wrangling
- Exploratory data analysis
- Data pre processing
- Implementing algorithms
- Conclusions

Introduction

A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system. Rental Bike Sharing is the process by which bicycles are procured on several basis- hourly, weekly, membership-wise, etc.

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.

Data summary

	Date	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_speed	Visibility	DPT	Solar_Radiation	Rainfall	Snowfall	Seasons	Holiday	Functioning_Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

- Given Dataset contains 8760 lines and 14 columns.
- One Datetime feature ‘Date’.
- Out of which Three are categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.
- Numerical variables such as temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall which tells the environment conditions at that particular hour of the day.

Feature summary

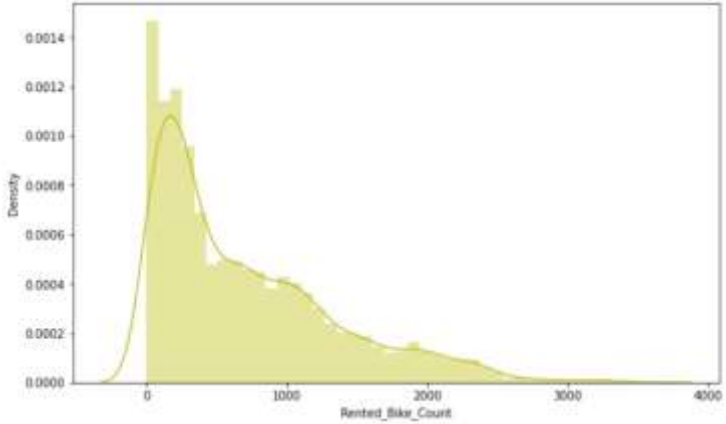
- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour.
- Hour - Hour of the day
- Temperature-Temperature in Celsius.
- Humidity - Humidity in the air during the booking.
- Wind speed - Speed of the wind while booking (m/s)
- Visibility - Visibility to the eyes during driving in “m”.
- Dew point temperature – the temperature the air needs to be cooled at to achieve a relative humidity of 100%.
- Solar radiation - solar radiation during ride booking (MJ/m2).
- Rainfall - The amount of rainfall during bike booking (mm).
- Snowfall - Amount of snowing in cm during the booking in cm.
- Seasons - Winter, Spring, Summer, Autumn.
- Holiday - If the day is holiday period or not and there are 2 types of data that is holiday and no holiday .
- Functional Day - If the day is a Functioning Day or not and it contains object data type yes and no. NoFunc(Non Functional Hours), Func(Functional hours)

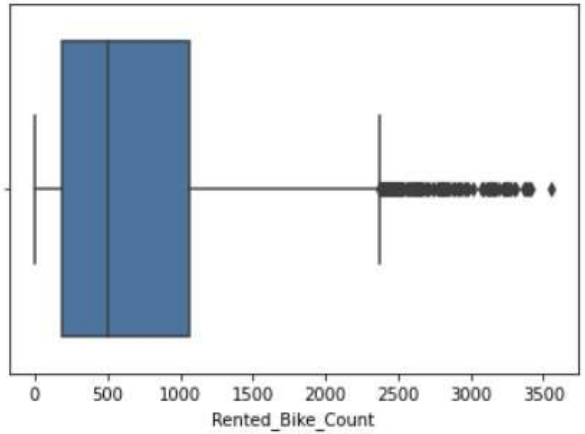
Data wrangling

- We Convert date column to date-time and extract day, month and year from it, since the data given to us is for one year only i.e. 2017-2018 there is nothing much we can do with year therefore we drop year column, and use month and day
- Day column is converted to weekend column as knowing whether the day belongs to weekdays or weekends gives us more insights than normal days.
- Outlier detection was done with the help of box plot.
- There were no Null values or Duplicate values in dataset.

EDA (EXPLORATORY DATA ANALYSIS)

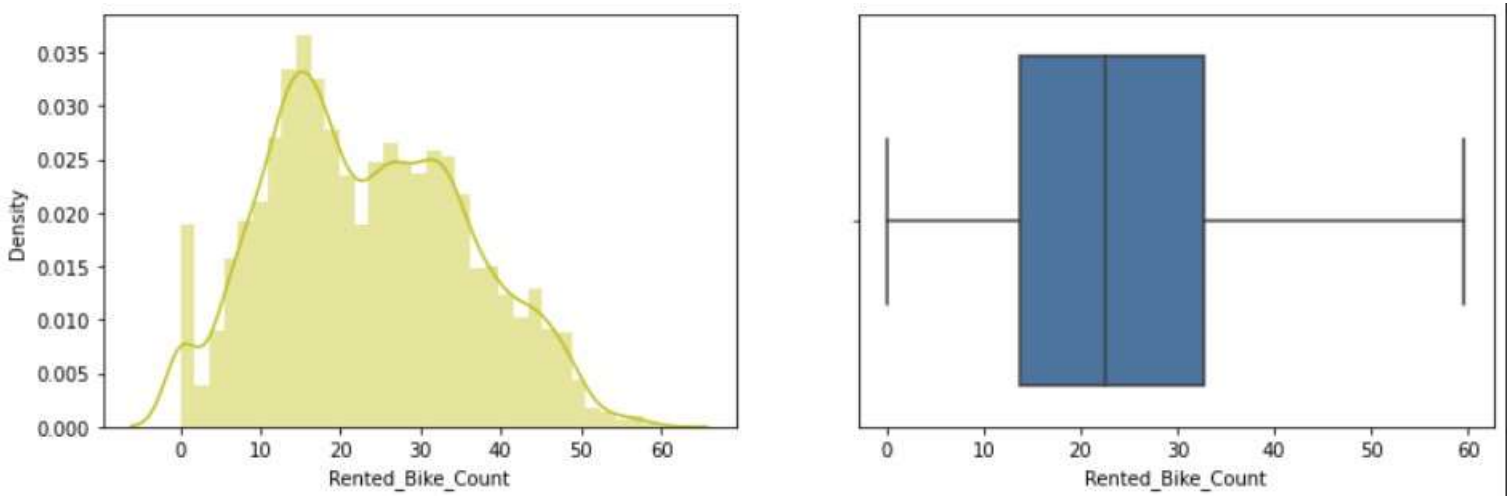
- Dependent Variable





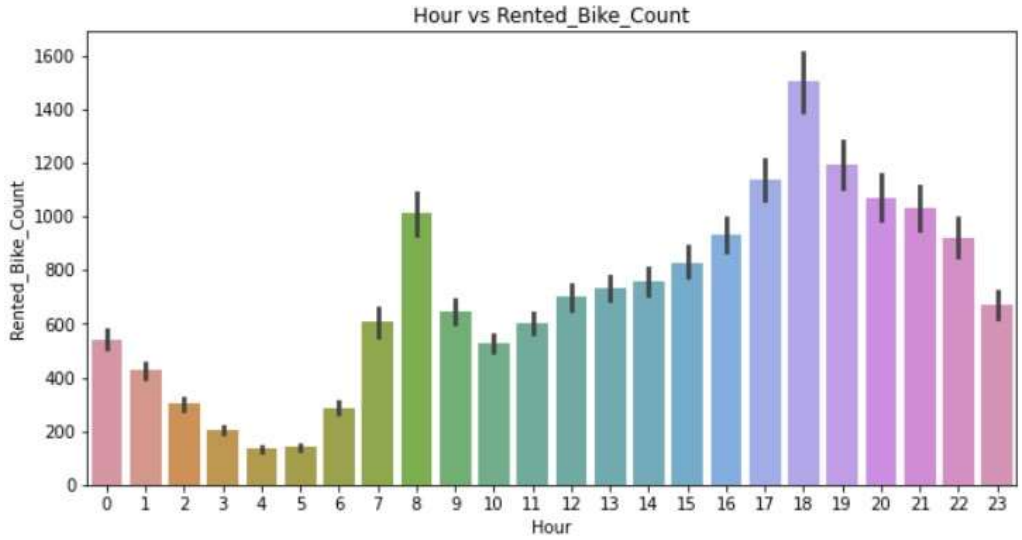
- The above graph shows us that Rented Bike Count is moderately right skewed .
- Also the boxplot shows that we have numerous outlier in Rented Bike Count column.
- Since the assumption of linear regression is that 'the distribution of dependent variable has to be normal', thus we transform it by using Square root operation to make it normally distributed.

EDA (Dependent Variable)



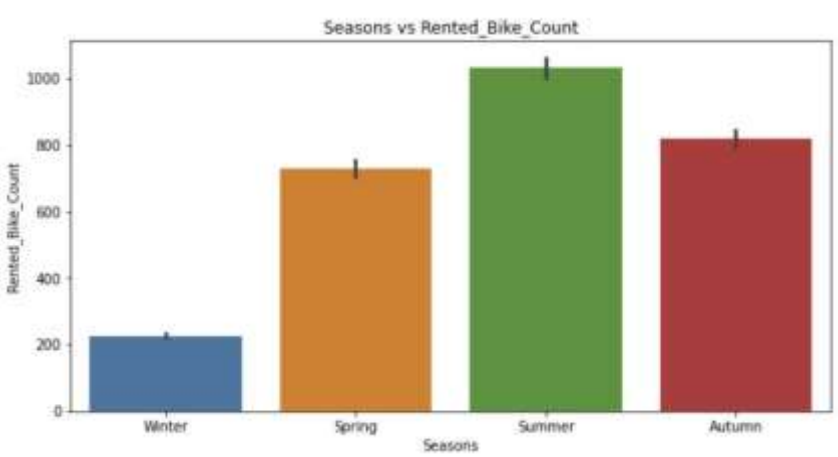
- Here in above graph we can see that after applying the square root operation now the outliers are not present and also graph is kind of moderately distributed.

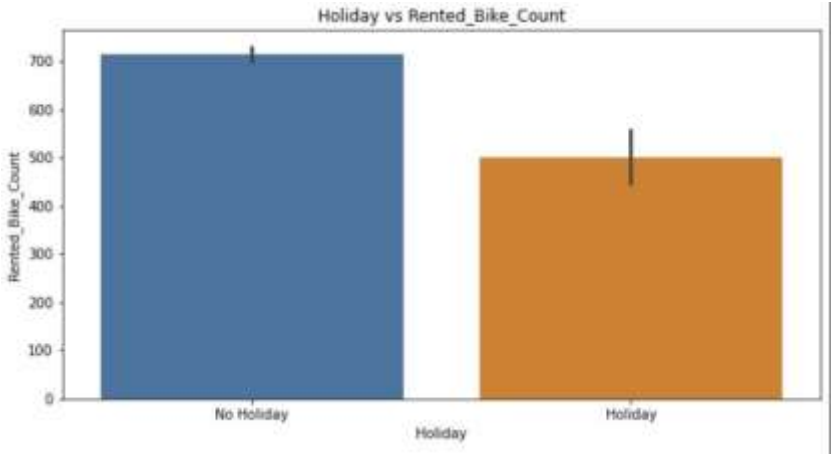
Categorical Variable vs. Rented_Bike_Count



- This plot shows the use of rented bike according the Hour of day.
- Generally people use rented Bikes mostly for transits from 7am to 9am and 5pm to 7pm.

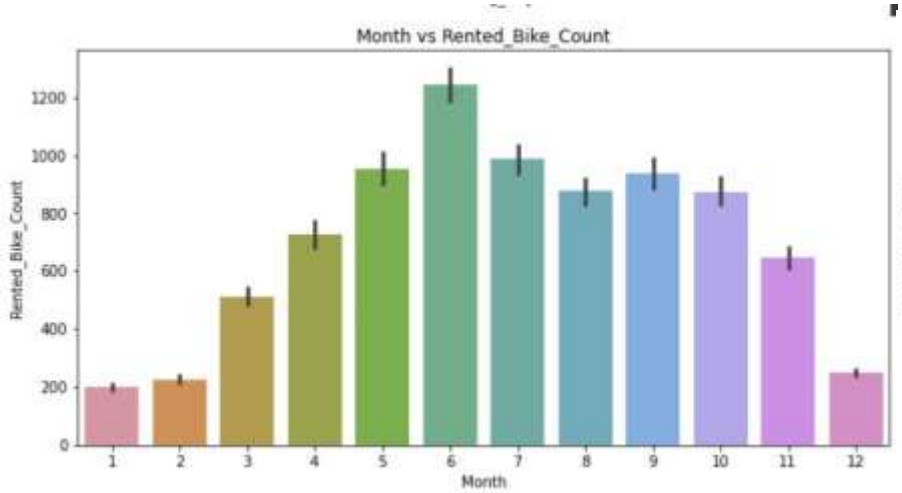
Categorical Variable vs. Rented_bike_count (cont.)

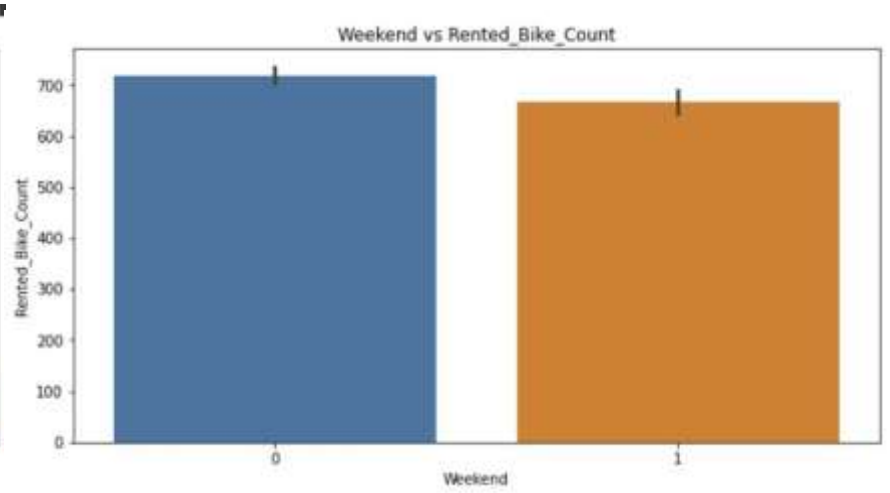




- Summer season was the peak (2208) of all the activity with the most number of Rented bike count. Whereas; Winter was the least (2160) popular season with the minimum bike counts recorded.
- And we can clearly see that that peoples love to ride bike in summer seasons and autumn season.
- But in winter season people didn't take any rented bike because of snowfall.

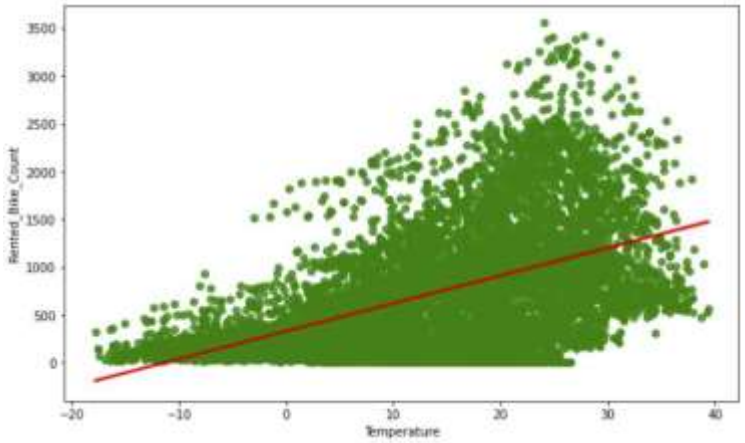
Categorical Variable vs. Rented_bike_count (cont.)

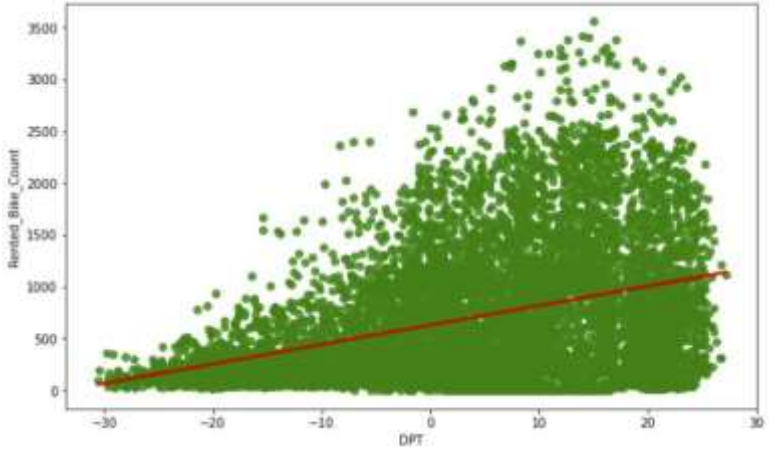




- In above bar plot we can clearly see that from the month 5 to 10th demand of the rented bike is high as compare to other months.
- 5^{th} to 10^{th} is when Summer (June to August) and Autumn (September to November) season occur.

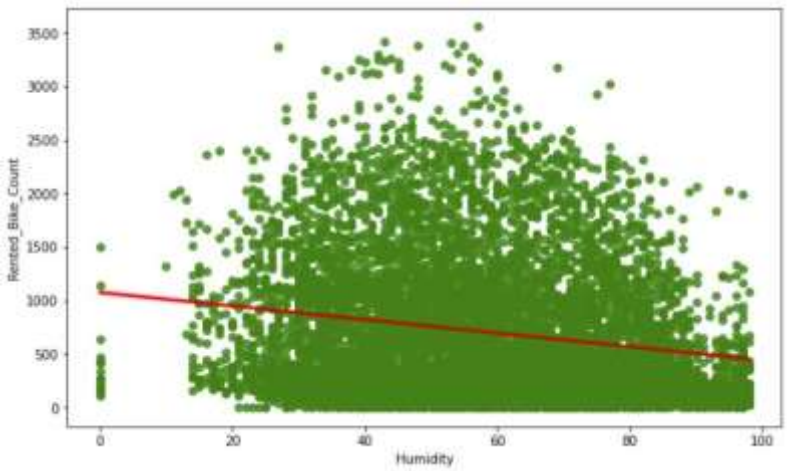
Numerical variable Vs. Rented_Bike_Count

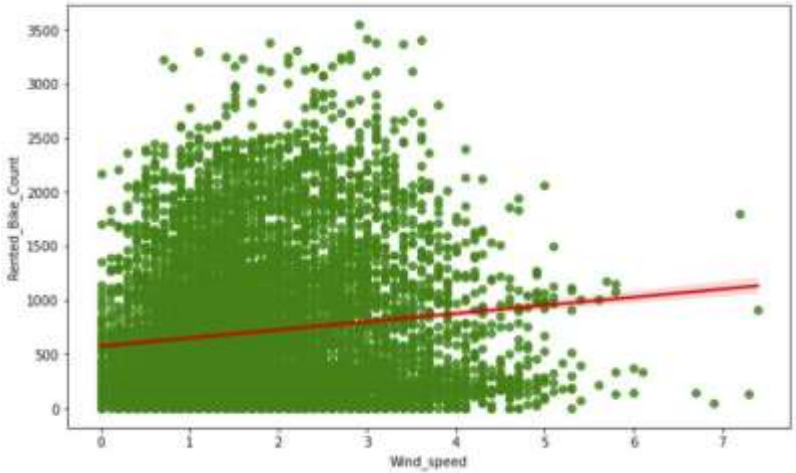




- From the above regression plot, we see that the columns 'Temperature', 'Dew_point_temperature', are positively related to the target variable.

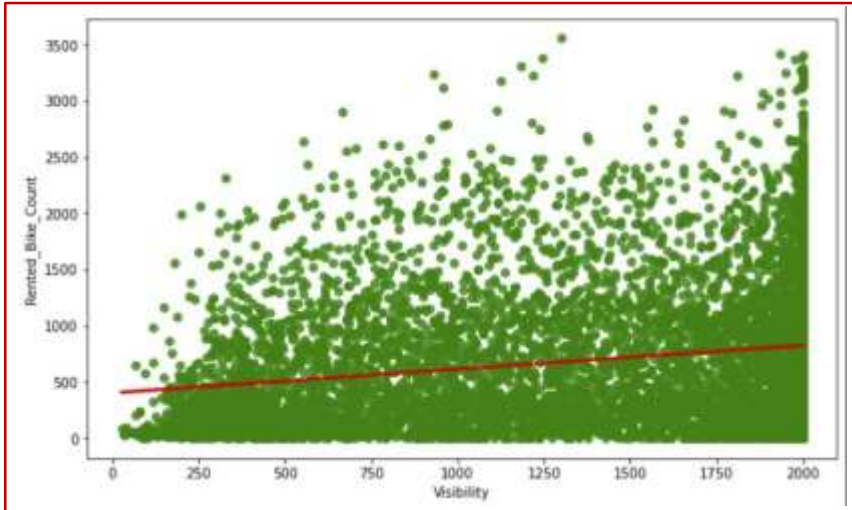
Numerical variable Vs. Rented_Bike_Count (cont.)

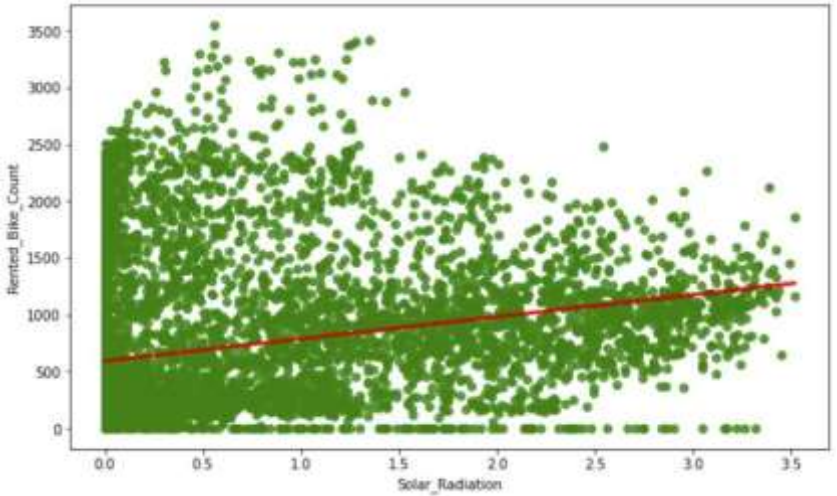




- 'Humidity' is negatively related and 'Wind_speed' is positively related with the target variable which means the rented bike count decreases when 'Humidity' increase and vice versa.
- 5-7m/s is considered light breeze

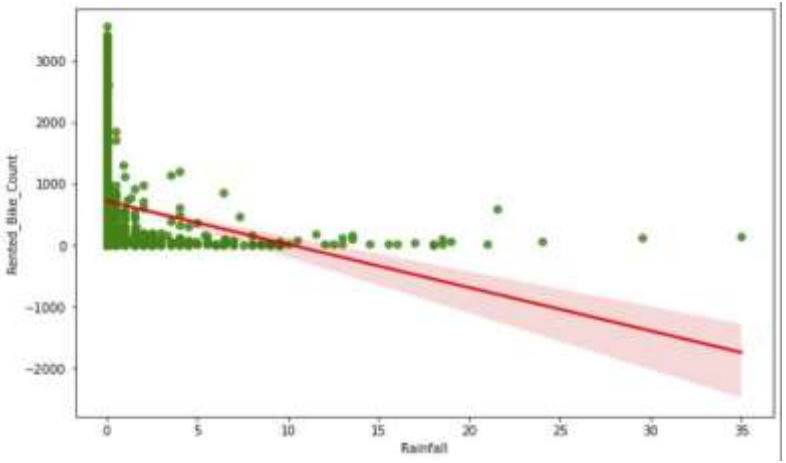
Numerical variable Vs. Rented_Bike_Count (cont.)

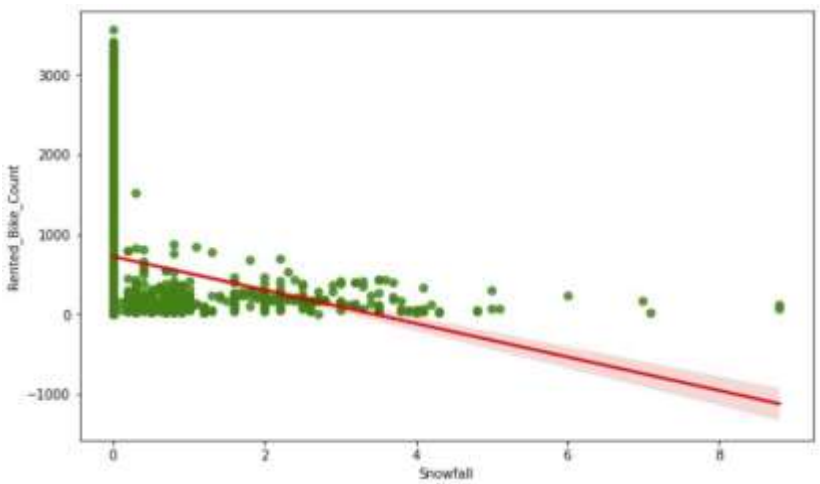




- From the above regression plot of all numerical features we see that the columns 'Visibility', 'Solar_Radiation' are positively relation to the target variable.

Numerical variable Vs. Rented_Bike_Count (cont.)





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

Multicollinearity:

Heatmap

Here we can see that Temperature and DPT (dew point Temperature) are strongly correlated. Therefore we will drop DPT column

Dew point temperature:
We know that the dew point temperature is related to temperature and can be approximated as follows:
Td = T - ((100 - RH)/5.)
Where
Td - Dew point temperature (in degrees Celsius)
T - Observed temperature (in degrees Celsius)
RH - Relative humidity (in percent)

Dummy Variable

- Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form
- The one-hot encoding creates one binary variable for each category.

The problem is that this representation includes redundancy. For example, if we know that [1, 0, 0] represents "*blue*" and [0, 1, 0] represents "*green*" we don't need another binary variable to represent "*red*", instead we could use 0 values for both "*blue*" and "*green*" alone, e.g. [0, 0].

This is called a dummy variable encoding

- Our Data contains categorical data, we must encode it to numbers before you can fit and evaluate a model.
- Here we apply dummy encoding to ['Hour', 'Seasons', 'Holiday', 'Functioning_Day', 'Weekend', 'Month'] .

VIF (Variance inflation factor)

- Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.
- After applying VIF to our data we found that there were few variable who had high VIF value.
- To deal with it we dropped “Season” column and found that the VIF values are now in range that are tolerable.

Linear Regression

Evaluating the model on train set

```
MSE : 35.07751288189292
RMSE : 5.9226271942350825
MAE : 4.4740240929968005
R2 : 0.7722101548255267
Adjusted R2 : 0.7675375426168196
```

Evaluating the model on test set

```
MSE : 33.27533089591942
RMSE : 5.768477346399084
MAE : 4.41017847531819
R2 : 0.7893518482962673
Adjusted R2 : 0.7850308605690113
```

Lasso regression

Evaluating the model on train set

```
MSE : 41.644306346281205
RMSE : 6.45323998827575
MAE : 4.9686375505438365
R2 : 0.7295660576917784
Adjusted R2 : 0.7240186947726355
```

Evaluating the model on test set

```
MSE : 40.02113965431951
RMSE : 6.326226336001543
MAE : 4.912430986283005
R2 : 0.7466477756861858
Adjusted R2 : 0.7414508069823127
```

Ridge regression

Evaluating the model on train set

```
MSE : 35.077544951265374
RMSE : 5.922629901594846
MAE : 4.474058341284995
R2 : 0.7722099465702434
Adjusted R2 : 0.7675373300896331
```

Evaluating the model on test set

```
MSE : 33.27591023508452
RMSE : 5.76852756213269
MAE : 4.4103019496797735
R2 : 0.7893481808128479
Adjusted R2 : 0.7850271178551627
```

Decision tree

Hyperparameter tuning

```
# Maximum depth of trees
max_depth = [6,8,10]

# Minimum number of samples required to split a node
min_samples_split = [50,100,150]

# Minimum number of samples required at each leaf node
min_samples_leaf = [80,90,100]

# Hyperparameter Grid
param_1 = {
            'max_depth' : max_depth,
            'min_samples_split' : min_samples_split,
            'min_samples_leaf' : min_samples_leaf}
```

set

```
# Hyperparameter tuning

decisionTree = DecisionTreeRegressor()

gridSearch_decisionTree=GridSearchCV(decisionTree,param_1,scoring='r2',cv=10)

gridSearch_decisionTree.fit(X_train_Scaled,y_train)   # Fitting Decision Tree to the Training set

best_DecisionTree=gridSearch_decisionTree.best_estimator_
```

Evaluating the model on train set

```
MSE : 37.795986207339595
RMSE : 6.14784402919752
MAE : 4.580488330713205
R2 : 0.7545566621164104
Adjusted R2 : 0.7249226817827028
```

Evaluating the model on test

```
MSE : 42.579547975448605
RMSE : 6.525300604221127
MAE : 4.741437695066741
R2 : 0.7304518741086786
Adjusted R2 : 0.7249226817827028
```

Random forest

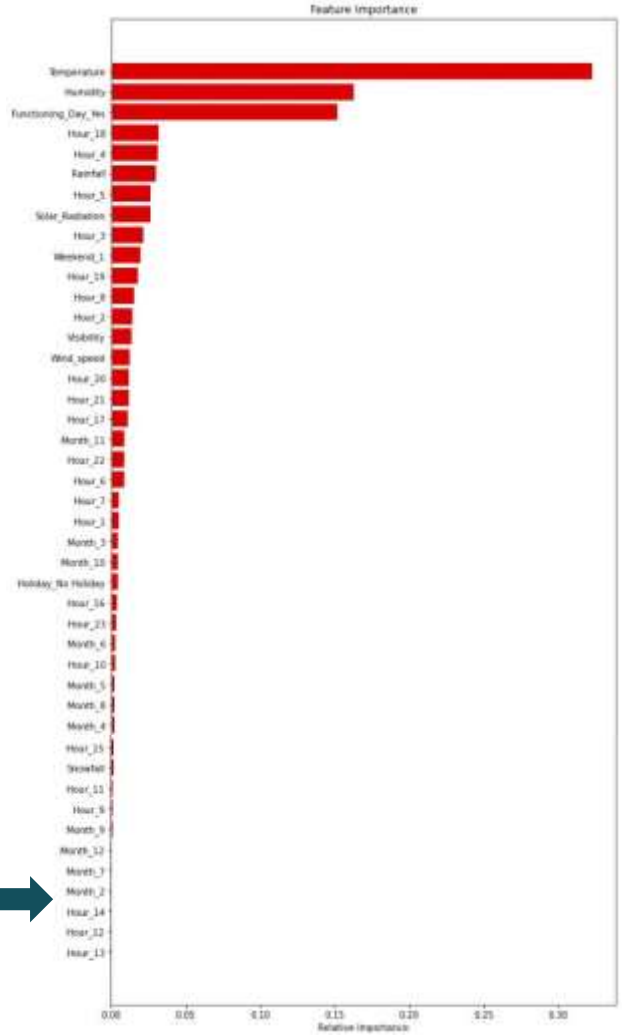
Evaluating the model on train set

```
MSE : 1.6784001995042013
RMSE : 1.29553085625322
MAE : 0.8228178871206258
R2 : 0.989100637697058
Adjusted_R2 : 0.9888770610344336
```

Evaluating the model on test set

```
MSE : 12.99177684171366
RMSE : 3.60441074819639
MAE : 2.246826528678882
R2 : 0.9177560761870608
Adjusted R2 : 0.9160690213396159
```

Feature Importance



Conclusion

- Peak hours for rented bike is between 7 to 9 am in the morning and 5 to 7 pm in the evening which suggest that the bikes are rented mostly by the office going people
- Demand of rented bikes is high during no holiday and functioning day.
- Bike Demand was high during springs and summer and autumn owning to the beautiful weather.
- We used different type of regression algorithms to train our model like, Linear Regression, Regularized linear regression (Ridge and Lasso),Random Forest Regressor and Decision tree ,where we tuned the parameters of Decision tree. Out of them Random forest regression gave the best result.