Bike Sharing demand Prediction

Submitted by:
Mukti Verma
Cohort- Hardeol, Alma Better-2022 Batch

**Recommended to view the story in slideshow mode for better experience*

Agenda

1. Problem Understanding
2. Approach Overview & Data Dictionary
3. Univariate, Bivariate & Multivariate Analysis
4. Model Implementation and comparison
5. Recommendation & Conclusion

Image source: https://inhabitat.com/6-awesome-bike-sharing-schemes-from-around-the-world/bike-shares-bicing-inhabitat-top-6/

Business Problem Understanding

Context & 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. Hence, it is imperative for the business to predict the demand of bikes at hourly basis for the stable supply of rental bikes.

Objective

- ***Explore and analyze the data to discover key factors affecting shortage of bikes and time delay in availing bike.***
- ***Build a model to predict number of bikes required per hour, so that business can -***
  - Maximize the bike availability
  - Minimize the waiting period
  - Provide sufficient pick and drop stations
  - Ensure stable supply of bikes

Methodology used:

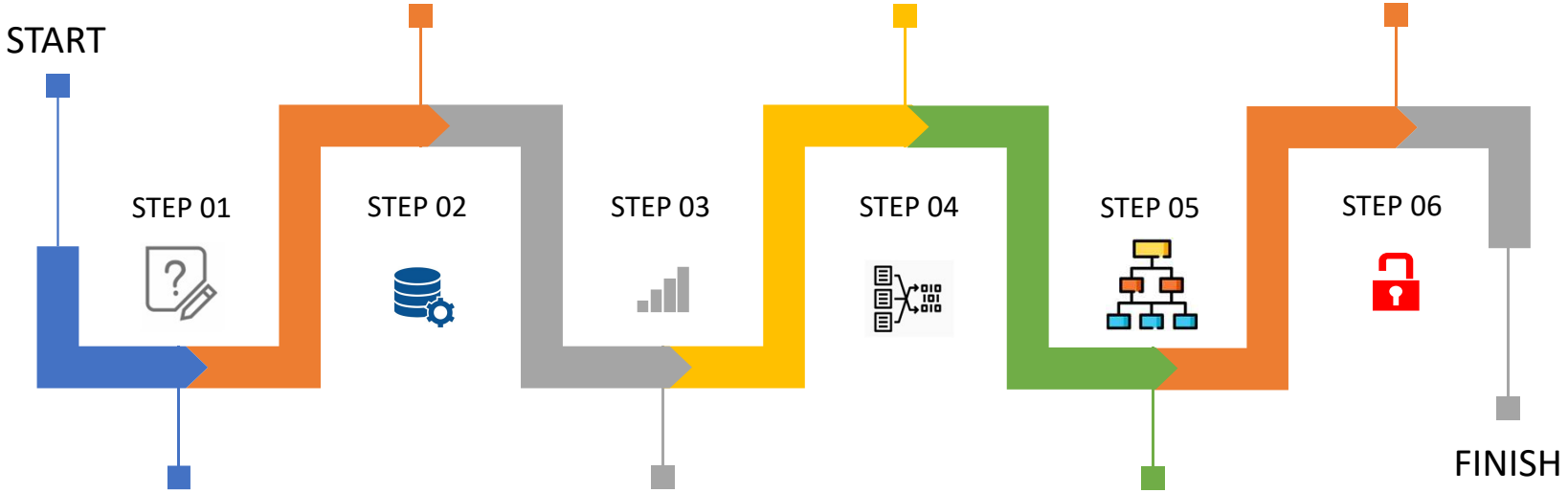
- Exploratory data analysis: Techniques - Univariate Analysis, Bivariate Analysis, Correlation heatmap, pairplot
- Model building: Linear Regression, Ridge, Lasso, ElasticNet, , Decision Tree, Random Forest, Gradient Boosting, Gradient Boosting Gridsearchcv

Approach Overview

Data Sanity Check & Cleansing
Identifying data types of variables, Null value checks

Variable Transformation
Encoding of Categorical variables

Selecting the best model
Model evaluation based on key metrics- Accuracy, MAE, MSE & Rsquared error



Understanding Problem Statement

Descriptive Analytics (EDA)
Univariate, Bivariate & Multivariate Analysis

Building Predictive Model
Predictive Model to predict the bike demand per hour

Data Dictionary

14 variables – 1 target variable with 2 categorical, 2 Boolean and 10 numerical variables. Each data record indicates bike rented per hour

Date :
The date of the day, during 365 days from 01/12/2017 to 30/11/2018, formatting in DD/MM/YYYY, we need to convert into datetime format.

Hour:
The hour of the day, starting from 0 to 23.

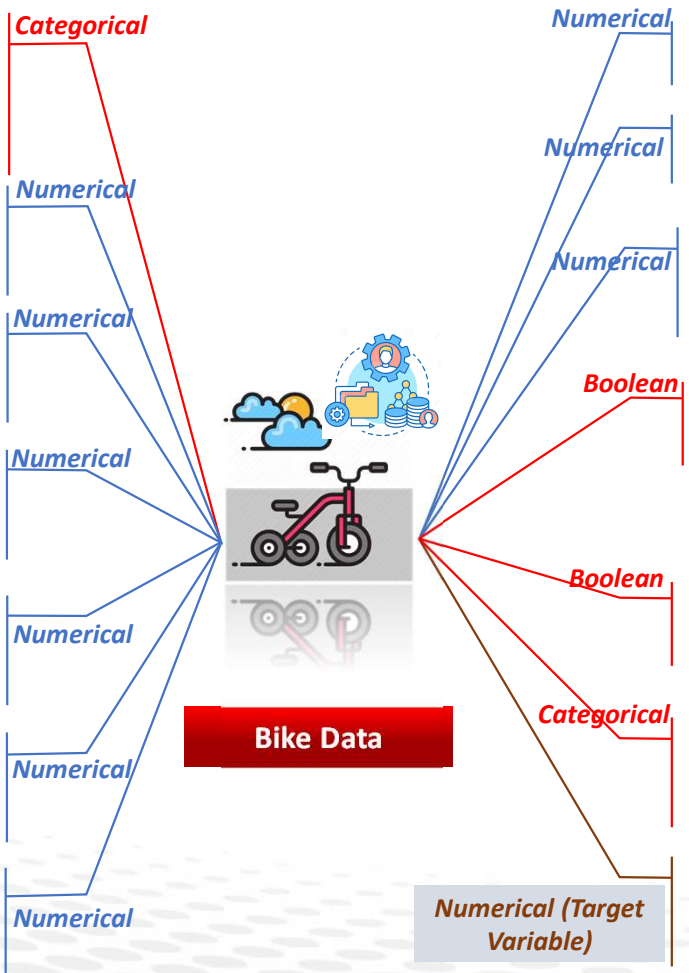
Temperature(°C): Temperature in Celsius

Humidity(%): Humidity in the air in %

Wind speed (m/s) : Speed of the wind in m/s

Visibility (10m): Visibility in 10m

Dew point temperature(°C): Temperature at the beginning of the day



Solar Radiation (MJ/m2): Sun contribution

Rainfall(mm): Amount of raining in mm

Snowfall (cm): Amount of snowing in cm

Holiday: If the day is holiday period or not

Functioning Day: If the day is a Functioning Day or not

Seasons: Season of the year, there are only 4 season's in data

Rented Bike Count :
Number of rented bikes per hour, this is the dependent variable

Image source: https://www.iconfinder.com/icons/4060621/bike_cycle_tricycle_icon

Univariate Analysis

Insights from EDA: Data Summary

No null values in any of the variables and no duplicated values in the record

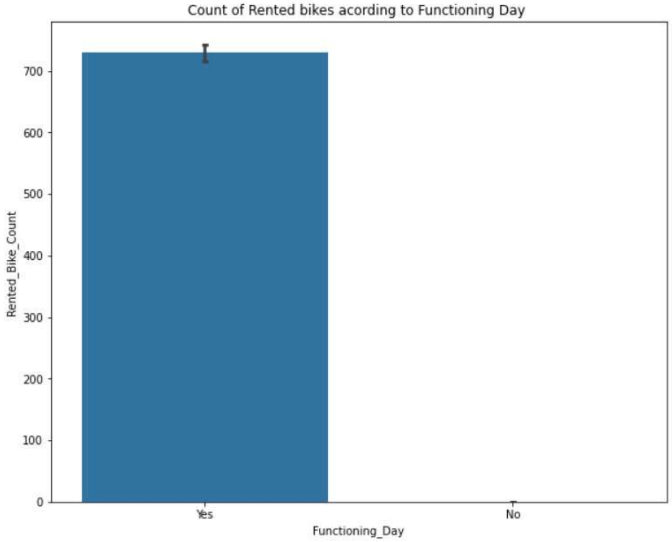
	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
count	8760	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760	8760	8760
unique	365	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4	2	2
top	01/12/2017	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Spring	No Holiday	Yes
freq	24	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2208	8328	8465
mean	NaN	704.602055	11.500000	12.882922	58.226256	1.724909	1436.825799	4.073813	0.569111	0.148687	0.075068	NaN	NaN	NaN
std	NaN	644.997468	6.922582	11.944825	20.362413	1.036300	608.298712	13.060369	0.868746	1.128193	0.436746	NaN	NaN	NaN
min	NaN	0.000000	0.000000	-17.800000	0.000000	0.000000	27.000000	-30.600000	0.000000	0.000000	0.000000	NaN	NaN	NaN
25%	NaN	191.000000	5.750000	3.500000	42.000000	0.900000	940.000000	-4.700000	0.000000	0.000000	0.000000	NaN	NaN	NaN
50%	NaN	504.500000	11.500000	13.700000	57.000000	1.500000	1698.000000	5.100000	0.010000	0.000000	0.000000	NaN	NaN	NaN
75%	NaN	1065.250000	17.250000	22.500000	74.000000	2.300000	2000.000000	14.800000	0.930000	0.000000	0.000000	NaN	NaN	NaN
max	NaN	3556.000000	23.000000	39.400000	98.000000	7.400000	2000.000000	27.200000	3.520000	35.000000	8.800000	NaN	NaN	NaN

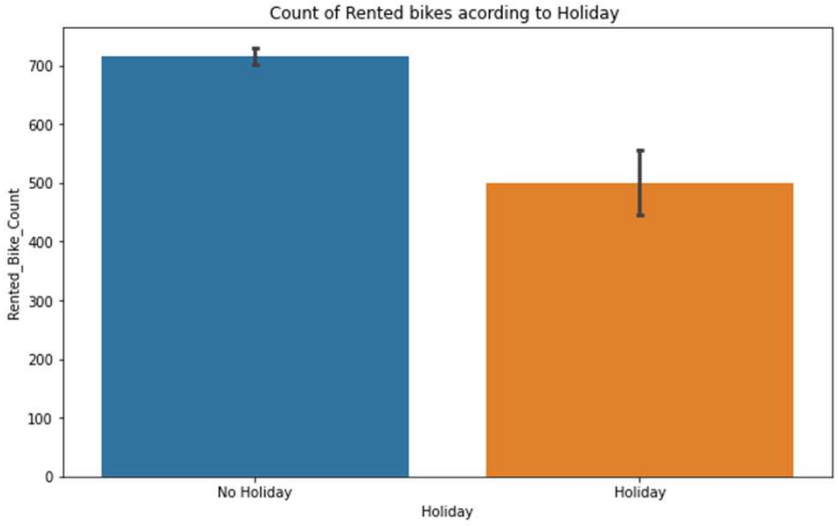
```
Date                          0
Rented Bike Count             0
Hour                          0
Temperature(°C)               0
Humidity(%)                   0
Wind speed (m/s)              0
Visibility (10m)              0
Dew point temperature(°C)     0
Solar Radiation (MJ/m2)       0
Rainfall(mm)                  0
Snowfall (cm)                 0
Seasons                       0
Holiday                       0
Functioning Day               0
dtype: int64
```

- Above table shows the data summary of all the variables. There are 2 categorical variables, 2 Boolean variables & 10 numerical variables
- Temperature at that place varies from -17.8 to max 39.4**°C, with average temp 12.88.**
- Minimum humidity is 0 whereas maximum have been observed 98% with an average of 58.22%.
- Average wind speed is 1.72 m/s, with the range of 0 – 7.4 m/s.
- Visibility (10m) is varying from 27 to 2000 with considerable standard deviation of 608.299.
- Dew point temperature is varying from -30.60**°C** to 27.2**°C with an average of 4.07 °C.**
- Minimum solar radiation has been found 0, with std deviation in solar radiation 0.868 (MJ/m2), highest radiation has been recorded 3.5 (MJ/m2).
- Maximum rainfall is 35mm, average rain fall observed is 0.148 mm.
- Average snowfall is observed 0.075 cm with maximum 8.8 cm.

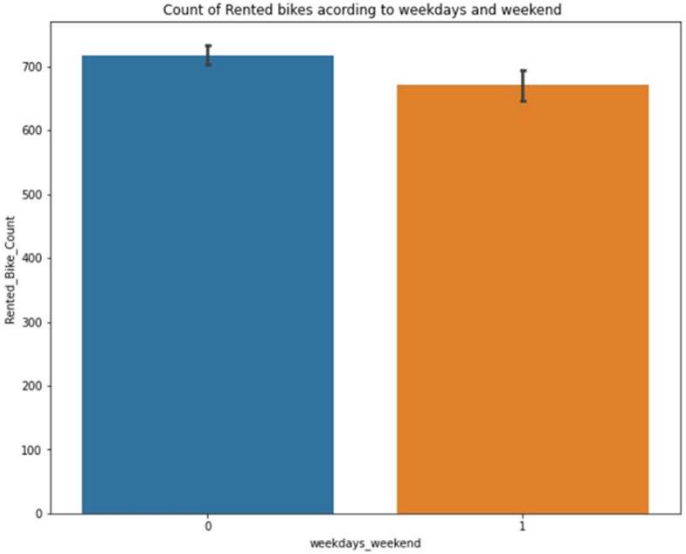
Insights from EDA: Univariate Analysis

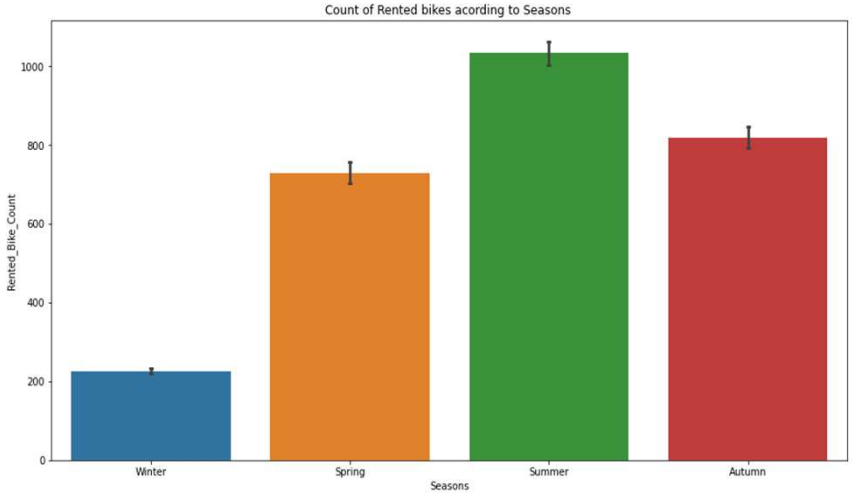
Categorical variables: Countplot of Target variable with respect to all categorical variable





- Functioning days (hours) have comparatively very high demand than non functioning day (hours)
- Bike rented on holidays are comparatively lesser
- **From both of these graphs it can be concluded bikes might be rented by working professionals**

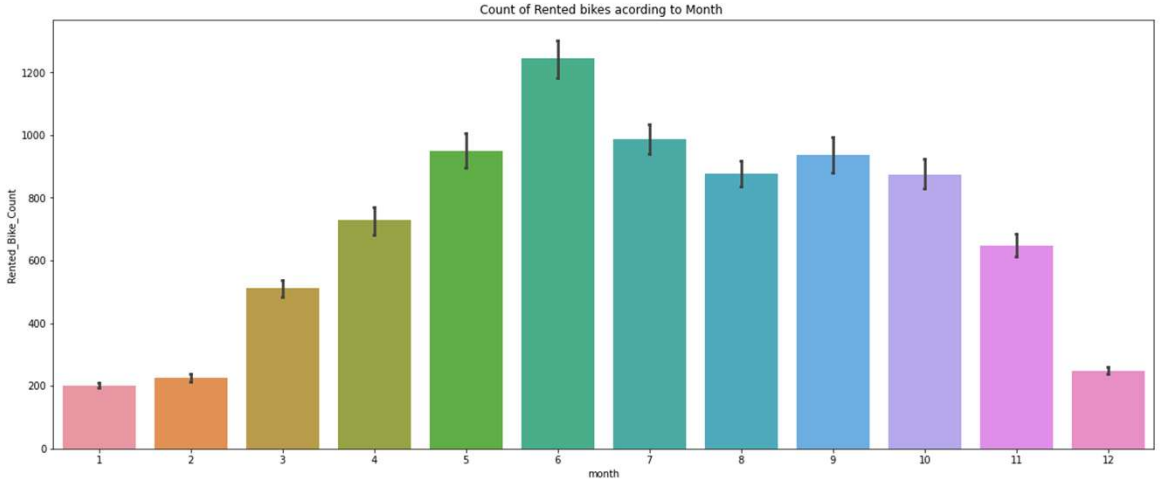


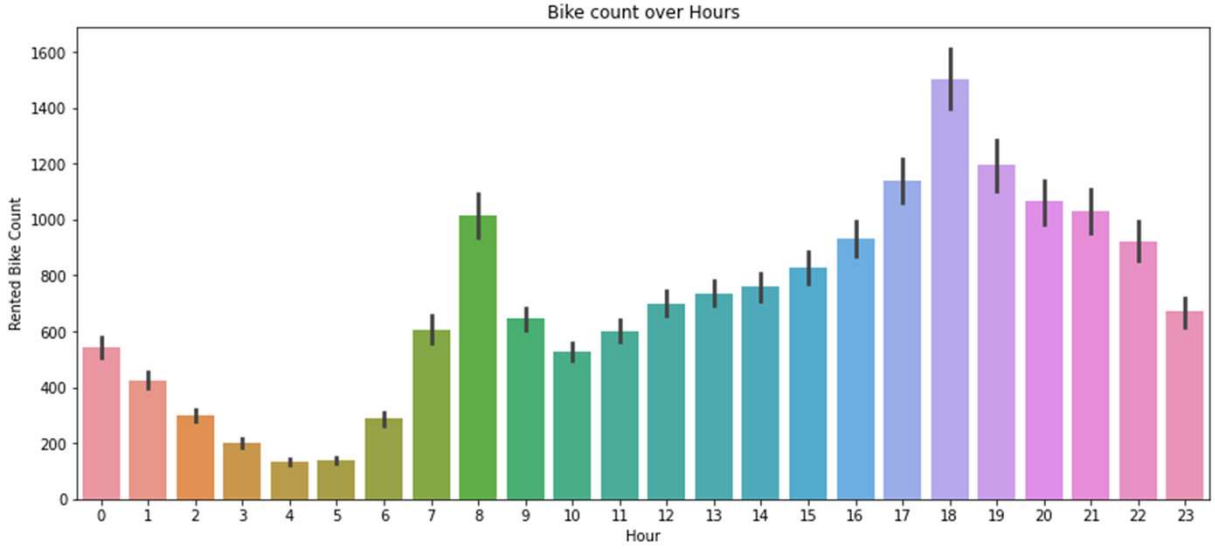


- However, bikes rented on weekdays and weekends is approximately same indicating some professionals might be even working on weekends
- Bike rented in summer is highest and least in winter indicating bike demand is affected by season

Insights from EDA: Univariate Analysis

Date time variables: Countplot of Target variable with respect to date time variables

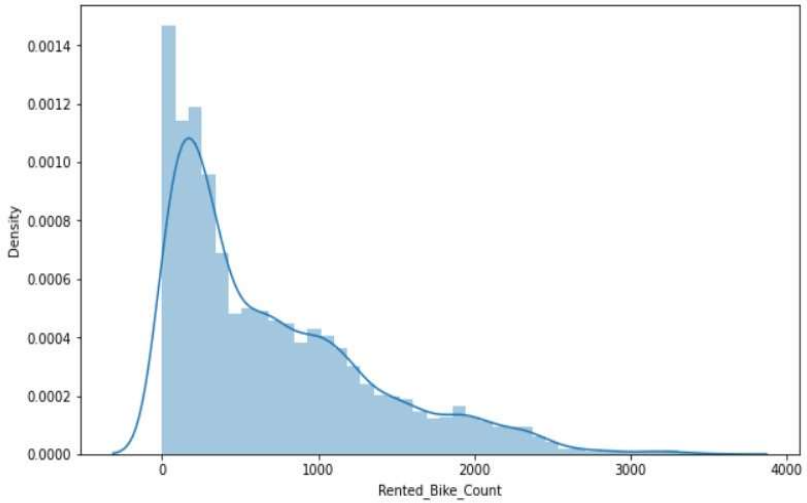


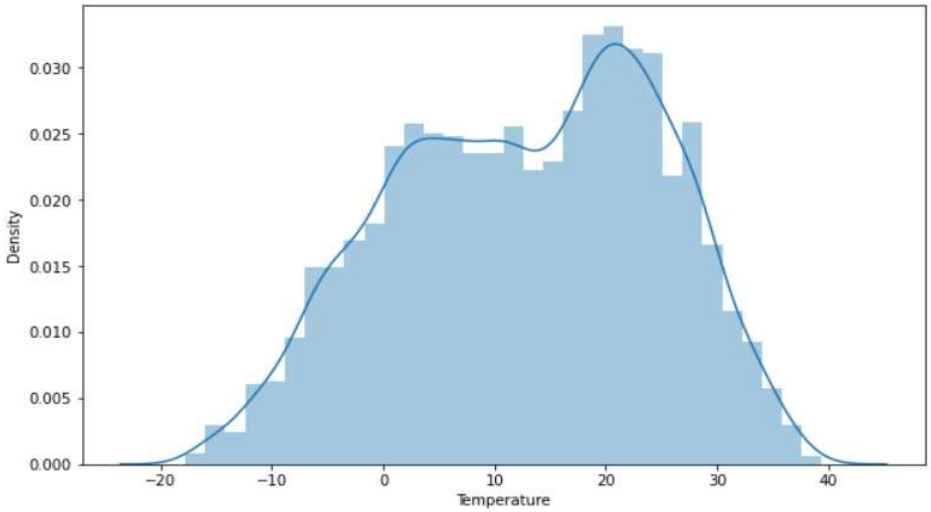


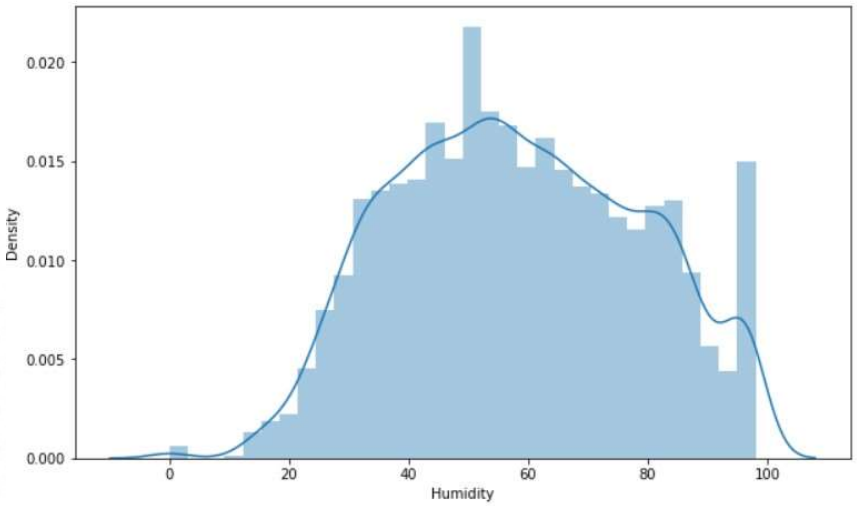
- Highest demand is observed between May to Sep, with peak in June
- Rented bike demand is highest during evening hours starting 17:00 hrs peaking at 18:00 hrs indicating majority of customers prefer to take rented bike back from office

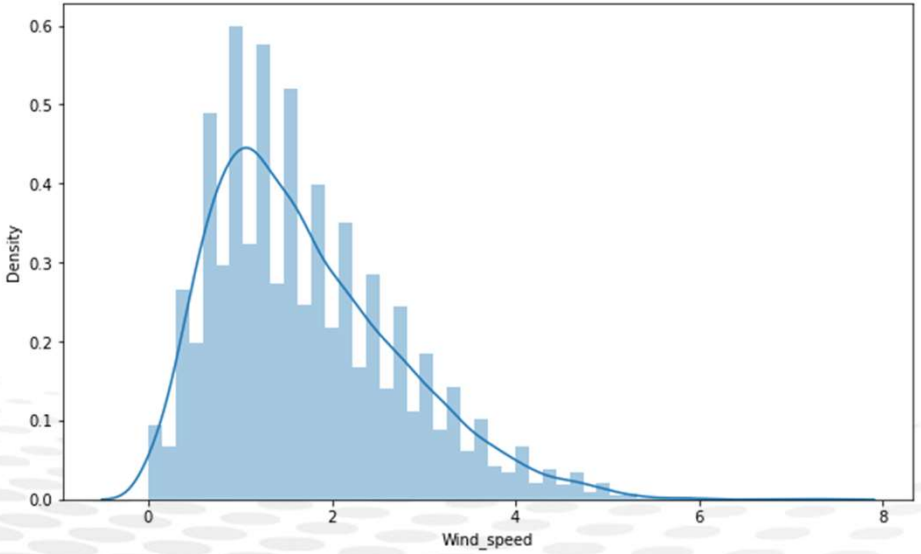
Insights from EDA: Univariate Analysis

Numerical variables: Distribution of numerical variables



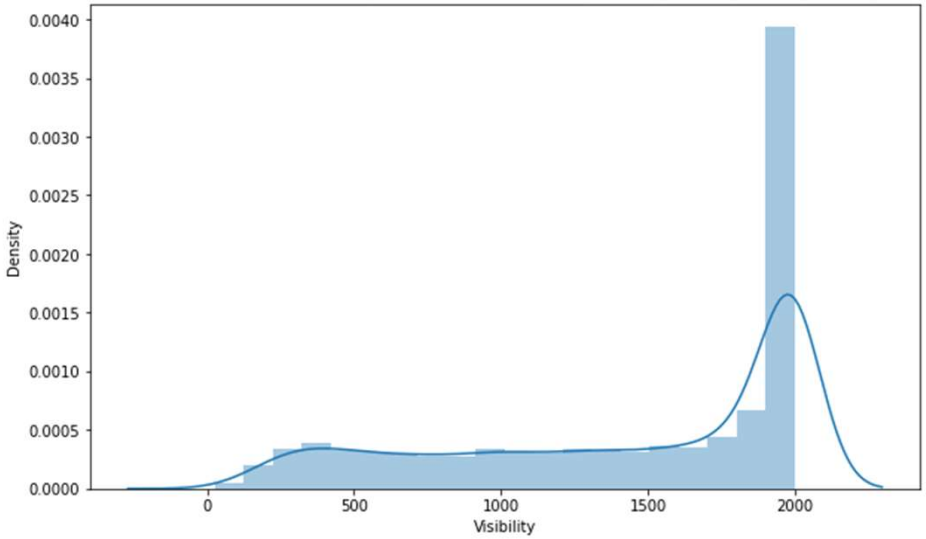


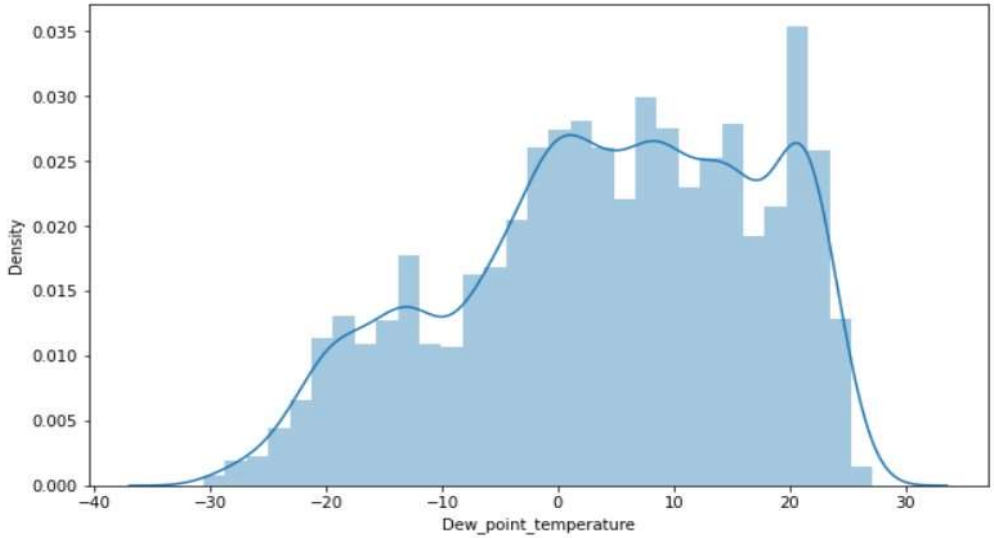


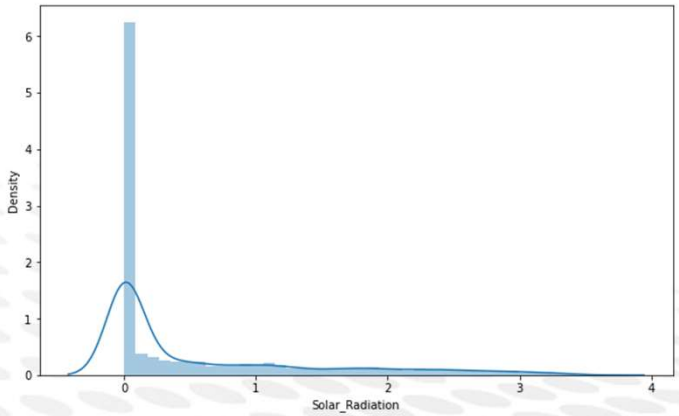


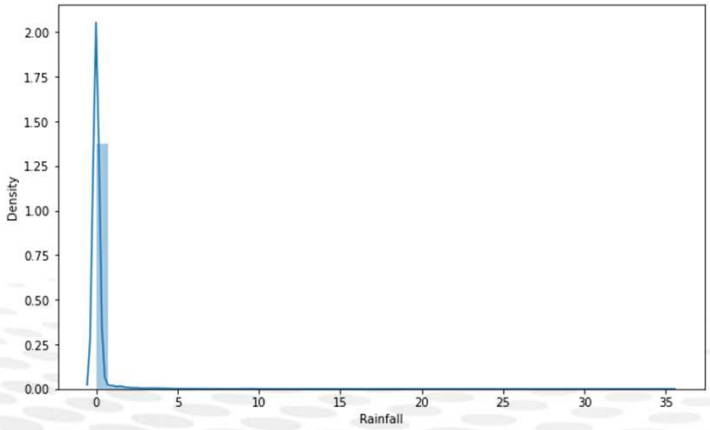
Insights from EDA: Univariate Analysis

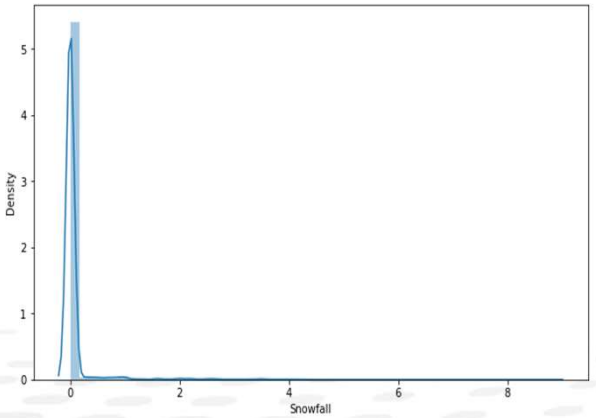
Numerical variables: Distribution of numerical variables





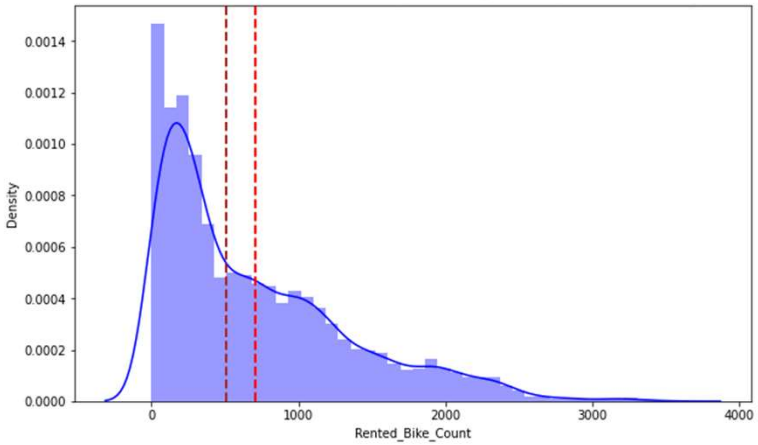


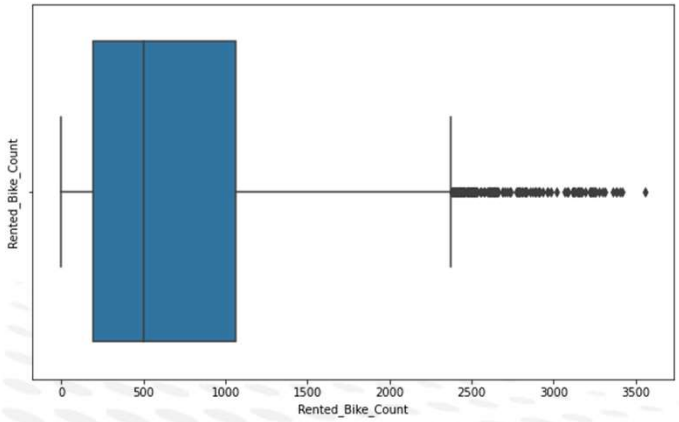


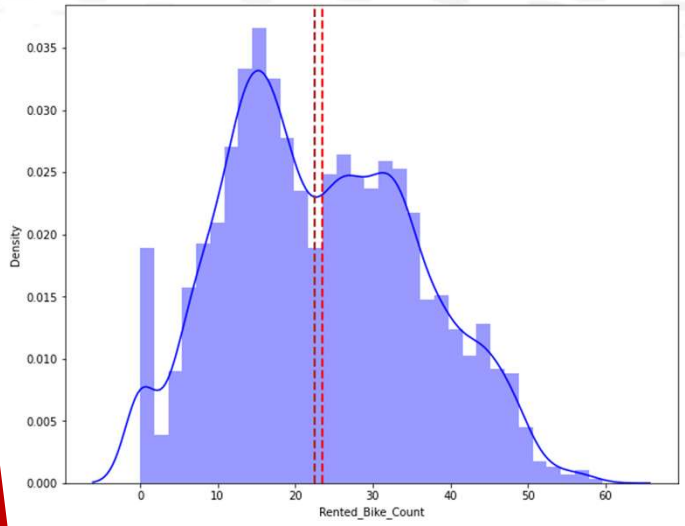


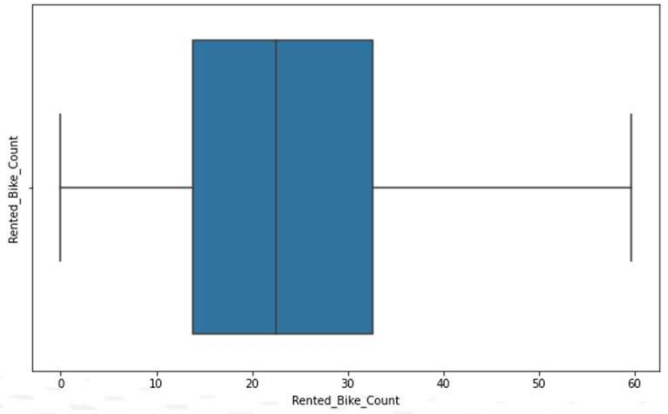
Insights from EDA: Univariate Analysis

Target variable: Improving skewness







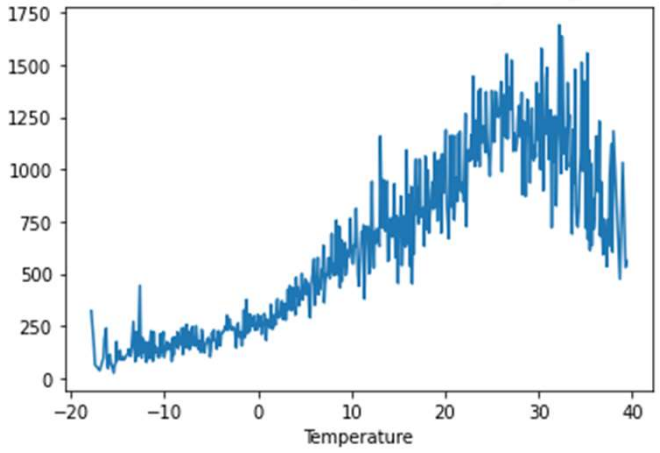


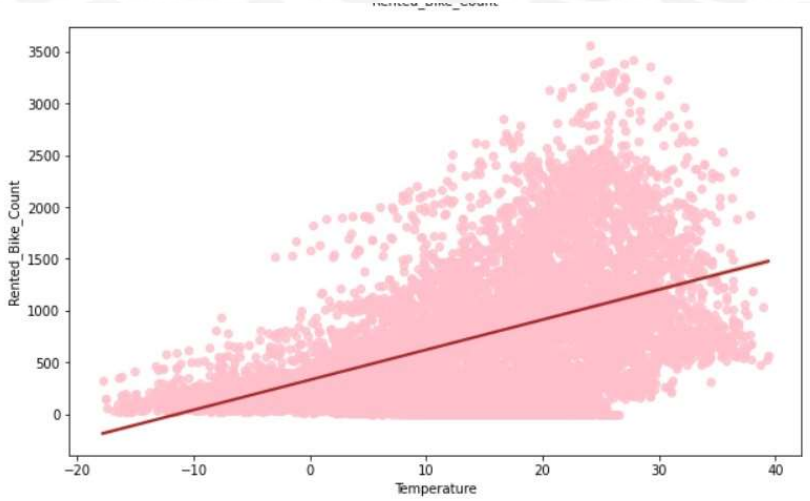
- Since many models assume the distribution of target variable to be normalized, however in our case the target variable is right skewed
- Also, there are many outliers in the target variables.
- Hence we have tried to improve the skewness by applying square root function which also resulted in outlier correction

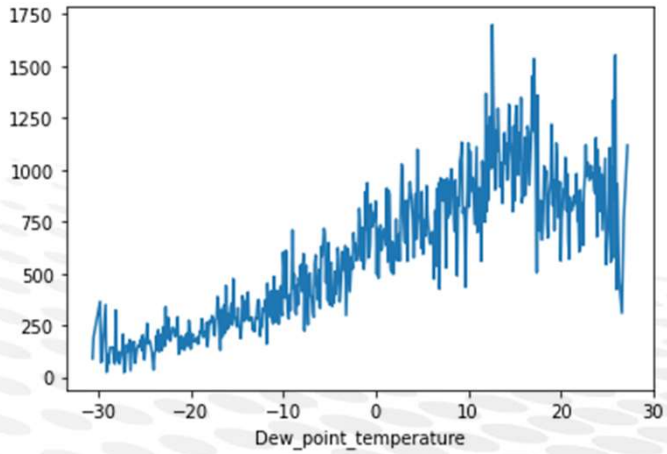
Bivariate Analysis

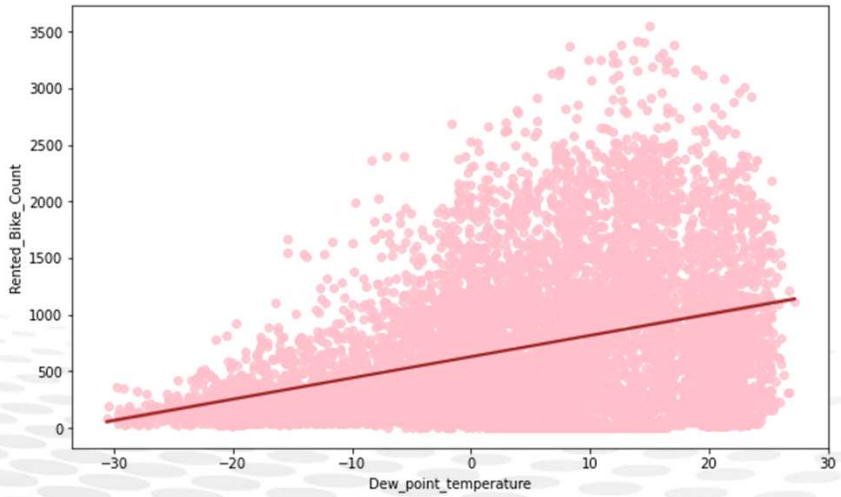
Insights from EDA: Bivariate Analysis

Numerical variables: Relationship with Target variable





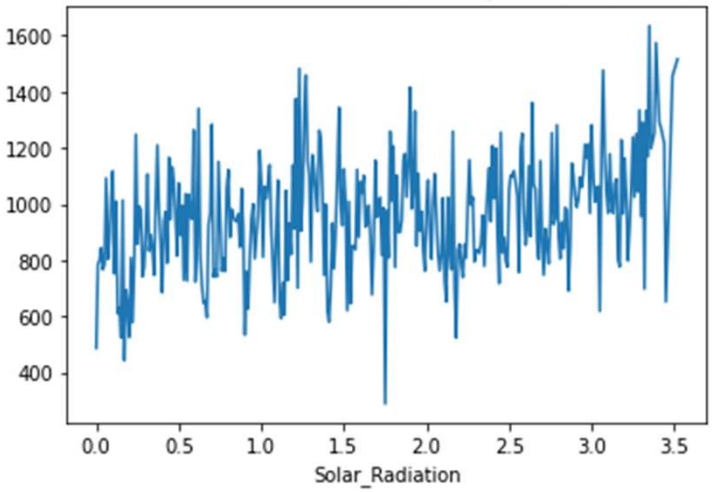


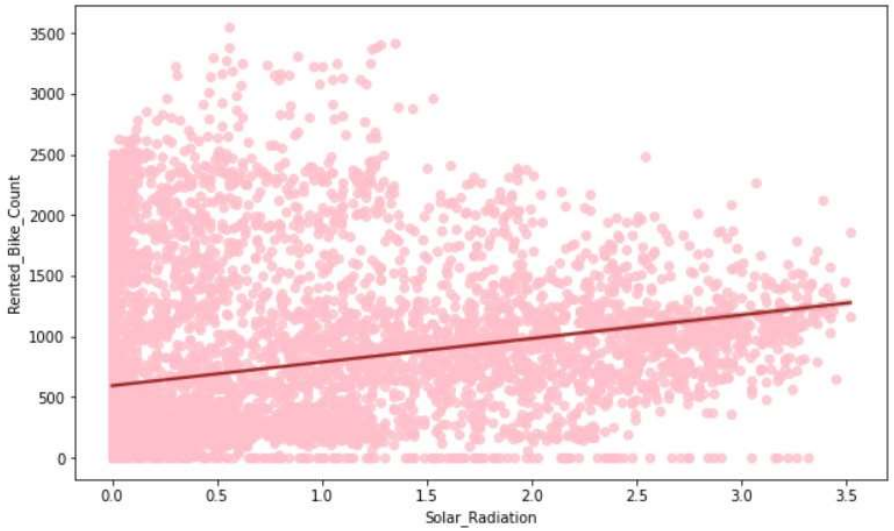


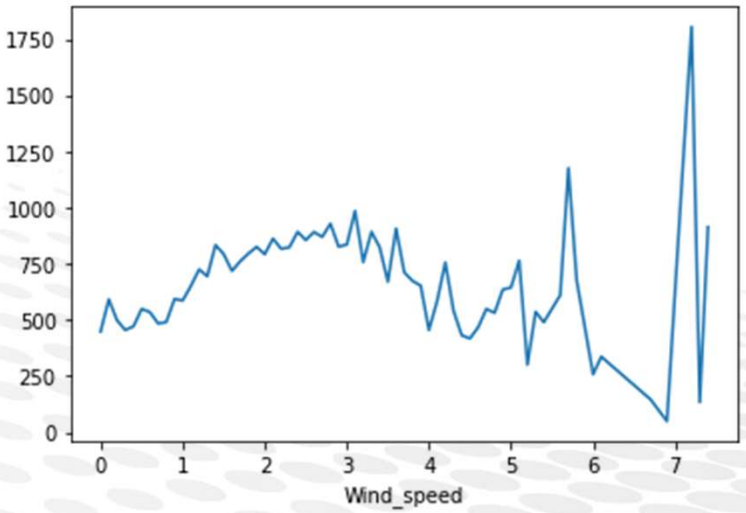
- Rented Bike count significantly increases with increase in Temperature & Dew point temperature

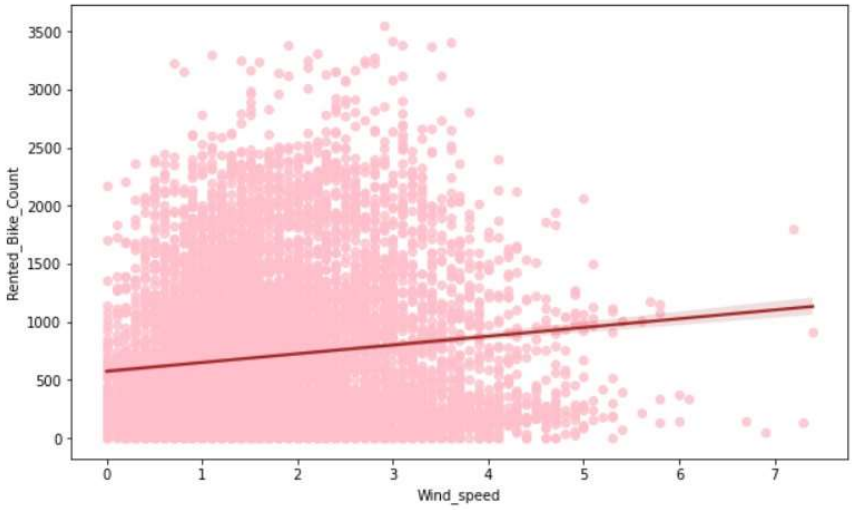
Insights from EDA: Bivariate Analysis

Numerical variables: Relationship with Target variable





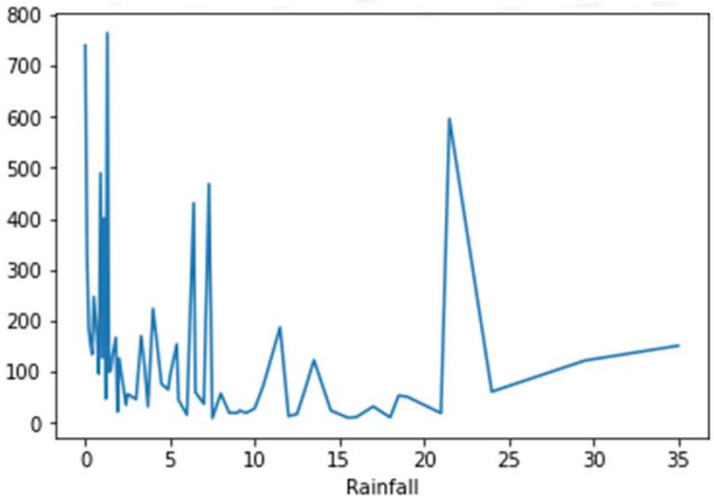


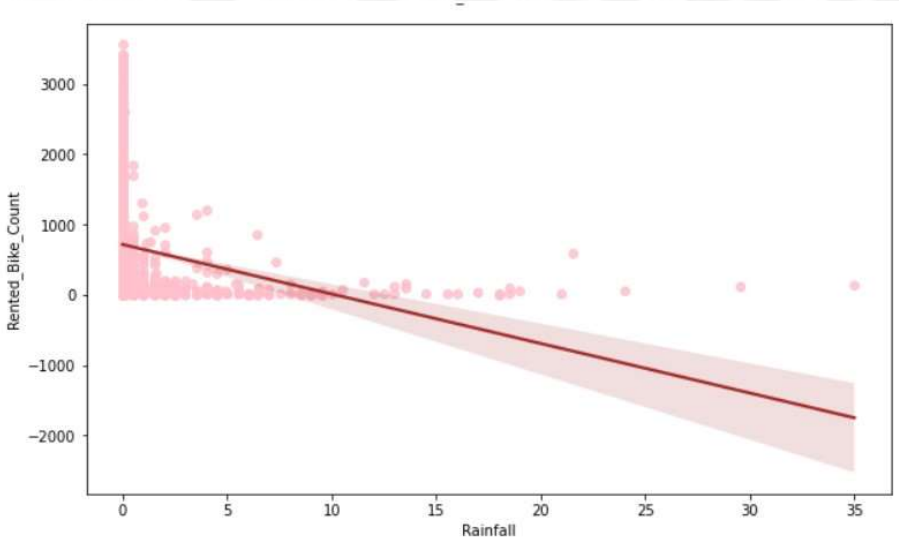


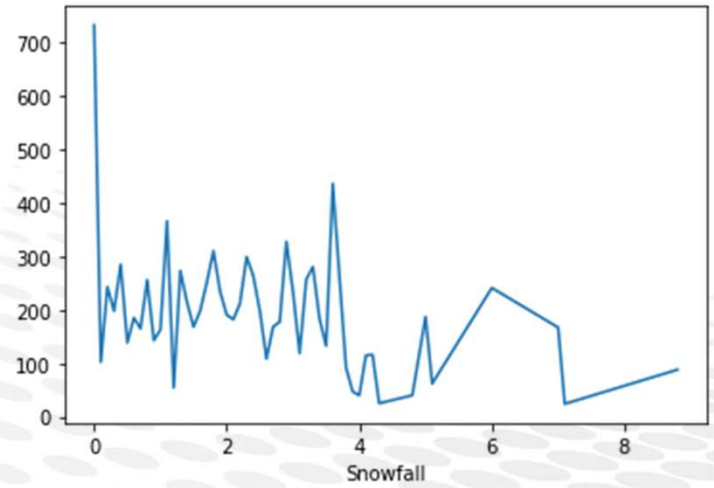
- Rented Bike count increases with Solar radiation & windspeed

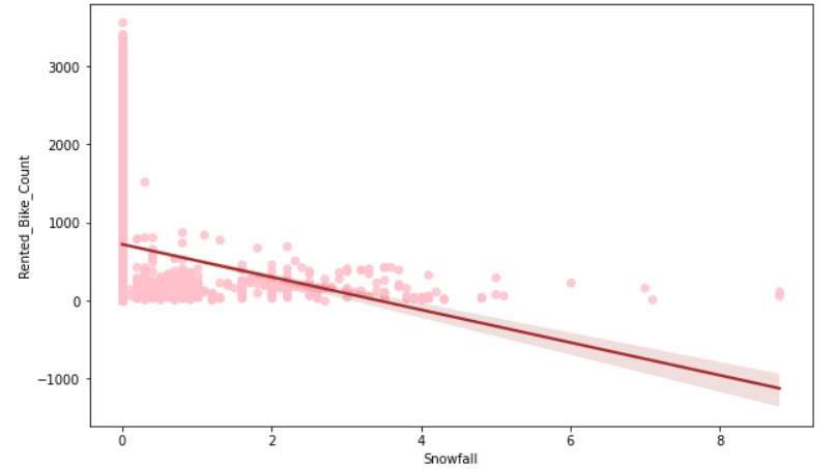
Insights from EDA: Bivariate Analysis

Numerical variables: Relationship with Target variable





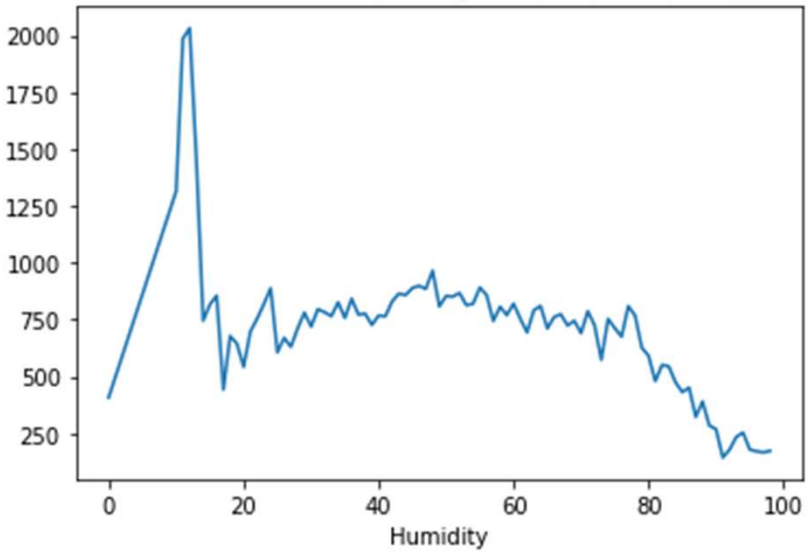


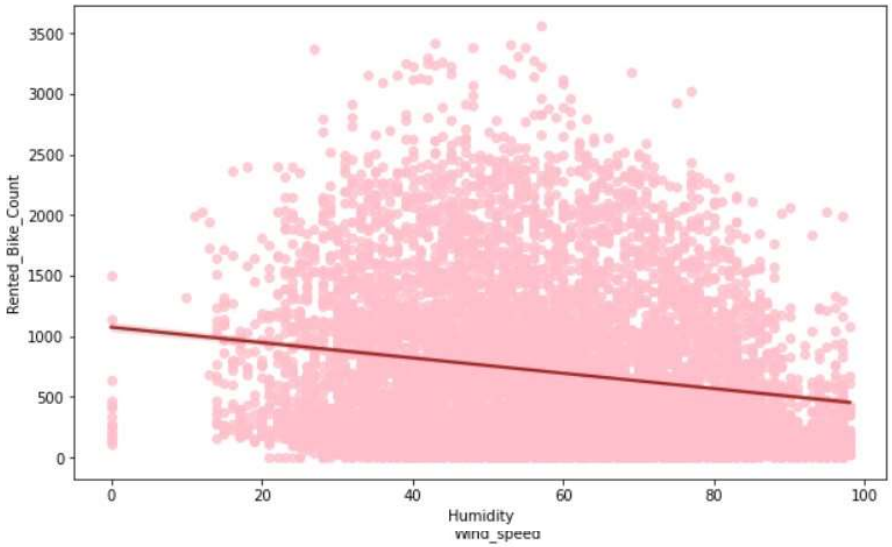


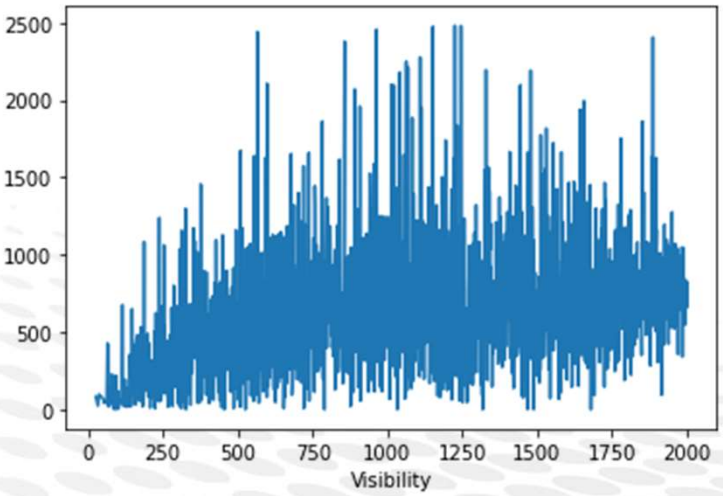
- Rented Bike count decreases with increase in Rainfall & Snowfall

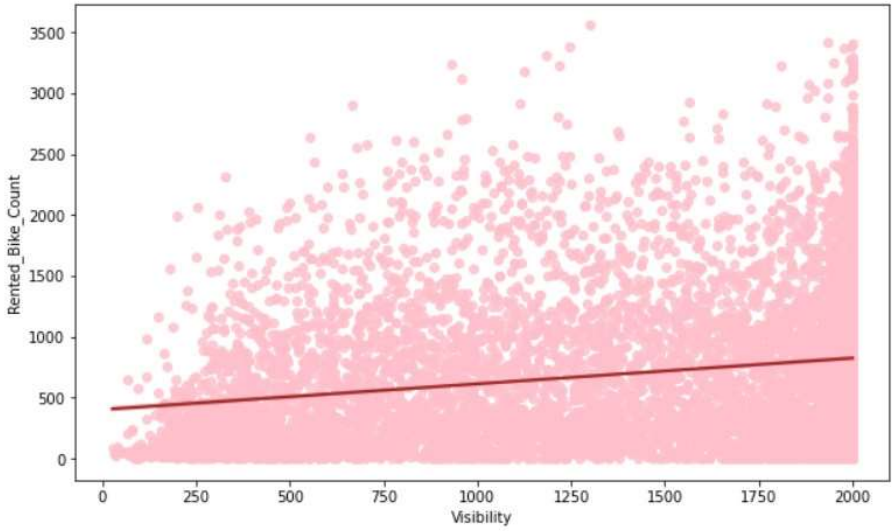
Insights from EDA: Bivariate Analysis

Numerical variables: Relationship with Target variable





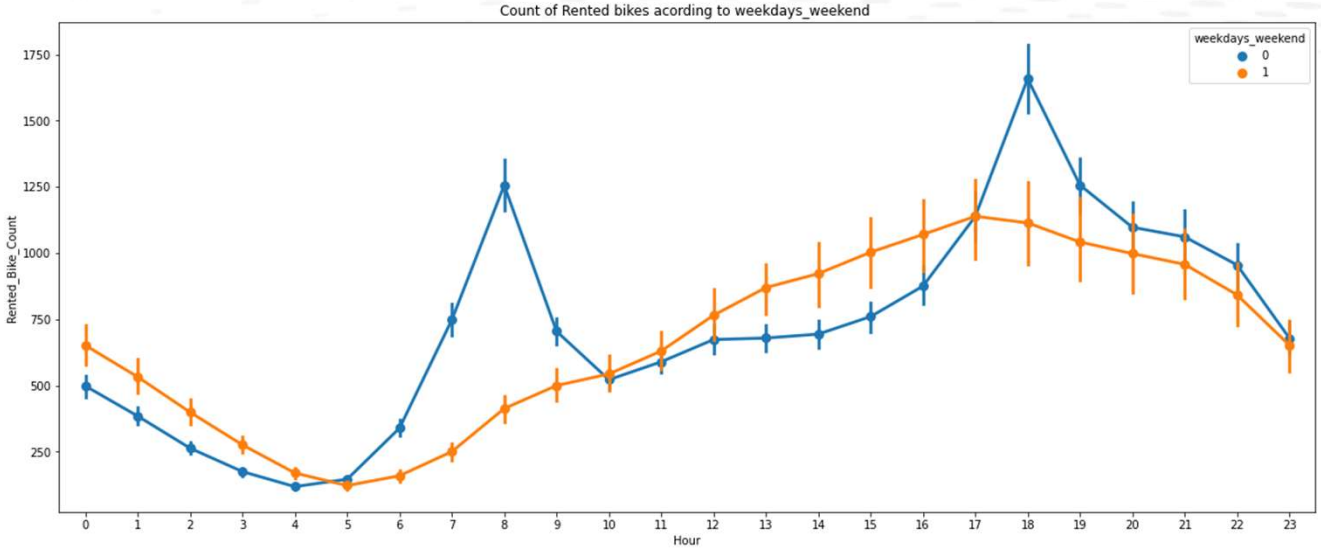


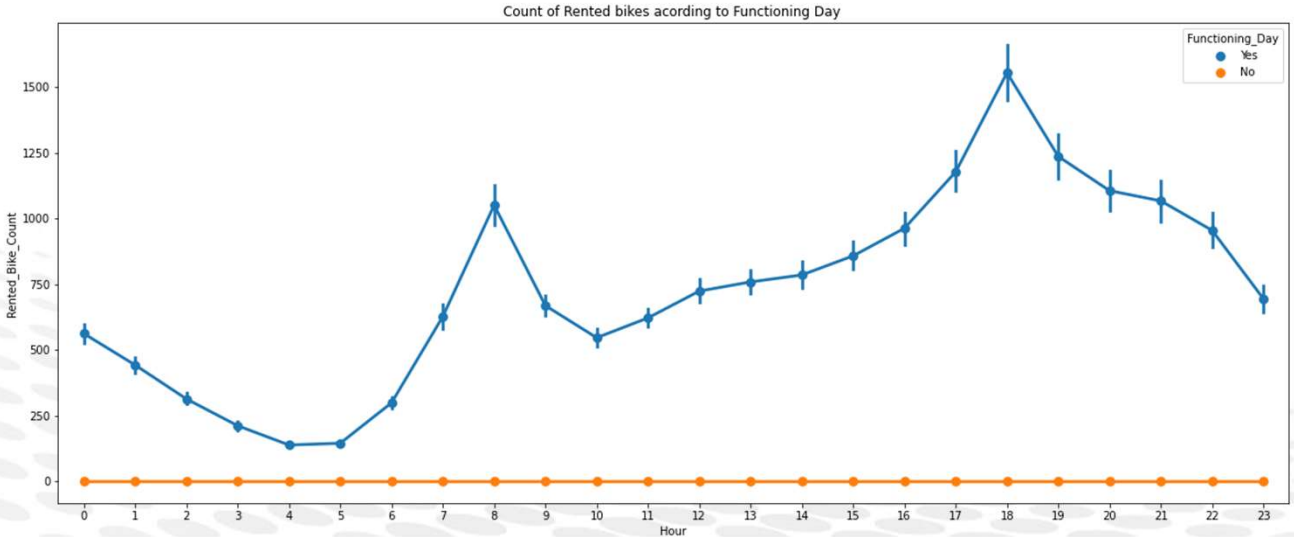


- Rented Bike count decreases with increase in Humidity but increases with increase in Visibility

Insights from EDA: Bivariate Analysis

Categorical variables: Relationship with Target variable

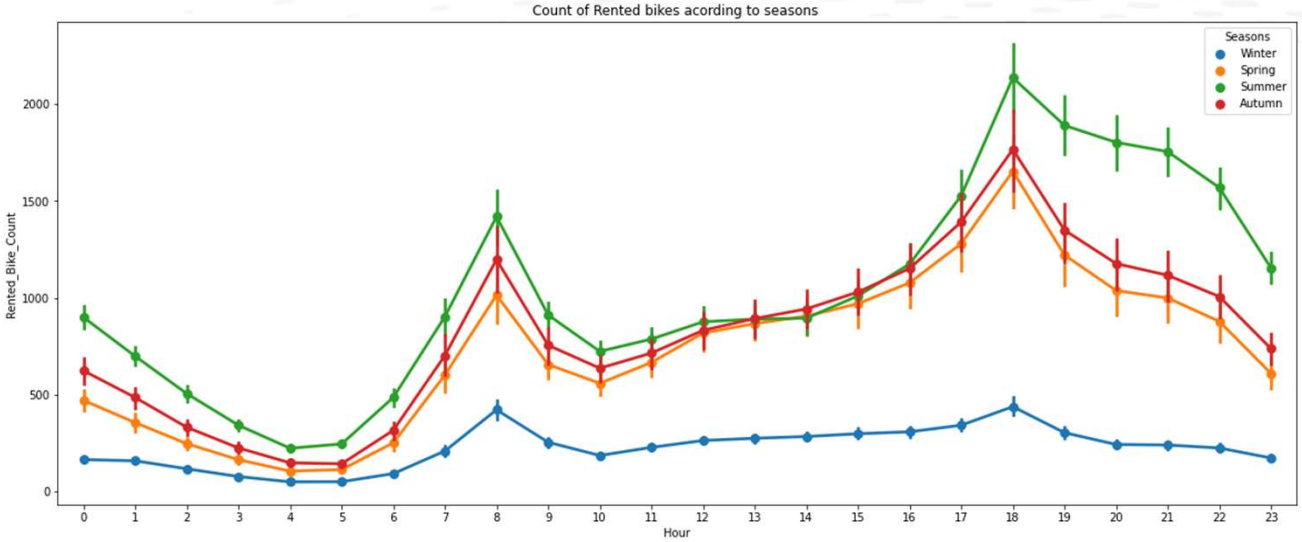


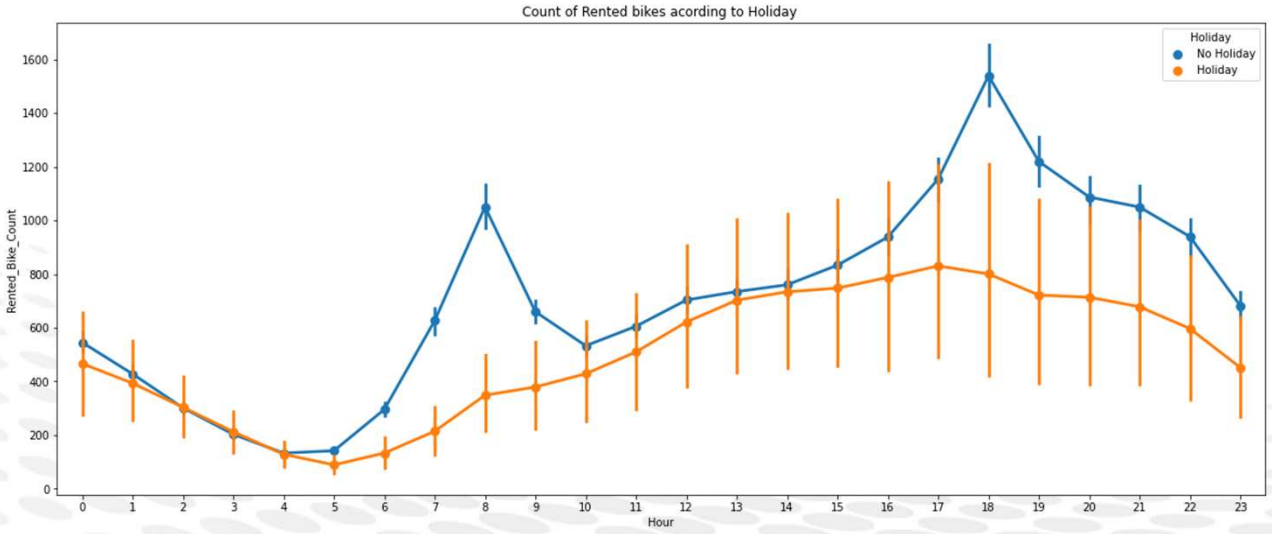


- Demand for bike is peak at 8:00 morning and 18:00 hrs in evening during weekdays. Same trend is not observed during weekends indicating majority of customers are using it for office commute.
- No demand during non functioning days/hrs.

Insights from EDA: Bivariate Analysis

Categorical variables: Relationship with Target variable



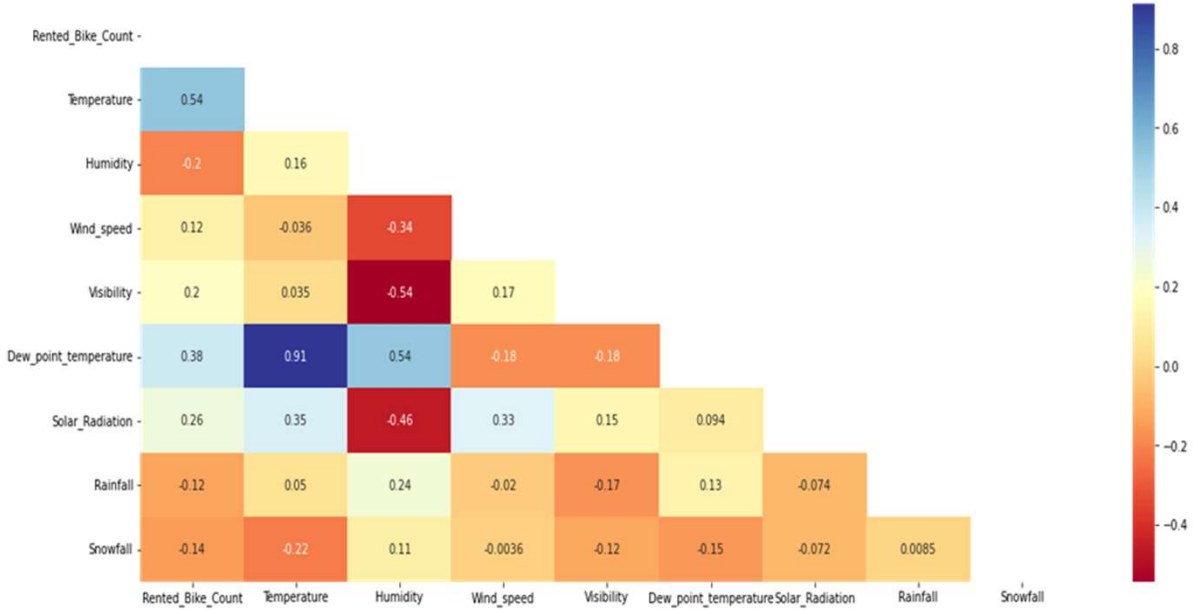


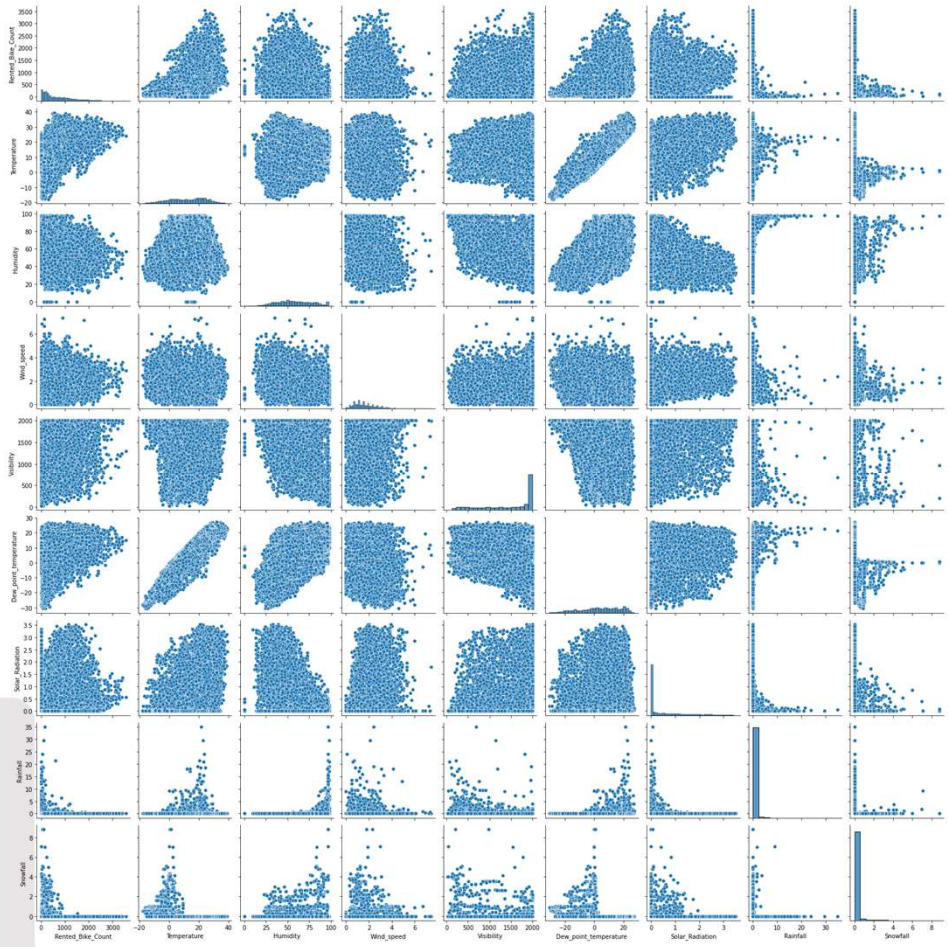
- Demand for bike is peak at 8:00 morning and 18:00 hrs in evening irrespective of season. However, this is not true for holidays
- Comparatively very less demand is observed during winter season
- Spread of demand is very high during holiday seasons indicating unstable demand during holidays

Multivariate Analysis

Insights from EDA: Multivariate Analysis

Numerical variables: Correlation heat map analysis & pairplot





- High correlation between following variables-
  - Dew point temperature with Temperature (+ve), very close to 1
  - Dew point temperature with Humidity (+ve)
  - Rented Bike count with Temperature (+ve) & Dew point temperature (+ve)
  - Temperature with solar radiation (+ve) & snowfall (-ve)
  - Humidity with Solar radiation, Visibility & Windspeed (-ve)

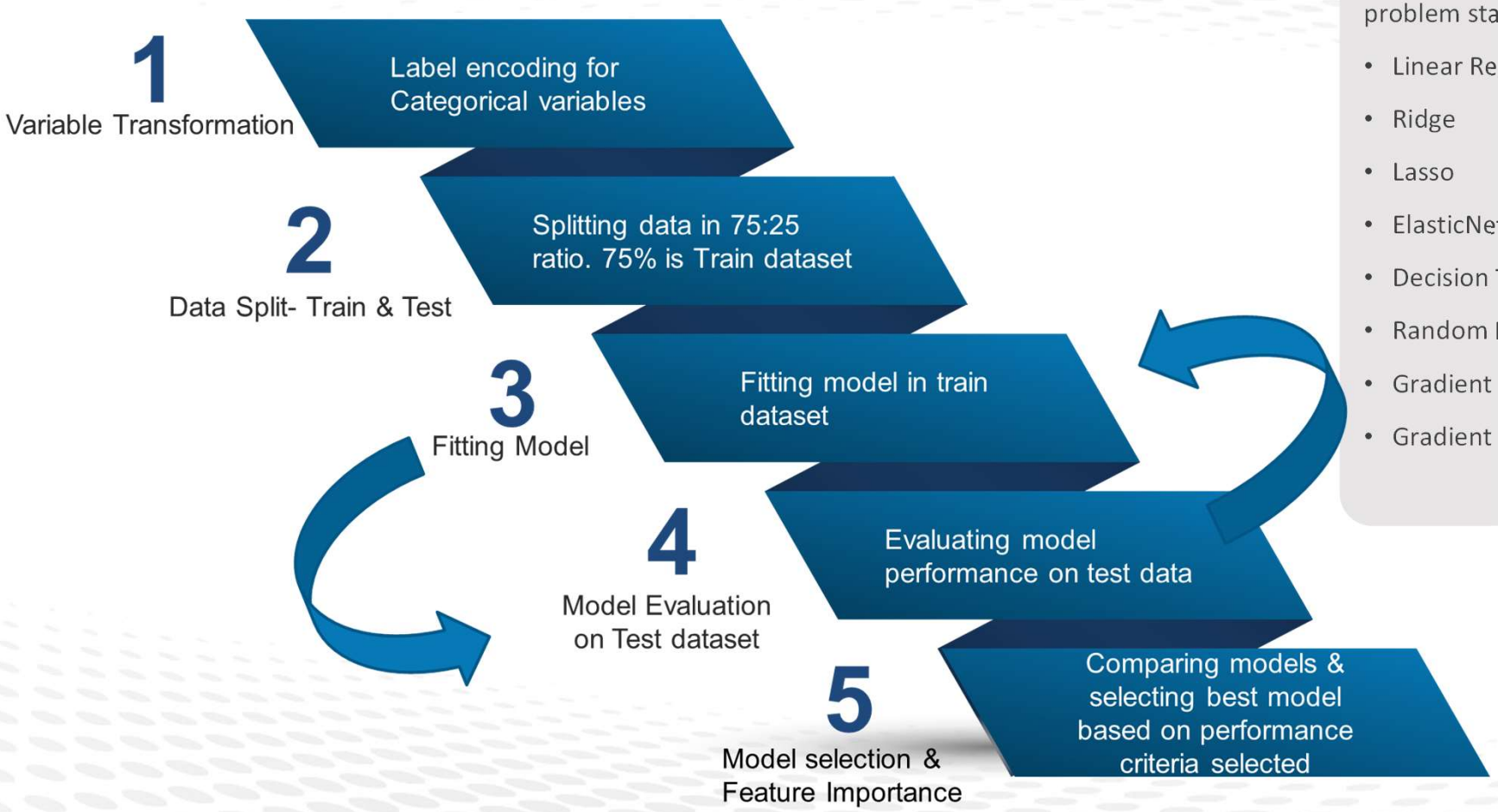
- Pairplot to visualize the correlation between variables

Variable Transformation

Variable Transformation

- From Date variable below variables are derived which could be driving factors for bike demand. We have observed the impact on bike demand due to these variables using Univariate & Bivariate analysis
    - Month
    - Weekend & Weekday
- Since some of the categorical variables are available as string. For e.g.- Hour, Month, Seasons, Holiday & Functioning Day. These needs to be transformed into numerical so that Machine learning algorithms can establish the numerical relationship with respect to the Target variable.
- For transformation, One Hot Encoding is used.
- Post variable transformation, data is split into train & test data set.
- Further, model is developed and tested on test dataset.
- Using feature importance, key factors impacting bike demand is analysed.

Predictive Modelling Approach



Predictive Models leveraged for the given problem statement-

- Linear Regression
- Ridge
- Lasso
- ElasticNet
- Decision Tree
- Random Forest
- Gradient Boosting
- Gradient Boosting Gridsearchcv

Model Results

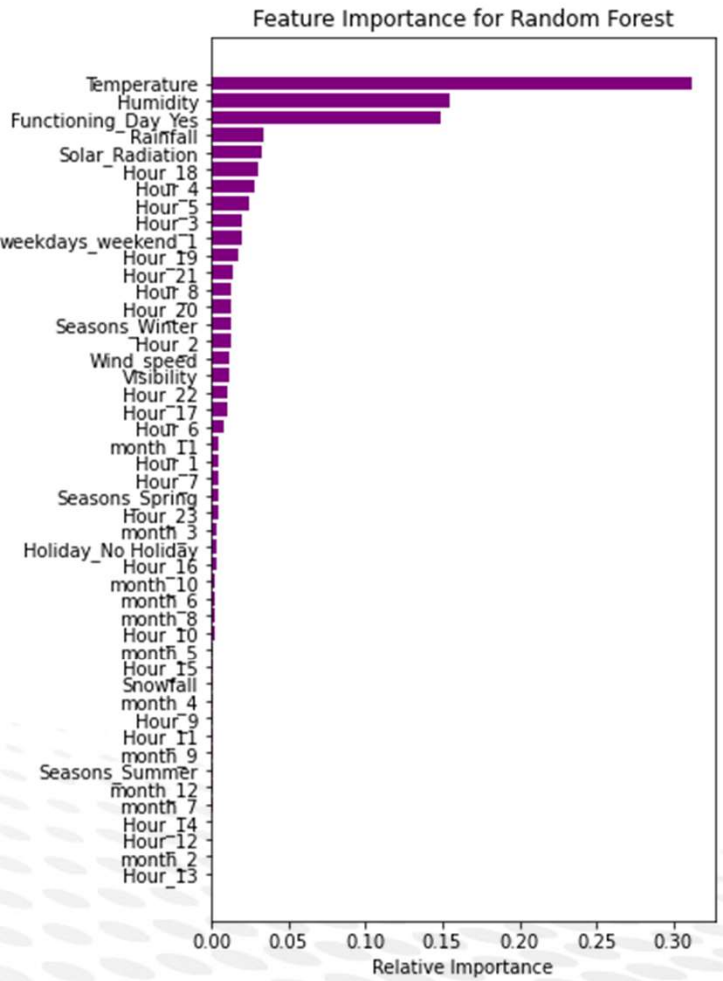
		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	4.474	35.078	5.923	0.772	0.77
	1	Lasso regression	7.255	91.594	9.570	0.405	0.39
	2	Ridge regression	4.474	35.078	5.923	0.772	0.77
	3	Elastic net regression	5.792	57.574	7.588	0.626	0.62
	4	Decision tree regression	6.229	76.262	8.733	0.505	0.49
	5	Random forest regression	0.807	1.628	1.276	0.989	0.99
	6	Gradient boosting regression	3.269	18.648	4.318	0.879	0.88
	7	Gradient Boosting gridsearchcv	1.849	7.455	2.730	0.952	0.95

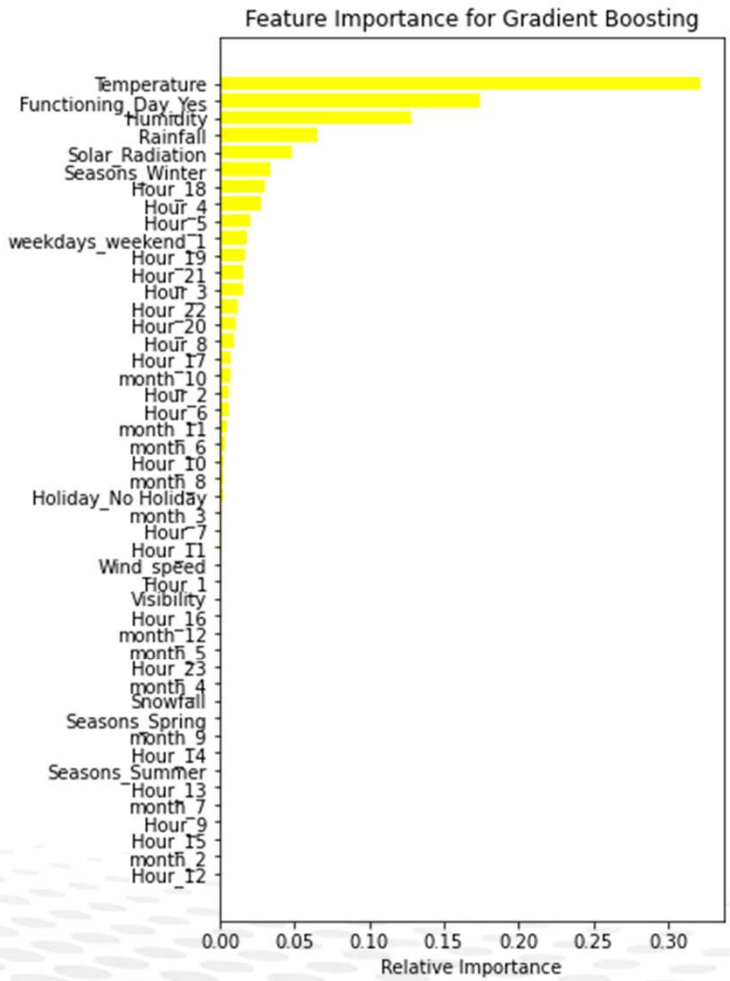
		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Test set	0	Linear regression	4.410	33.275	5.768	0.789	0.78
	1	Lasso regression	7.456	96.775	9.837	0.387	0.37
	2	Ridge regression	4.410	33.277	5.769	0.789	0.78
	3	Elastic net regression Test	5.874	59.451	7.710	0.624	0.62
	4	Decision tree regression	6.520	84.349	9.184	0.466	0.45
	5	Random forest regression	2.231	12.749	3.571	0.919	0.92
	6	Gradient boosting regression	3.493	21.289	4.614	0.865	0.86
	7	Gradient Boosting gridsearchcv	2.401	12.393	3.520	0.922	0.92

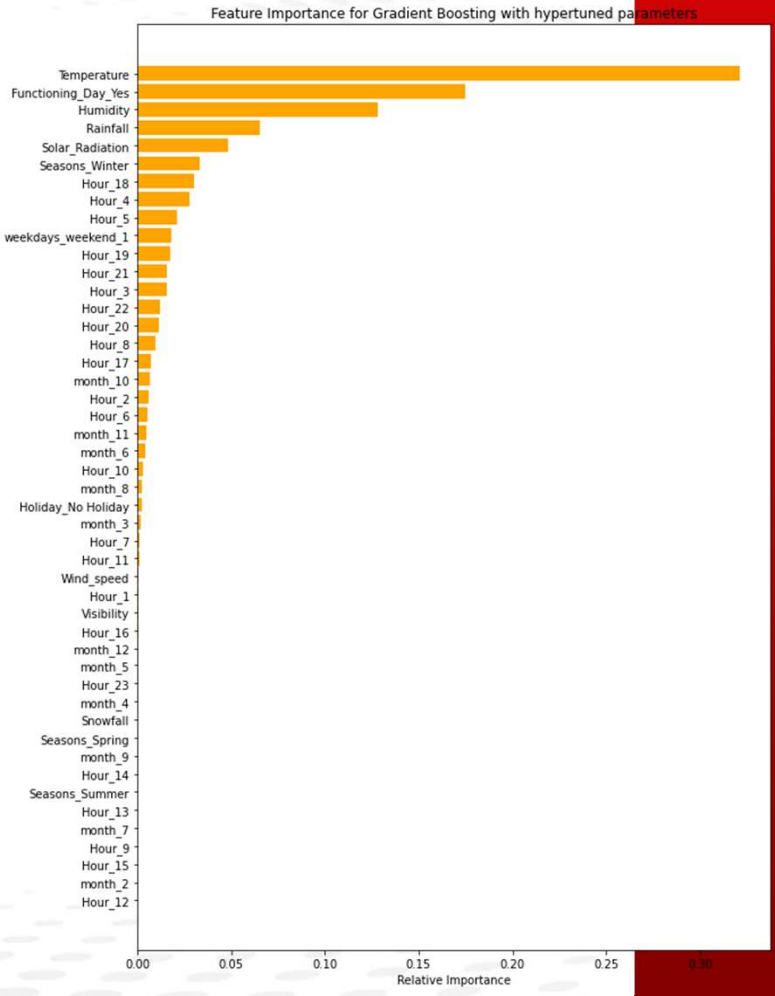
- Adjusted R2 is used as key performance metric along with MAE, MSE, RMSE and R2
- Random Forest & Gradient Boosting with hyper tuned parameters proves to be best model in terms of performance metrics.
- Random Forest is comparatively overfit as compared to Gradient Boosting with hyper tuned parameters model.

Feature Importance

Top 5 key features across all 3 models with best performance metrics is same- Temperature, Functioning day, Humidity, Rainfall & Solar Radiation







Recommendations

Recommendations

- **Gradient Boosting with hyper tuning or Random Forest model suits to be best model** for predicting the Bike demand based on given dataset. Recommended to review the model every 3-4 months for accuracy & with new data.

- **4 out of Top 6 features** in each of the best performing model **is weather related** (Temperature, Humidity, Rainfall & Solar radiation). Thus, **weather forecast plays very important role in predicting the bike demand**. Further to enhance the model, weather forecast data can also be considered as one of the input.

- Peak demand is observed at 18:00 hrs especially during functioning days and non holidays. This indicates mostly demand is for users returning from office/school etc. Hence, **availability of bikes can be planned near the office zones**. Such zones can be earmarked and more bikes should be made available during these peak hours in these areas.

- High demand is in summer while demand in winter is very less. **Annual maintenance of bikes can be planned during winters.**

- Low demand during 02:00 hrs to 05:00 hrs, **daily maintenance can be planned during these hours.**

- Rainfall, Snowfall & Humidity are only 3 variables which negatively impacts the bike demand.

- Further insights can be developed by **analysing location specific data** so that availability of bikes can be maximized & waiting times can be minimized.

- **Incentivize users through discount offers**, if they **drop the bike near peak zones before high demand hours** (18:00 hrs), thus enabling user driven bike availability.

- **Additional discounts can be offered to users, if they pick the bike before and after peak hours.** This will ensure spreading the demand to wider horizon (say instead of 18:00 hrs, demand can be spread in range of 16:00 hrs to 20:00 hrs).

Thank You

Image source: https://blog.google/products/maps/real-time-bikeshare-information-google-map/