Capstone Project : 2

Bike Sharing Demand Prediction

TEAM MEMBERS

Team Member

Atul Chouhan

Content

1. ❑ **Understanding Business Problem**
2. ❑ **Dataset Information**
3. ❑ **Feature Analysis**
4. ❑ **Exploratory Data Analysis**
5. ❑ **Data Pre-processing**
6. ❑ **Model Implementing**
7. ❑ **Challenges**
8. ❑ **Conclusions**

Understanding Business Problem

-> Bike rentals have became a popular service in recent years and it seems people are using it more often.

-> “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”

-> Target is to maximize the availability of bikes to the customer. And minimise the time of waiting to get a bike on rent

Dataset Information

This Dataset contains 8760 lines and 14 columns.
Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.
One Datetime features ‘Date’.
We have some numerical type 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

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	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

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 - %

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 : Non Functional day or Functional day

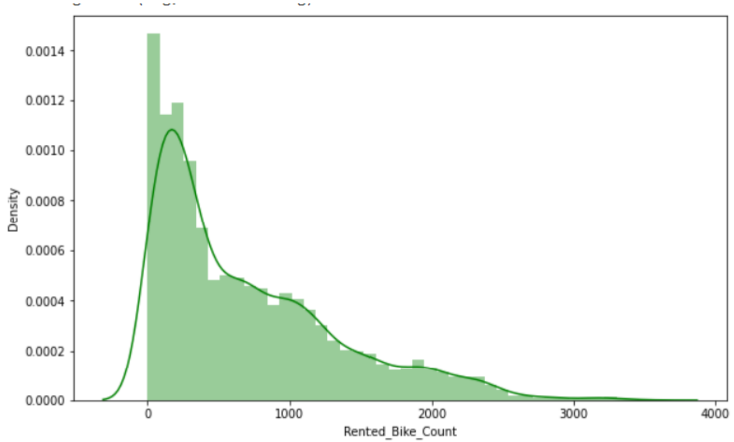
Insights From Our Dataset

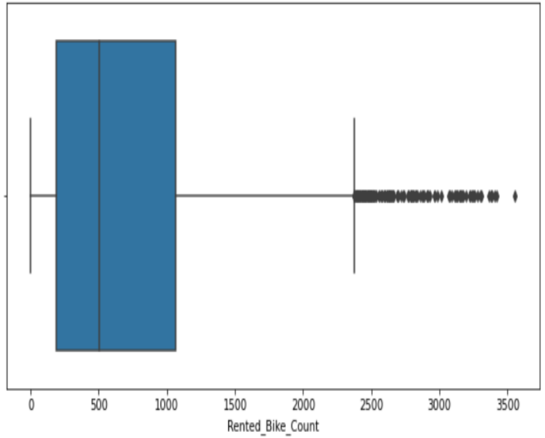
- ➢ There are No Missing Values present
- ➢ There are No Duplicate values present
- ➢ There are No null values.
- ➢ And finally we have 'rented bike count' variable which we need to predict for new observations
- ➢ The dataset shows hourly rental data for one year (1 December 2017 to 31 November(2018)(365 days).we consider this as a single year data
- ➢ So we convert the "date" column into 3 different column i.e. "year","month","day".

```
 #    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)
```

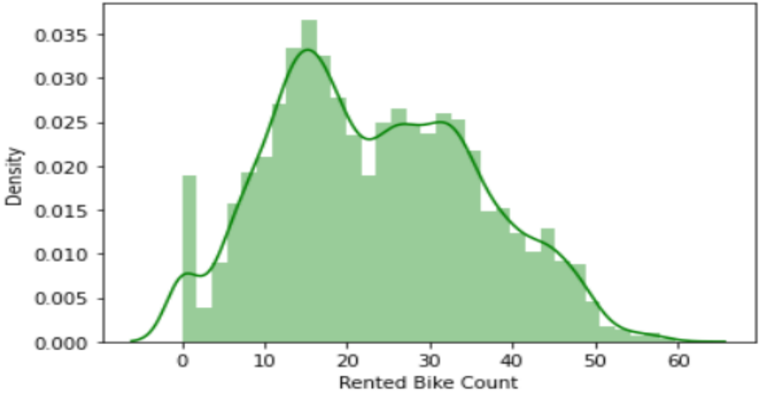
Data Distribution Of Dependent Variable

Original data → distribution

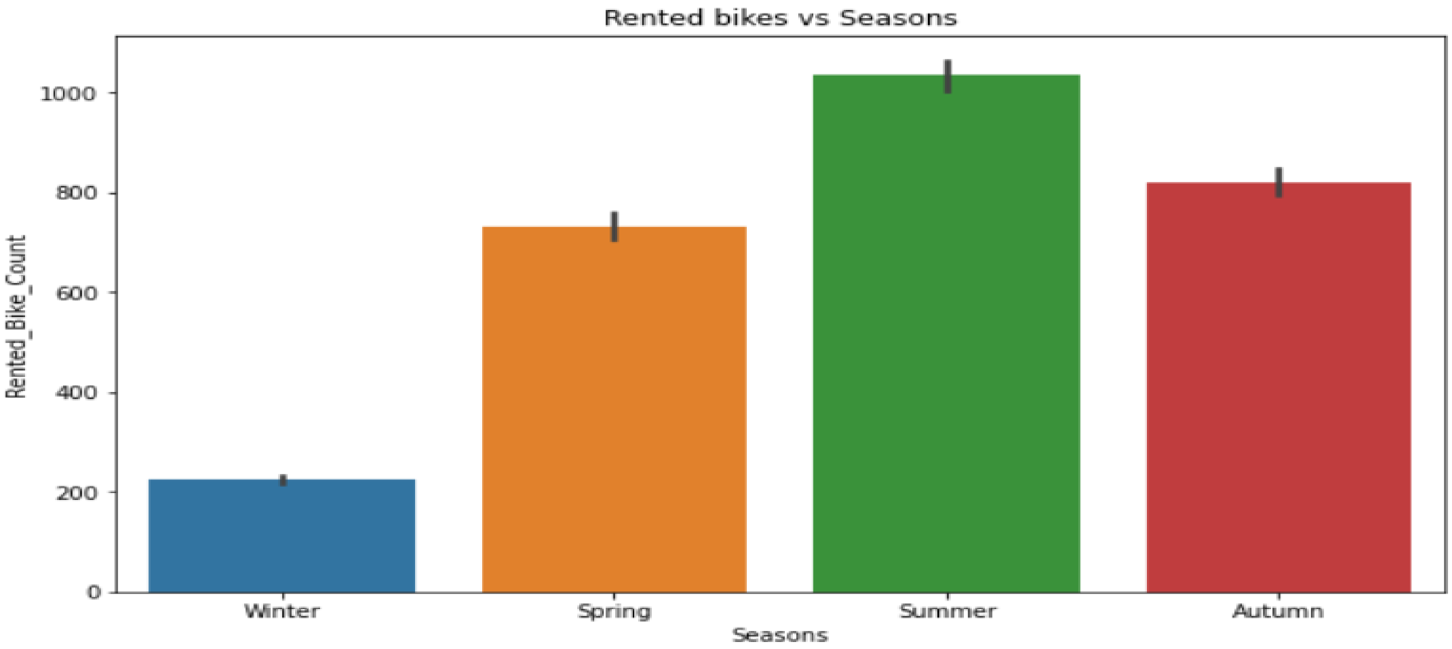




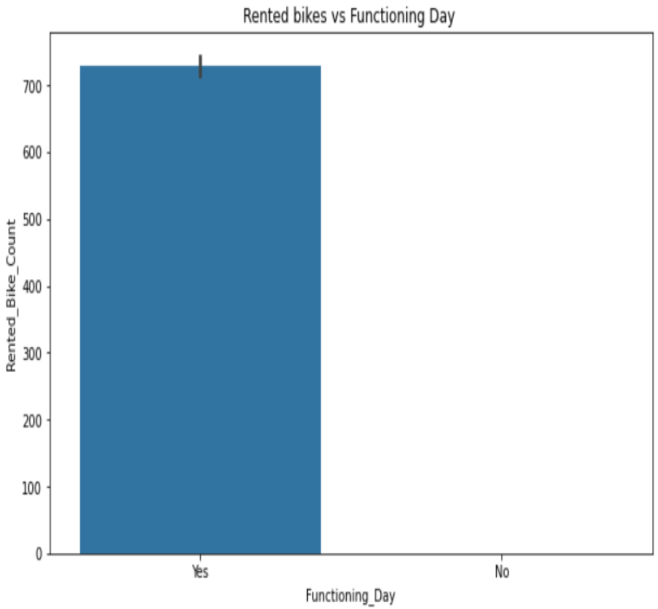
Transformed data→ distribution

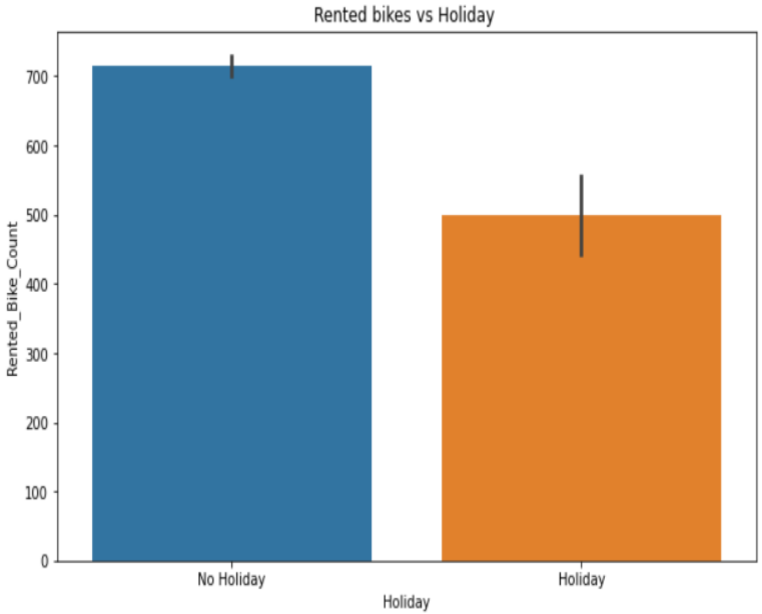


Analysis Of Season Variable

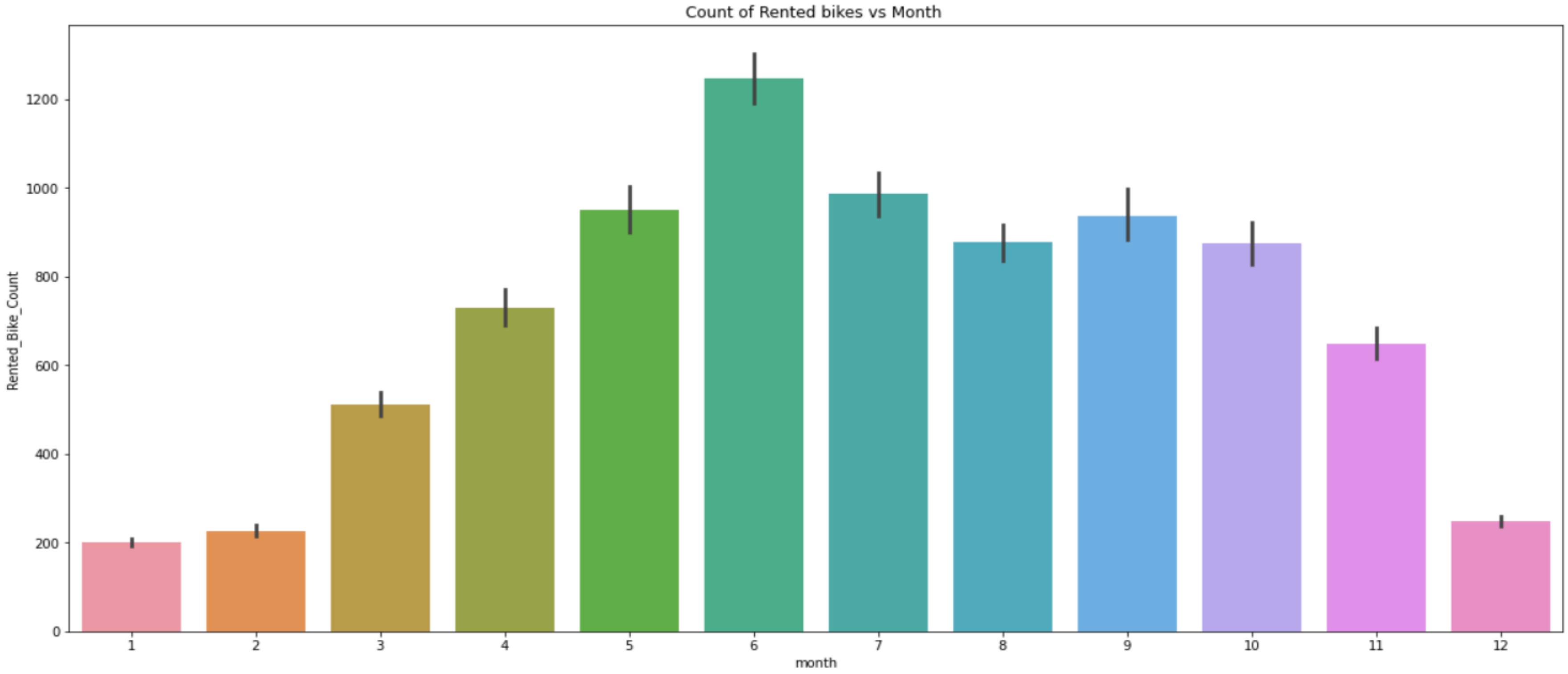


Analysis Of Function Day & Holiday

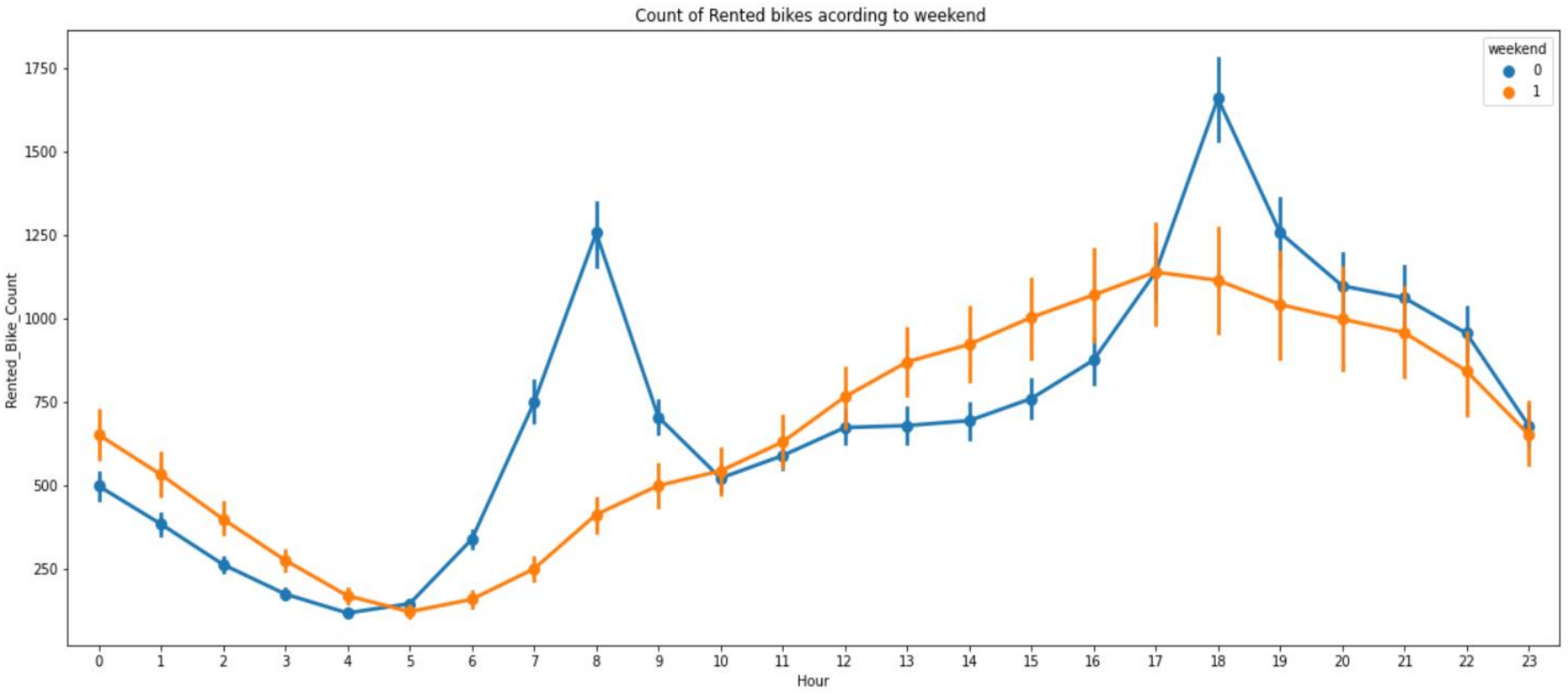




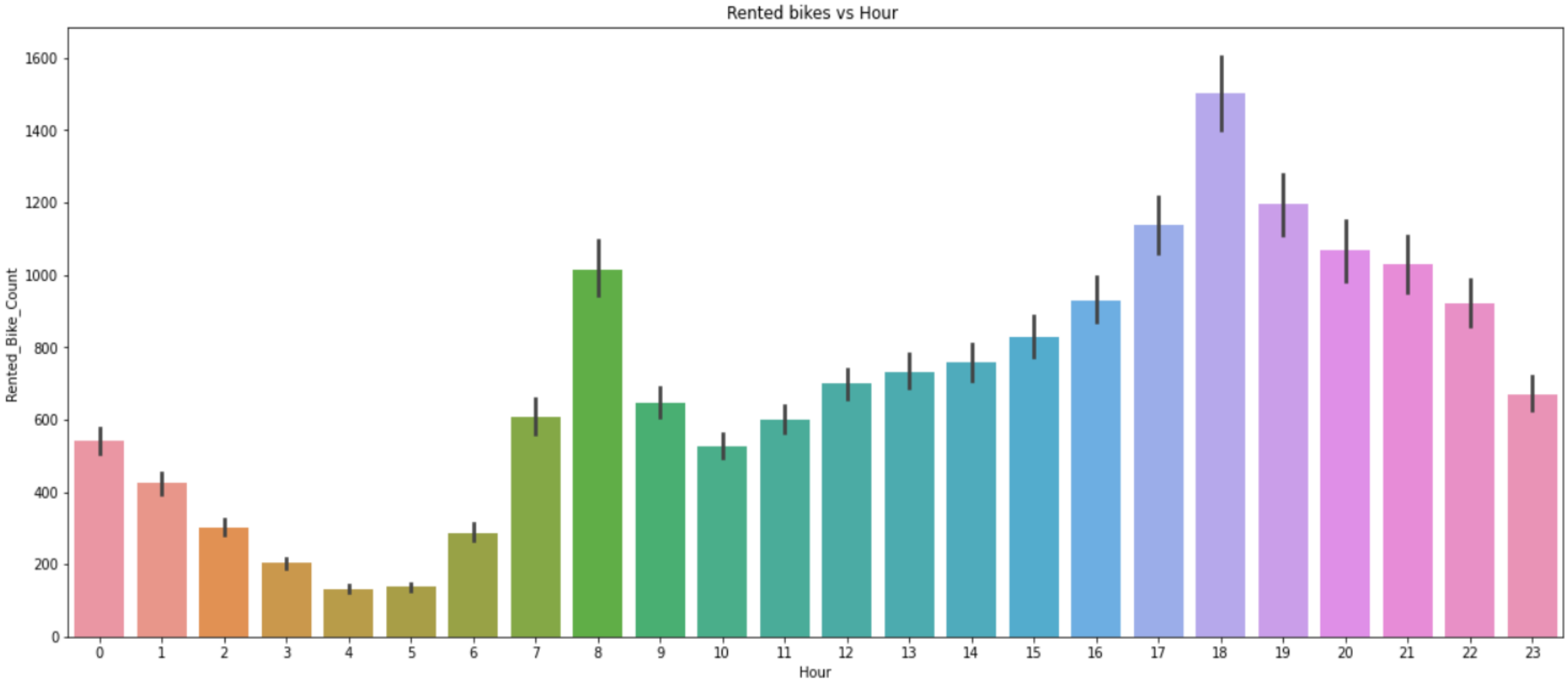
Analysis Of Month Variable



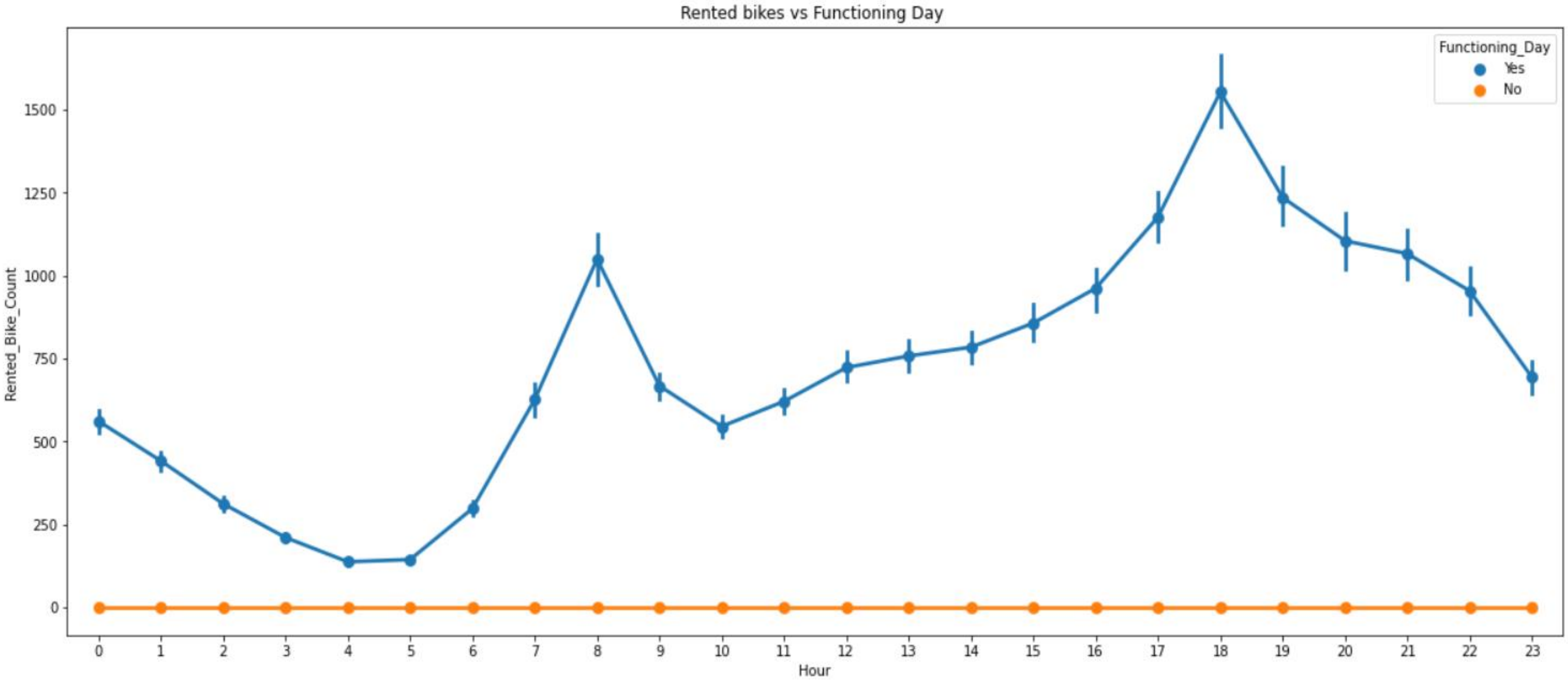
Analysis Of Weekdays & Weekend Variable



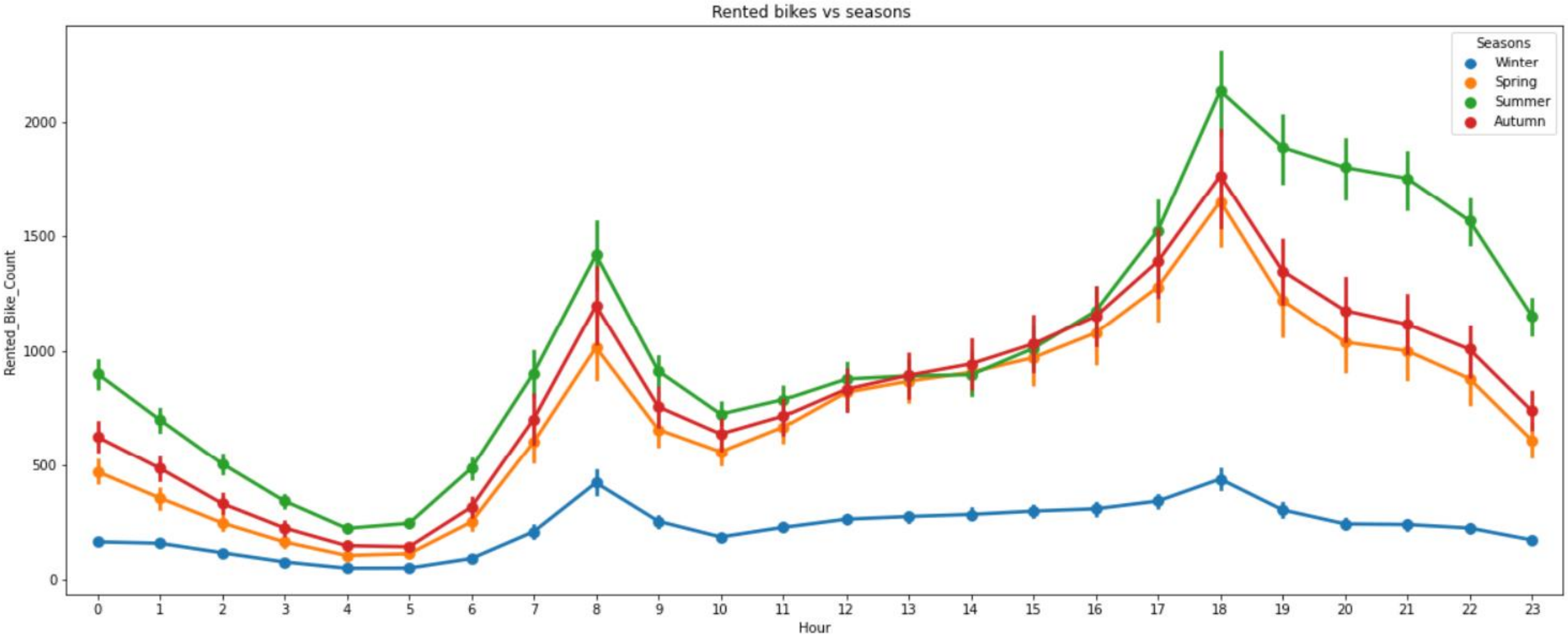
Analysis Of Hour Variable



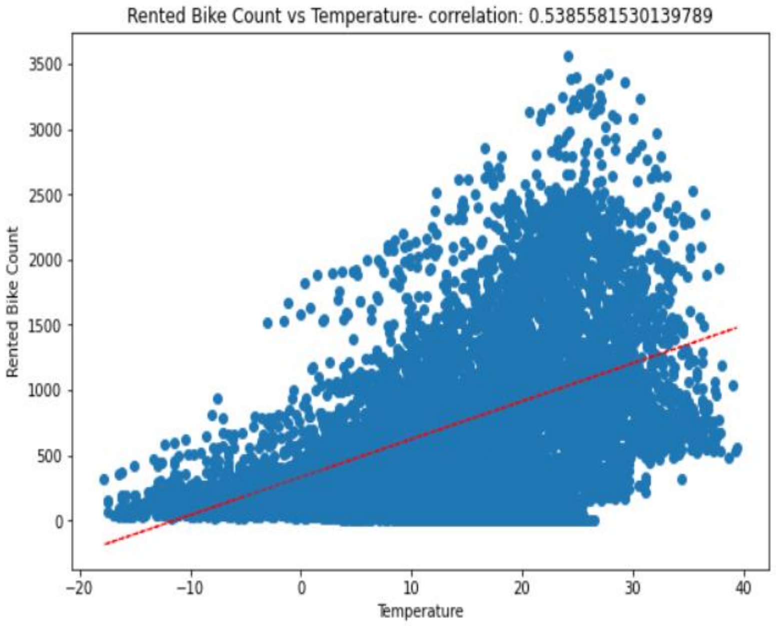
Analysis Of Functioning day Variable

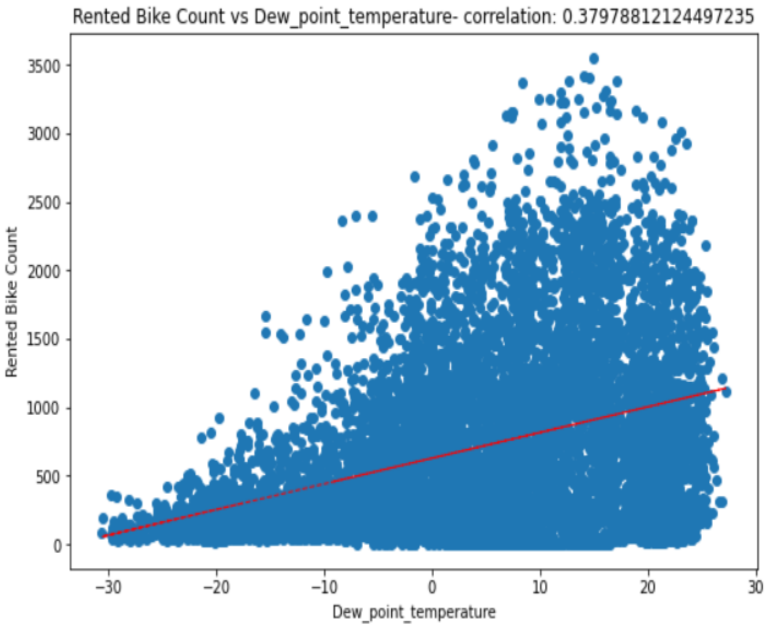


Analysis Of Season Variable



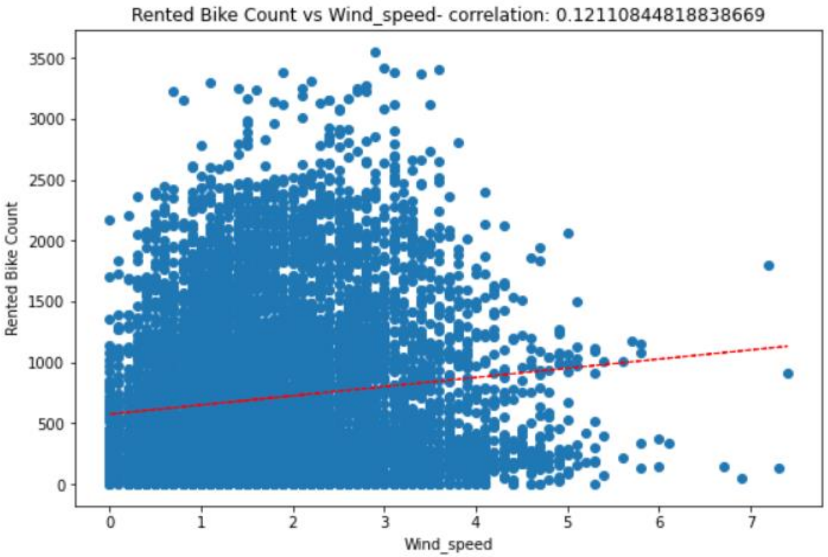
Regression Plot For Numerical Variable

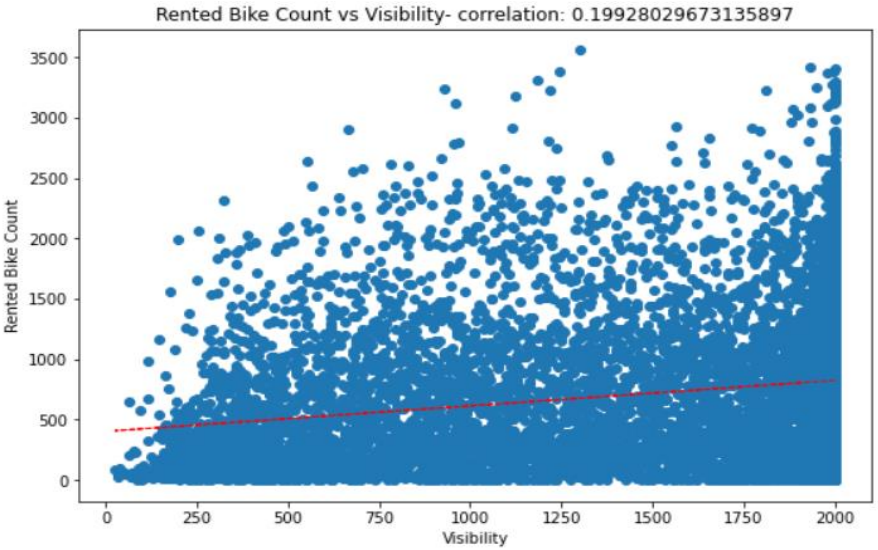




Positively correlated variable

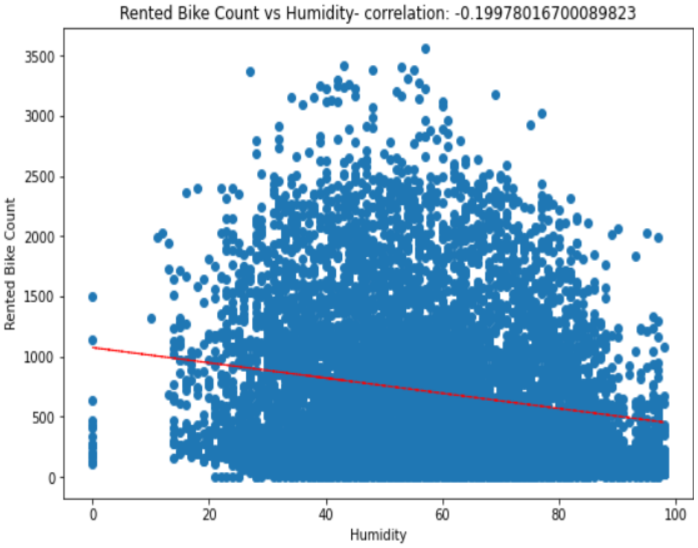
Regression Plot For Numerical Variable

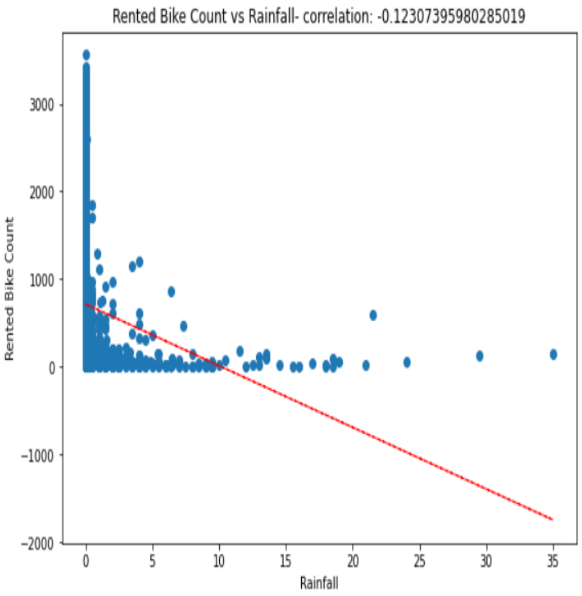


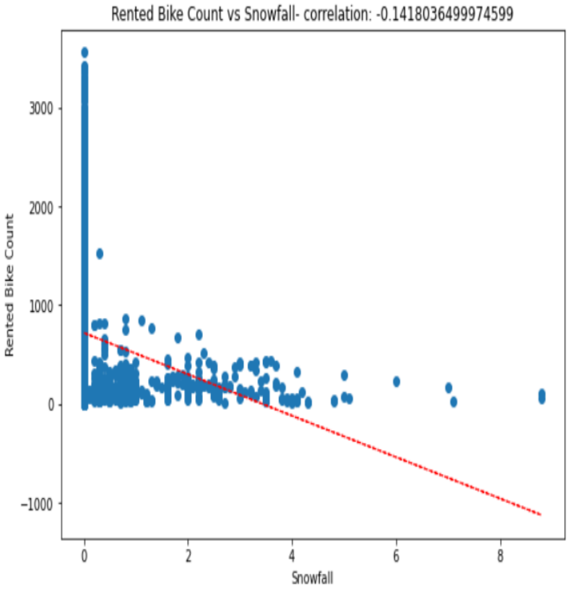


Positively correlated variable

Regression Plot For Numerical Variable

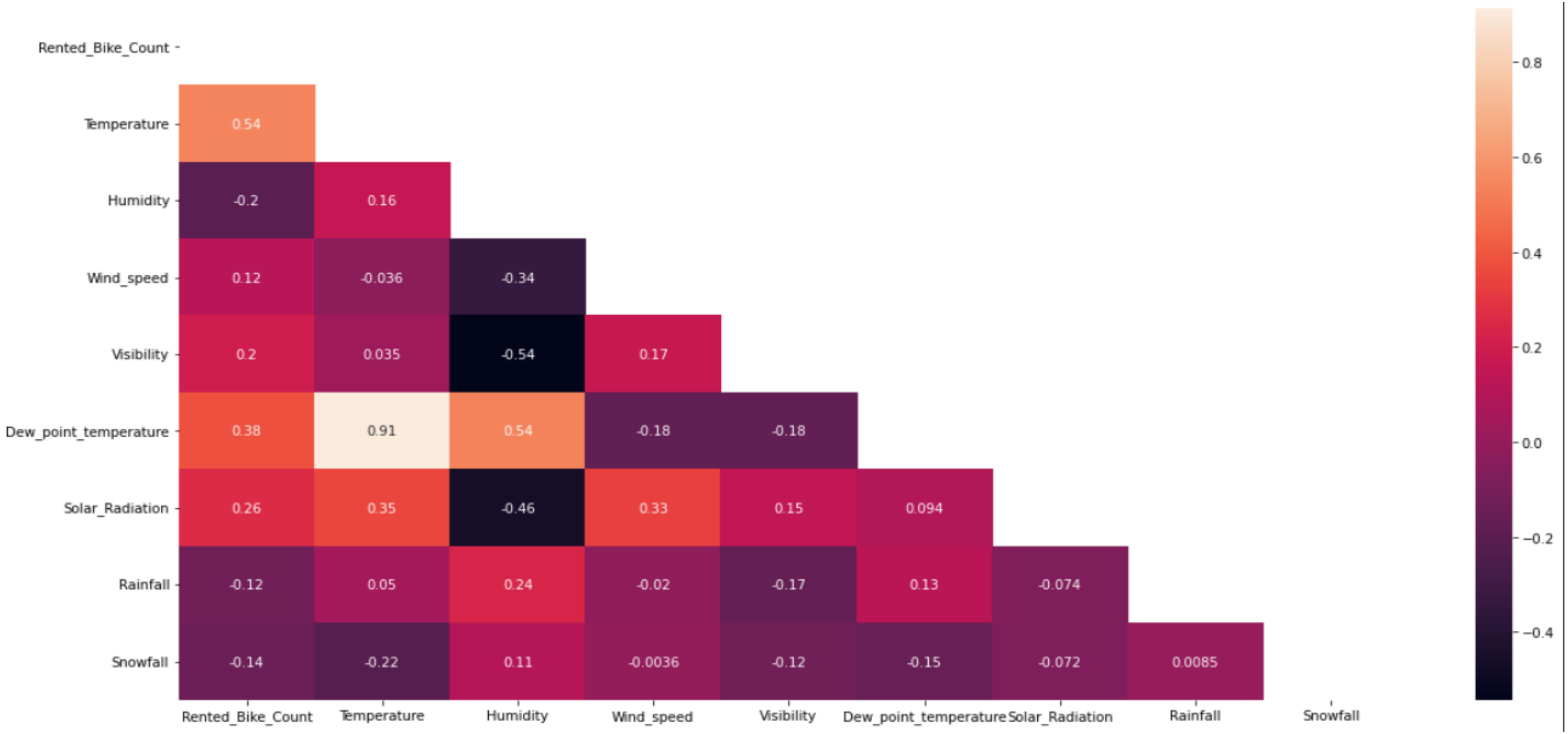






- Negatively correlated variable

Correlation Matrix (Heatmap)



Model Implementation

→Linear Regression

_step_1 → VIF range should be lower 10

	variables	VIF
0	Temperature	29.075866
1	Humidity	5.069743
2	Wind_speed	4.517664
3	Visibility	9.051931
4	Dew_point_temperature	15.201989
5	Solar_Radiation	2.821604
6	Rainfall	1.079919
7	Snowfall	1.118903

→

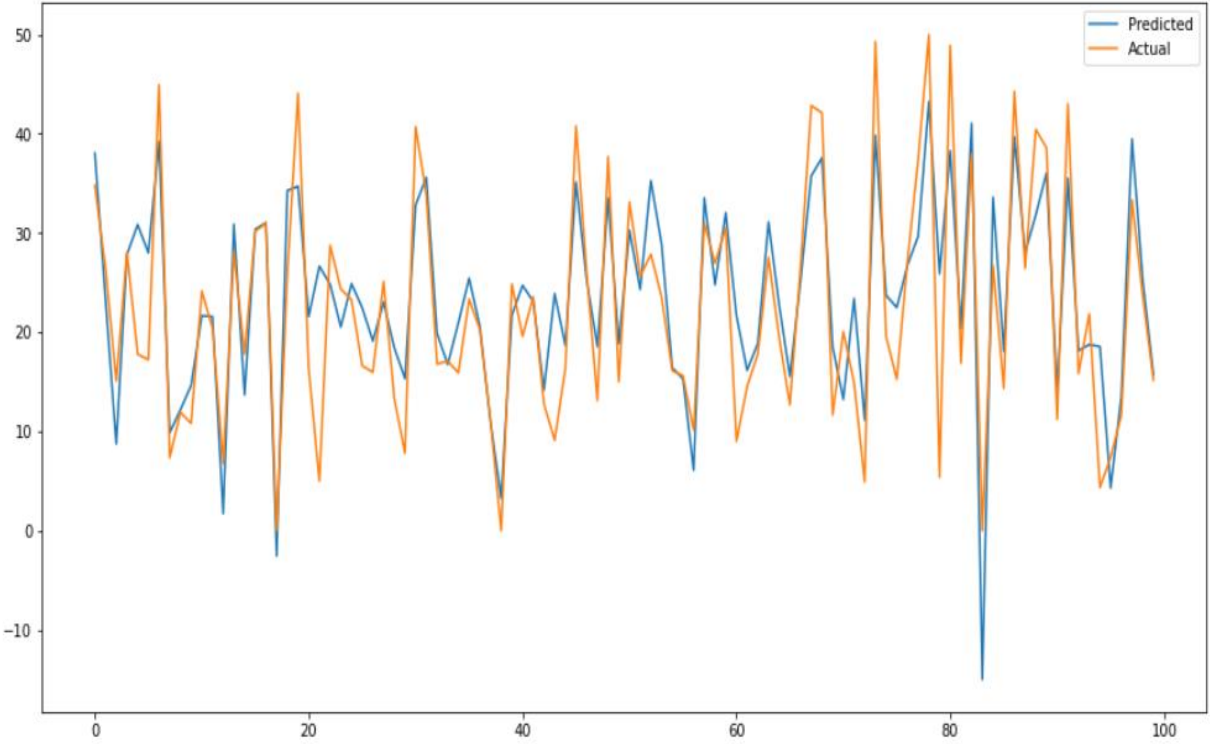
	variables	VIF
0	Temperature	3.166007
1	Humidity	4.758651
2	Wind_speed	4.079926
3	Visibility	4.409448
4	Solar_Radiation	2.246238
5	Rainfall	1.078501
6	Snowfall	1.118901

vif should be in range of 1 to 10 clearly we need to drop heighly correlated column

Linear Regression (continue)

Test dataset result

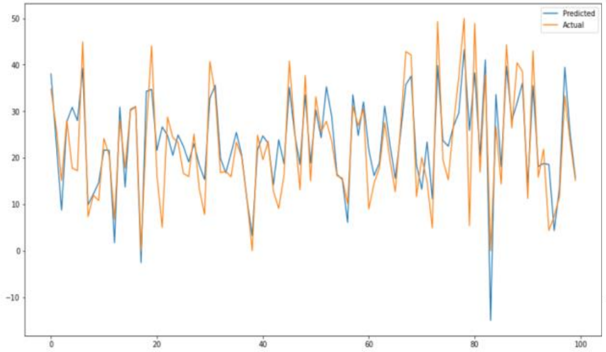
```
MSE : 35.077512998569425
RMSE : 5.922627204085145
MAE : 4.474042217333856
R2 : 0.772210154067841
Adjusted R2 : 0.766558065194056
```

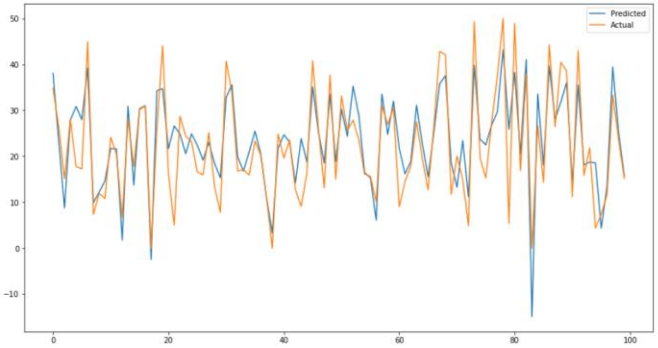


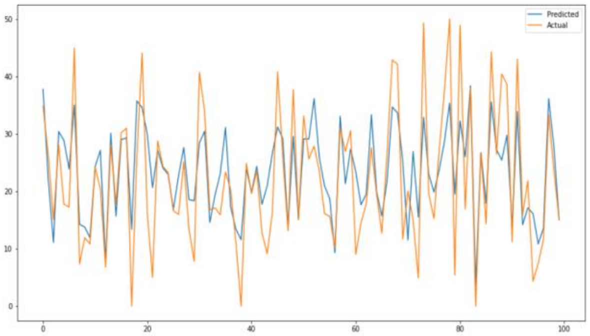
Lasso, Ridge and Elastic Net

Test dataset results

Lasso	Ridge	Elastic Net
MSE : 35.07752396970736	MSE : 35.078136174806644	MSE : 52.987831883019176
RMSE : 5.922628130290417	RMSE : 5.922679813632225	RMSE : 7.279274131602627
MAE : 4.474047285132368	MAE : 4.474831316083193	MAE : 5.607820103665487
R2 : 0.7722100828223736	R2 : 0.7722061072239539	R2 : 0.6559023422953525
Adjusted R2 : 0.7665579921807939	Adjusted R2 : 0.7665539179369079	Adjusted R2 : 0.6473643386163515







SVM

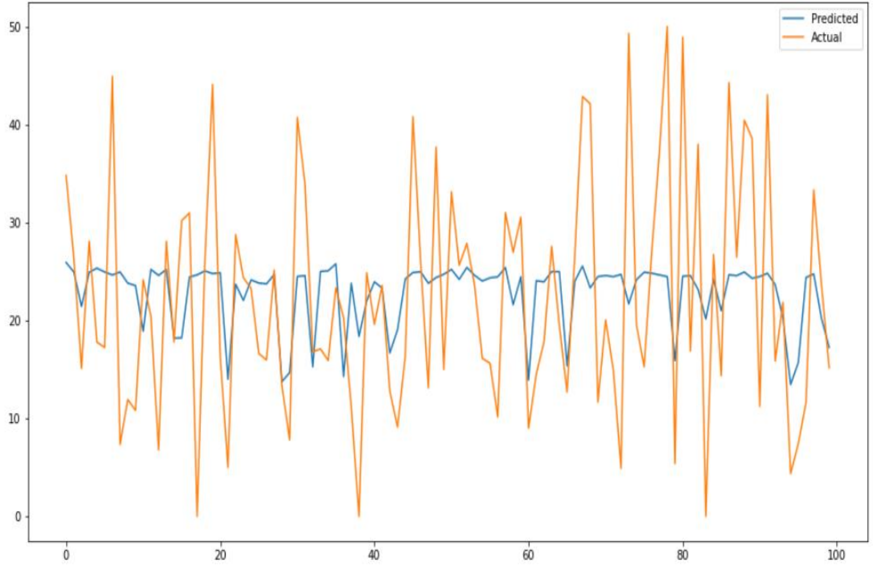
Test dataset result

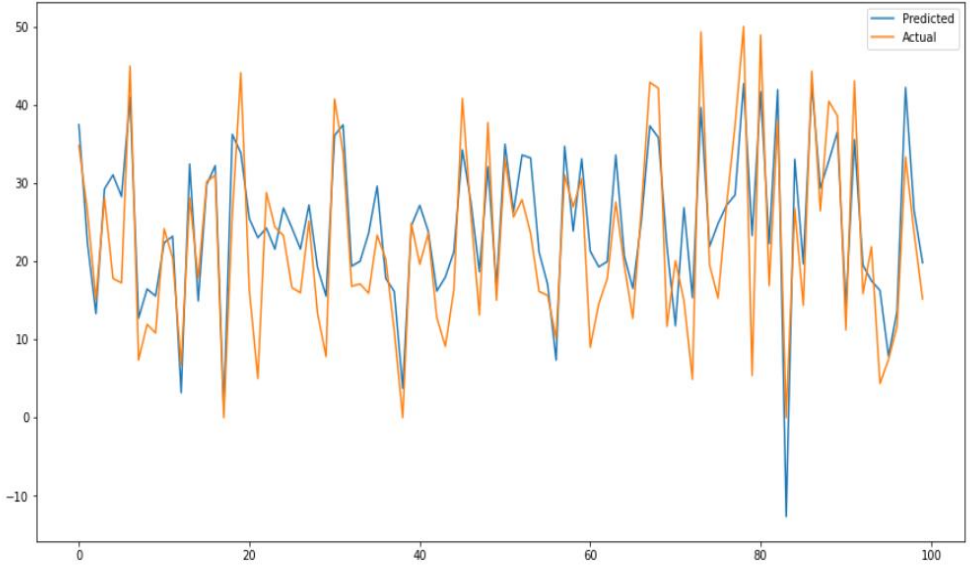
Kernel= ‘rbf’

MSE : 140.71554768323855
RMSE : 11.86235843680499
MAE : 9.80545492295386
R2 : 0.08620736799865203
Adjusted R2 : 0.06353367441434898

kernel=‘linear’

MSE : 40.93355714508648
RMSE : 6.397933818435955
MAE : 4.944313631955972
R2 : 0.7341815916107068
Adjusted R2 : 0.7275859101291373

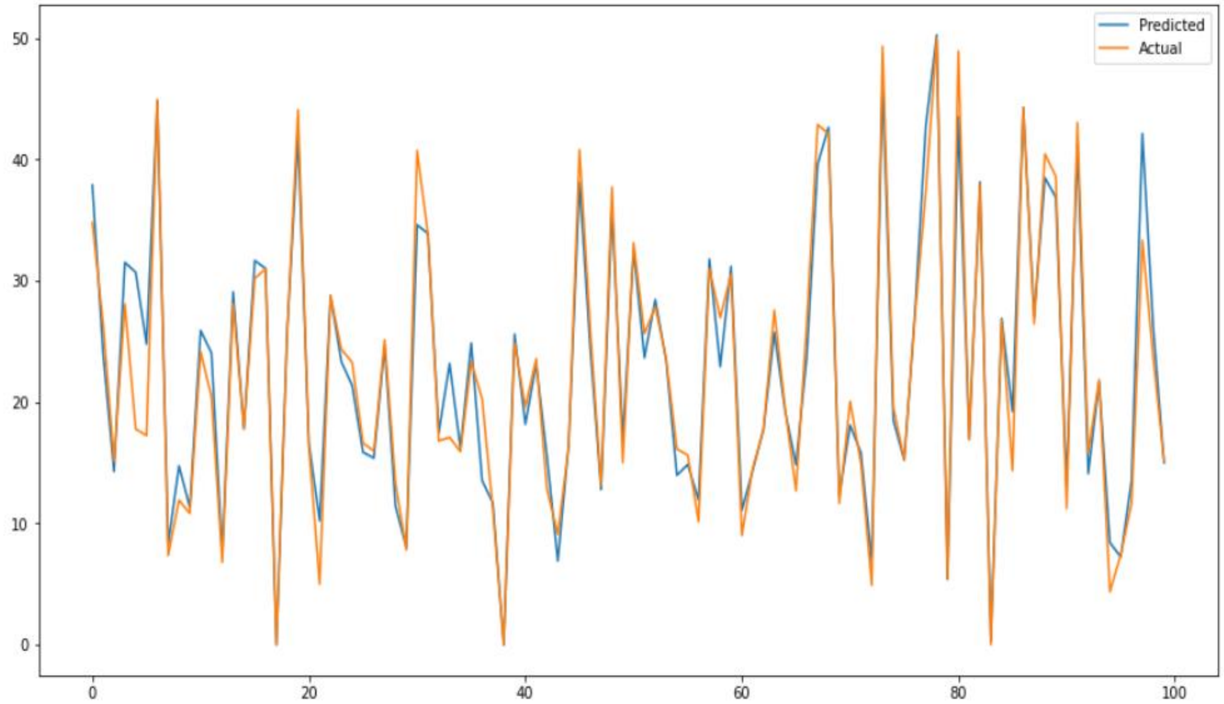




Random Forest

Test dataset result

```
MSE : 12.120575429745571
RMSE : 3.4814616800627824
MAE : 2.1724688159056256
R2 : 0.9232711818900445
Adjusted R2 : 0.9213673301298256
```



→ 92% accuracy

XGBoost

Test dataset result

```
Model Score: 0.9921200071811659
MSE : 1.2134454431020136
RMSE : 1.1015649972207784
MAE : 0.754253930308519
R2 : 0.9331907895417064
Adjusted R2 : 0.9315330703683499
```

Hyper parameter

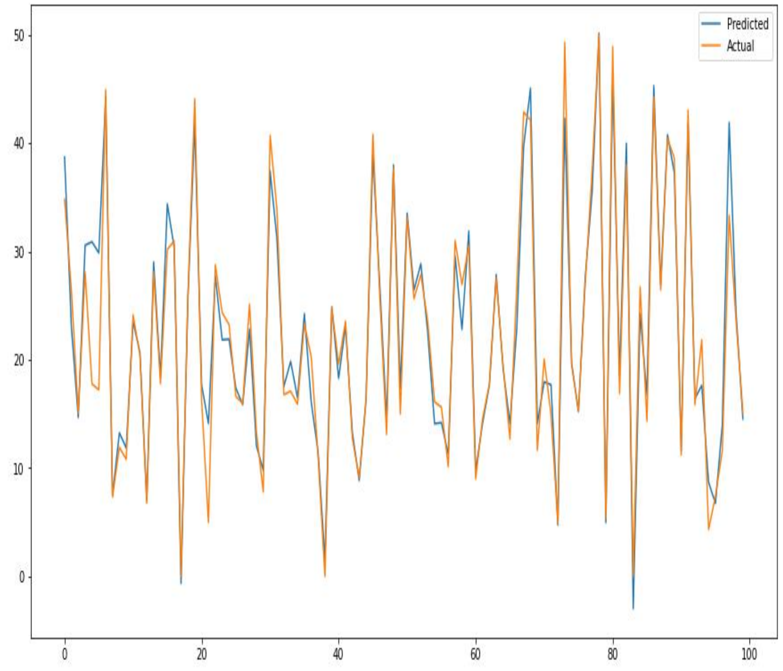
```
]  # Number of trees
   # Maximum depth of trees
   # Minimum number of samples required to split a node
   # Minimum number of samples required at each leaf node

   param_dict = {'n_estimators' : [100, 120],
                 'max_depth' : [8,9],
                 'min_samples_split' : [50,100],
                 'min_samples_leaf' : [40,50]}
```

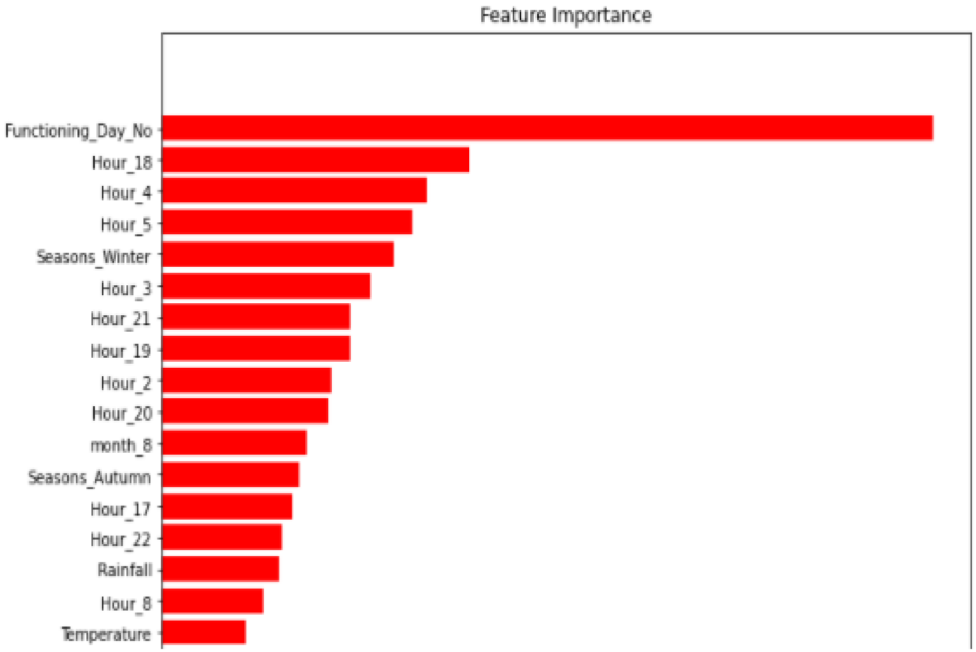
```
]  # Grid search
   grid = GridSearchCV(estimator=xgb,param_grid = param_dict,verbose=2)
```

Output and Important Features

Output Graph



Important Features



Summary

	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear regression	4.474	35.078	5.923	0.772	0.770
1	Lasso regression	4.474	35.078	5.923	0.772	0.770
2	Ridge regression	4.475	35.078	5.923	0.772	0.770
3	Elastic net regression	5.608	52.988	7.279	0.656	0.650
4	SVM (kernel)	4.944	40.934	6.398	0.734	0.650
5	Random forest regression	2.156	12.098	3.478	0.923	0.920
6	Gradient boosting regression	0.754	1.213	1.102	0.933	0.932

Random forest Regressor and XGBoost gives the highest R2 score of 92% and 93.2% respectively

Challenges

- Large Dataset to handle
- Need to analyze lot of variable
- Feature engineering
- Feature selection
- Optimizing the model
- Deciding the flow of the presentation

Conclusion

●'Functioning day' column holds the most important feature.
● Bike rental count is mostly correlated with the time of the day as it is peak at 8 am morning and 6 pm at evening.
● Bike rental count is high during working days than non working day.
● We see that people generally prefer to bike at moderate to high temperatures, and when little windy.
● It is observed that highest number bike rentals counts in Autumn & Summer seasons & the lowest in winter season.
● highest number of bike rentals on a clear day and the lowest on a snowy or rainy day.
● We observed that with increasing humidity, the number of bike rental counts decreases.
● When we compare the root mean squared error and mean absolute error of all the models, Random forest Regressor and XGBoost with gridsearchcv gives the highest R2 score of 92% and 93% respectively

Q&A

Thank you