Capstone Project

Seoul Bike Sharing Demand Prediction

Team Members

Mohammad Jibran

Siddhi Thakur

Problem Statement

Currently, rental bikes are being offered in several urban areas to improve transportation comfort.

It is crucial to make the rental bikes accessible to the general public at the appropriate time because it reduces waiting time.

Eventually, maintaining a steady supply of rental bikes for the city emerges as a top priority.

Predicting the number of bikes needed to maintain a steady supply of rental bikes at each hour's interval is essential.

Content

- Data Summary
- Exploratory Data Analysis
  - Categorical Features counts with Dependent variable
  - Rented Bike Count, Hour with Respect to different categorical Feature
  - Distribution of Data
  - Regression Plot of Data
- Correlation Analysis
- Normalize Dependent Variable
- Models Performed
  - Evaluation Matrix of All the models
  - Model Evaluation & Selection
- Challenges
- Conclusion

Data Summary

- Over the recent times, bike sharing has become more and more significant. People are migrating in greater numbers to lively, greener cities where activities like bike sharing are widely accessible. Sharing bikes has several advantages, including those for the environment. It is a sustainable mode of transportation.

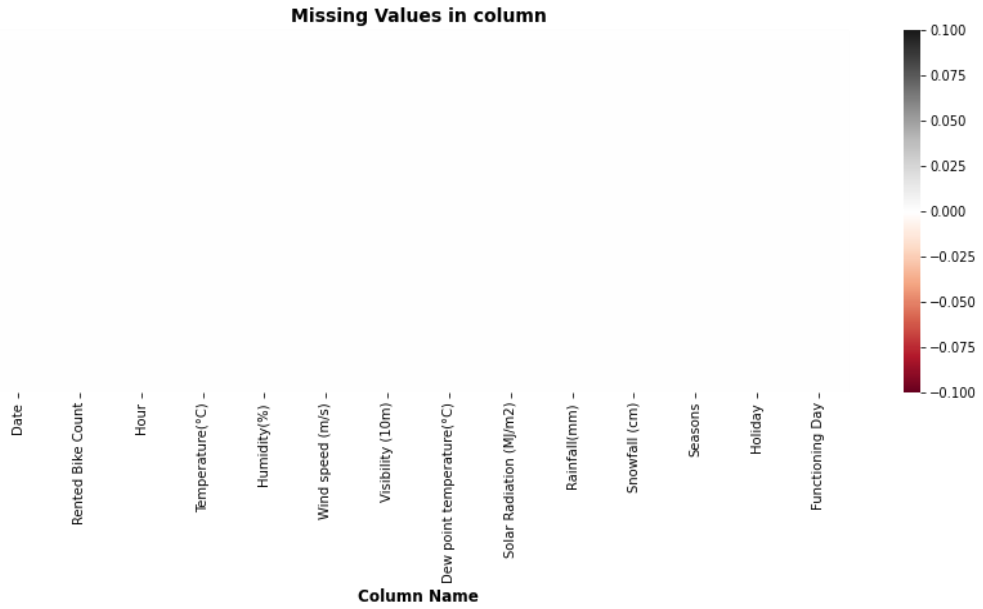
- The dataset includes rental rates for bikes per hour, date information, and weather parameters (including humidity, windspeed, visibility, dewpoint, solar radiation, snowfall, and rain).

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   Date                       8760 non-null   object 
 1   Rented Bike Count          8760 non-null   int64  
 2   Hour                       8760 non-null   int64  
 3   Temperature(°C)            8760 non-null   float64
 4   Humidity(%)                8760 non-null   int64  
 5   Wind speed (m/s)           8760 non-null   float64
 6   Visibility (10m)           8760 non-null   int64  
 7   Dew point temperature(°C)  8760 non-null   float64
 8   Solar Radiation (MJ/m2)    8760 non-null   float64
 9   Rainfall(mm)               8760 non-null   float64
 10  Snowfall (cm)              8760 non-null   float64
 11  Seasons                    8760 non-null   object 
 12  Holiday                    8760 non-null   object 
 13  Functioning Day            8760 non-null   object 
dtypes: float64(6), int64(4), object(4)
memory usage: 958.2+ KB
```

- This dataset includes the hourly and daily counts of rental bikes from the Seoul Bike Share programme for the years of 2017 and 2018, together with relevant weather and seasonal data. The dataset has 14 columns (the attributes that are being considered) and 8760 rows (every hour of every day for 2017 and 2018).

- The dataset does not contain any missing entries or duplicate values.



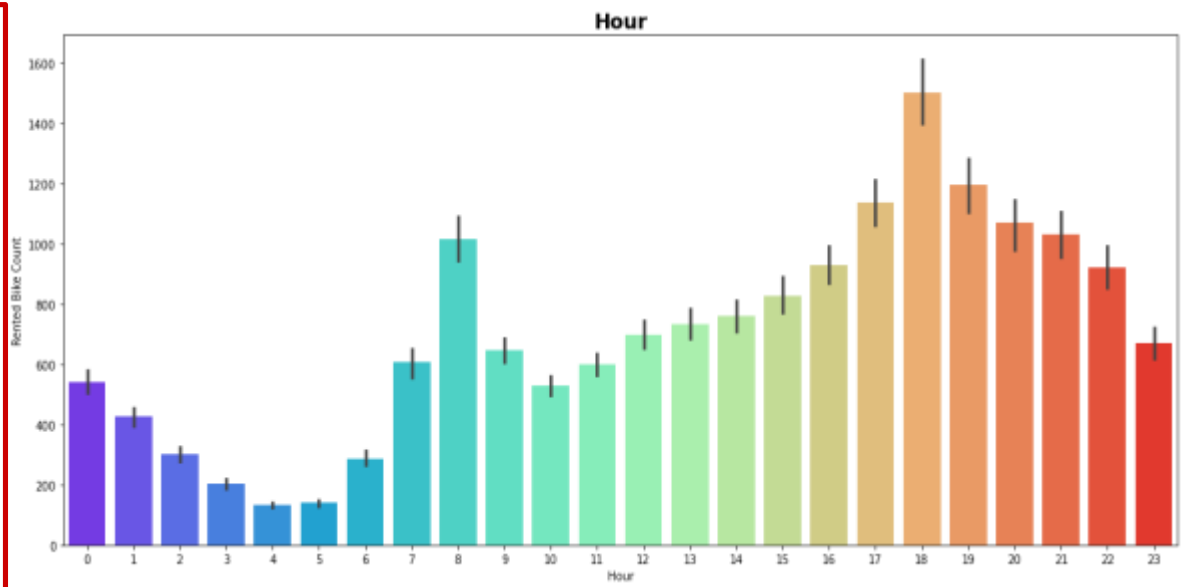
Exploratory Data Analysis

Categorical Features with Dependent

Hour

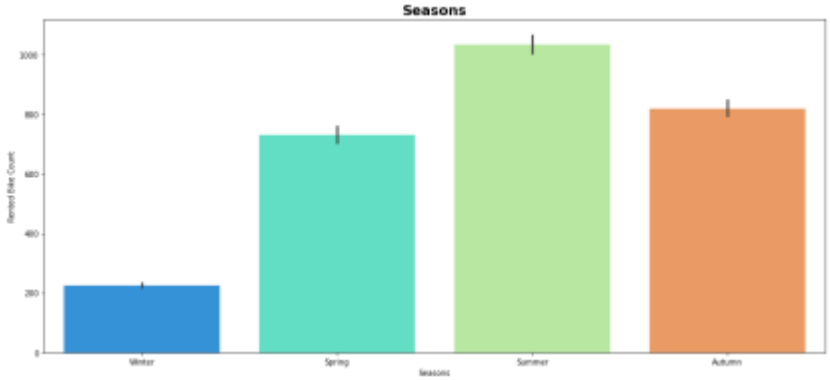
- The bar plot reveals that from 8 hours to 19 hours, the demands for the bikes are practically incremental before they start to decline.

- Additionally, the morning and evening peaks are typically those that mark the beginning and end of office hours.



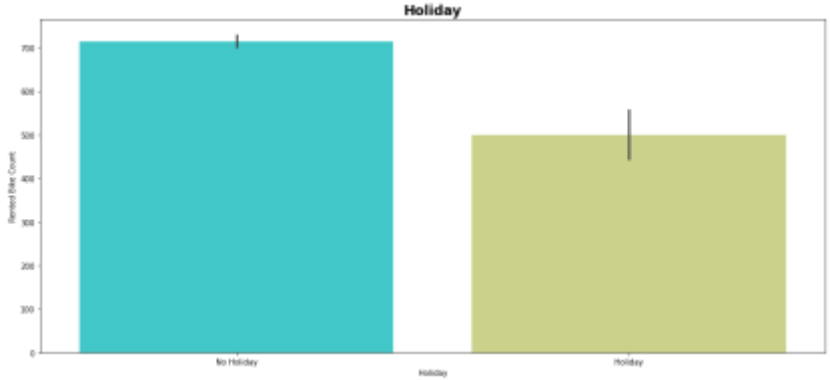
Seasons

- Bikes are most in demand throughout the summer, followed by the fall and spring.
- Winter is the least popular seasons for bike rental.



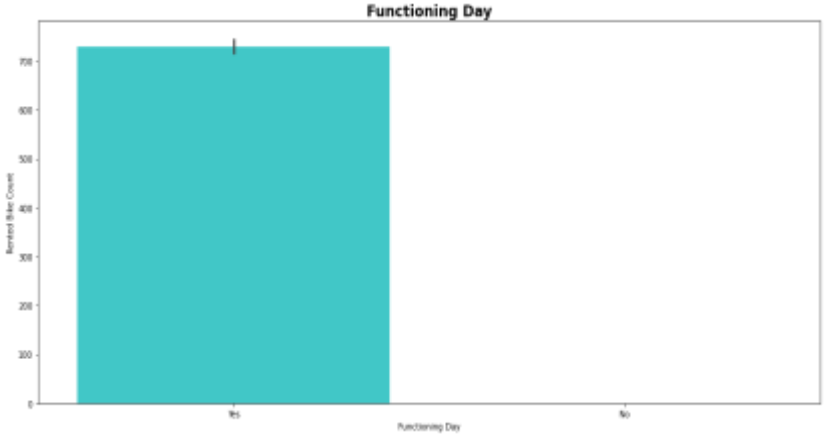
Holiday

- Slightly higher demand of bikes during non holidays.
- People going to their offices could be the reason for the same.



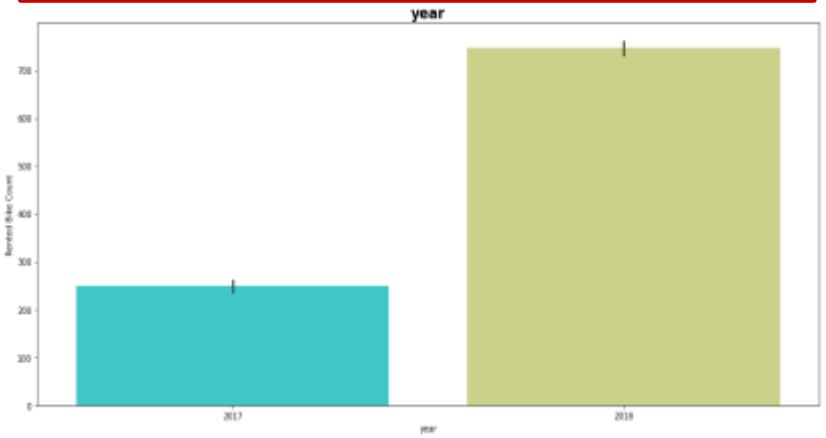
Functioning Day

- There is no demand on Non – Functioning day.



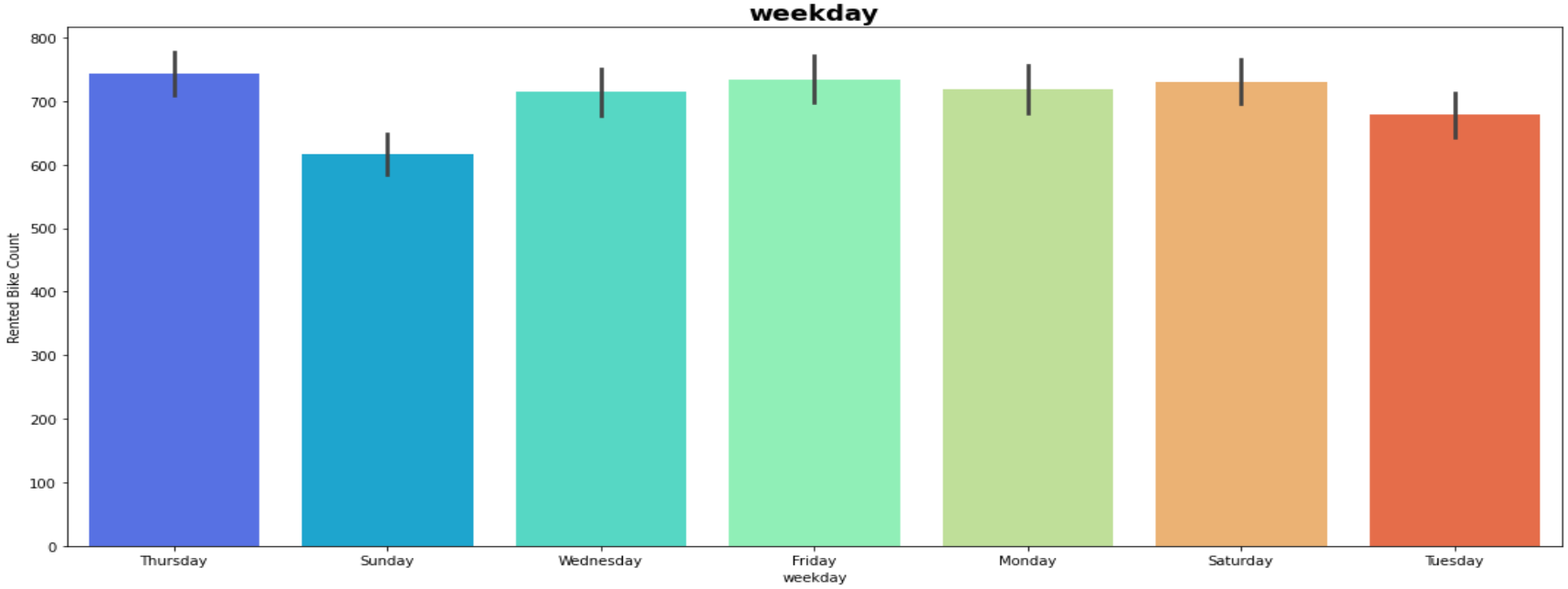
Year

- When compared to 2017, the demand for bike rentals is higher in 2018.
- It can be because the service is new in 2017.



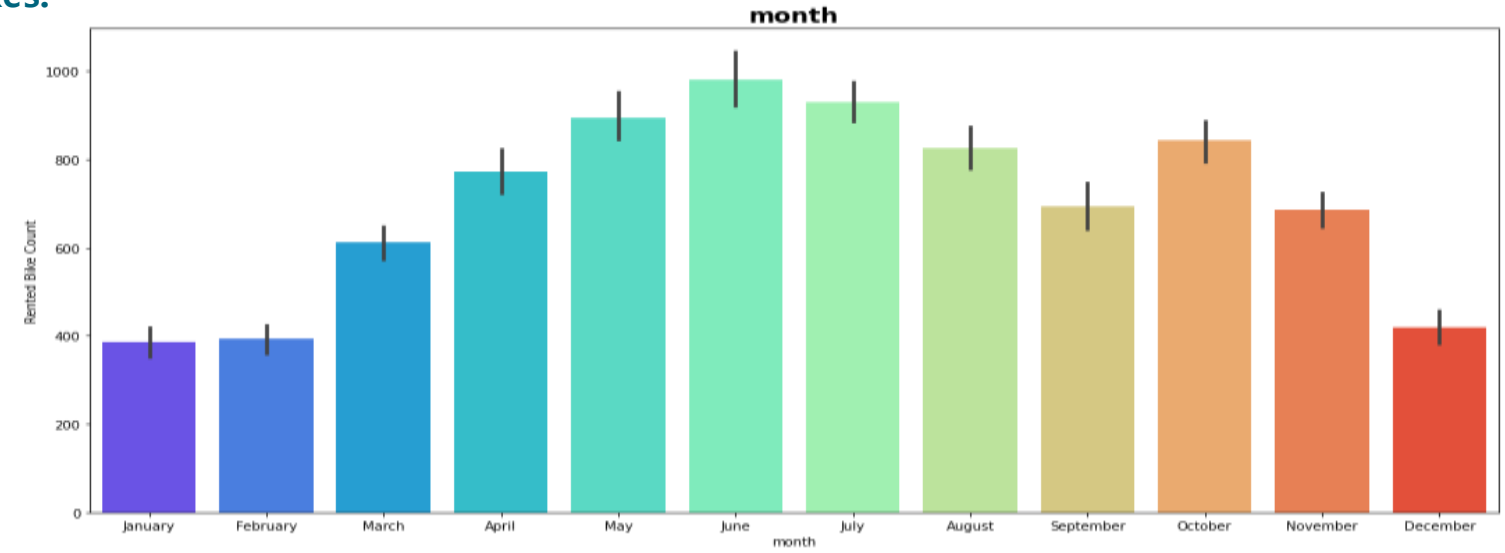
Weekday

- The demand is consistent throughout the week, with a slight decline on Sundays i.e holiday.



Months

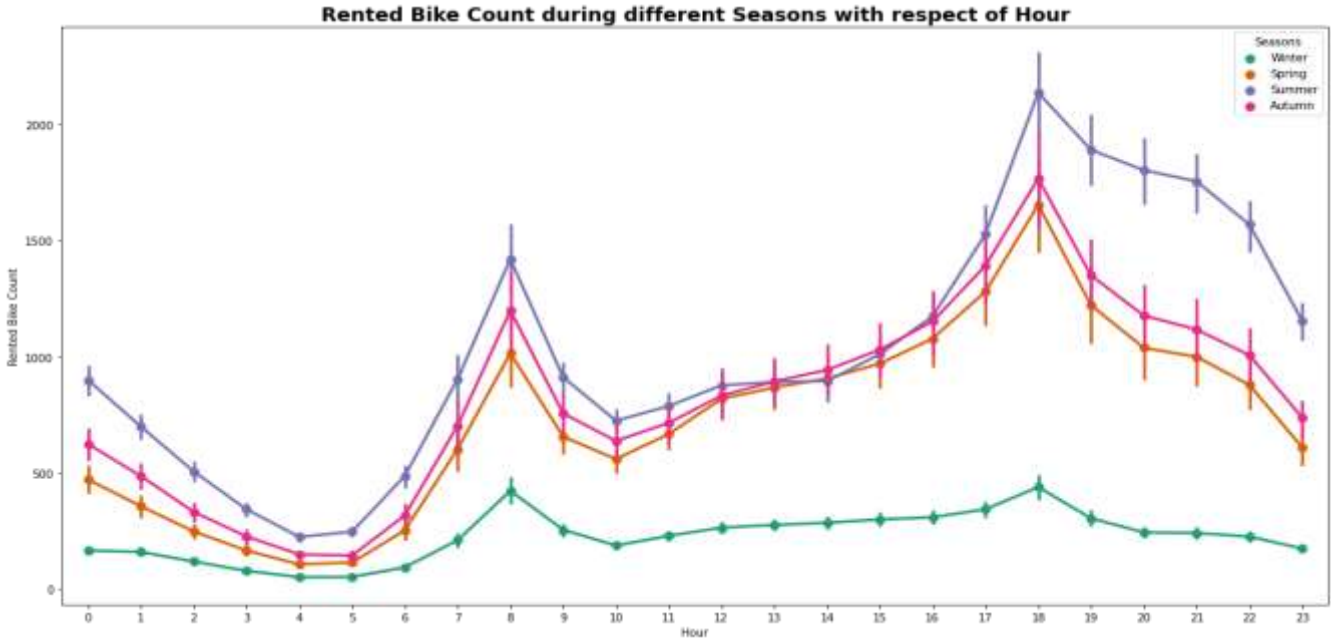
- We can see that the winter months of December, January, and February have lower demand for rented bikes.
- Additionally, the summer months of April, May, June, and July see the highest demand for rental bikes.



Rented Bike Count and Hour with Respect to different categorical Feature

★ During different Seasons

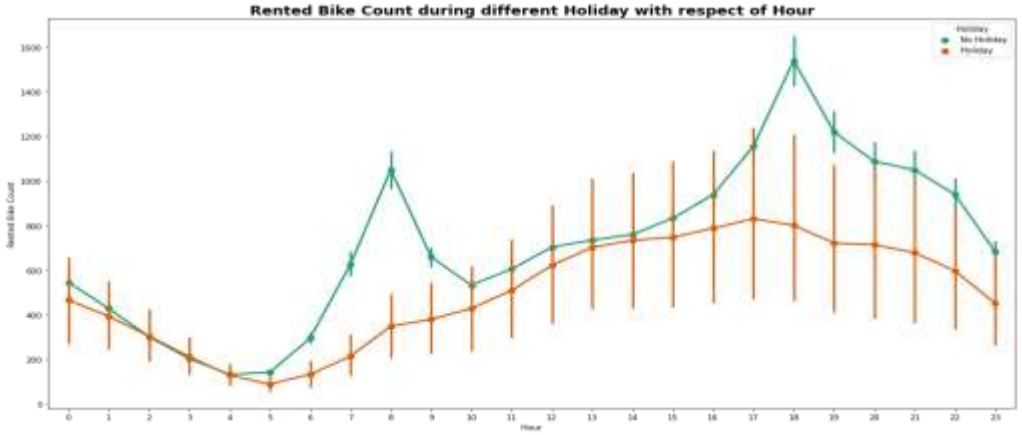
- In the season column, we are able to understand that the demand is low in the winter season.



★During different Holiday

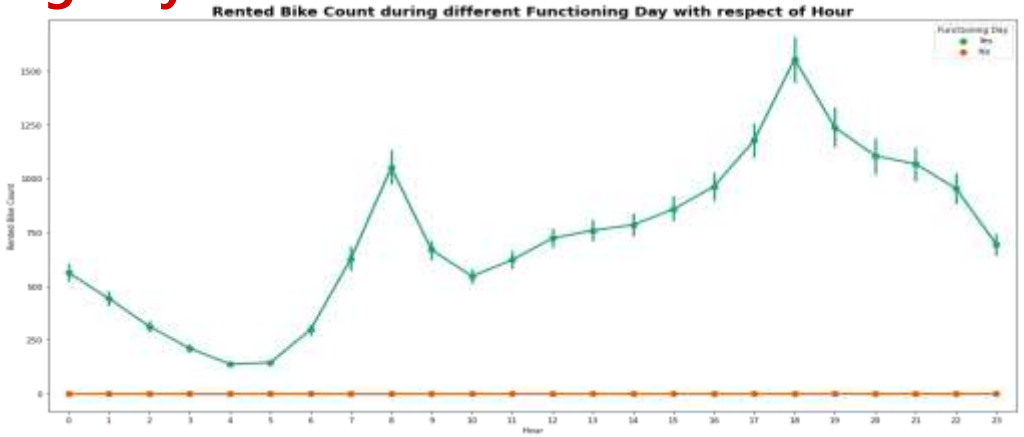
- In the Holiday column,

the demand is low during holidays, but in no holidays the demand is high, it may be because people use bikes to go to their work.



★During different Functioning day

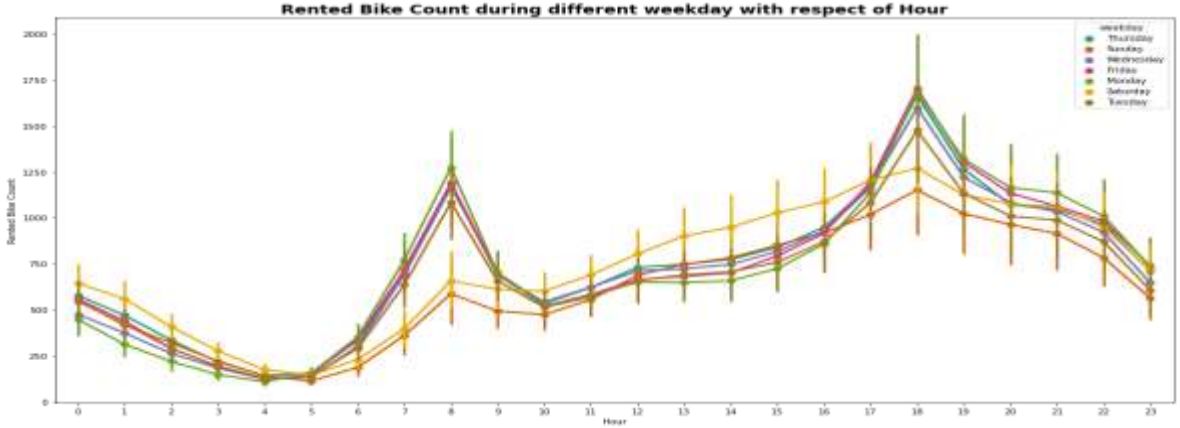
- In the Functioning Day column, if there is no Functioning Day then there is no demand.

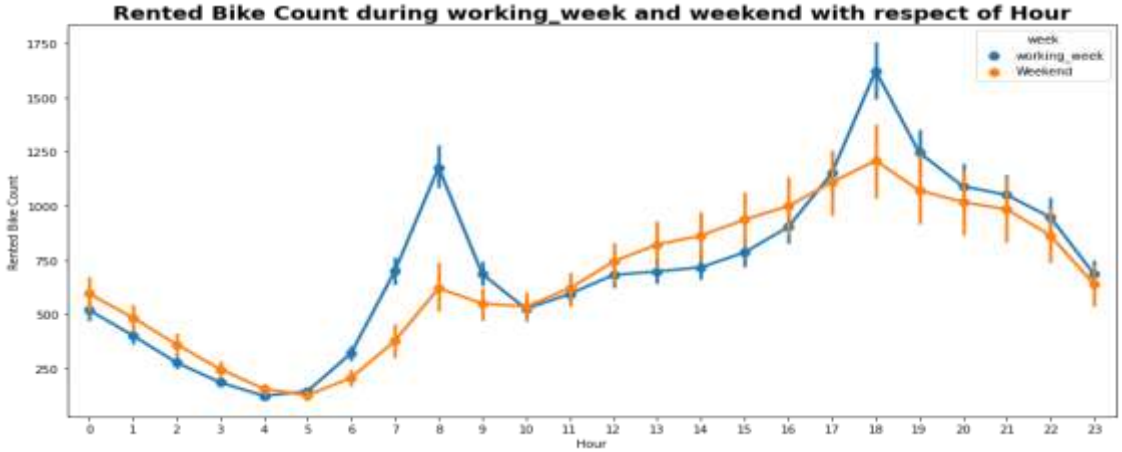


★ During different days of Week

- In the Days of week column, We can observe from this column that the pattern of weekdays and weekends is different, in the weekend the demand becomes high in the afternoon. While the demand for office timings is high during weekdays, we can further change this column to weekdays and weekends.

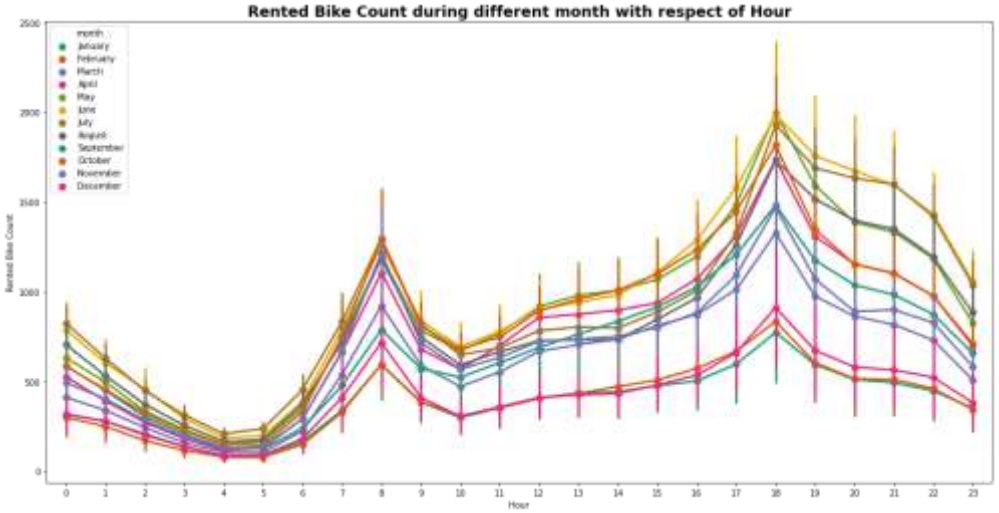
- Demand is high in the afternoon on the weekend, While there is more demand during office hours in weekdays.





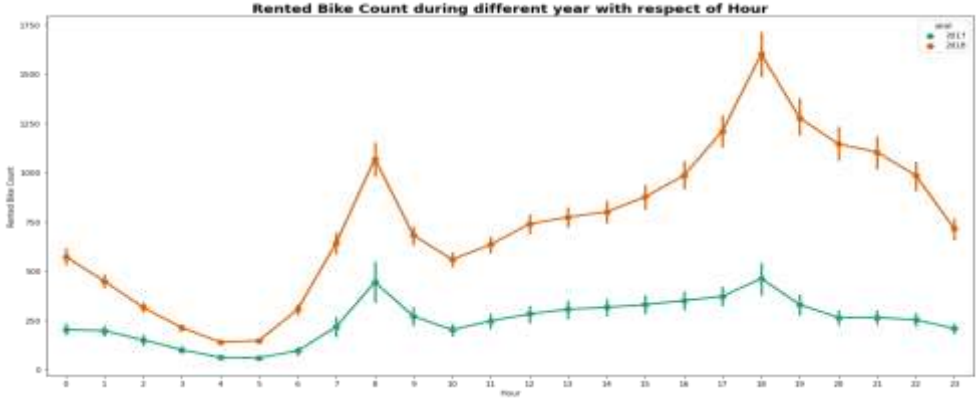
★ During different Months

- In the month column, We can clearly see that the demand is low in December January & February, It is cold in these months and we have already seen in season column that demand is less in winters.

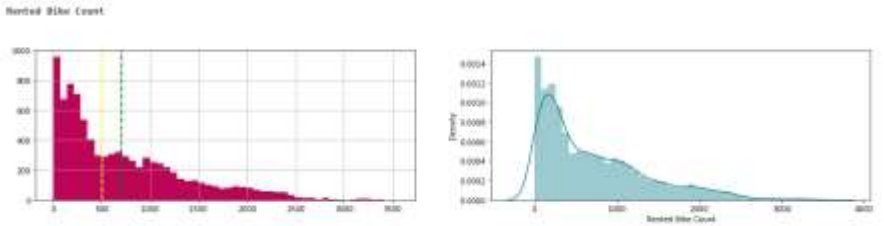


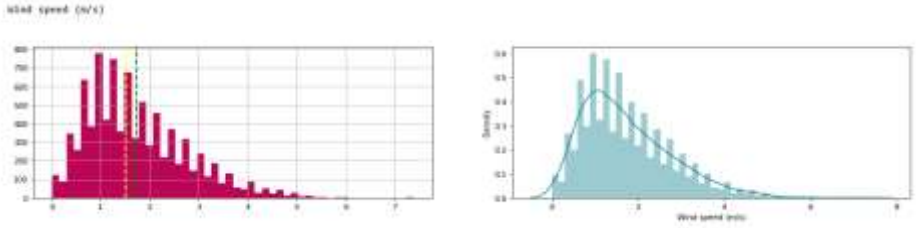
★ During different Years

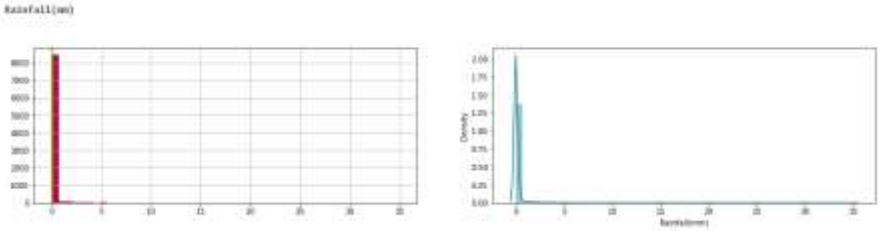
- The demand was less in 2017 and higher in 2018, it may be because it was new in 2017 and people did not know much about it.

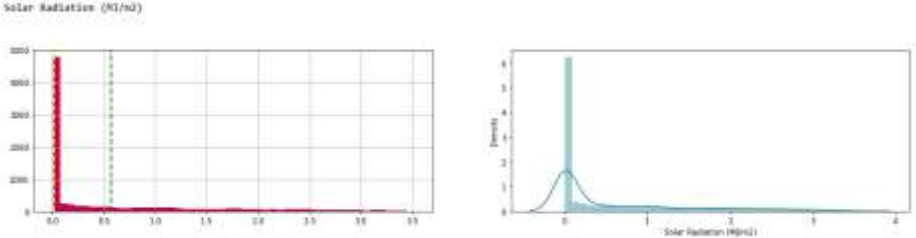


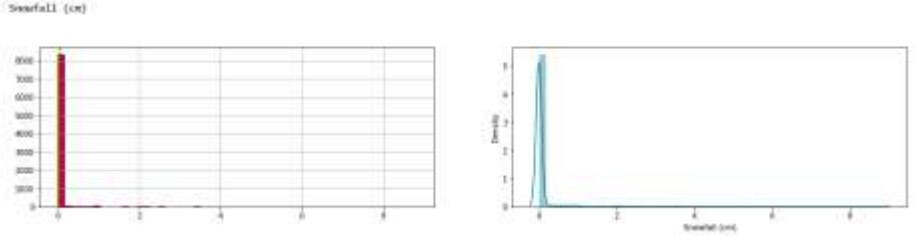
Visualizing Distribution of Data











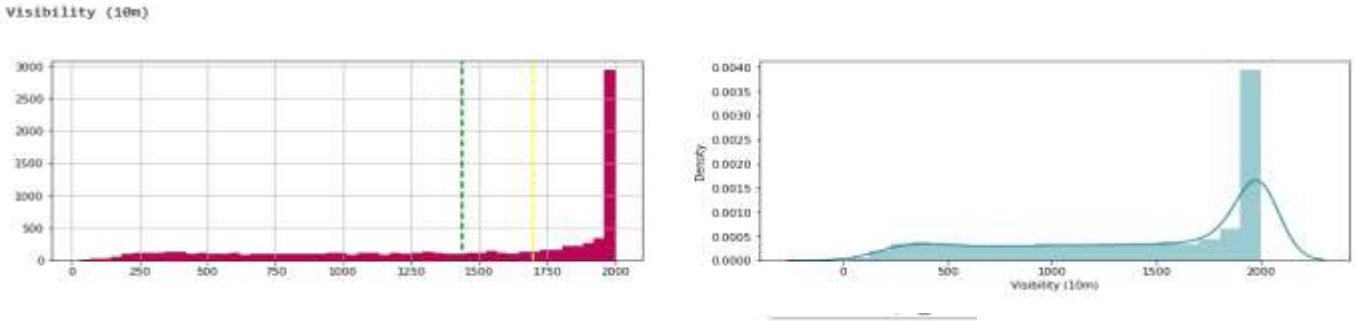
Right skewed columns are -

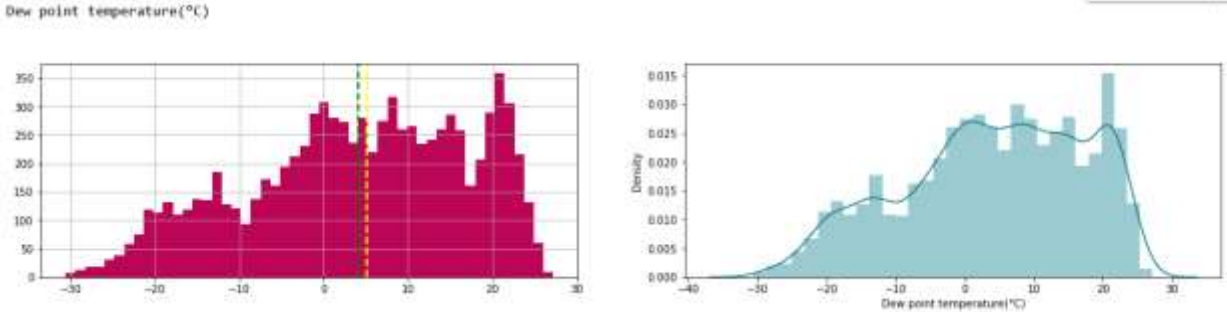
Rented Bike Count (Its also our Dependent variable), Wind speed (m/s), Solar Radiation (MJ/m2), Rainfall(mm), Snowfall (cm),

Left skewed columns are -

Visibility (10m), Dew Point Temperature (°C)

Also, from the Histogram we are coming to know that the features which are skewed, their mean and the median are also skewed.

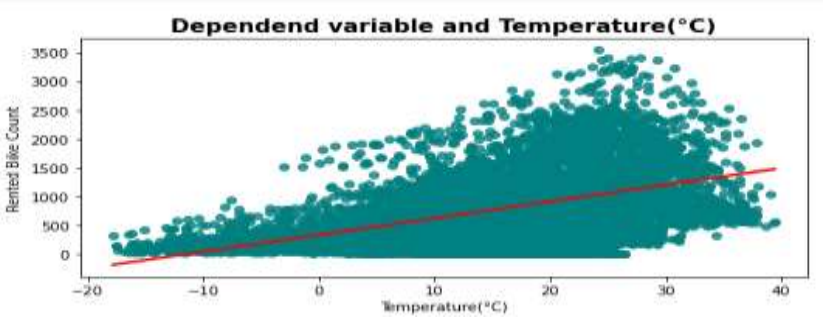


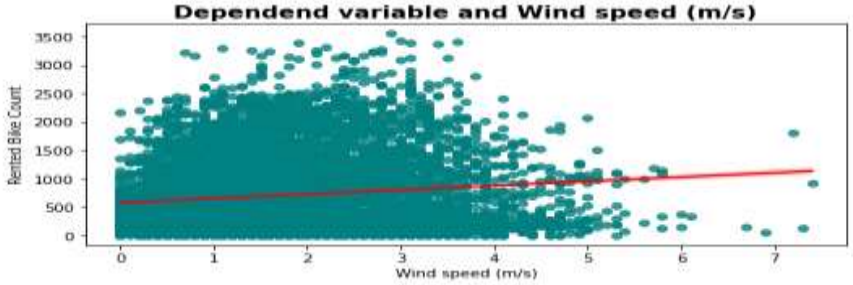


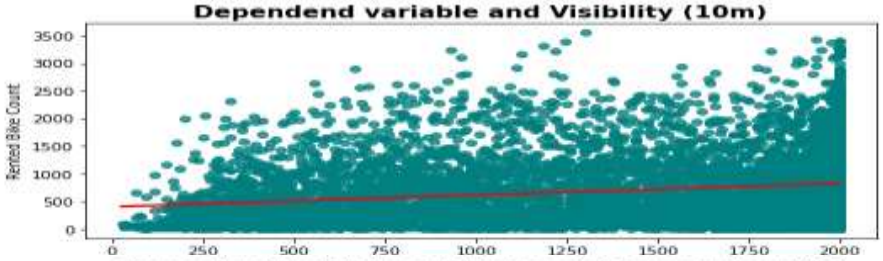
Regression Plot of Data

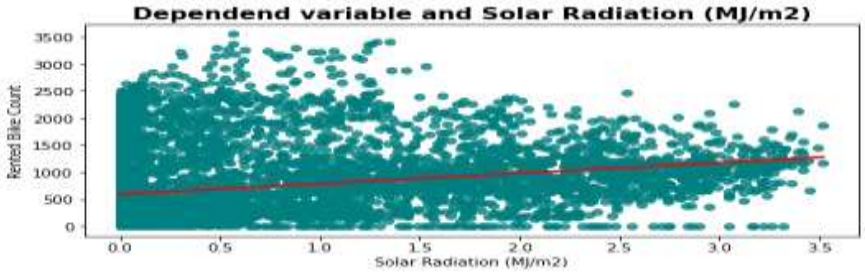
Positive Linear columns are –

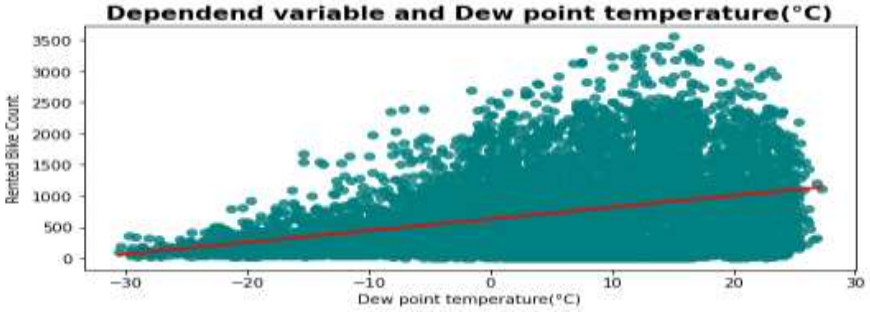
- Temperature, Visibility, Wind Speed, Dew Point and Solar Radiation's regression plots shows positive linear in relation to our target variable.





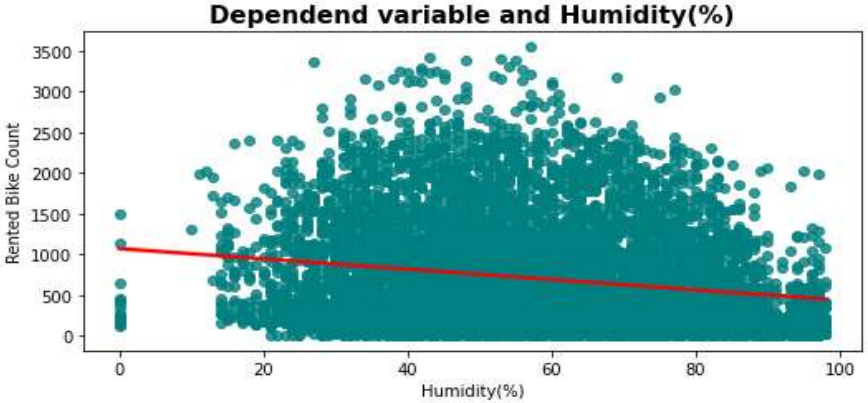


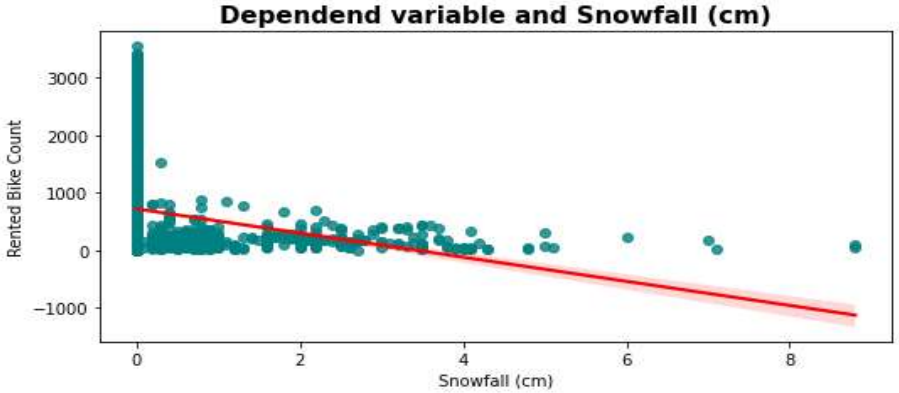


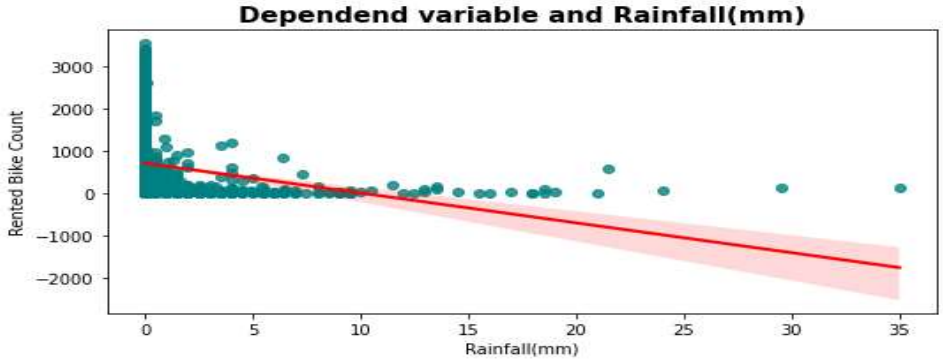


Negative Linear columns are –

- Humidity, Snowfall and Rainfall's regression plots shows Negative linear in relation to our target variable.

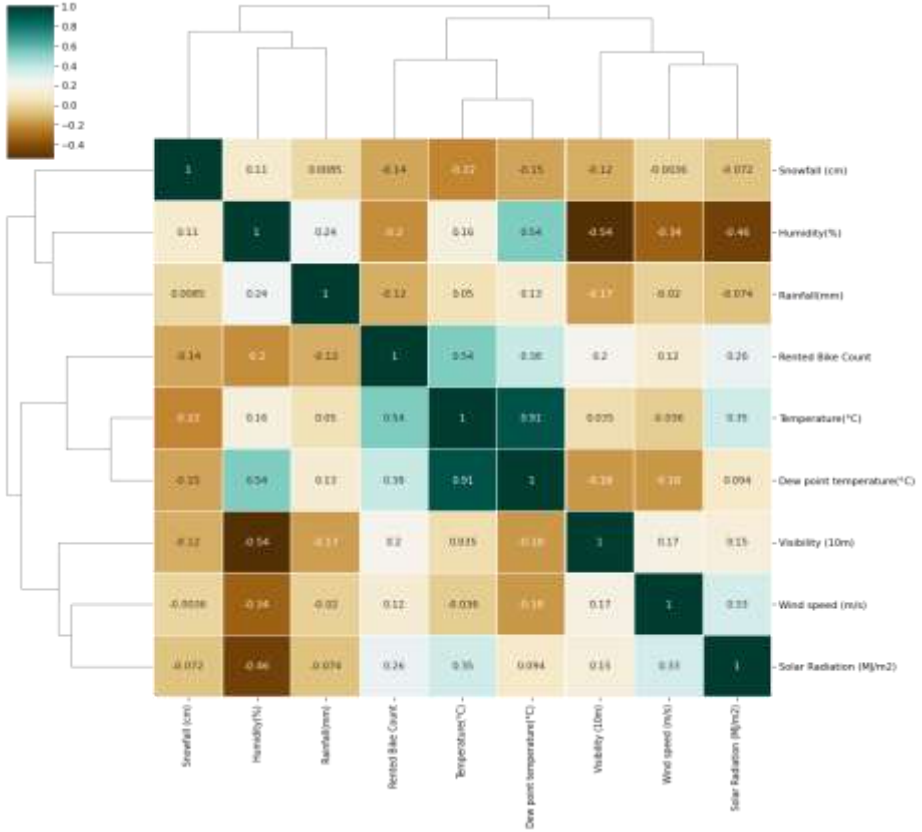






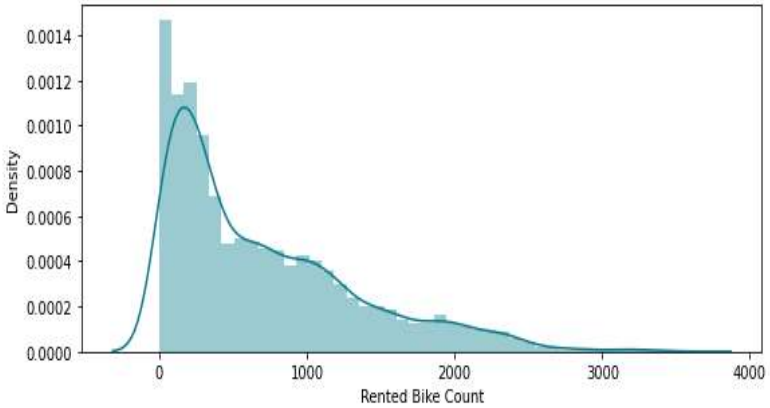
Correlation Analysis

- We observed a strong association between Dew Point temperature & Temperature from the correlation graph.
- Then we looked at VIF and came to the conclusion that these two features are also affecting VIF score, so we made the decision to remove one of these features.
- To do this, we looked at which feature was least correlated with the dependent variable and found that it was Dew point temperature, so we removed it.

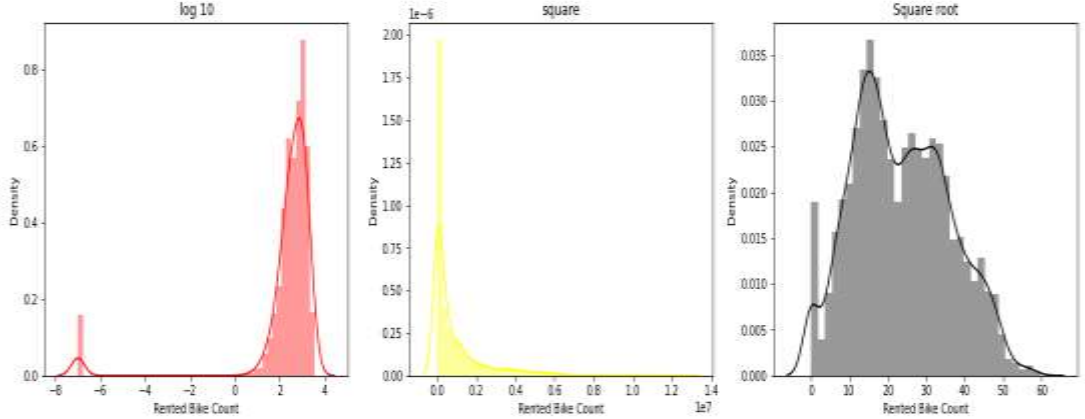


Normalize Dependent Variable

BEFORE TRANSFORMATION



AFTER TRANSFORAMTION



BEFORE TRANSFORMATION - Our dependent variable was right Skewed.

AFTER TRANSFORMATION - Our dependent variable in Black plot is normalized to some extent : So we will go with square root on our dependent variable.

Models Performed

List of Models

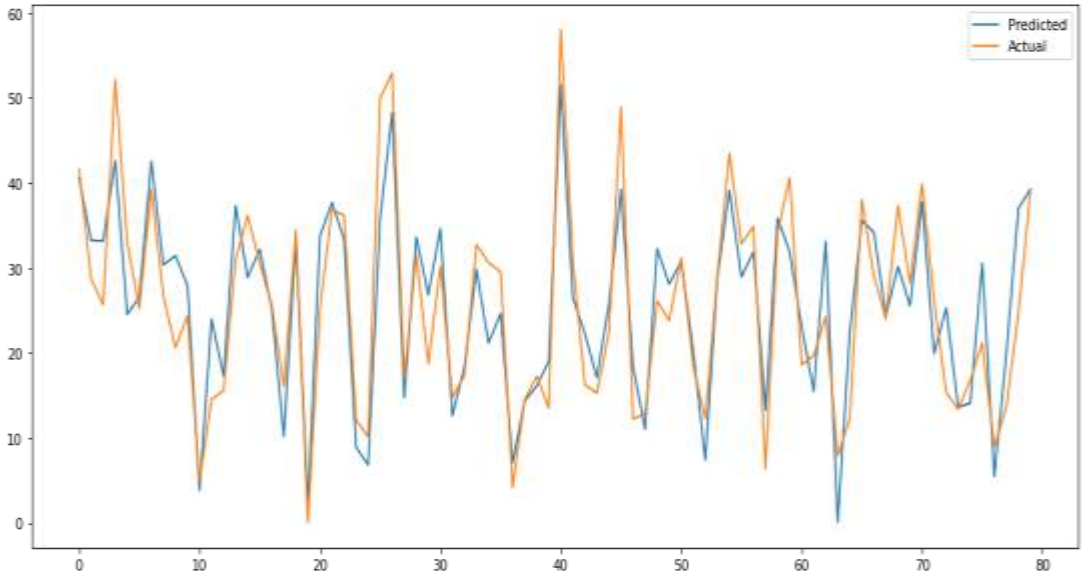
- Linear Regression with regularizations (Lasso & Ridge)
- Polynomial Regression
- Decision Tree
- Random Forest Regressor

Linear Regression with Regularization

- **Simple Linear Regression is giving us 79% Accuracy on Training dataset while on Test dataset R Square score is 0.762 and Adjusted R Square 0.756.**

```
matrix_score(Linear,X_train,X_test,y_train,y_test)
```

```
Training score  = 0.7902269075496007

Best Parameters & Best Score

None

Evalution Matrix

MAE : 210.6768562260802
MSE : 97363.04473354678
RMSE : 312.0305189136902
R2 : 0.7620654672586054
Adjusted R2 :  0.756731110616108
```



Lasso

```
matrix_score(L1,X_train,X_test,y_train,y_test)
```

```
Training score  = 0.7902086120456122

Best Parameters & Best Score

The best parameters found out to be :{'alpha': 0.0014}
where model best score is:  0.7863509105820757

Evalution Matrix

MAE : 210.98378286630376
MSE : 97743.31929081825
RMSE : 312.6392798271168
R2 : 0.761136157279183
Adjusted R2 :  0.7557809660364931
```

Ridge

```
matrix_score(L2,X_train,X_test,y_train,y_test)
```

```
Training score  = 0.7902221237981176

Best Parameters & Best Score

The best parameters found out to be :{'alpha': 0.5}
where model best score is:  0.7862861209197668

Evalution Matrix

MAE : 210.78508169126954
MSE : 97501.6328762928
RMSE : 312.2525146036342
R2 : 0.7617267873716111
Adjusted R2 :  0.7563848377190363
```

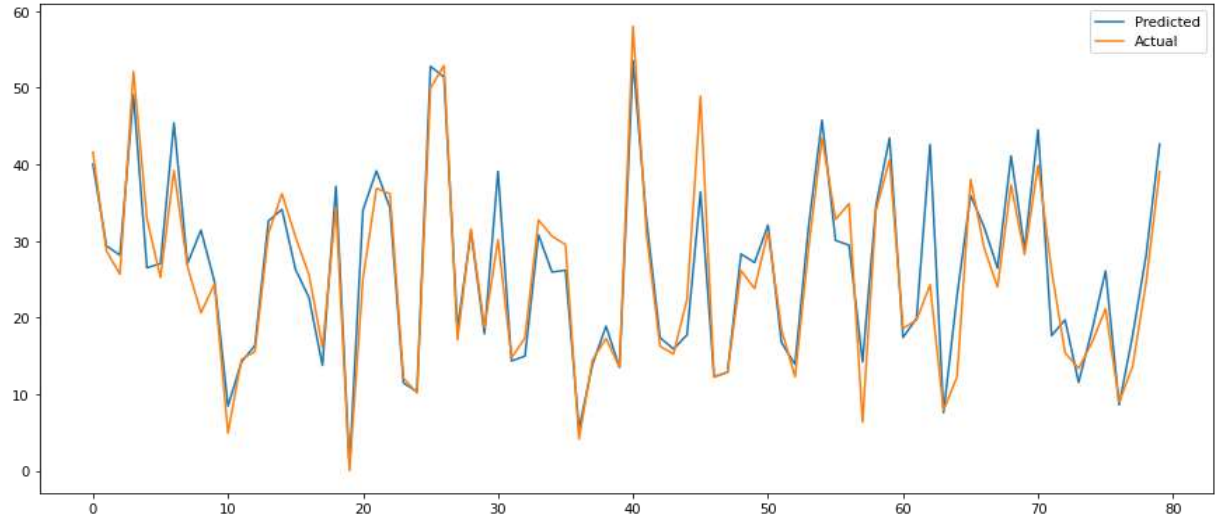
- **With this we can conclude that simple Linear regression as well as Linear Regression with Regularization (Lasso & Ridge) providing us almost similar results which we can visualize with Evaluation matrix.**

Polynomial Regression

```
matrix_score(Linear, poly_X_train,poly_X_test,y_train,y_test)
```

```
Training score  = 0.9212305400305827

Best Parameters & Best Score

None

Evalution Matrix

MAE : 135.61254379577196
MSE : 50477.57812303929
RMSE : 224.67215698221105
R2 : 0.8766435561101101
Adjusted R2 :  0.7198887389263808
```

- Polynomial of degree 2 is giving us 92% accuracy on testing datasets while R Square and Adjusted R Square are 0.87 and 0.71 respectively. Lowest Adjusted R2 among all.

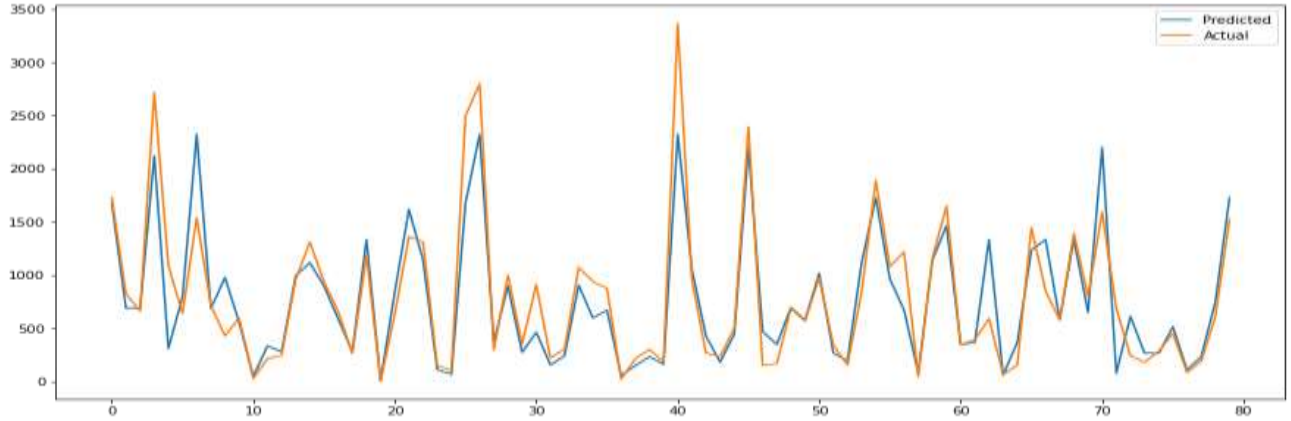


Decision Tree

- **Decision Tree Model providing us 88% accuracy with Training datasets and best parameters (max depth- 25, criterion- MSE, min sample leaf- 5 and min sample split- 35) with Best Model score of 0.82.**

```
matrix_score(Dt_grid_search,X_train,X_test,y_train,y_test)
```

```
Training score  = 0.8807688497254179

Best Parameters & Best Score

The best parameters found out to be :{'criterion': 'mse', 'max_depth': 25, 'max_features': 'auto', 'min_samples_leaf': 5, 'min_samples_split': 35}
where model best score is:  0.825471774962945

Evalution Matrix

MAE : 166.4147006664265
MSE : 72640.25604207265
RMSE : 269.5185634461431
R2 : 0.8224826942616074
Adjusted R2 :  0.8185028574211389
```



Random Forest Regressor

- From Evaluation matrix, we can see Random Forest regressor giving us 91% accuracy with Training datasets and best parameters are (max depth- 25, max features- Auto, min sample leaf- 2, min sample split- 20 and estimators – 100) with Best Model score of 0.86.

```
matrix_score(Ranom_forest_Grid_search,X_train,X_test,y_train,y_test)
```

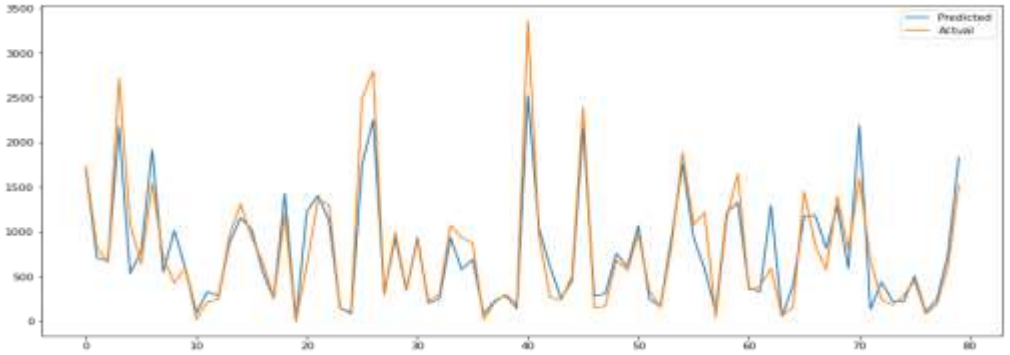
```
Training score  = 0.9177937685104807
```

Best Parameters & Best Score

```
The best parameters found out to be :{'max_depth': 25, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 20, 'n_estimators': 100}
where model best score is:  0.8619391165071102
```

Evalution Matrix

```
MAE : 144.97433432853114
MSE : 56589.28788304904
RMSE : 237.88503080910542
R2 : 0.8617078398947986
Adjusted R2 :  0.8586074084678721
```

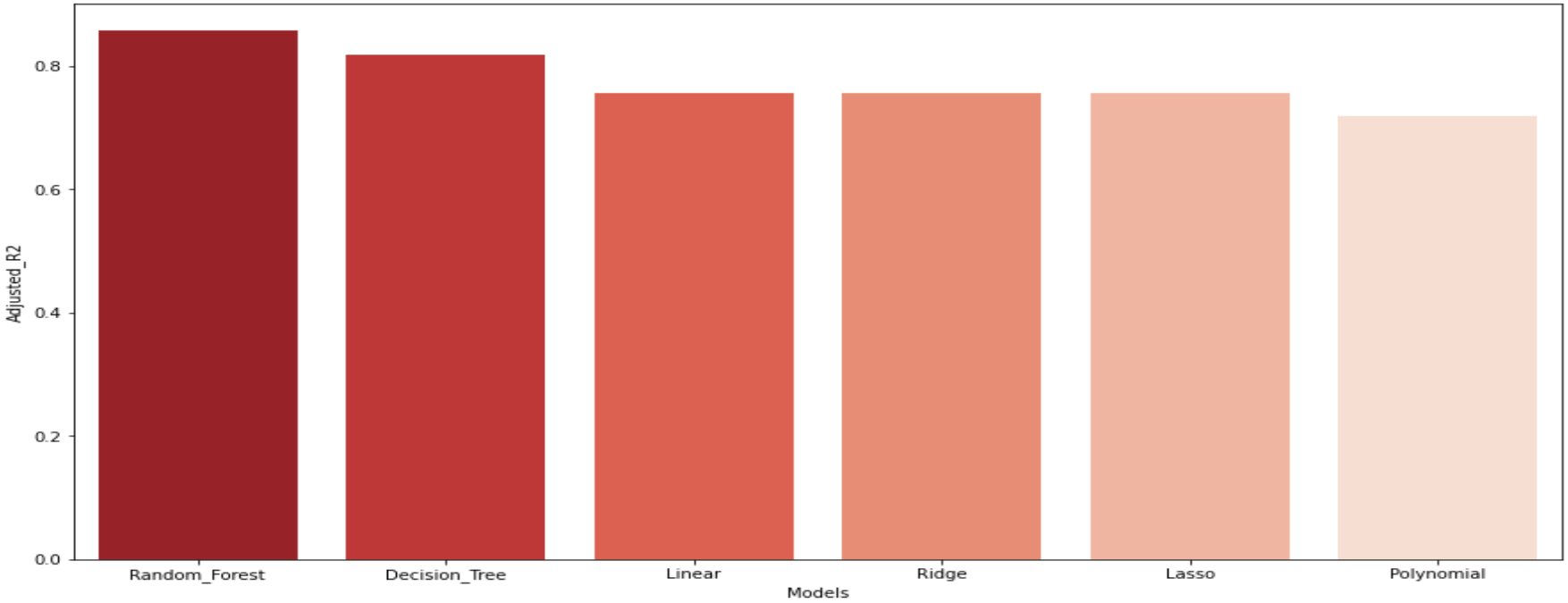


All Model Evaluation with Hyperparameter Tuning

	Models	Mean_Absolute_error	Mean_square_error	Root_Mean_square_error	Training_score	R2	Adjusted_R2
0	Random_Forest	144.974334	56589.287883	237.885031	0.917794	0.861708	0.858607
1	Decision_Tree	166.414701	72640.256042	269.518563	0.880769	0.822483	0.818503
2	Linear	210.676856	97363.044734	312.030519	0.790227	0.762065	0.756731
3	Ridge	210.785082	97501.632876	312.252515	0.790222	0.761727	0.756385
4	Lasso	210.983783	97743.319291	312.639280	0.790209	0.761136	0.755781
5	Polynomial	135.612544	50477.578123	224.672157	0.921231	0.876644	0.719889

- As seen in the Model Evaluation Table Random Forest is the best Model followed by Decision Tree.
- Linear Regression with Ridge and Lasso providing almost same results.
- Polynomial Regression is good with Training set but not good with Testing datasets.

Adjusted R-Square Score with respect to Models



Challenges

- A significant amount of data had to be handled when working on the project, and even the smallest inferences had to be taken into consideration.
- Due to the size of the data collection, there was a significant increase in computing time.

Conclusion

- We observed that the number of bicycle rentals is higher on weekdays than on weekends.
- The peak rental bike counts occur at 8 AM in the morning and 6 PM in the evening. We can observe an upward trend from 5 AM to 8 AM; the graph reaches its top at 8 AM and subsequently experiences a decrease. After that, the demand gradually rises until 6 o'clock, peaks at that time, and then gradually declines till midnight.
- We observed that people prefer to rent bikes at moderate to sunny temperature, and even when it is little windy.
- It has been noted that the summer and autumn seasons have the highest rates of bicycle rentals, while the winter season has the lowest rates.
- We found that the number of bike rentals is highest on clear days and lowest on days when it is snowing or raining.
- When RMSE and Adjusted R2 of all the models are compared, Random Forest Regressor has the greatest score, with an R2 score of 0.86 and a Training score of 0.92, making it the best model for predicting the daily number of bike rentals.

Thank You!