Machine Learning

Assignment – 1 – PS9

Seoul Bike Sharing Demand

Assignment Contributor –

1. Ashique Zzaman (2021SC04612)
2. TIMLO LOUIS G (2021SC04286)
3. SUBHASHINI J A J (2021sc04554)
4. Shushank Yadav (2021SC04520)

Synopsis –

Introduction:

Bike sharing system is an innovative transportation strategy that provides individuals with bikes for their common use on a short-term basis for a price or for free. Over the last few decades, there has been a significant increase in the popularity of bike-sharing systems all over the world. This is because it is an environmentally sustainable, convenient and economical way of improving urban mobility. In addition to this, this system also helps to promote healthier habits among its users and reduce fuel consumption.

Paper Contribution:

We referred to a Research paper from ResearchGate Portal. From the search paper, received the idea about the below things –

1. Data Cleaning: Learned the process of Data cleaning by identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.
2. Exploratory data analysis (EDA): Learned how EDA helps in analyzing data sets to summarize their main characteristics, using statistical graphics and other data visualization methods.
3. Feature Engineering: Learned how Feature engineering is the act of converting raw observation into desired features using statistical or machine learning approaches.
4. Feature Coding: Learned how to encode categorical data in both encoder and check accuracy of encoders - One Hot Encoder Data and Label Encoder Data
5. Outlier Detection: Learned about Outliers in data and its influence and detection.
6. ML Algorithms: Regression – Learned about various ML Regression Algorithms.
7. Evaluation Metrics: Learned about various Evaluation Metrics.

Data Pre-processing:

Data Set Information:

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.

The observations in the dataset were recorded during a span of 365 days, from December 2017 to November 2018.

This dataset contains information about the total count of rented bikes at each hour as well as the data of observation and weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

The Seoul Bike Dataset contains the following information:

1. Date - The date of each observation in the format 'year-month-day'
2. Hour - Hour of the day
3. Temperature - Temperature recorded in the city in Celsius (°C).
4. Humidity - Relative humidity in %
5. Wind speed - Speed of the wind in m/s
6. Visibility - measure of distance at which object or light can be clearly discerned in units of 10m
7. Dew point temperature - Temperature recorded in the beginning of the day in Celsius(°C).
8. Solar radiation - Intensity of sunlight in MJ/m^2
9. Rainfall - Amount of rainfall received in mm
10. Snowfall - Amount of snowfall received in cm
11. Seasons - Season of the year (Winter, Spring, Summer, Autumn)
12. Holiday - Whether the day is a Holiday or not (Holiday/No holiday)
13. Functional Day - Whether the rental service is available (Yes-Functional hours) or not (No-Nonfunctional hours)

Machine Learning Activity:

Below are the Feature engineering activities used –

1. Scaling Numerical data - Min Max Scaler: Transform features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g., between zero and one.
2. Encoding Categorical data - One hot encoding: Transform a categorical feature having string labels into K numerical features in such a manner that the value of one out of K (one-of-K) features is 1 and the value of rest (K-1) features is 0. Each category value is converted into a new column and assigned a 1 or 0 (notation for true/false) value to the column.
3. Finding best value of n_neighbors parameter: Provides best value of n_neighbors parameter.
4. yeo-johnson transformation: Power Transformation which inflates low variance data and deflates high variance data to create a more uniform dataset.
5. Square root Algorithm for removing Skewness: Square root (e.g., sqrt(x)) is a transformation that has a moderate effect on distribution shape. It is generally used to reduce skewed data.
6. Model explainability with LIME, SHAP and ELI5: Model explainability refers to the concept of being able to understand the machine learning model.

Below are the ML Algorithms used –

Classification Algorithms:

1. Logistic Regression Classification
2. KNeighbors Classifier
3. Support Vector Machine
4. Naive Bayes Classification
5. Decision Tree Classification

Ensemble ML for Classification Algorithms:

1. Random Forest Classifier
2. Gradient Boosting Classifier
3. Ada Boost Classifier

Regression Algorithms:

1. Linear Models
    a. Linear Regression
    b. Regularization Techniques
        i. Lasso
        ii. Ridge
        iii. Polynomial
    c. Stochastic Gradient Descent Regressor
2. Tree based model - Decision Tree

Ensemble ML for Regression Algorithms:

1. Random Forest Regression
2. Gradient Boosting Regressor
3. Extreme Gradient Boosting Regressor
4. Bagging Regressor
5. Stacking Regressor

Result analysis with metrics used from paper:

Classification and Ensemble Classification

Evaluation Metrics Used:

1. Confusion Matrix: A table showing correct predictions and types of incorrect predictions.
2. Precision: The number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier's exactness. Low precision indicates a high number of false positives.
3. Recall: The number of true positives divided by the number of positive values in the test data. The recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier's completeness. Low recall indicates a high number of false negatives.
4. F1-score: The weighted average of precision and recall.
5. Support: It is the total entries of each class in the actual dataset.

Result:

When compared the test accuracy, f1-score, precision and recall of all the models, KNeighbors Classifier and Random Forest Classifier gives the best Score with proper fit in Classification and Ensembled model respectively. Test accuracy and f1-score is 0.90 for KNeighbors Classifier model and Test accuracy and f1-score is 0.85 for Random Forest Classifier. So, these two models are the best for predicting the bike rental count on daily basis.

Regression and Ensemble Regression

Evaluation Metrics Used:

1. Mean Absolute Error (MAE): MAE is a very simple metric which calculates the absolute difference between actual and predicted values.
2. Mean Square Error (MSE): MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.
3. Root Mean Square Error (RMSE): As RMSE is clear by the name itself, that it is a simple square root of mean squared error.
4. Training Score: How the model generalized or fitted in the training data. If the model fits so well in a data with lots of variance then this causes over-fitting. This causes poor result on Test Score. Because the model curved a lot to fit the training data and generalized very poorly. So, generalization is the goal.
5. R2: R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.
6. Adjusted R2: The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases.

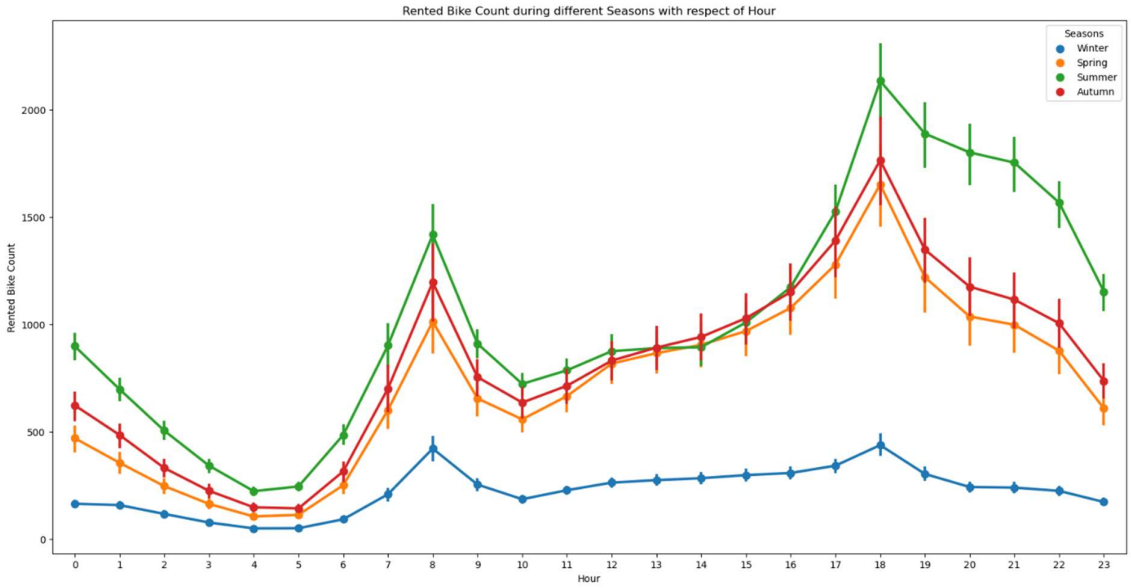
Result:

When compared the RMSE and Adjusted R2 of all the models, Stacking Regressor gives the highest Score where Adjusted R2 score is 0.90 and Training score is 0.95 so this model is the best for predicting the bike rental count on daily basis.

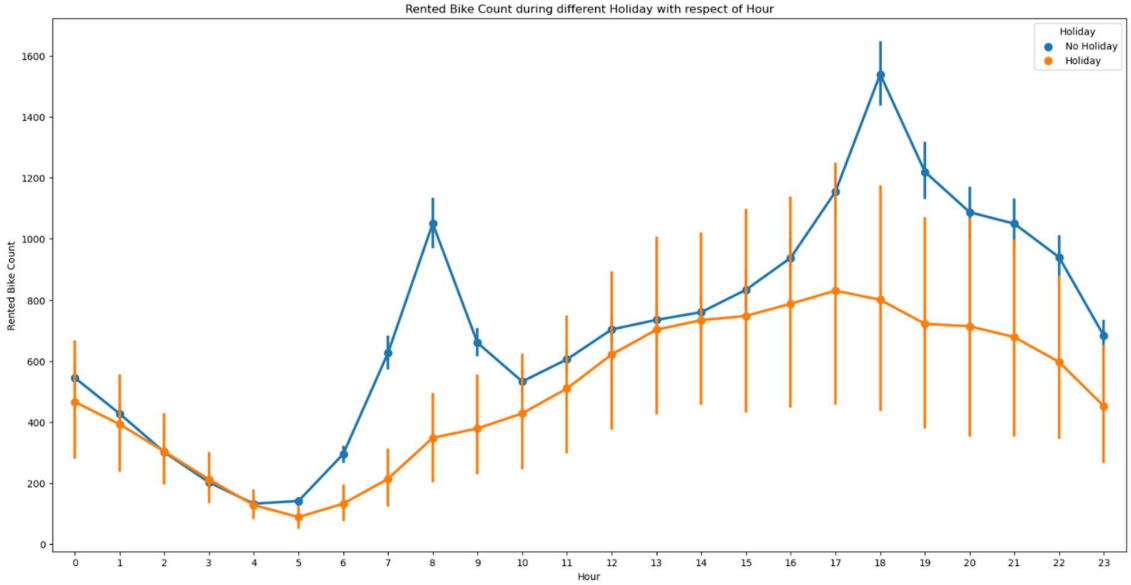
Exploratory Data Analysis / Visualization:

Below are few observations from EDA –

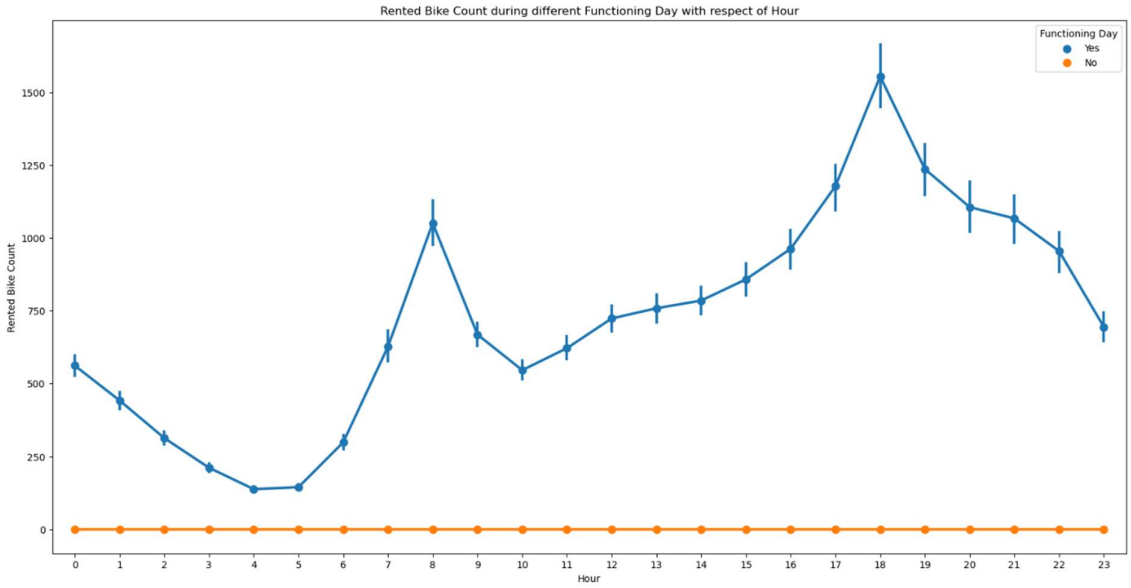
1. **Seasons -** The demand is low in the winter season.



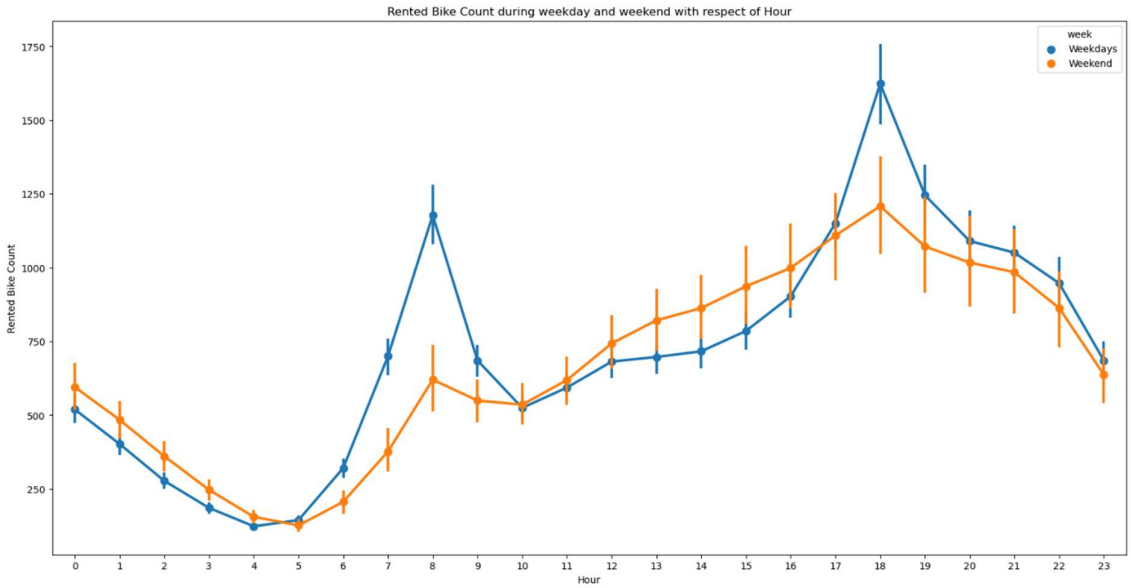
2. **Holiday -** The demand is low during holidays, but in non-holidays the demand is high, it may be due to reason that people use bikes to go to their workplace.



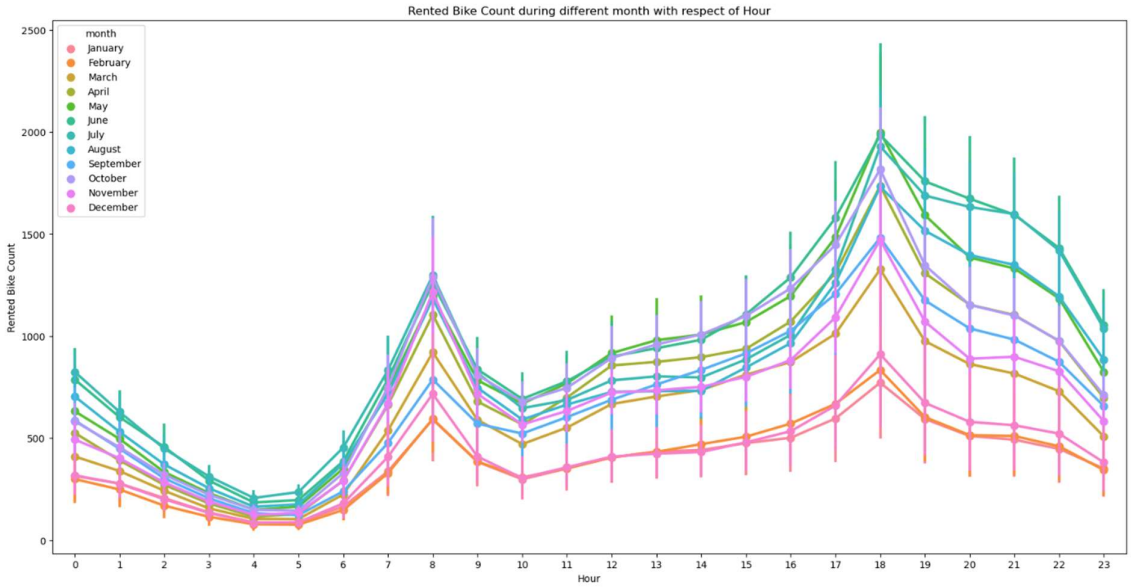
3. **Functioning Day -** There is no demand in non-functioning Day



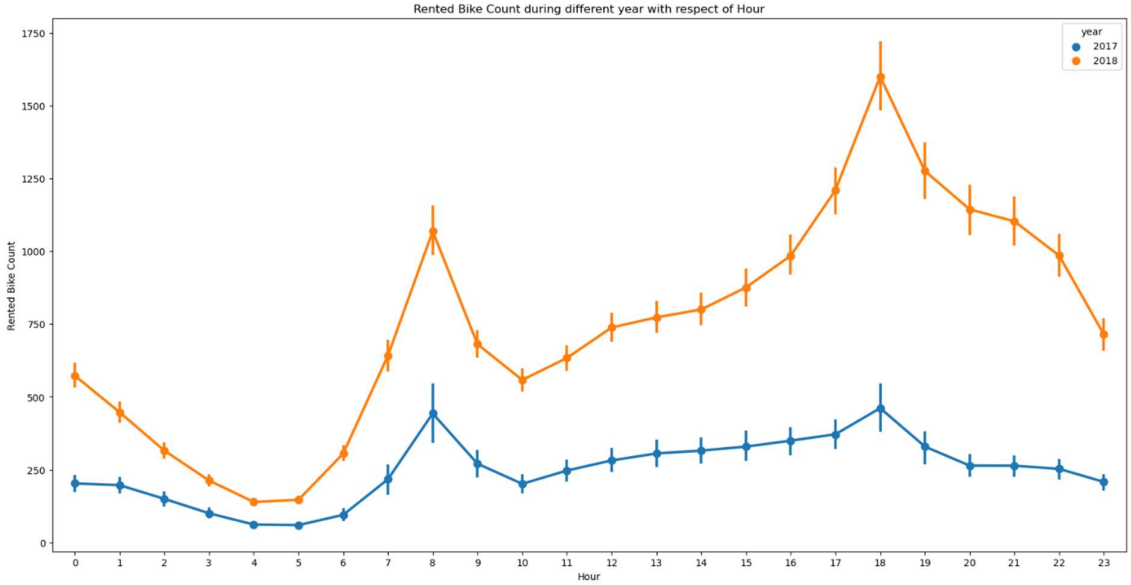
4. **Day of week -** There is different pattern for weekdays and weekends, in the weekend the demand becomes high in the afternoon. While the demand for office timings is high during weekdays, hence, we can change this column to weekdays and weekends.



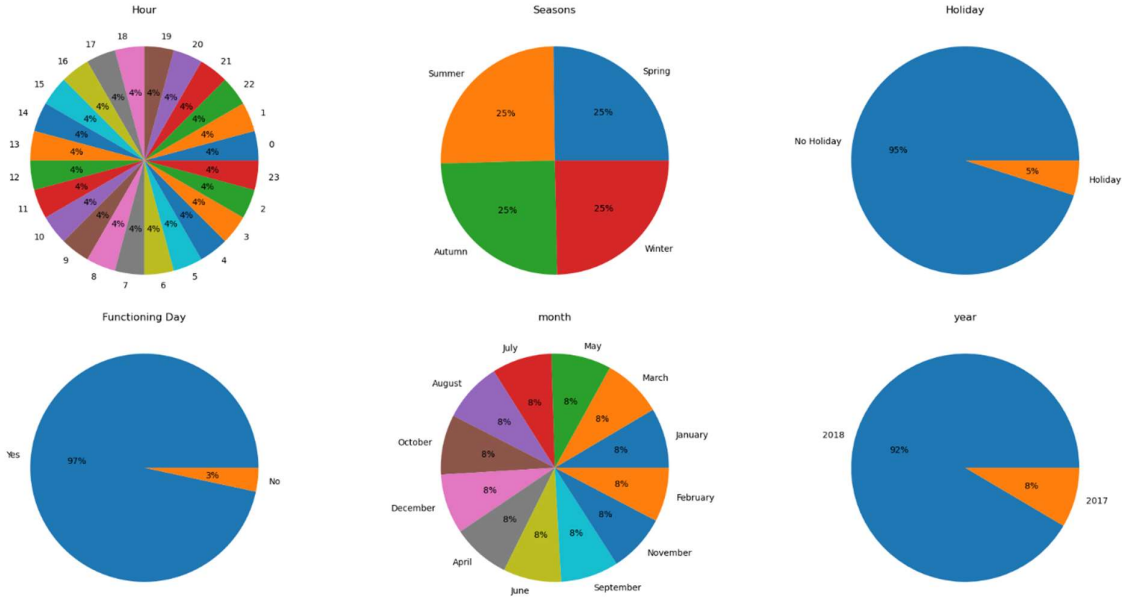
5. **Month -** The demand is low in December January & February; it is cold in these months and we have already seen in season column that demand is less in winters.



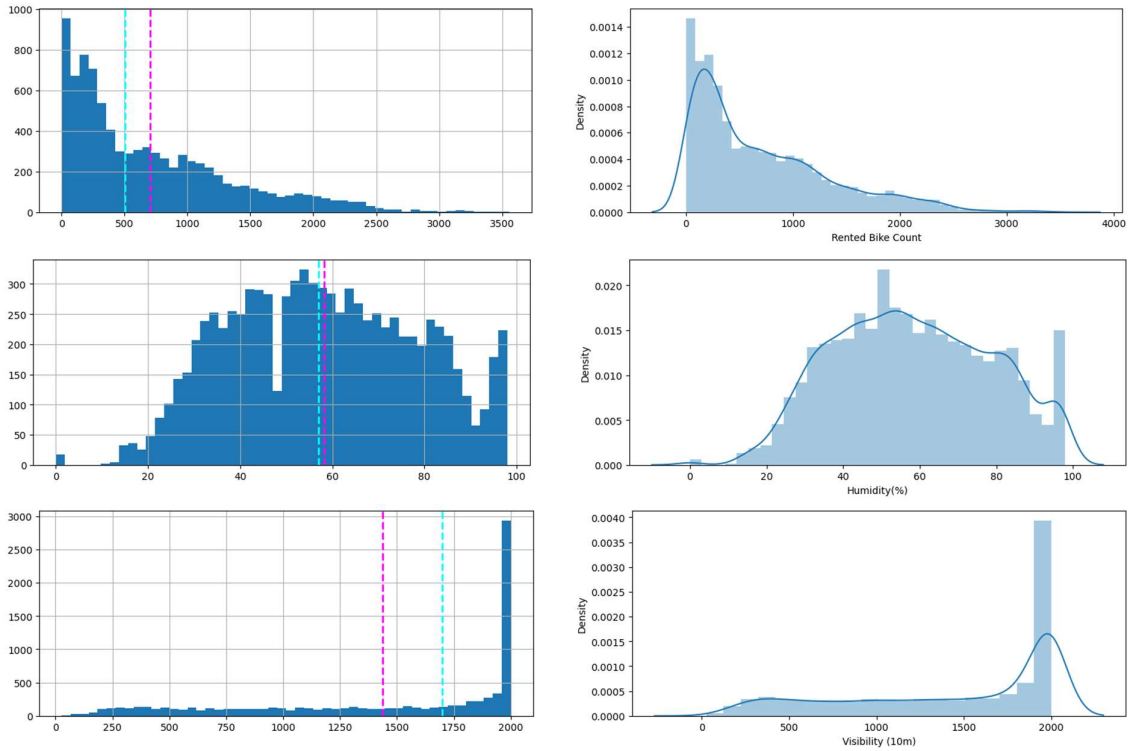
6. **Year -** The demand was less in 2017 and higher in 2018, it may be due to reason that it was new in 2017 and people did not know much about it.



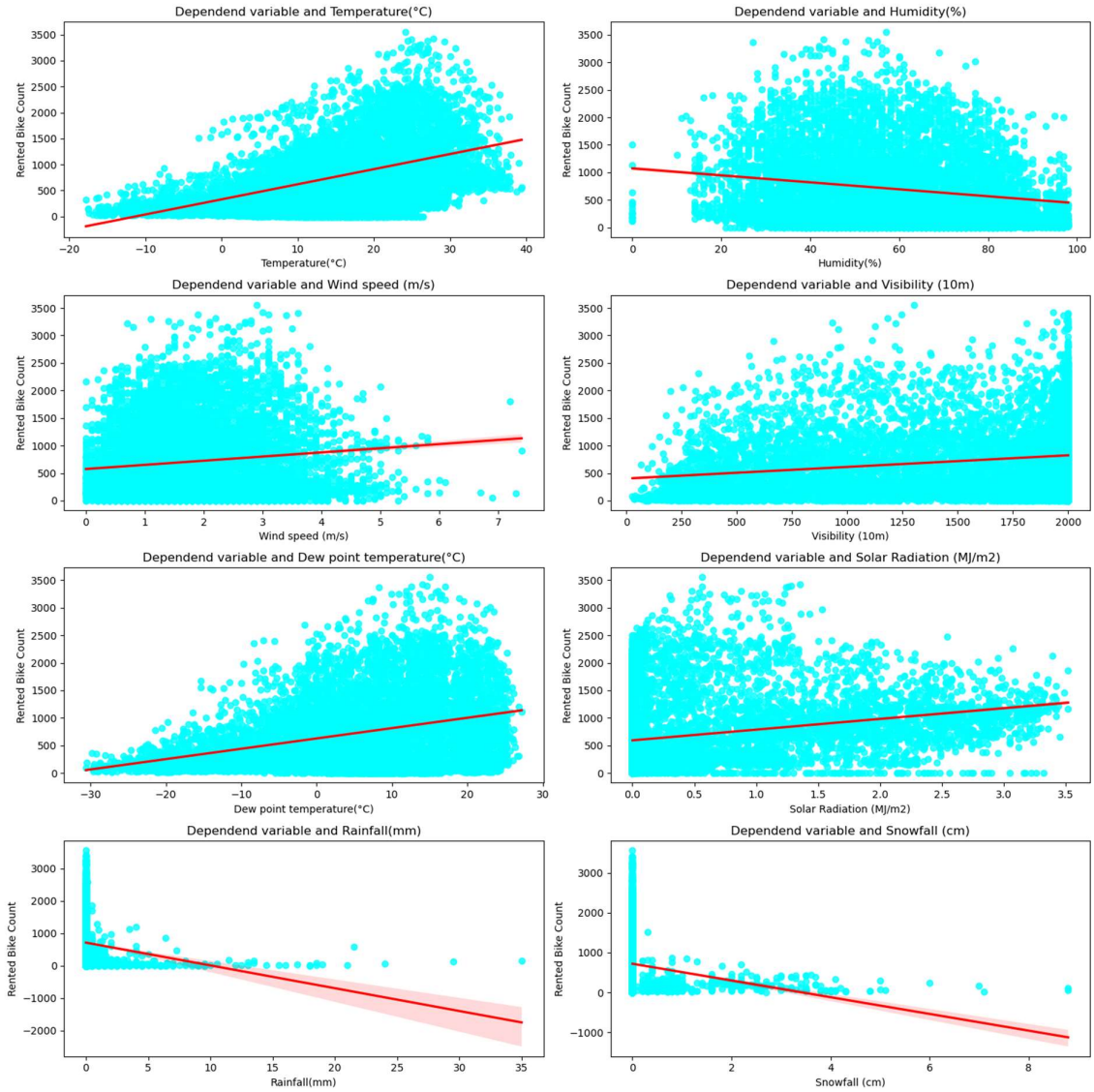
7. **Categorical features -** Creating Pie plot of all categorical features of the data set



8. **Numerical Features -** Plotting Histogram with mean and median, and Dist-plot of all the numeric features of the dataset to observe skewness of our numerical features



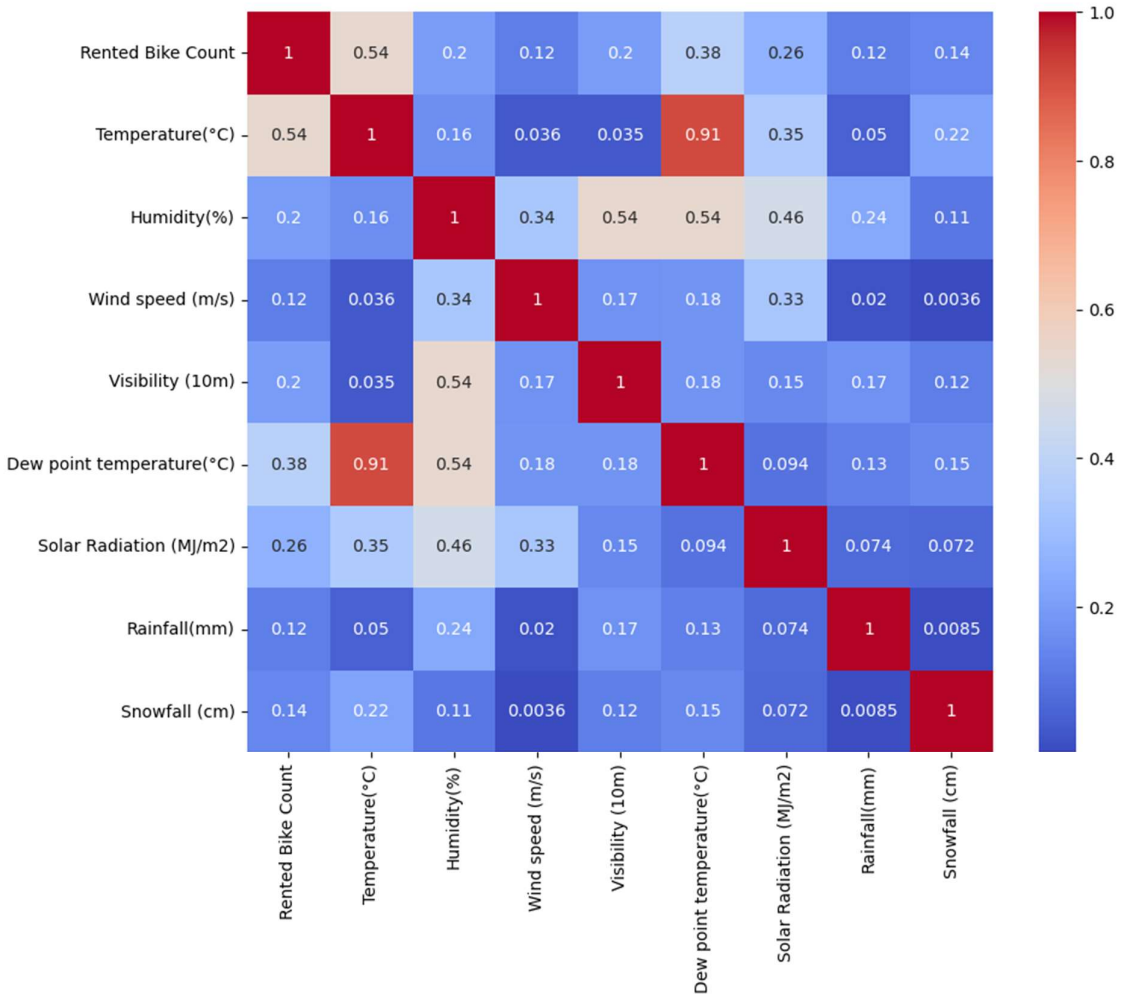
9. Regression plot to know relation of numerical features with dependent variable



10. Correlation –

I. Temperature and Hour have a strong correlation with the count of rented bikes.

II. Dew point temperature is highly positively correlated to the Temperature.



11. **Pandas Profiling** – It performs an automated Exploratory Data Analysis. It automatically generates a dataset profile report that gives valuable insights.

```
In [199]: profile.to_notebook_iframe()
```

Overview

Overview | Alerts 14 | Reproduction

Dataset statistics

Number of variables	14
Number of observations	8760
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	958.2 KiB
Average record size in memory	112.0 B

Variable types

Categorical	3
Numeric	10
Boolean	1