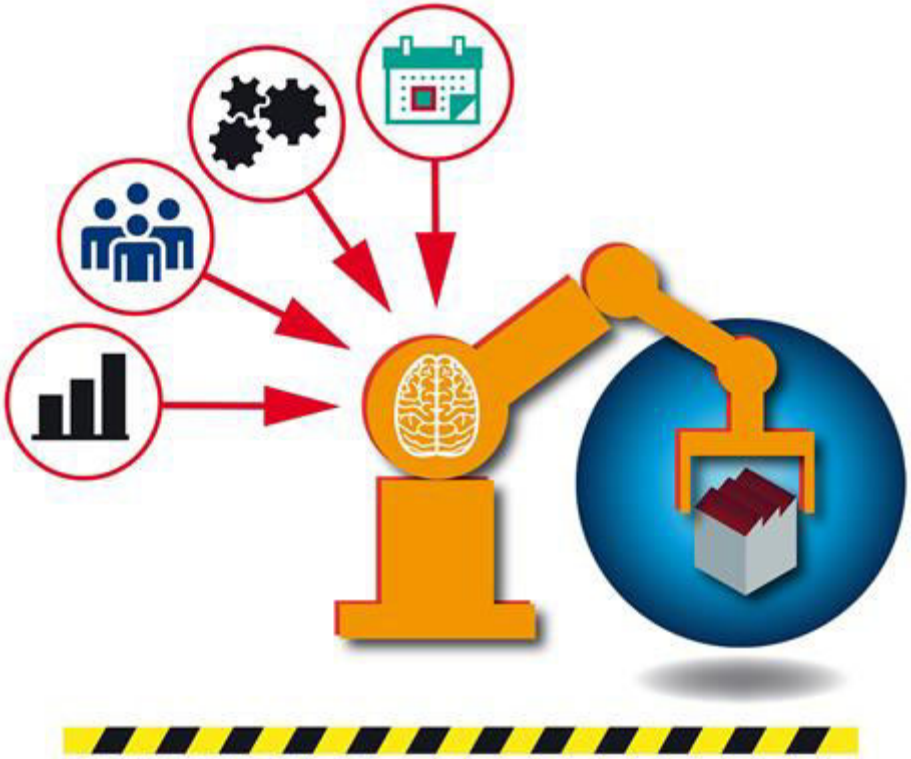
Capstone Project

Bike Sharing Demand Prediction

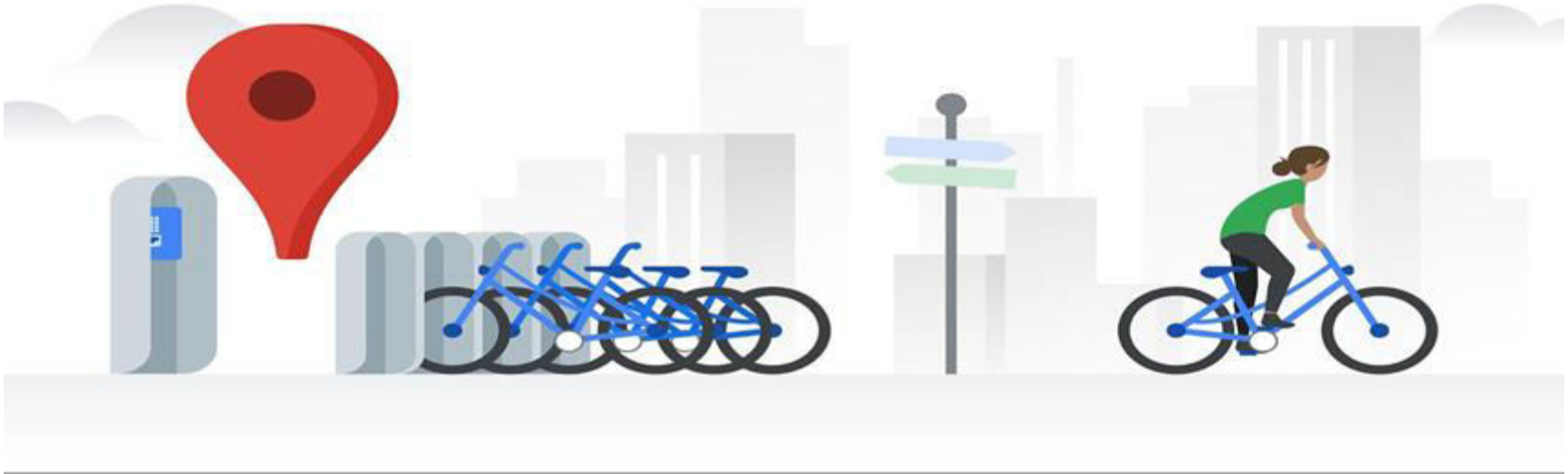
Content

1. Problem Statement
2. Introduction
3. Exploratory Data Analysis
4. Data Summary
5. Hypothesis
6. Model Building
7. Evaluation
8. Challenges
9. Conclusion
10. Q&A



Problem Statement

Predict the demand of bike rent based on the historical usage over different factors such as seasons, weather, temperature, humidity etc. where there is hourly rental data for one year 2017-2018.



Introduction

- Prediction of bike sharing demand can help bike sharing companies to allocate bikes better and ensure a more sufficient circulation of bikes for customers.
- This presentation proposes a real-time method for predicting bike renting based on historical data, weather data, and time data.
- This demand prediction model can provide a significant theoretical basis for management strategies and vehicle scheduling in public bike rental system.
- We evaluate the model inter alia with the Root Mean Squared Error and show that the prediction of proposed model outperforms that of other regression models by comparing their RMSE.

Data Summary

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

Let's Look At our Dataset

- The original dataset has 14 columns and 8760 rows.
- The data types of various columns are Object,Float and Integer.
- Dependent variable being Rented Bike Count and all other variable are our feature or independent variables like Hour,Temperature etc.

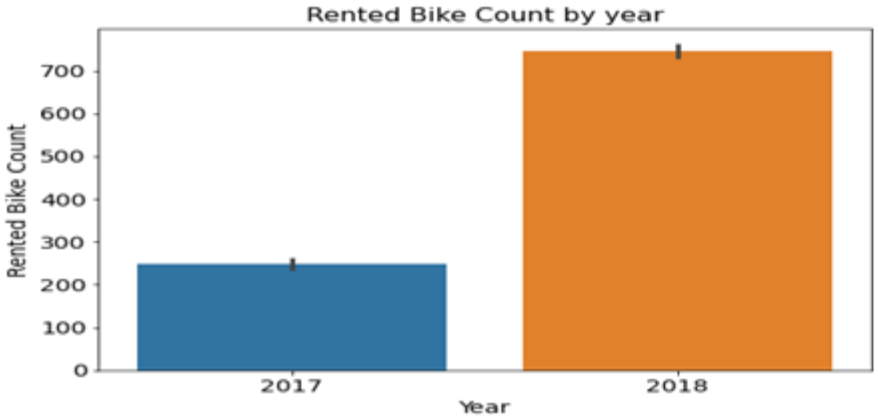
```
[ ] # Lets look at he first 15 rows for dataframe to explore column names, indexes better.
    df.head(24)
```

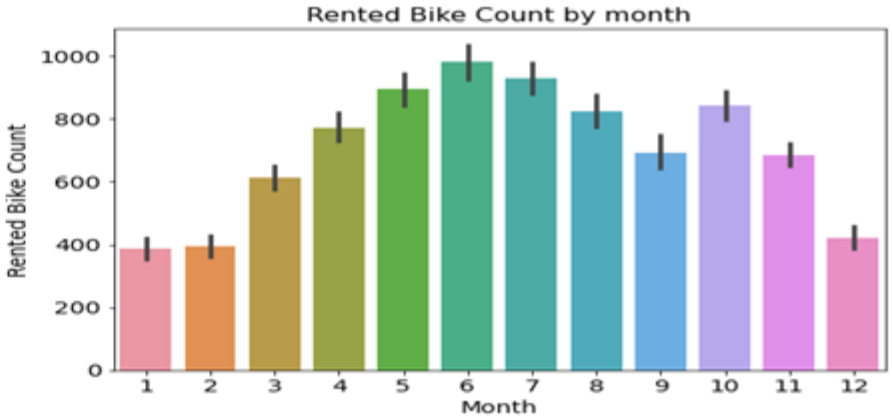
	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.00	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.00	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.00	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.00	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.00	0.0	0.0	Winter	No Holiday	Yes

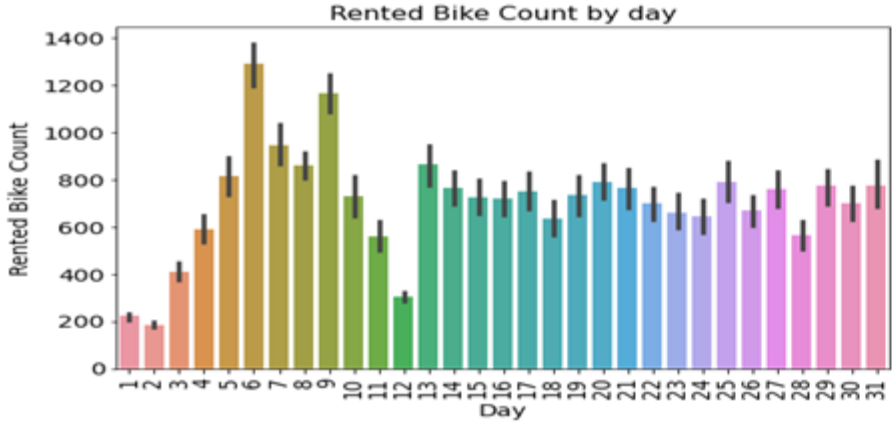
Hypothesis

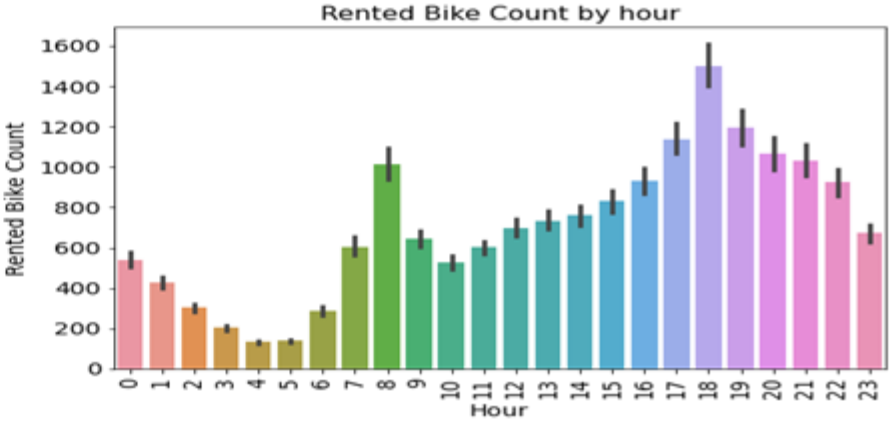
1. The dataset shows hourly rental data for one year (1 December 2017 to 31 November(2018)(365 days).We are required to predict the total count of bikes rented during each hour.
2. There is high demand during peak hours. This high demand might be due to office hours(i.e. 8am to 6pm) and there is low demand during 10:00 pm to 4:00 am.
3. In general, temperature has negative correlation with bike demand.
4. There is high demand for rented bikes on non holiday and low demand on holidays.
5. There is zero demand on Non Functioning Day.

Date Wise Trend



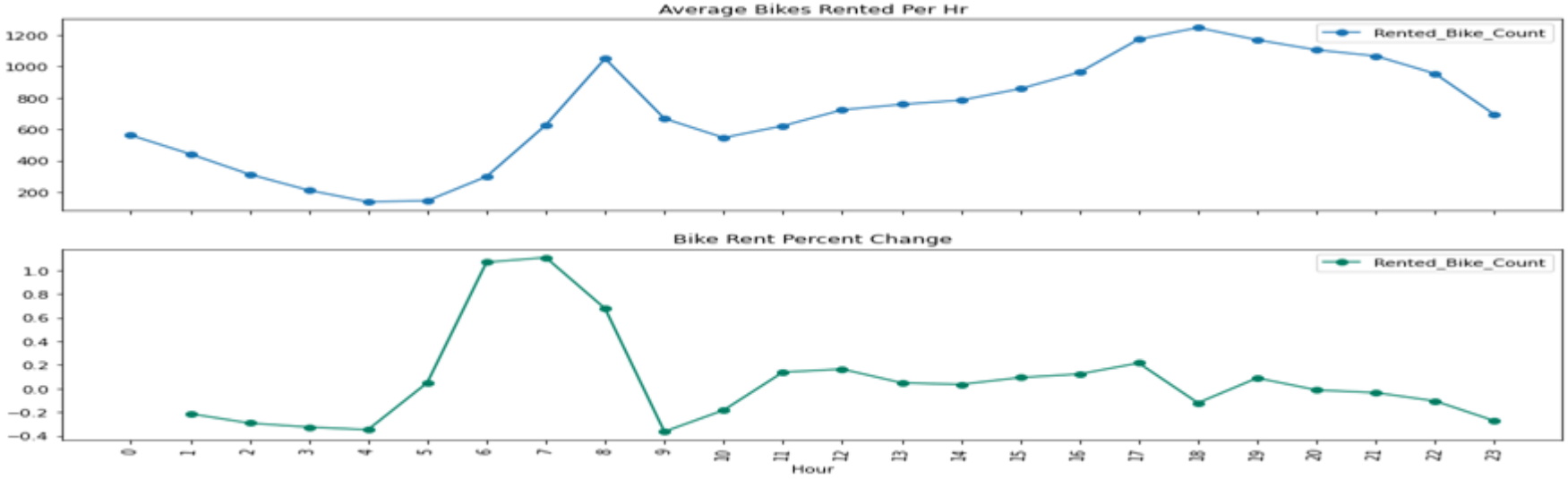






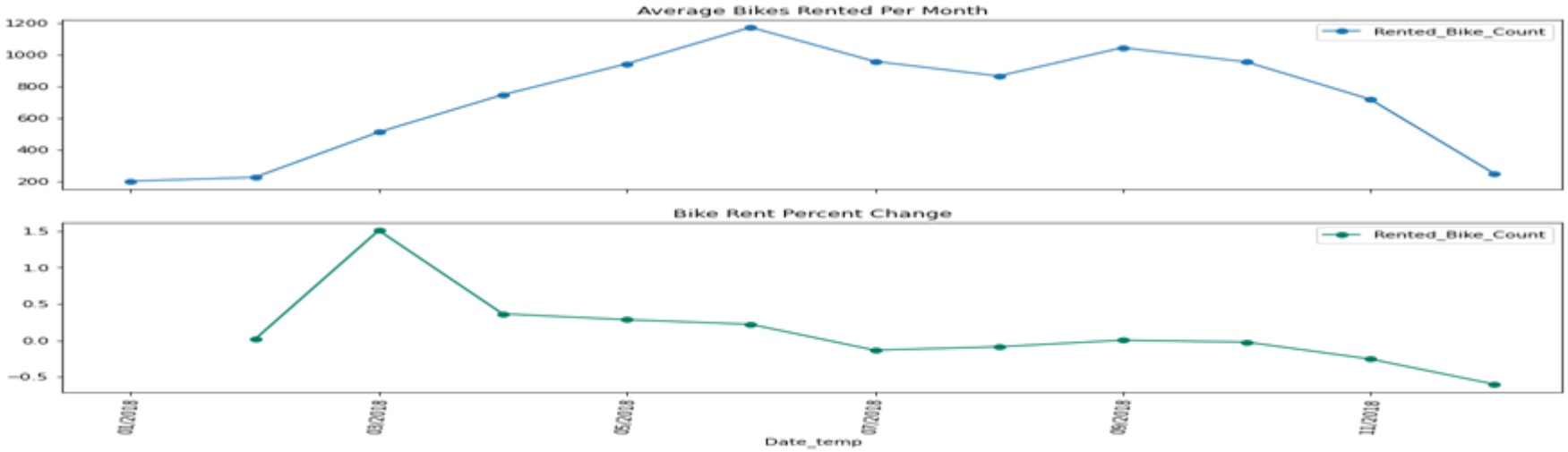
Exploratory Data Analysis: “Hour” Column

- Demand for rented bikes is less during night hours between 12 am to 4 am.
- Demand drastically increased on 8 a.m. It seems this increased due to office hours.(We can’t reject Hypothesis)

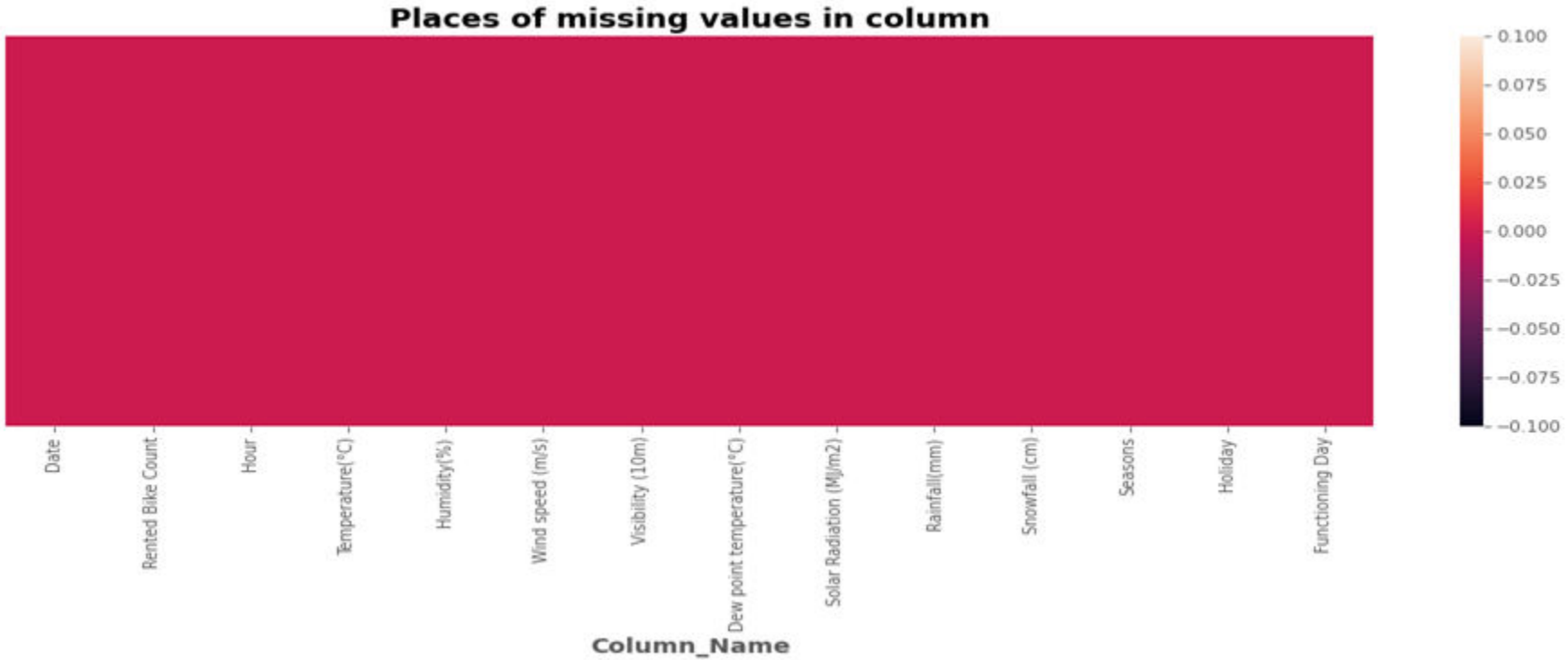


Analysis Of Rented Bike Count By Month

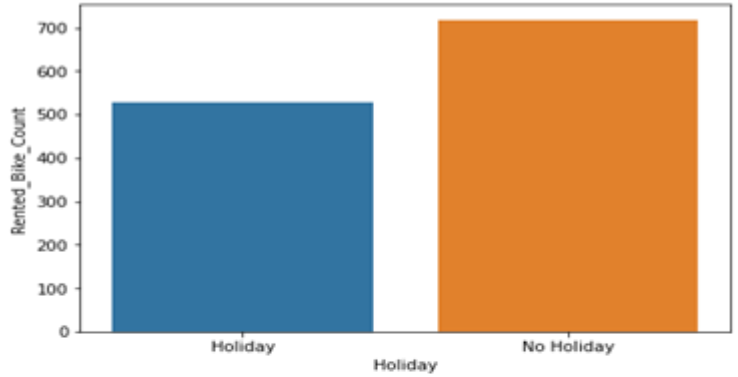
- We can see increasing trend from 1st month to 6th month of 2018. The demand for rented bikes is highest during 6th month of 2018.
- The declining trend continues after 6th month as evident from following graph.

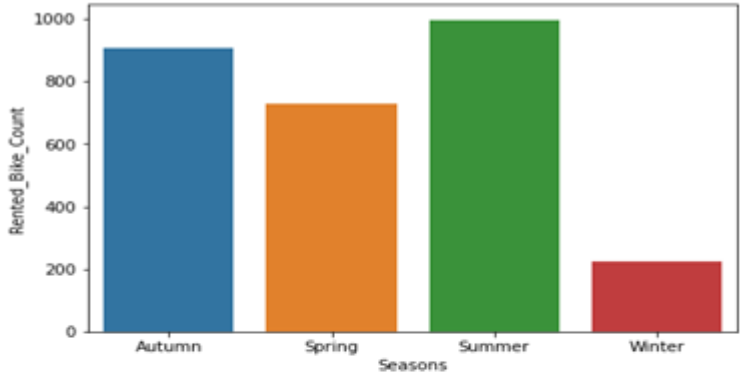


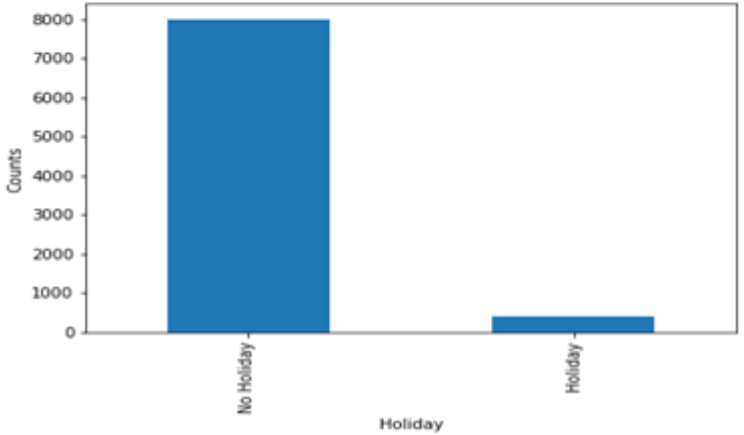
No Missing Values in Dataset

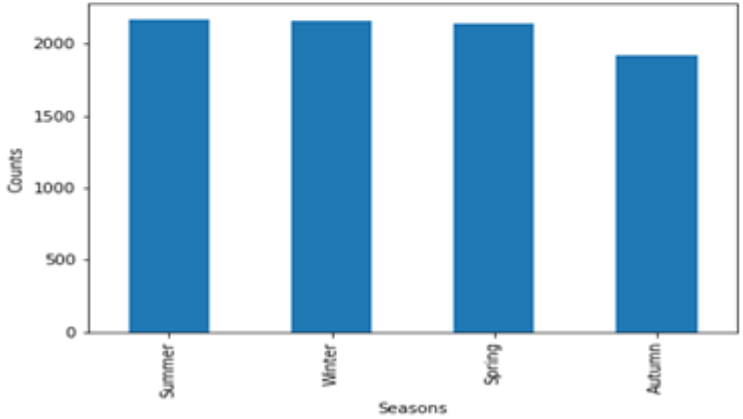


Exploratory Data Analysis

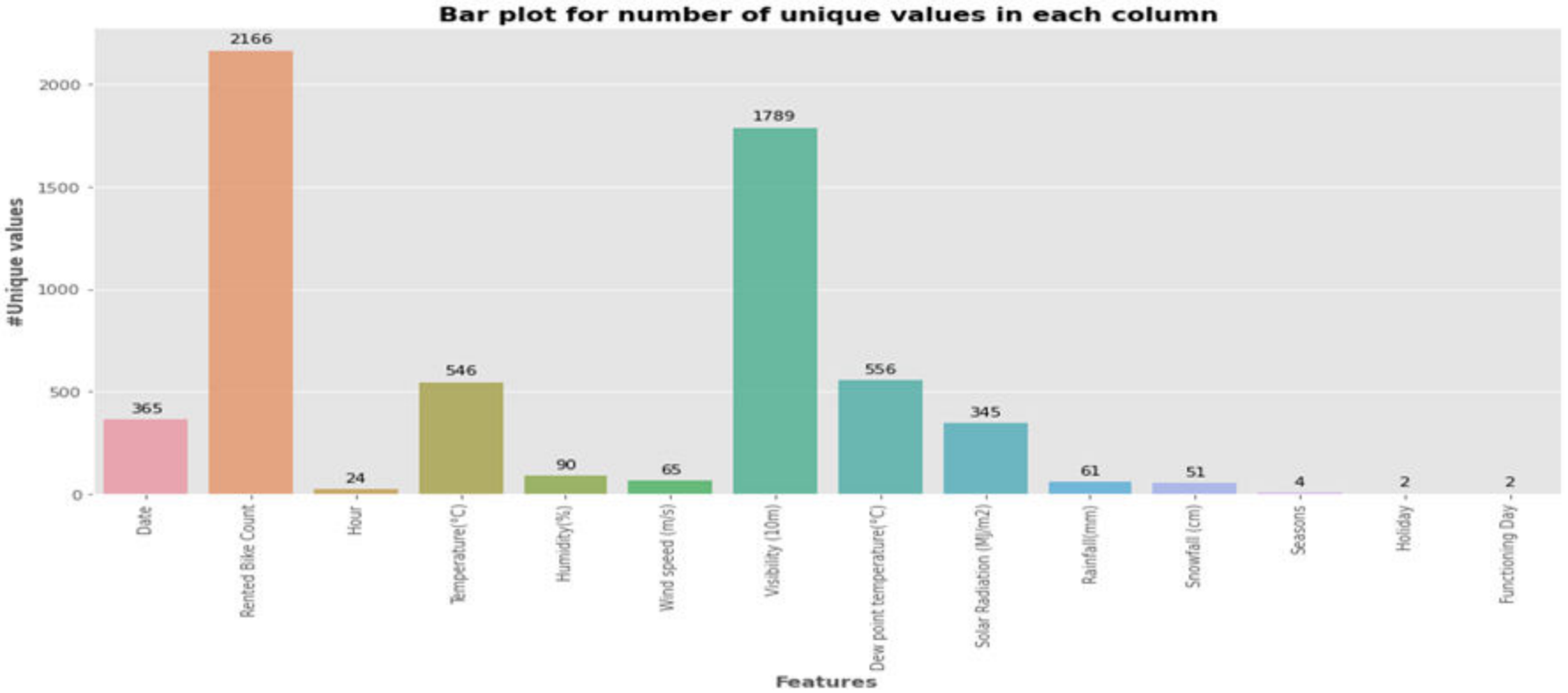






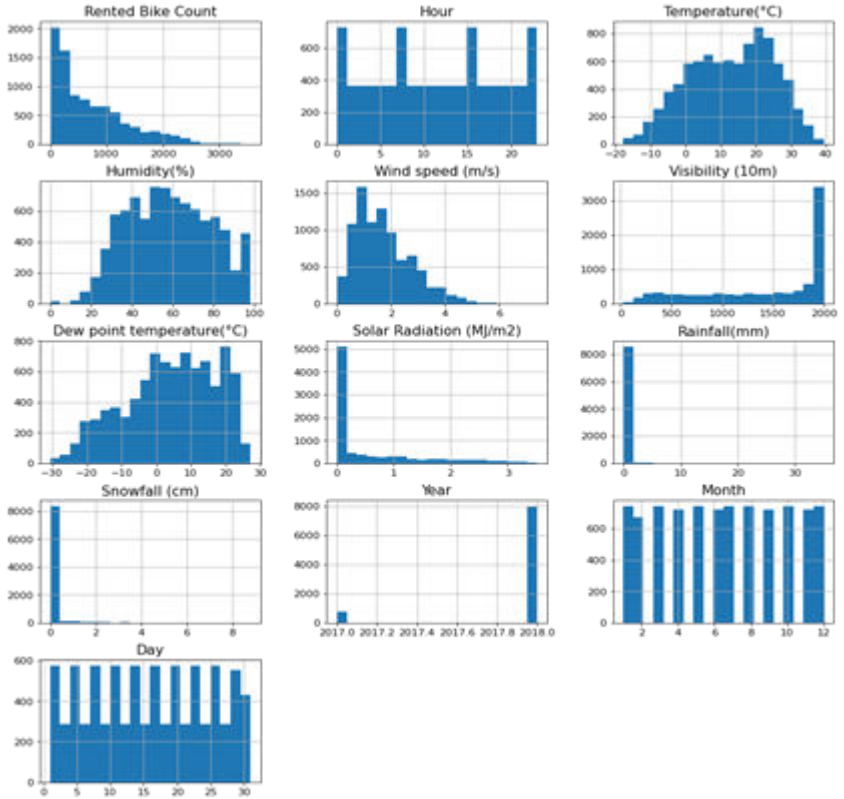


Unique Values in Our Dataset

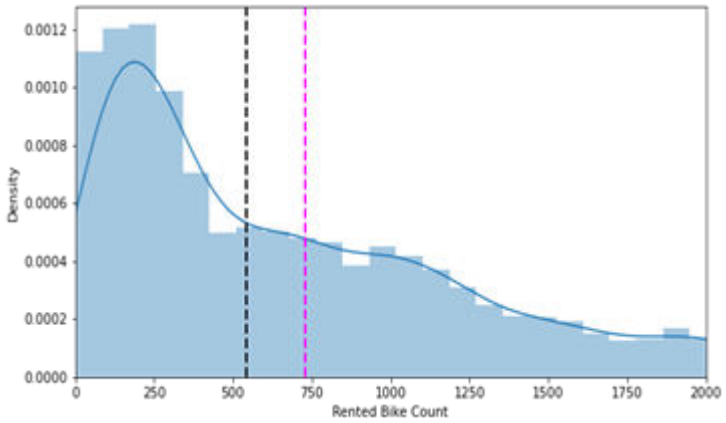


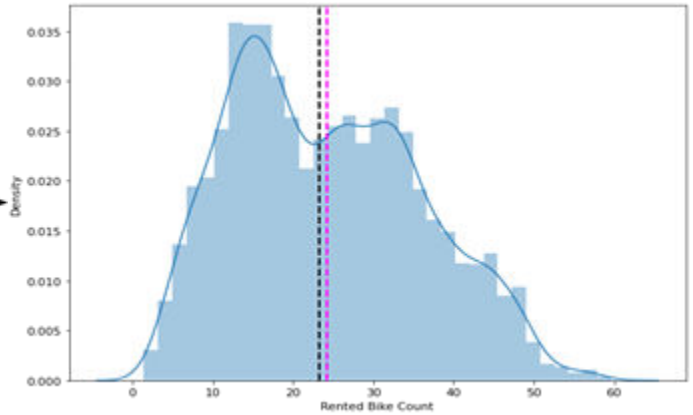
Distribution Of Features

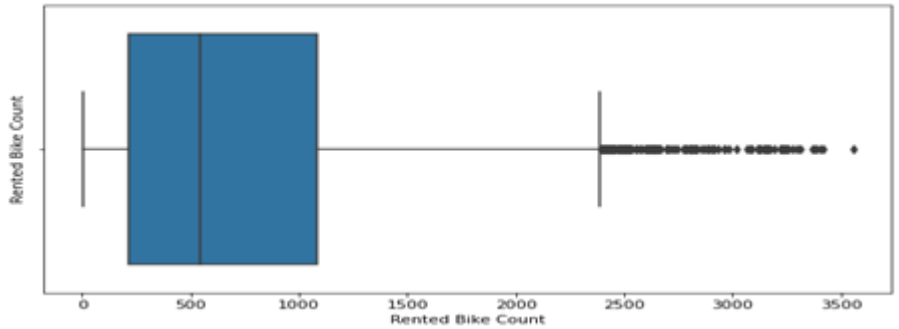
- “Temperature” and “Humidity” columns follows uniform distribution.
- “Wind Speed”, “Solar Radiation”, “Rainfall” and “Snowfall” are having positively skewed distribution.
- “Dew Point Temperature” and “Visibility” are negatively skewed.

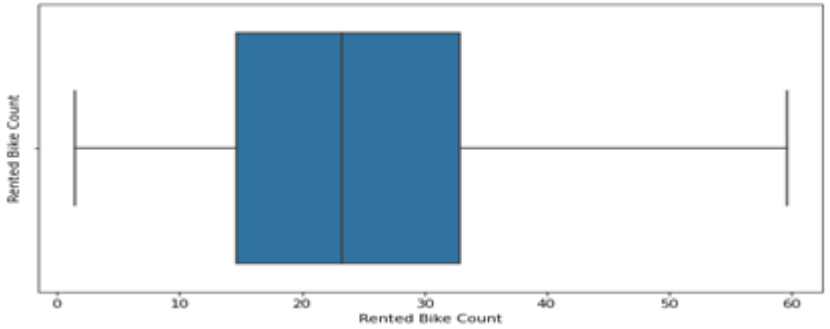


Distribution Of Dependent Variable

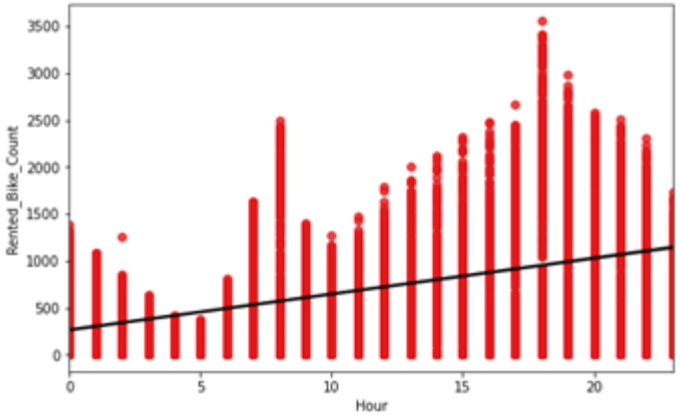


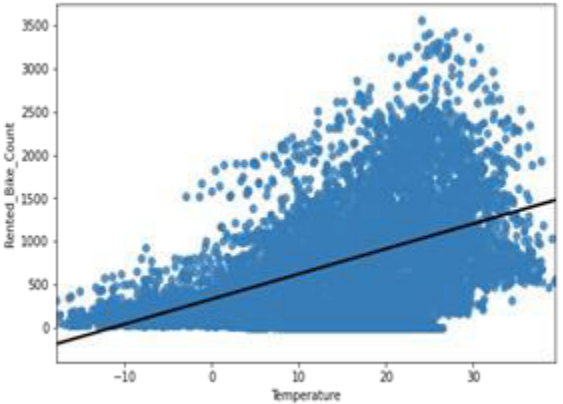


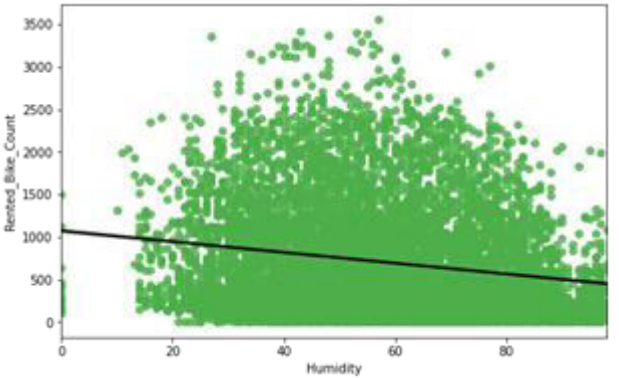


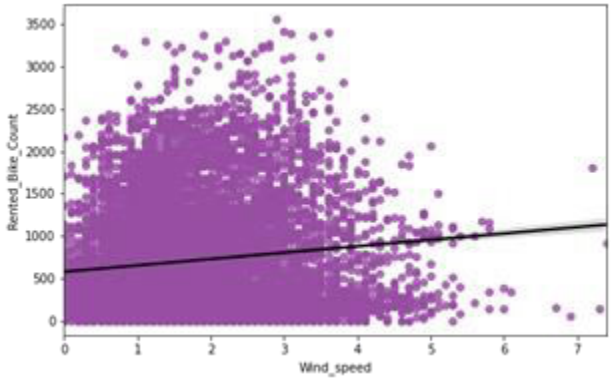


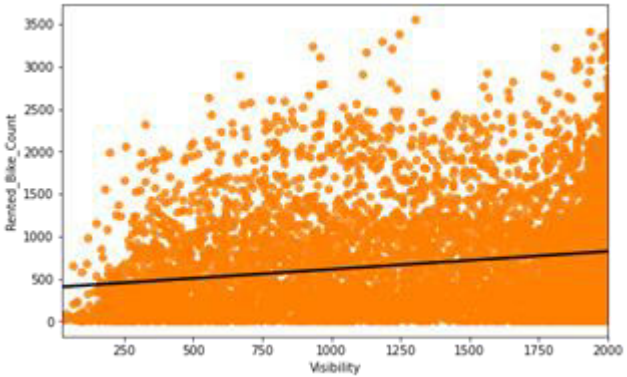
Data Preprocessing: Assumptions Check

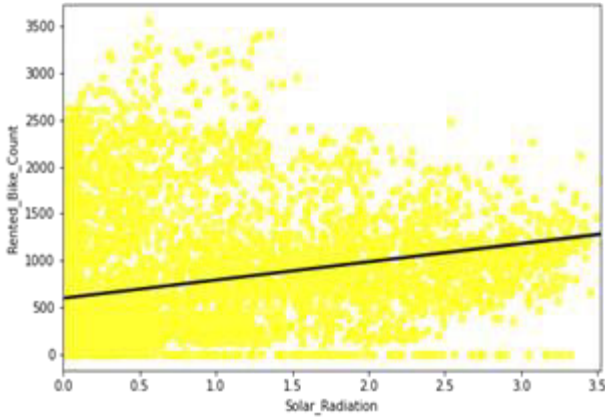




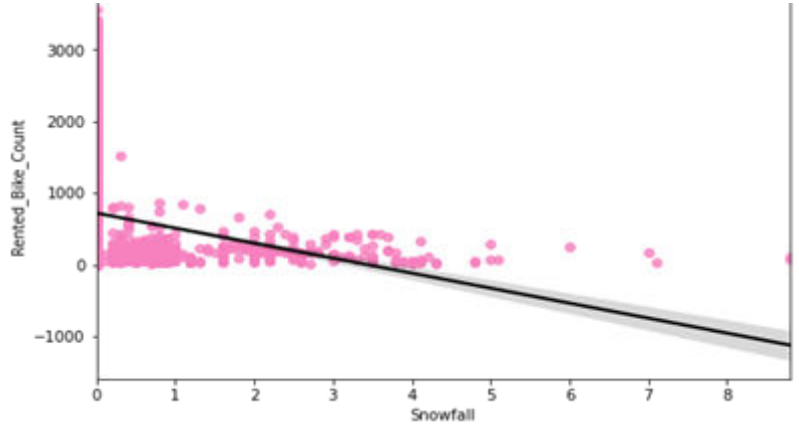


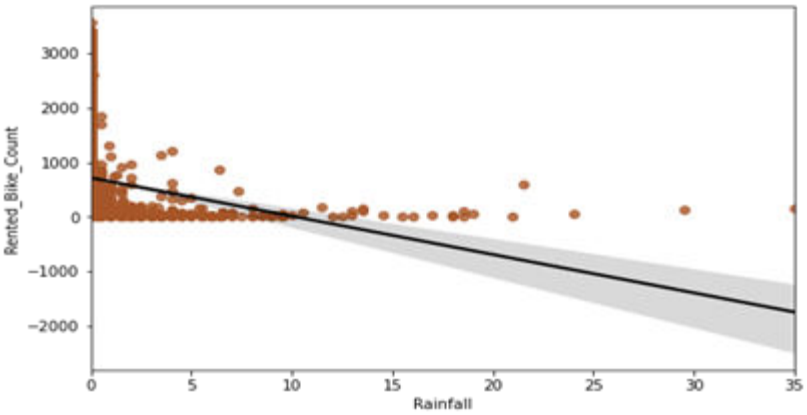


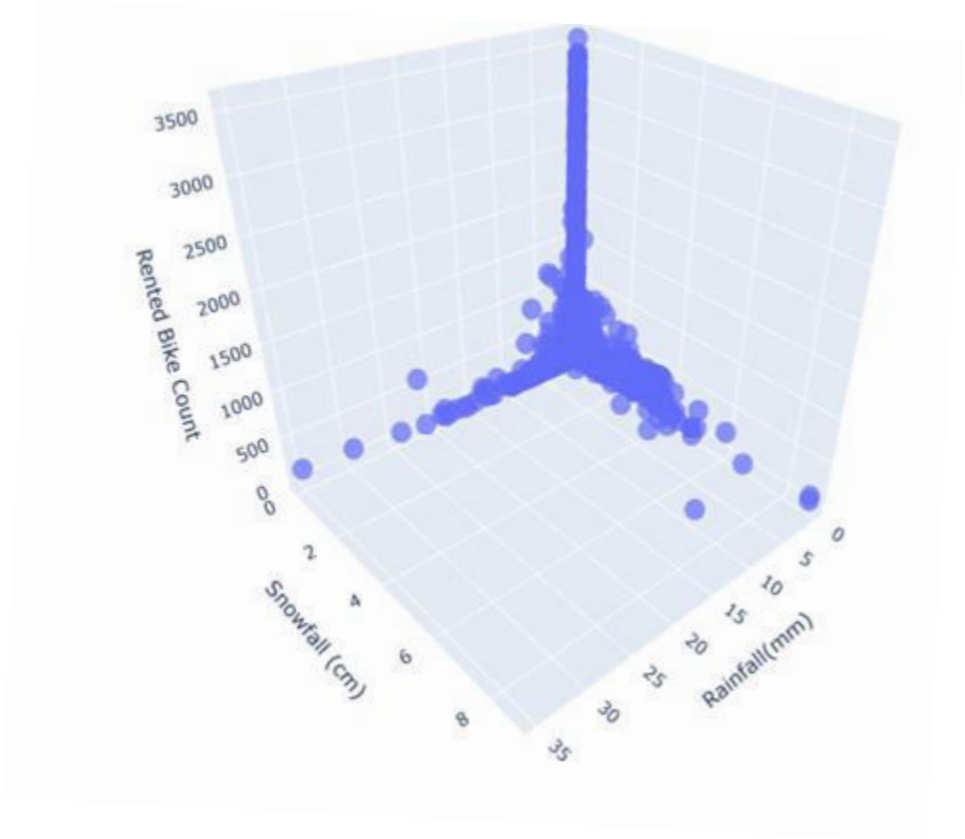




….Assumptions Check

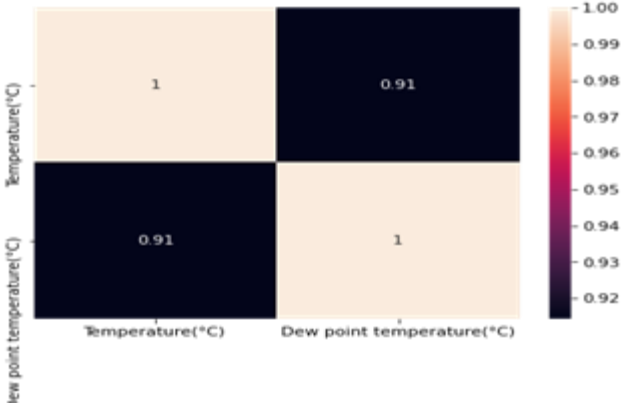


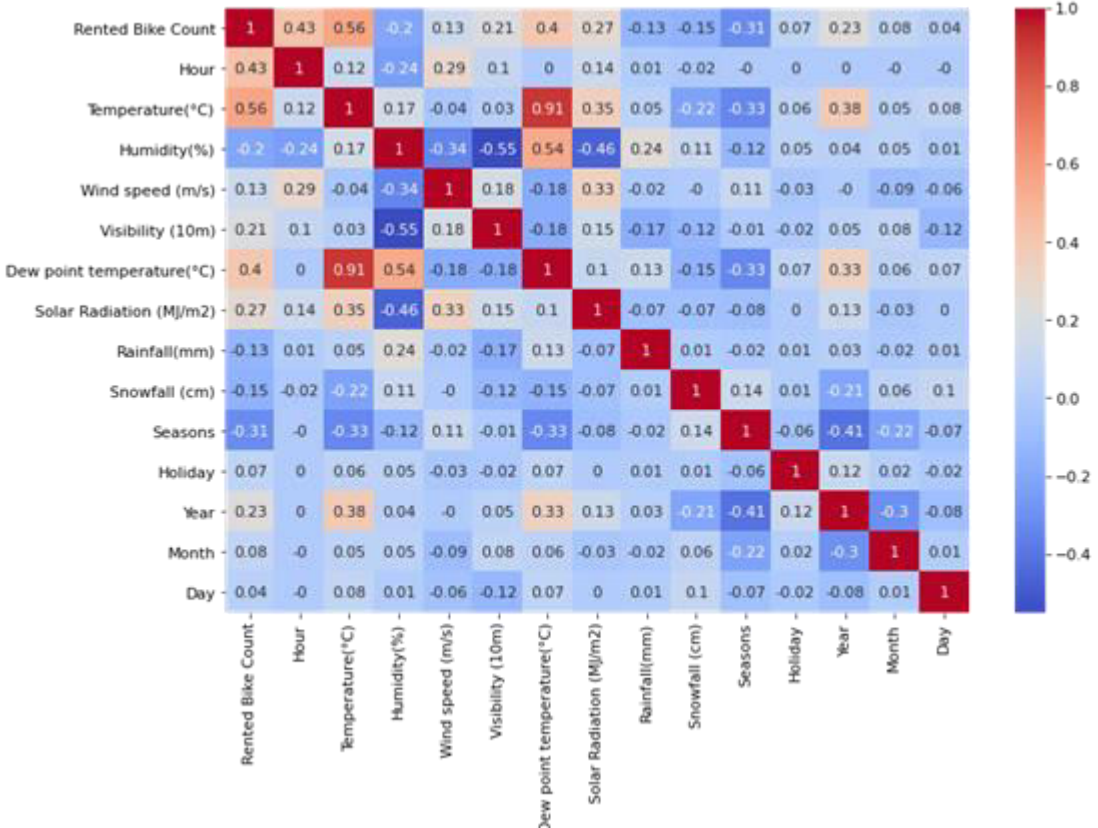




Multicollinearity

- **Variables like Dew Point Temperature, and Temperature are highly correlated.**





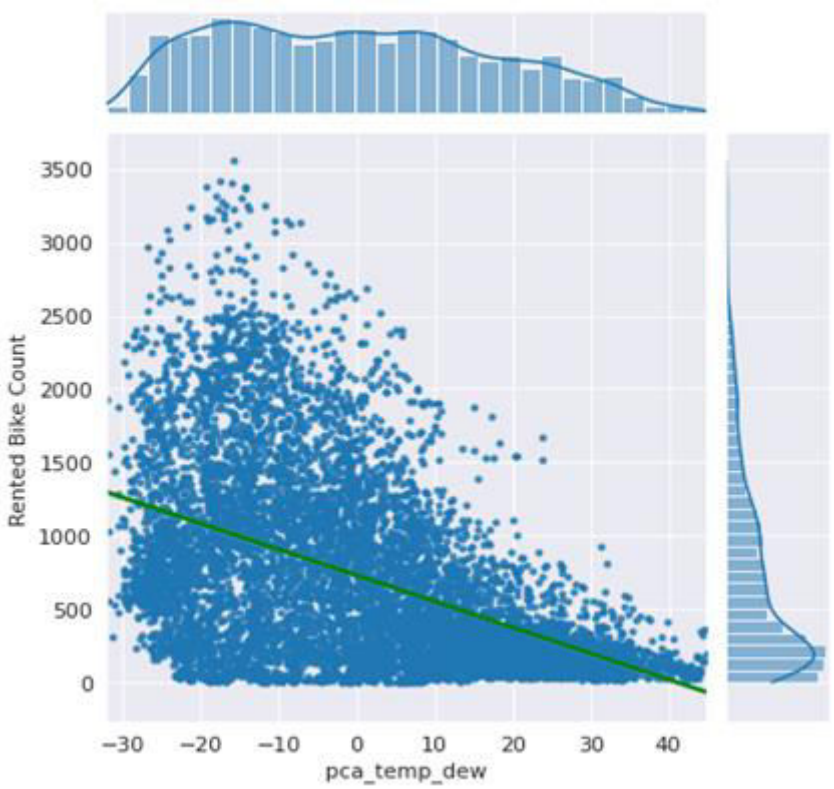
Data Pre-processing

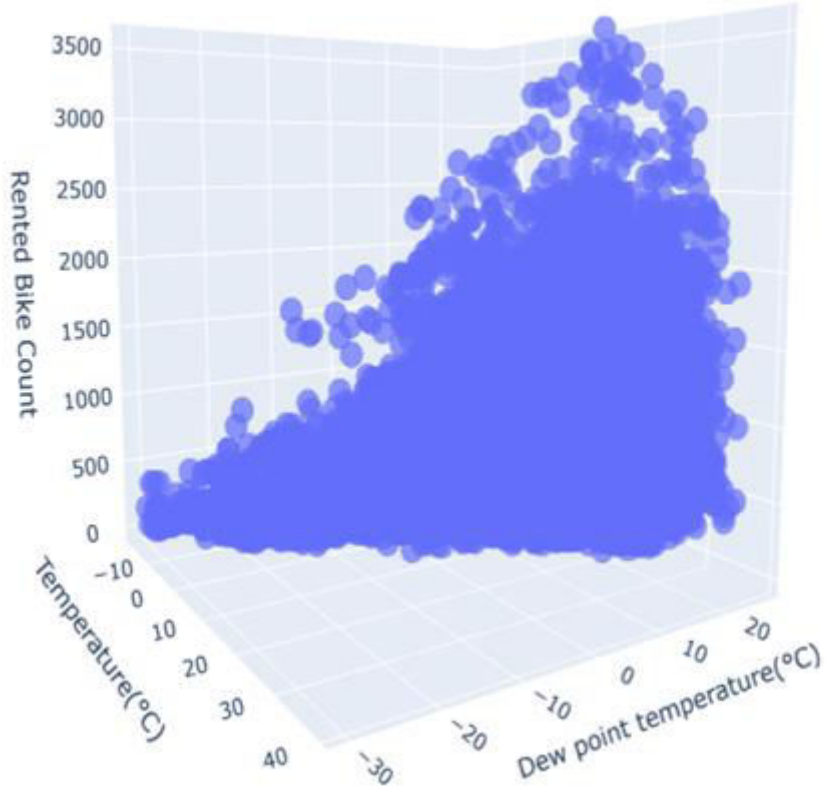
- Implemented PCA “Dew Point Temp” and “Temperature” as they were highly correlated.

- Also removed observations where it was “Nonfunctional Day” and bike rent was zero and removed the column as well.

- We have also dropped “Date” column as it would not help in giving good prediction for model.

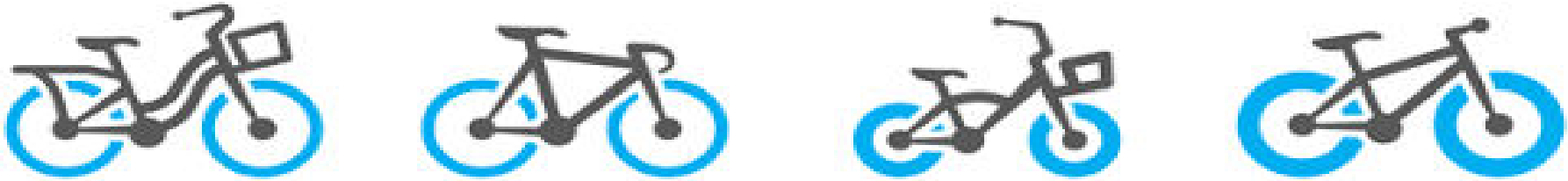
PCA For Temperature and dew point temp





Model Building

1. Linear Regression - Lasso, Ridge
2. Decision Trees
3. Random Forest
4. Gradient Boosted DT
5. Cat Boost
6. XGBoost



Best Parameters using Bayesian optimisation for the Xgboost

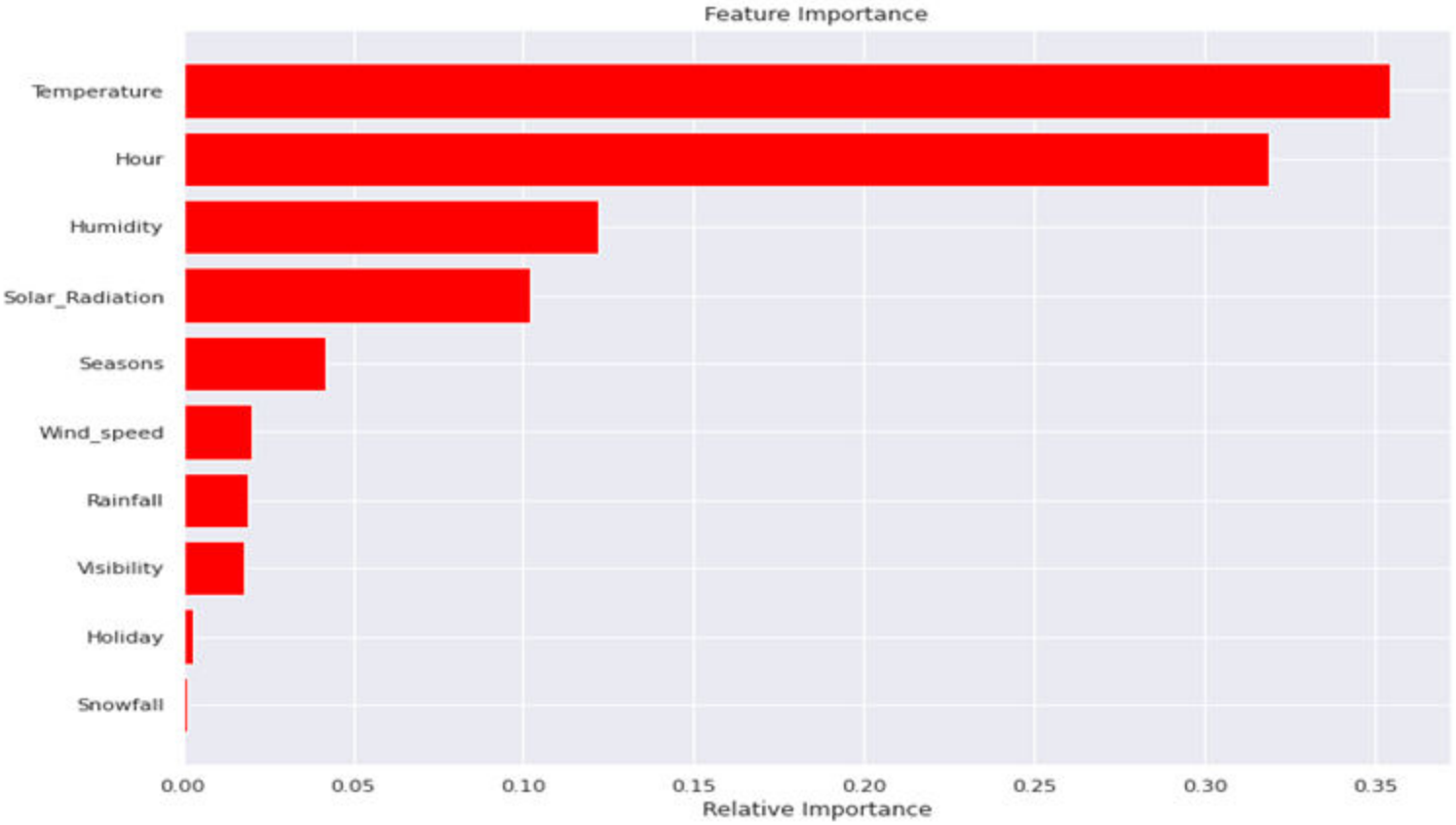
```
'base_score': 0.5,          'max_depth': 9,
'min_child_weight': 20,          'n_estimators': 100,
'objective': 'reg:linear', 'reg_alpha': 0,
'random_state': 0,          'scale_pos_weight': 1,
'reg_lambda': 1,                    'subsample': 1,

'booster': 'gbtree',        'colsample_bylevel':            1,
      'colsample_bynode': 1,     'colsample_bytree': 0.7,
'eta': 0.004,               'gamma': 0,
      'importance_type':'gain',  'learning_rate': 0.1,
```

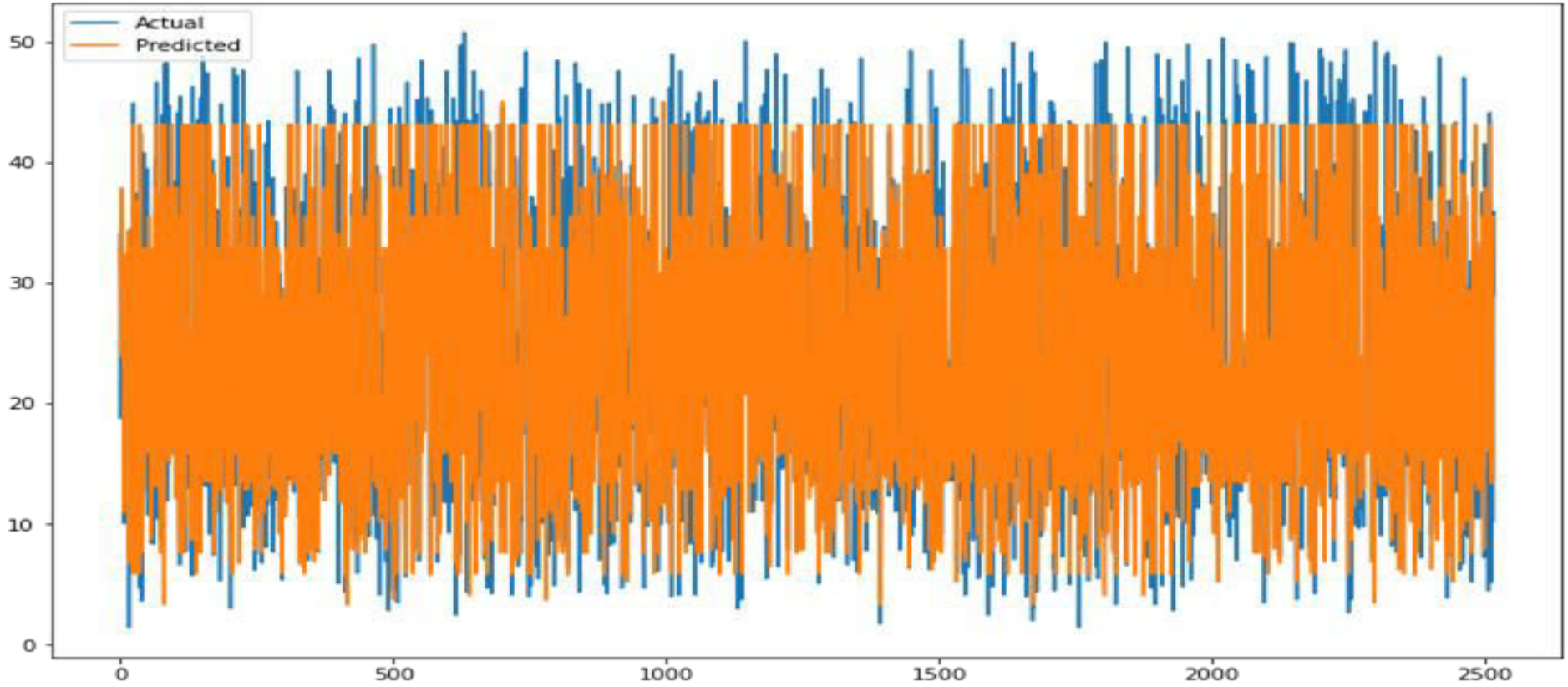
Bayesian optimisation parameters Table

iter	target	colsam...	gamma	max_depth
1	-4.18	0.7168	0.5418	3.979
2	-3.788	0.7322	0.6974	5.94
3	-4.374	0.351	0.253	3.475
4	-3.683	0.9	1.0	7.0
5	-3.705	0.8951	0.07094	6.993
6	-3.894	0.3	1.0	7.0
7	-3.702	0.8995	0.6334	6.777
8	-3.698	0.9	0.0	6.16

Feature Importance

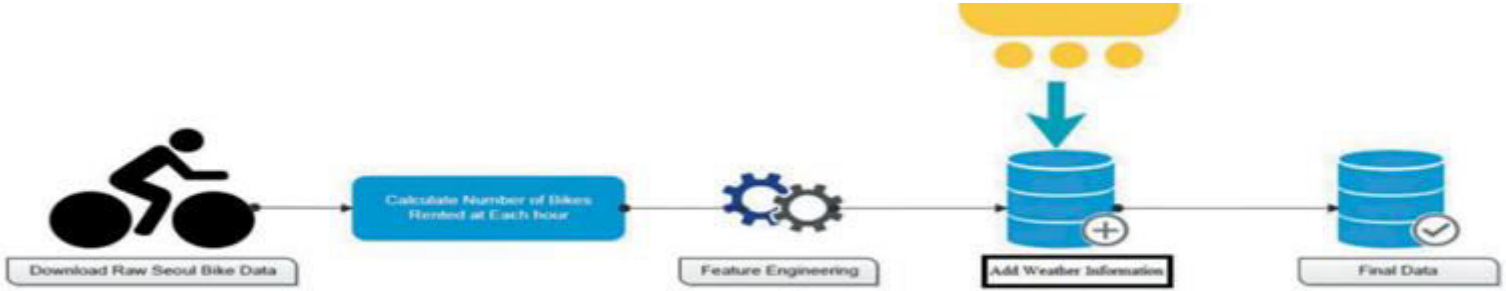


Actual Vs Predicted



Challenges

- Feature engineering
- Feature selection
- Model Training and performance improvement



Conclusion

- When we compare the root mean squared error and mean absolute error of all the models, the XGBoost model has less root mean squared error and mean absolute error, ending with the R-squared of 94% . So, finally this model is best for predicting the bike rental count on daily basis.
- As we can see the total amount of bike rentals increases with the temperature per month. Whereas it seems that the rentals are independent of the wind speed and the humidity, because they are almost constant over the months. This also confirms on the one hand the high correlation between rentals and temperature and on the other hand that nice weather could be a good predictor. So people mainly rent bikes on nice days and nice temperature. This could be important of planning new bike rental stations.

Q & A