Capstone Project-2

Seoul Bike Sharing Demand Prediction

By :- Viral Bhatu Shewale

Content :

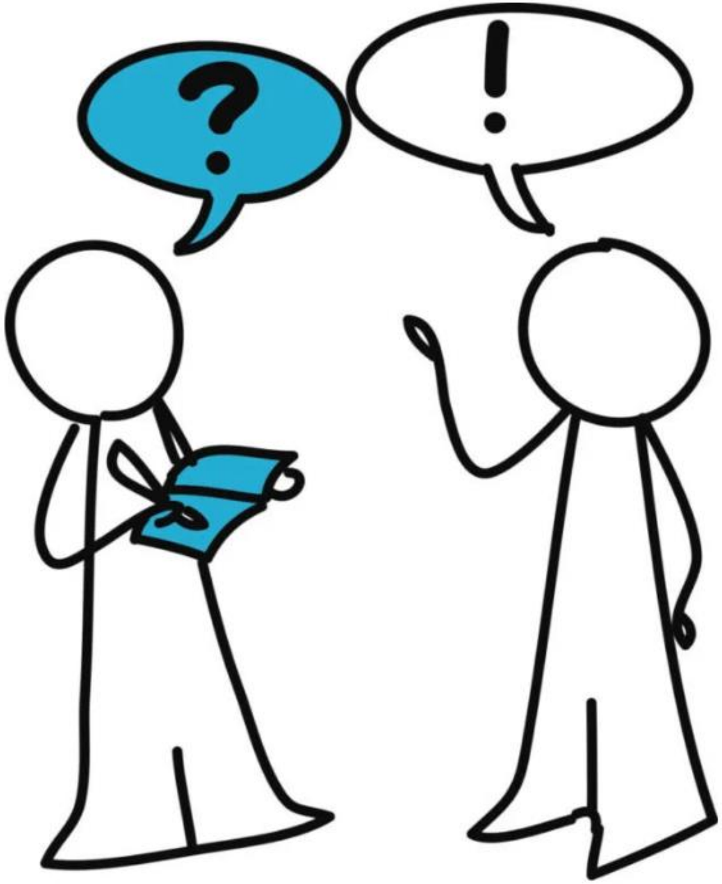
- Problem Statement
- Data Summary
- Feature Engineering
- Exploratory Data Analysis
- Modelling Approach
- Predictive Model
- Model Comparison
- Conclusion

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 and provides the city with a stable supply of rental bikes.

- The goal of the project is to build an ML model that is able to predict the demand for rental bikes in the city of Seoul.



Data Summary:

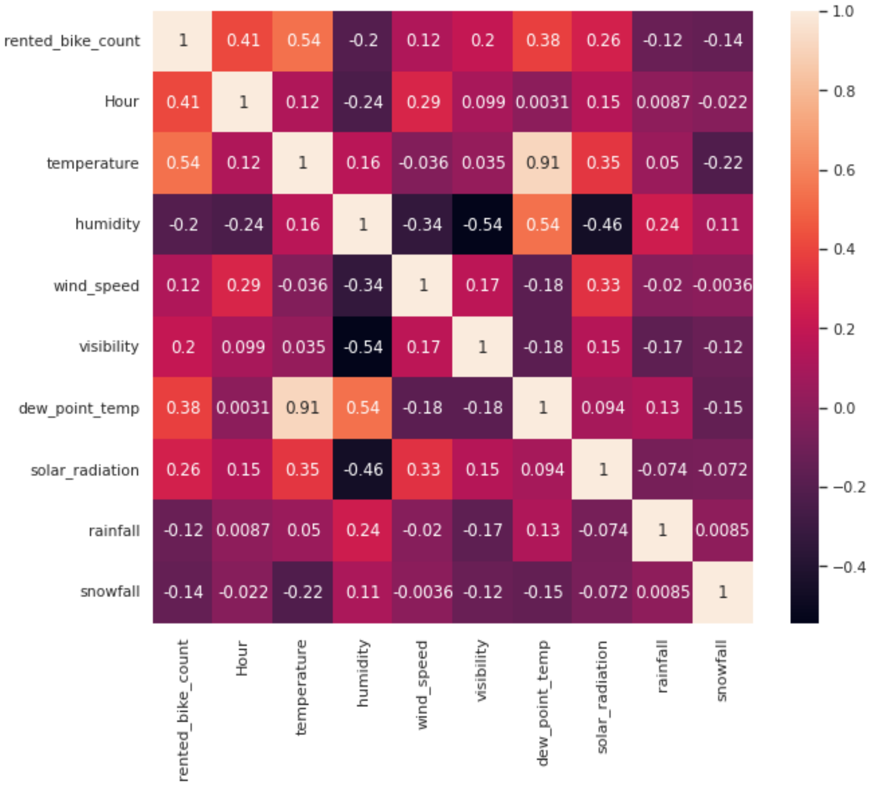
There are **8760** rows and **14** columns (Attribute) in the dataset.

- 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 – No Func(Non-Functional Hours), Fun(Functional hours)

Feature Engineering:

- The correlation matrix shows that **dew point temperature** and **temperature** are highly correlated(0.91). Hence we can drop the column from the dataset since it will not increase the accuracy of prediction and will only increase the model complexity.

- There are **no missing values** in the dataset.



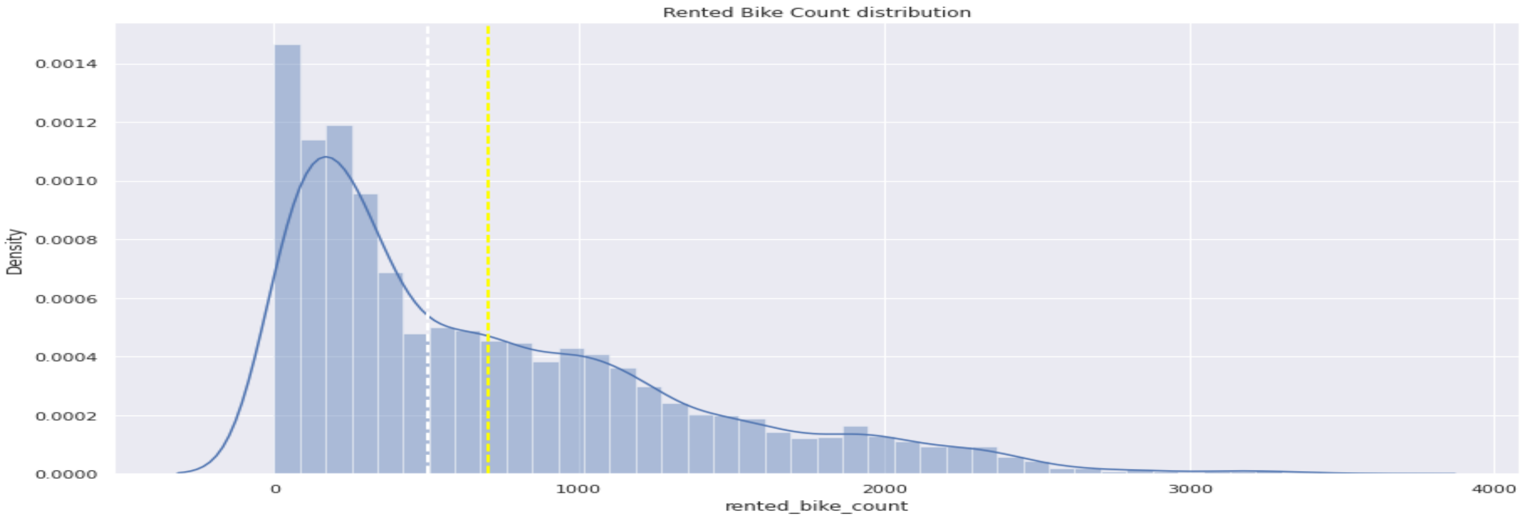
Exploratory Data Analysis:

Distribution of Dependent variable: -

- As we can see **rented bike count** (Dependent variable) is positively skewed.

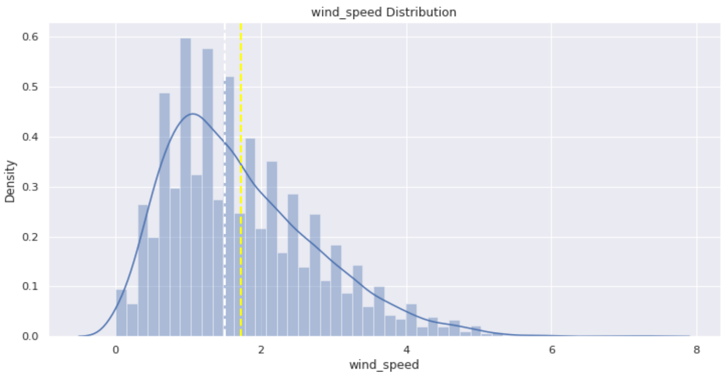
Rented bike count distribution: -

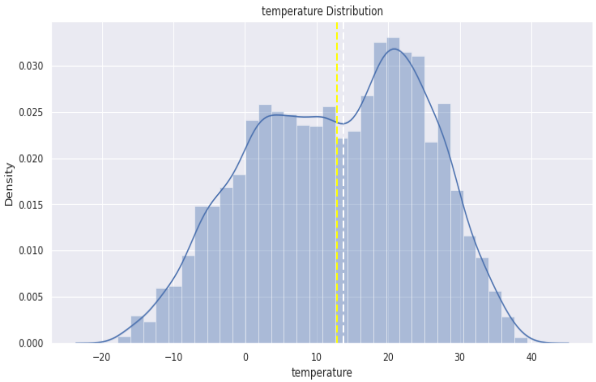
mean (yellow line) & **median** (white line)

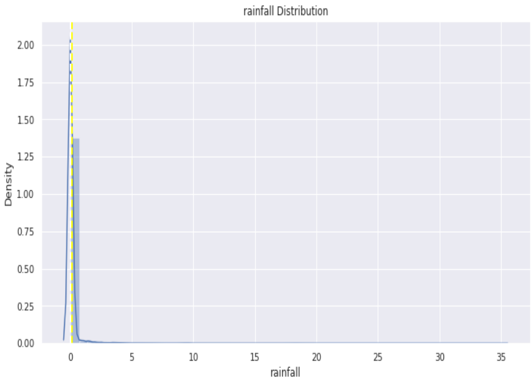


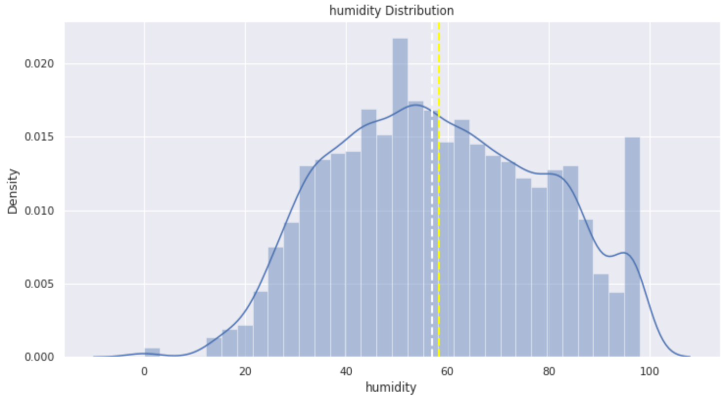
Distribution of Attributes: -

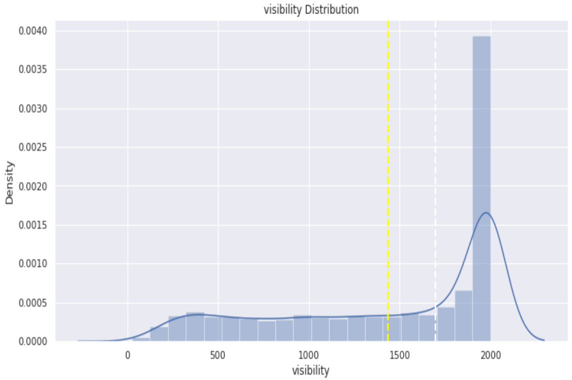
- **Normally distributed attributes**:- temperature and humidity
- **Positively skewed attributes**:- solar radiation, snowfall, rainfall and wind.
- **Negatively skewed attributes**:- visibility

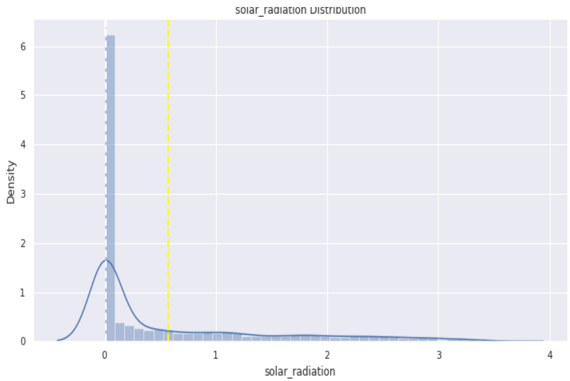






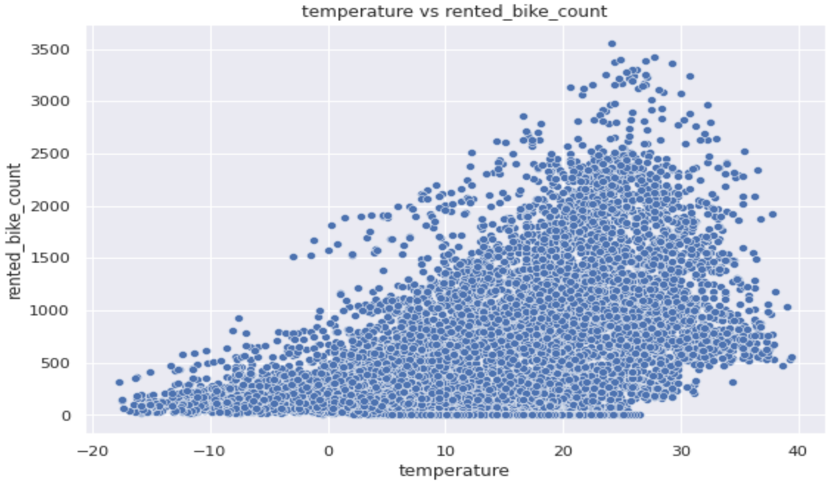


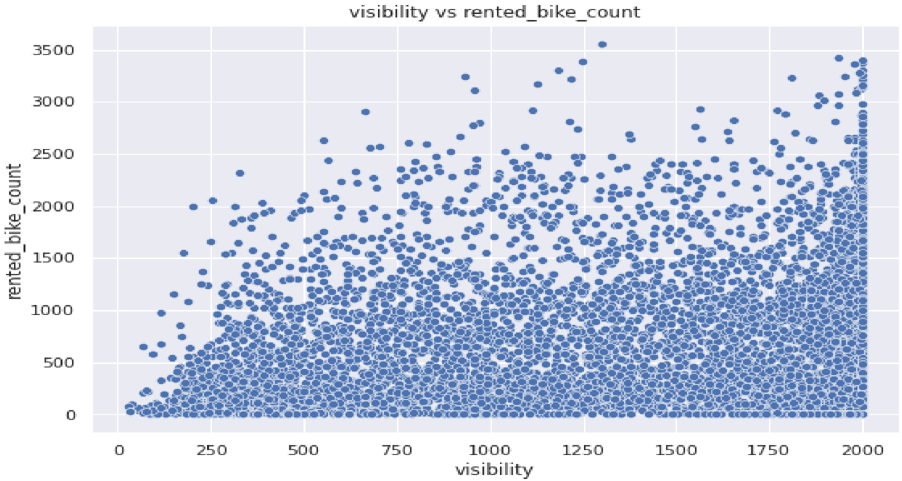


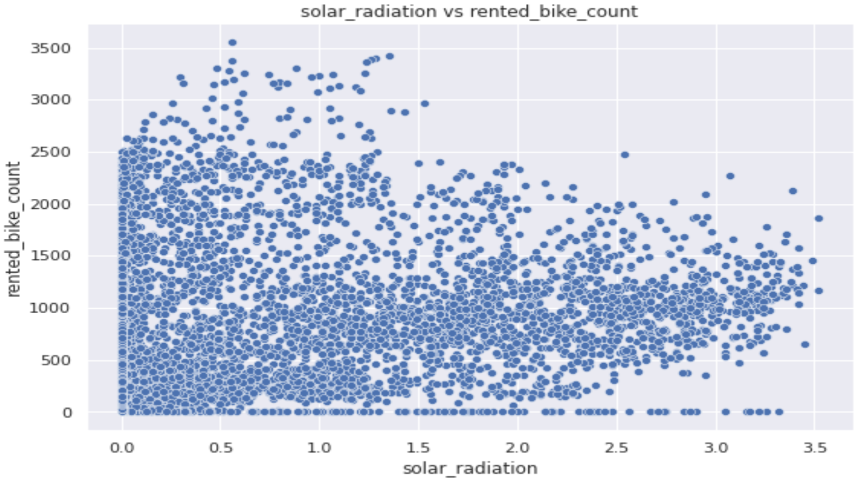


Relation between Continuous variable and Dependent variable (EDA):

- The temperature and visibility are **positively correlated** with the dependent variable (rented bike count).
- The demand for the rental bike is less for a day with low temperature and less visibility.
- Higher the solar-radiation lower the rental bike demand.

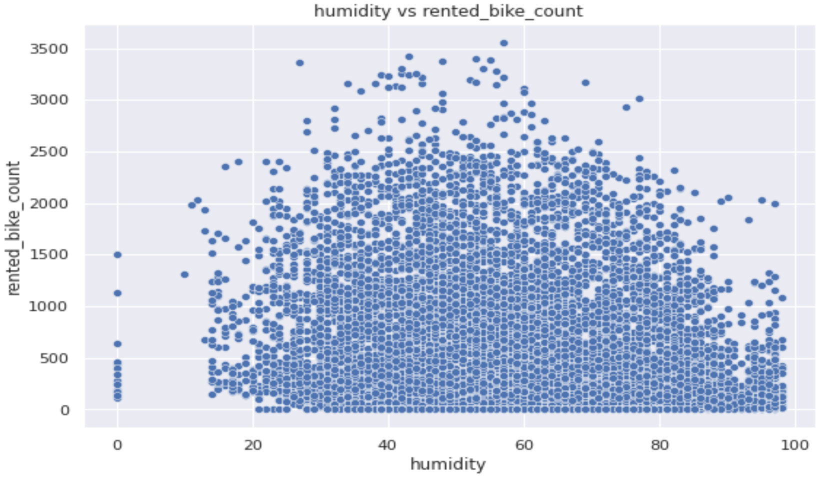


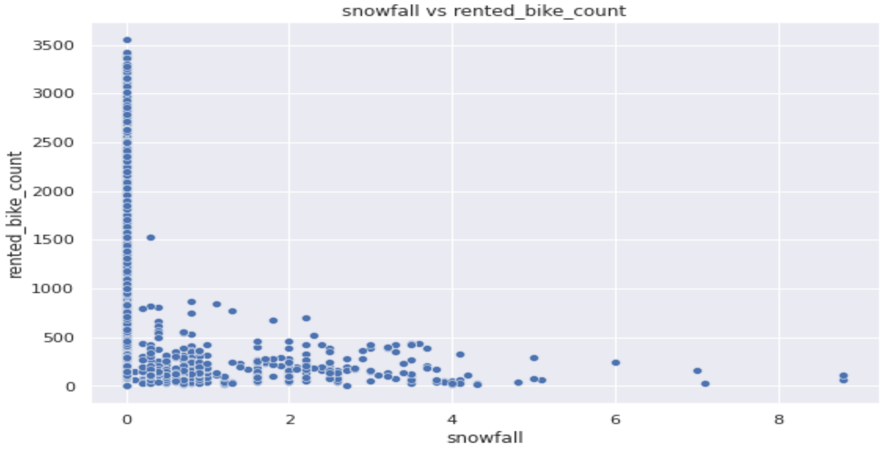


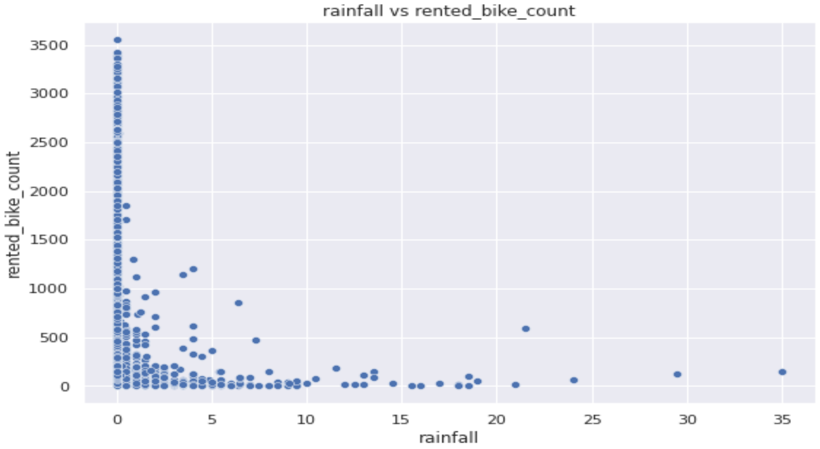


Relation between Continuous variable and Dependent variable (EDA):

- Snowfall, Rainfall and humidity **are negatively correlated** with Rented bike count.
- The demand for the rental bike is typically lower when there is rainfall and snowfall.
- Higher humidity lowers the rental bike demand.



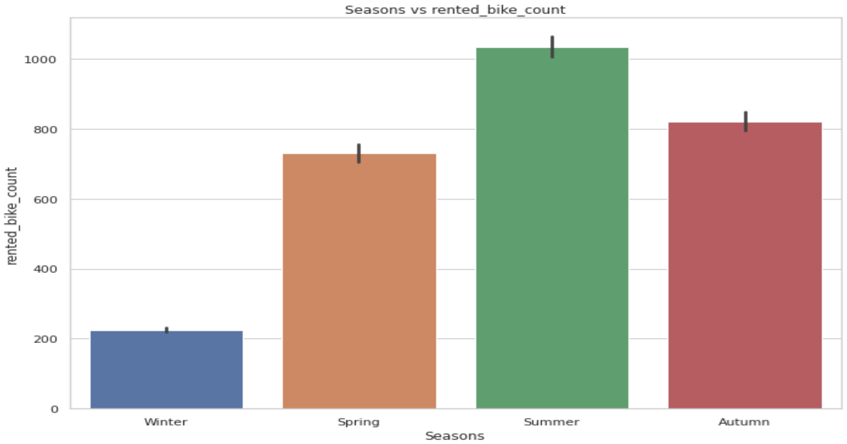


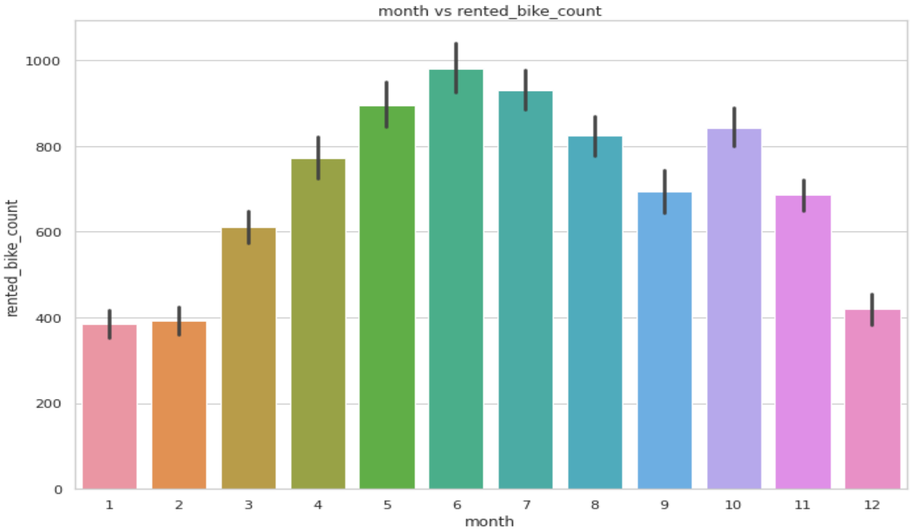


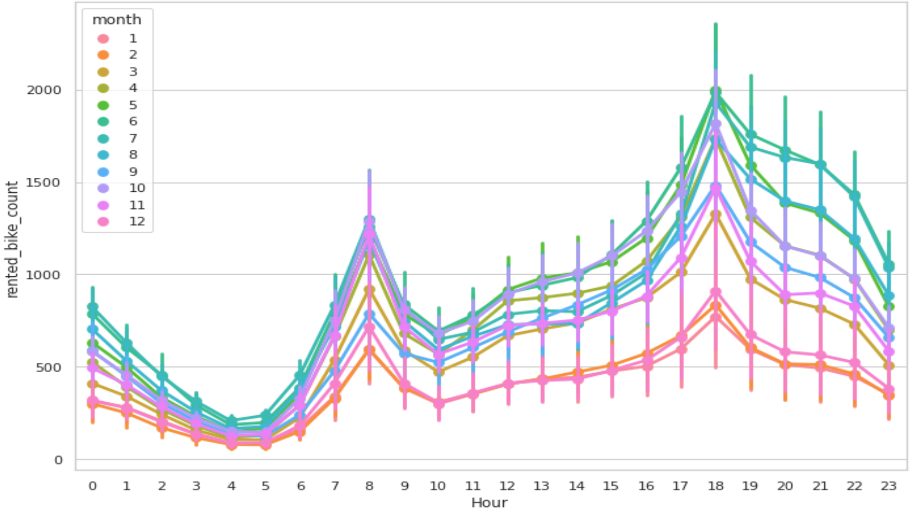
Relation between Categorical variable and Dependent variable (EDA):

- In the **Summer** season (May, June and July) demand for the rental bike is at its peak.

- In the **Winter** season (Dec, Jan, Feb) the rental bike demand is low.

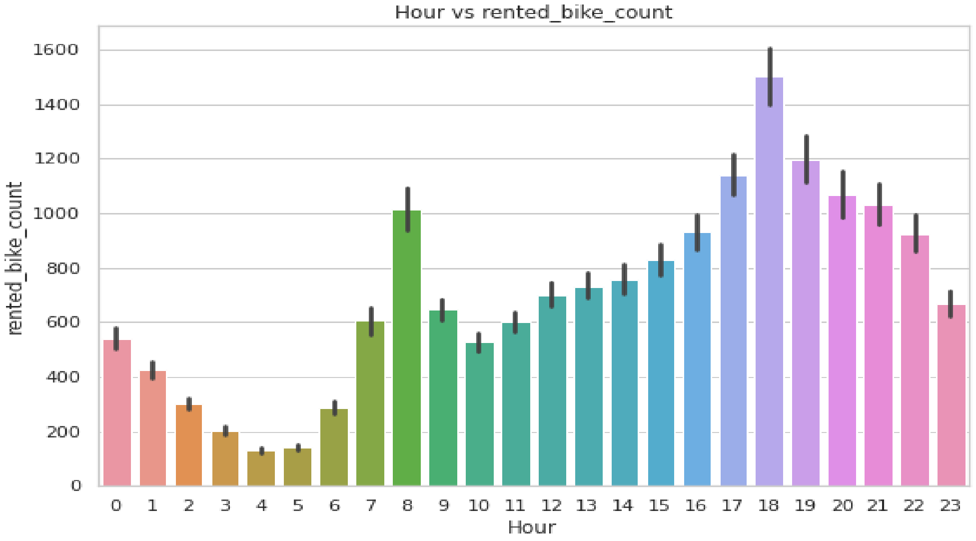


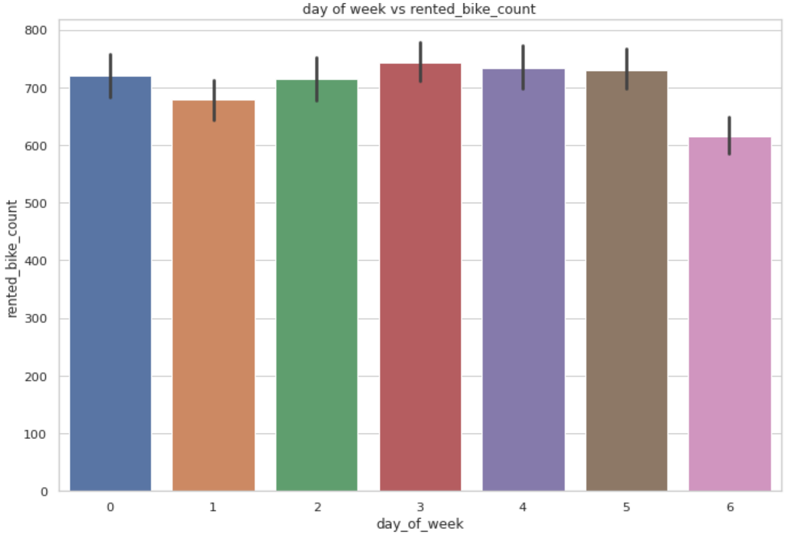


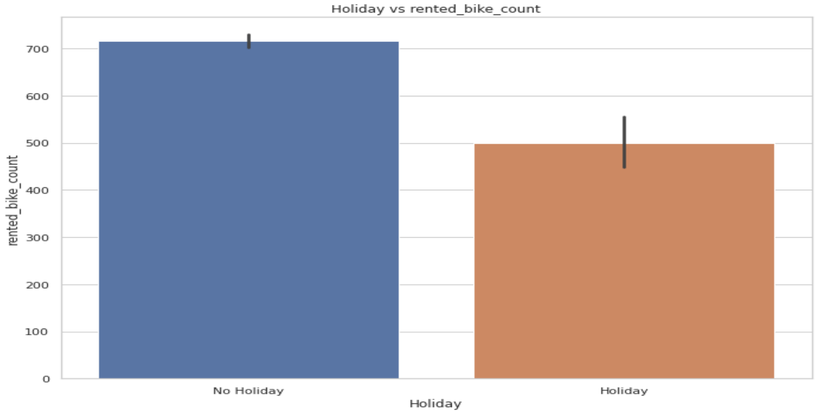


Relation between Categorical variable and Dependent variable (EDA):

- The rented bike count on average was constant from Monday to Saturday. Demand was lower on Sunday (weekend) and holidays.
- which means the majority of clients are **working professionals.**
- There is a surge in demand for rental bike count during rush hours (**4 pm** to **9 pm**).

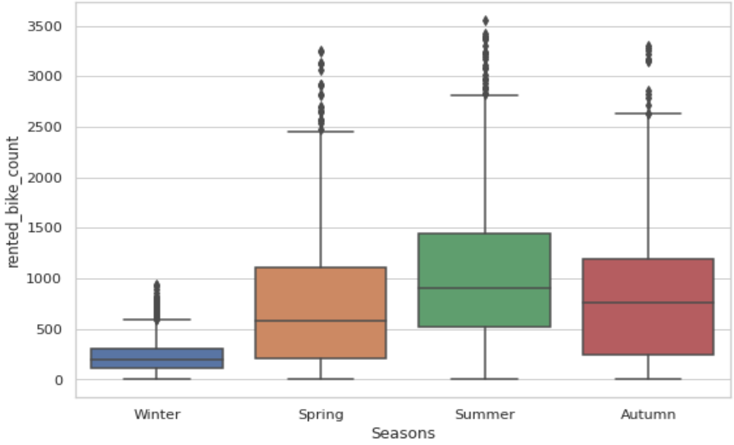


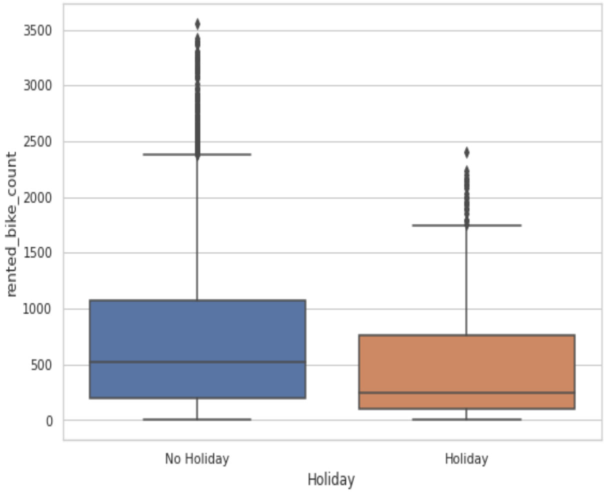


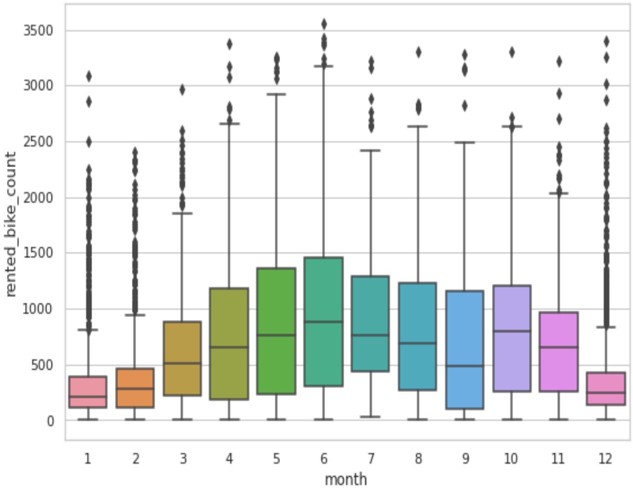


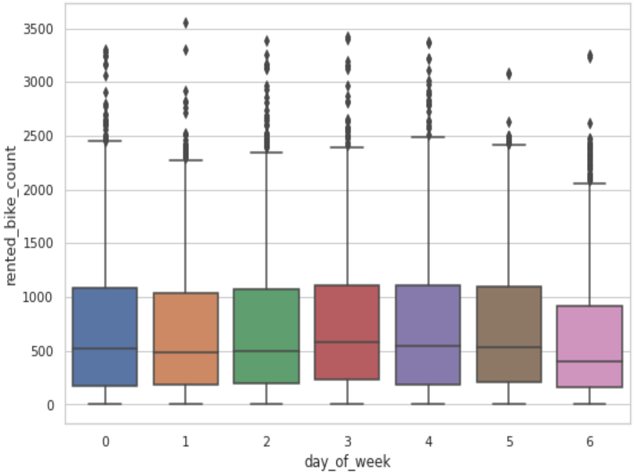
Checking for Outliers:

- As we can see, there are some outliers present in the dataset.
- We have to consider them at the time of model building. We didn't drop them because if we do so, we may loose out important trends/patterns in the data.









Modelling Approach:

- We are working on a dataset which contains outliers. Hence we have to choose a model which is less sensitive to outliers.

- A dataset with many categorical independent variables which are not linearly related to a dependent variable. Hence it is not advisable to use linear models to make predictions. We can use tree models instead.

- **List of Machine Learning algorithms which are less sensitive to outliers:**

    - Decision Tree

    - Random Forest

    - XG Boost

- Choose the model with the highest accuracy for deployment.

Modelling Approach:

- Choice of a split is taken as K-fold cross-validation, with k=5, because of the computational power available and to reduce overfitting.

- Evaluation metrics MAE is robust to outliers and chooses a model that can generalize the results for all points, including outliers.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| y_i - x_i \right|}{n}$$

where N is the number of data points, y(i) is the i-th measurement, and x(i) is its corresponding prediction.

- Hyperparameter tuning to prevent overfitting and the best parameters are chosen using GridsearchCV.

Decision Tree:

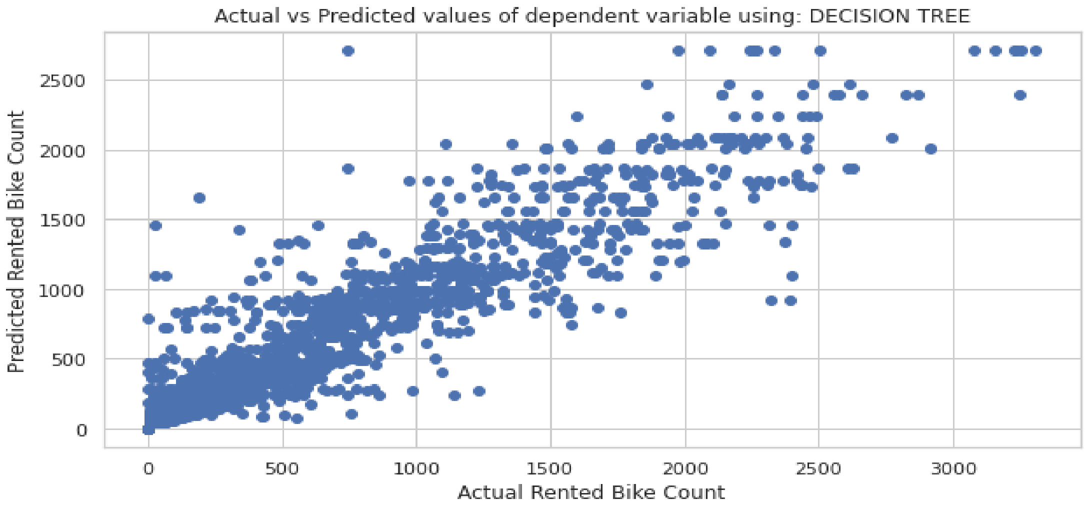
Parameters: -

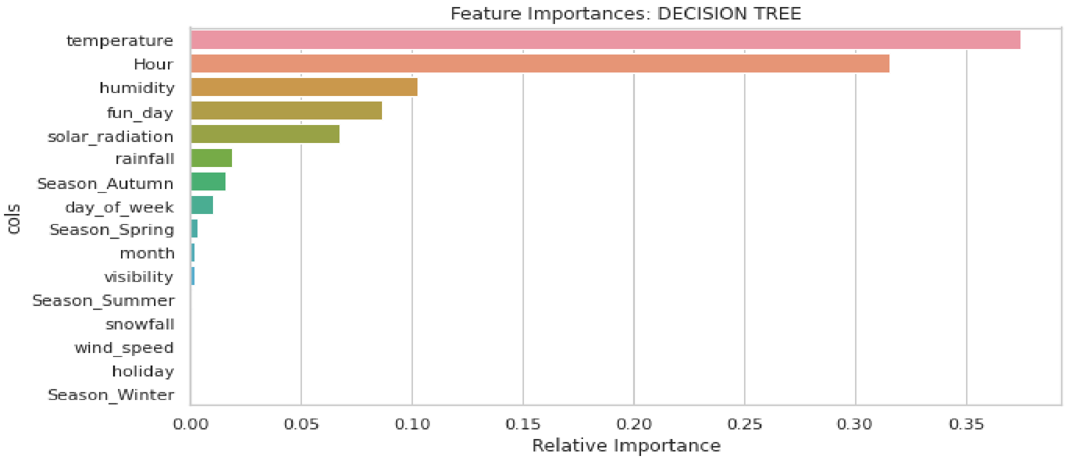
- ➢ max_depth = 18
- ➢ min_sample_leaf = 30

Evaluation metrics: -

- ➢ MAE = 165.71
- ➢ R2_test = 0.8386
- ➢ Adjusted R2 for test = 0.8371

Adjusted R Square is roughly the same as **R Square** meaning the model is quite robust.





Random Forest:

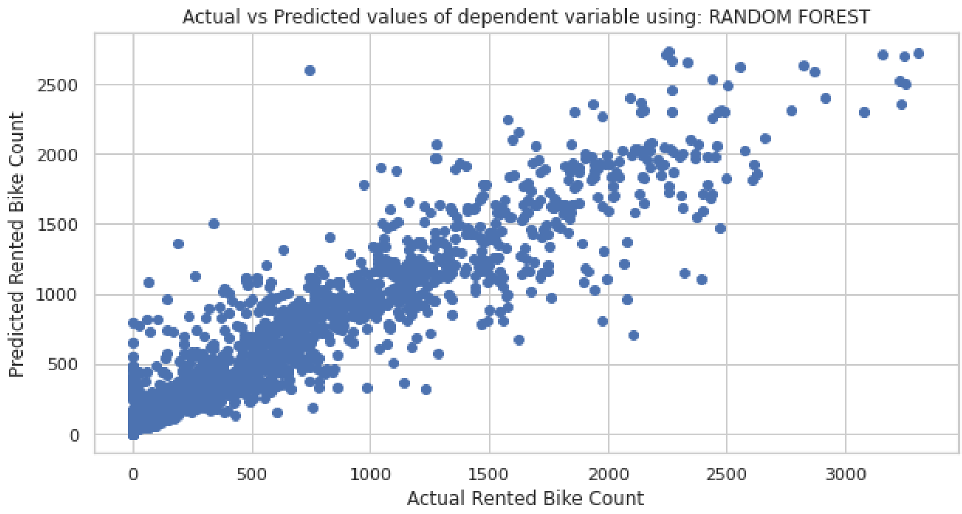
Parameters: -

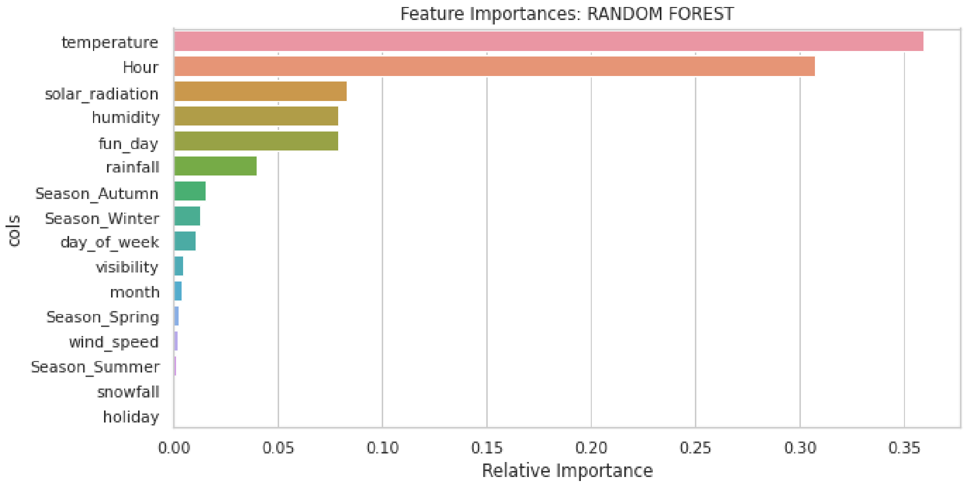
- n_estimators = 500
- min_sample_leaf = 20

Evaluation metrics: -

- MAE = 157.62
- R2_test = 0.856
- Adjusted R2 for test = 0.8550

Adjusted R Square is roughly the same as **R Square** meaning the <u>model is quite robust.</u>





XG Boost:

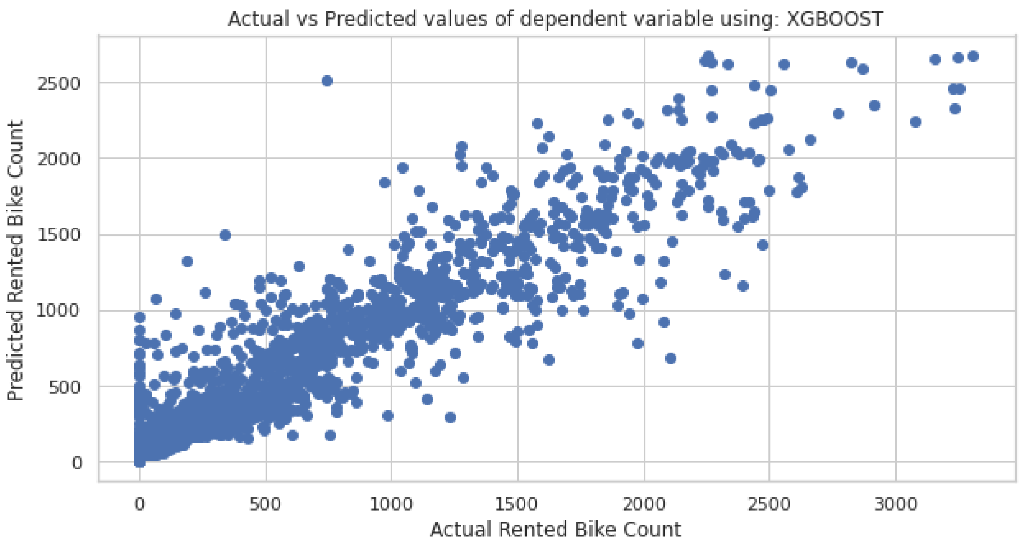
Parameters: -

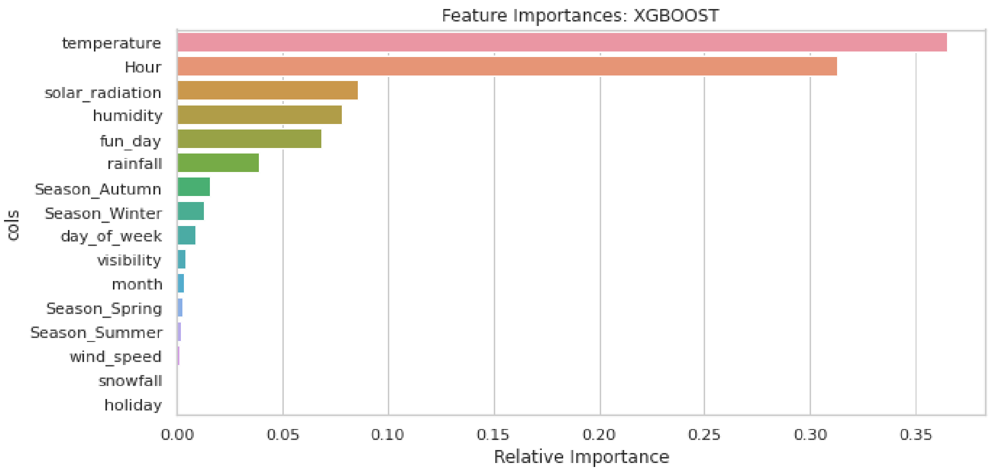
- n_estimators = 500
- min_sample_leaf = 25

Evaluation metrics: -

- MAE = 167.105
- R2_test = 0.8874
- Adjusted R2 for test = 0.8550

Adjusted R Square is roughly the same as **R Square** meaning the <u>model is quite robust.</u>





Model Comparison:

- The **XG Boost** model has the **lowest MAE** compared to others.

Sl. No.	Regression Model	test_MAE	train_MAE	Train R2 Score (%)	Test R2 Score (%)
1	Decision Tree	165.71218660977652	154.09613067576998	86.09009686242055	83.86768563110701
2	Random Forests	157.62253922081365	138.47699424532254	88.84561534385928	85.63540395467903
3	XG Boost	144.9873499474637	124.15145121783988	91.49361689298823	88.48309856060598

- **Lower the MAE better in model performance.**

Conclusion:

- The demand for rental bikes was **highest in the summer** season and **lowest in the winter** season.
- **May-July are peak months** to rent a bike. Dec-Feb is the least preferred month for bike renting.
- The rental bike demand was more on a weekday than on weekends. The majority of **clients belong to the working class.**
- The **temperature of 20-30 Degrees**, evening time 4 pm- 8 pm and the **humidity between 40%-60%** are the most favourable parameters where the Bike demand is at its peak.
- **Temperature, humidity, hour of day, solar radiation and functional day** are major driving factors for the bike rent demand.
- The **XG Boost** model has the **lowest** test **MAE**. A low MAE value indicates that the simulated and observed data are close to each other and show better accuracy. Thus **lower MAE is better for model performance**. (XG Boost model with an **accuracy of 88.48%**)