Capstone Project

Bike Sharing Demand Prediction

Nitin Gour
nitingour032@gmail.com

Problem statement

- The contents of the data came from a city called Seoul. A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free.The data had variables such as date, hour, temperature, humidity, wind-speed, visibility, dew point temperature, solar radiation, rainfall, snowfall, seasons, holiday, functioning day and rented bike count.The problem statement was to build a machine learning model that could predict the rented bikes count required for an hour, given other variables

Points to discuss

- Introduction
- Data description and summary
- Data Analysis Steps
- Scaling(types of scacling)
- Scaling Data and Model Building
- Handling outliers
- Regression plot
- Machine learning algorithms
- Conclusion

Introduction

A bike rental or bike hire business rents out motorcycles for short periods of time, Usually for a few hours. Most rentals are provided by bike shops
as a sideline to their main businesses of sales and service, but some shops specialize in rentals.

As with car rental, bicycle rental shops primarily serve people who do not have access to vehicles, typically travelers and particularly tourists.

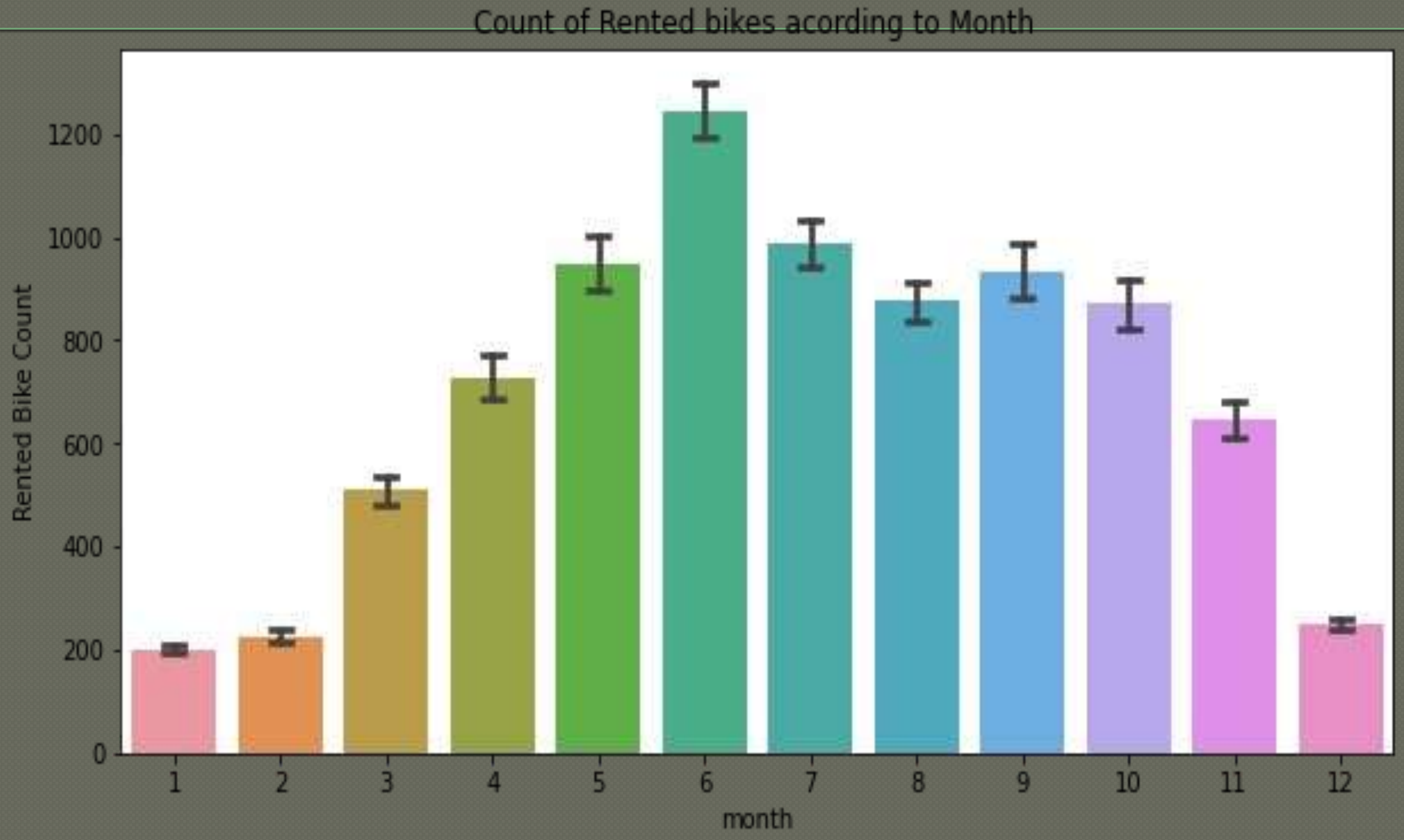
Bike rental shops rent by the day or week as well as by the hour, and these provide an excellent opportunity for those who would like to avoid shipping their own bikes but would like to do a multi-day bike tour of a particular area.

Data description and summary

- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

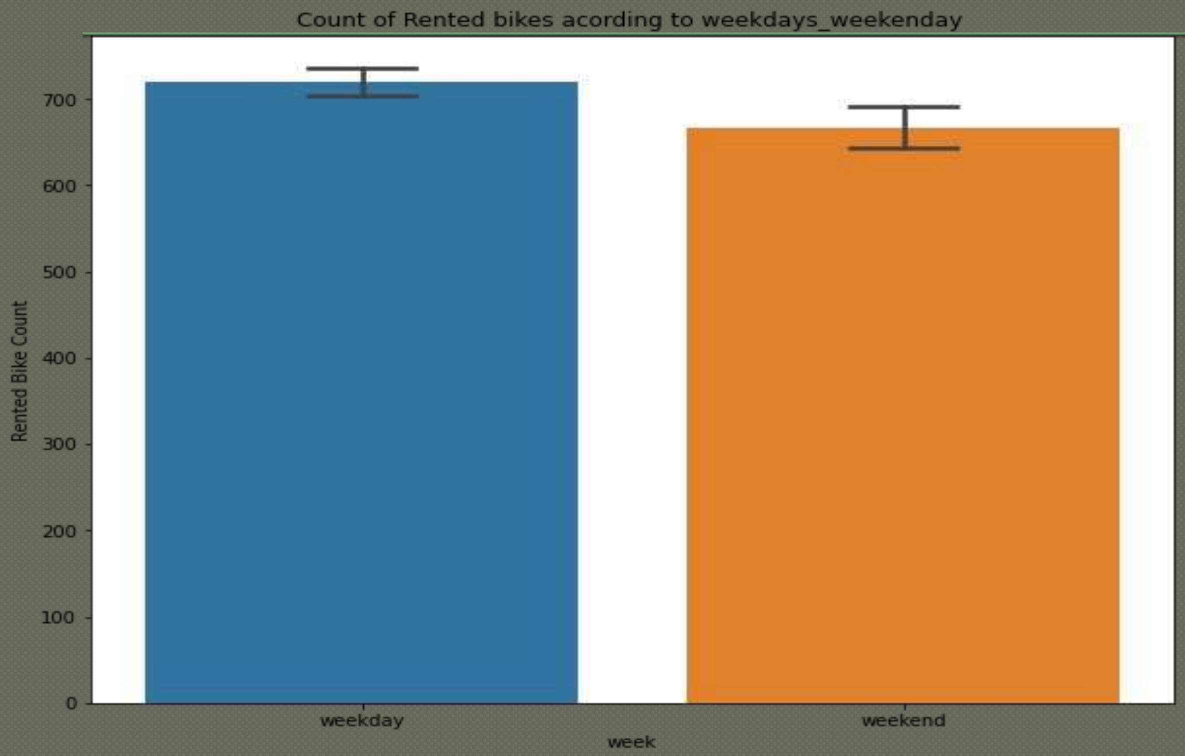
1. This dataset contains 8760 lines and 14 columns
2. Numerical variables - temperature, humidity,wind,visibility,dew point temp, solar radiation,rainfall,snowfall
3. Categorical variables -seasons,holiday and functioning day
4. Rented bike column -which we need to predict for new observations

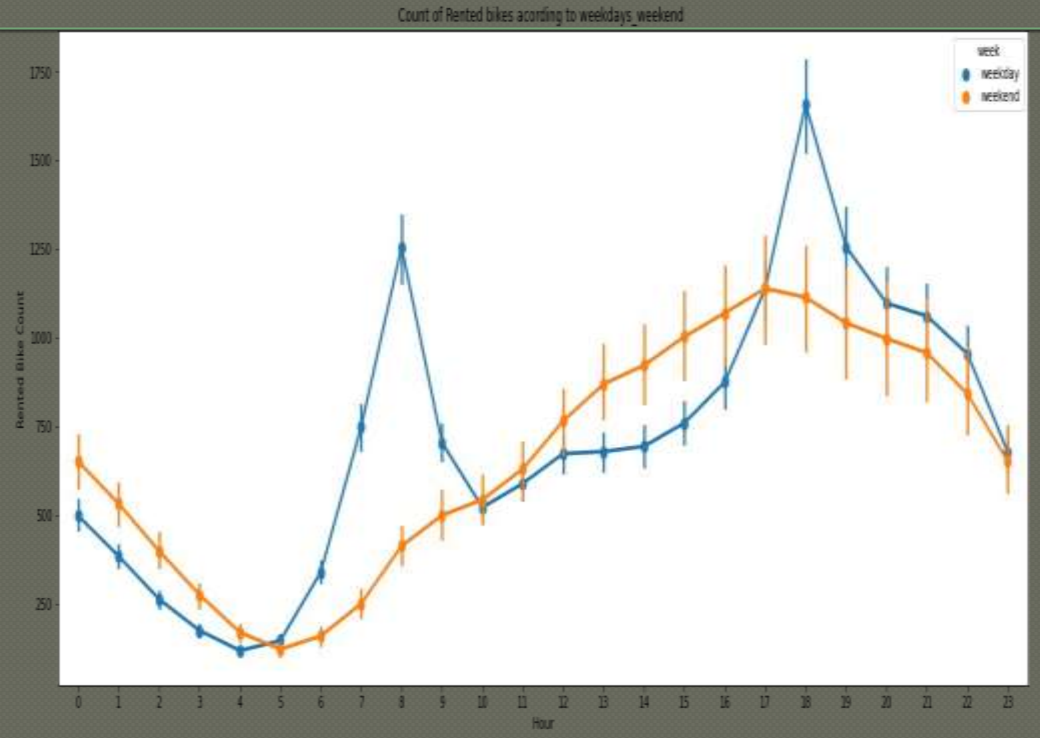
Month



The demand of the rented bike is high from the month 5 to 10

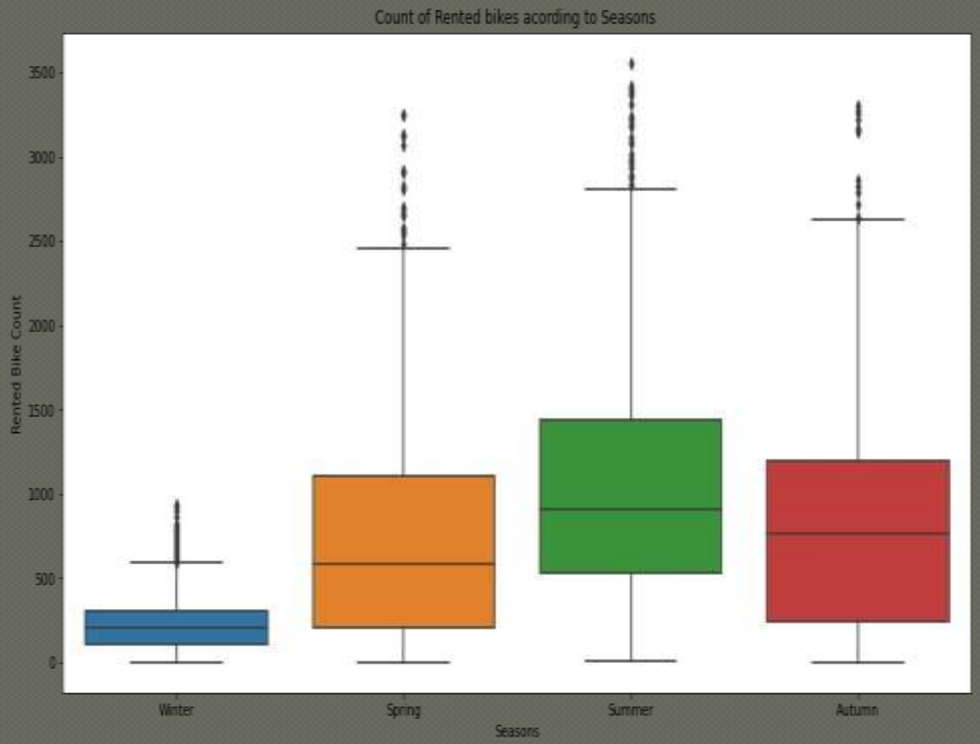
week

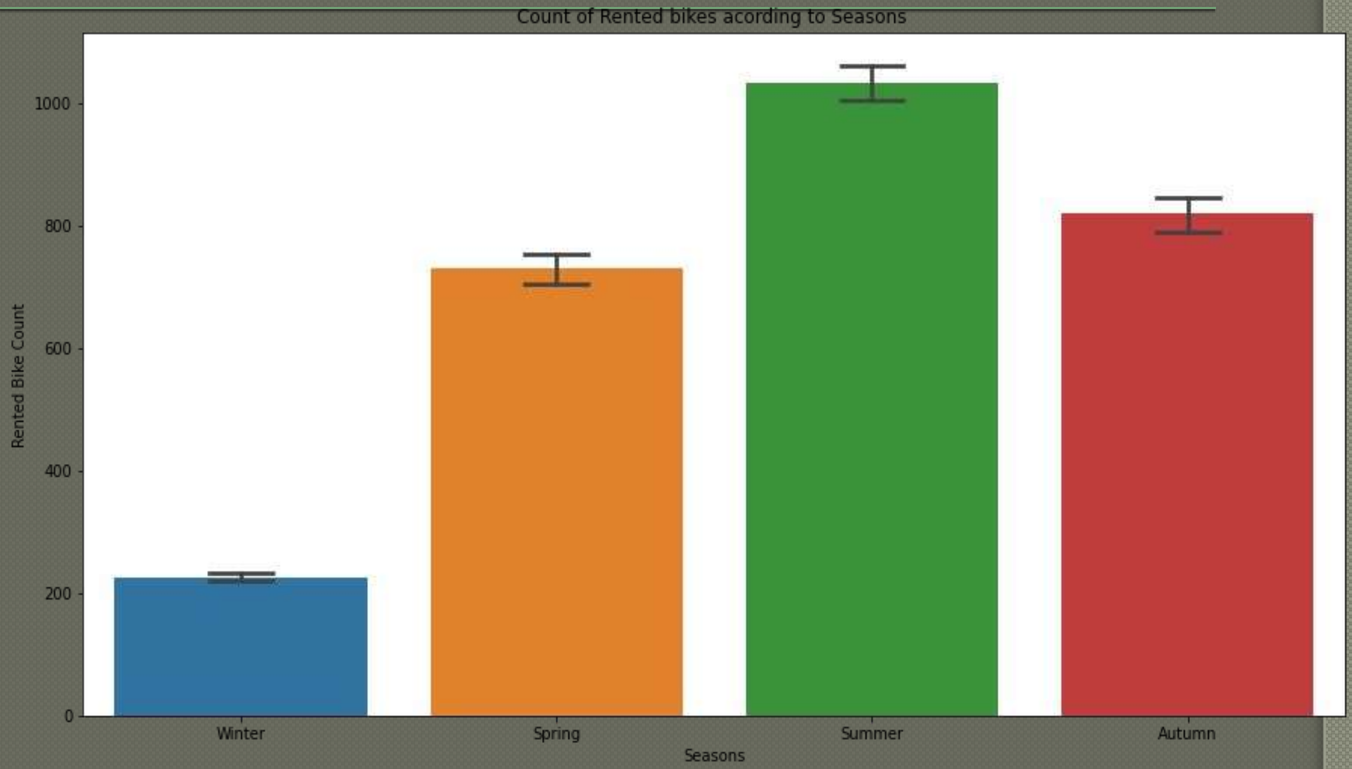




From the above point plot and bar plot we can say that in the week days which represent in blue colur show that the demand of the bike higher because of the office. Peak Time are 7 am to 9 am and 5 pm to 7 pm The orange colour represent the weekend days, and it show that the demand of rented bikes are very low specially in the morning hour but when the evening start from 4 pm to 8 pm the demand slightly increases

seasons

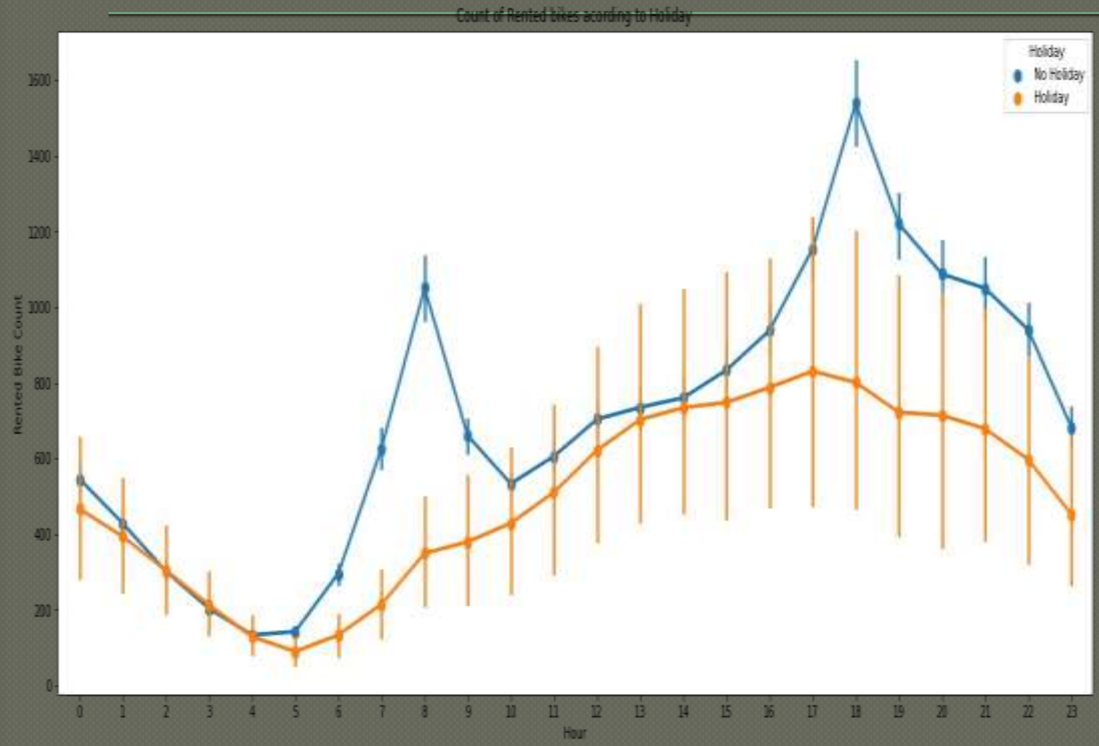


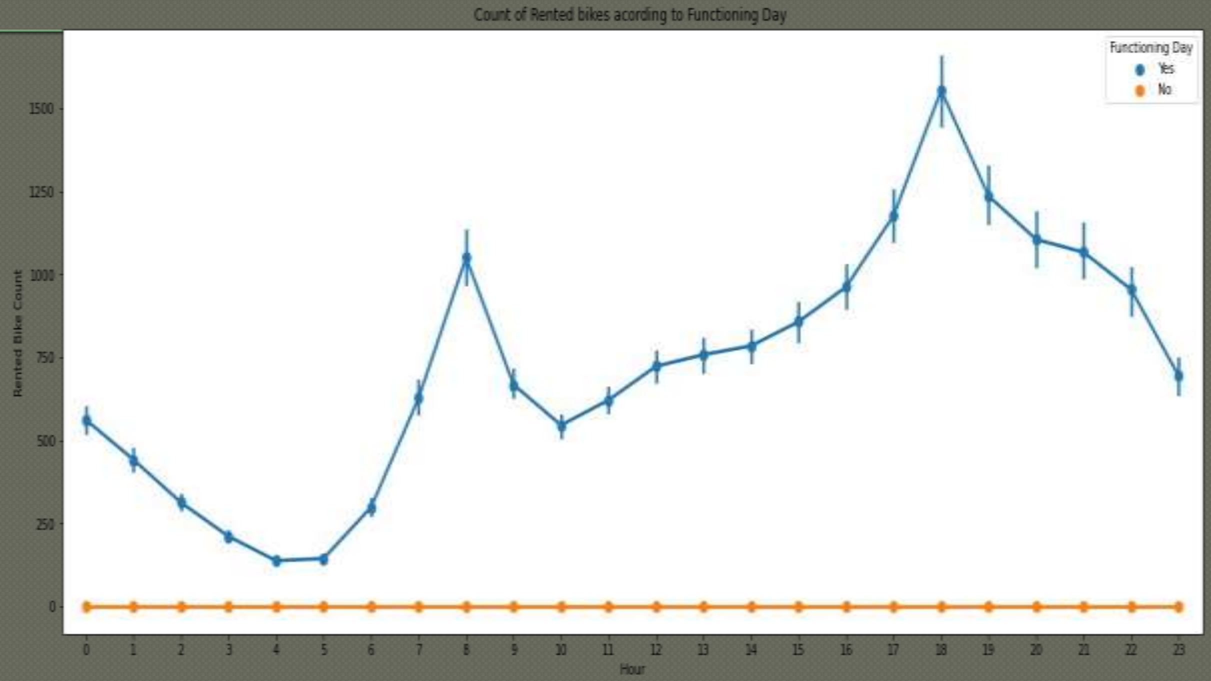


In the above box plot and bar plot which shows the use of rented bike in in four different seasons, and it clearly shows that, In summer season the use of rented bike is high In winter season the use of rented bike is very low because of snowfall.

Holiday

Functioning day





In the above point plot which shows the use of rented bike in a holiday, and it clearly shows that, plot shows that in holiday people uses the rented bike from 2pm-8pm

In the above point plot which shows the use of rented bike in functioning day or not, and it clearly shows that, Peoples dont use reneted bikes in no functioning day

Data Analysis Steps

Imported Libraries :-

In this part, we imported the required libraries NumPy, Pandas, matplotlib, and seaborn, to perform Exploratory Data Analysis and for prediction, we imported the Scikit learn library.

Descriptive Statistics :-

In this part, we start by looking at descriptive statistic parameters for the dataset. We will use describe() function to find out mean, median and standard deviation.

Missing Value Imputation :-

We will now check for missing values in our dataset. after checking non existed any missing values, In case there are any missing entries, we will impute them with appropriate values.

Encoded categorical data :-

Since machine learning models can only be trained with numeric data ,we used OneHot encoder and Label Encoder to change categorical data into numerical data

Scaling Data :-

We have used MinMax scalar and Standard Scale to scale our numeric data so that it becomes range bounded.

Spliting training and testing set :-

We split the dataset into a training and testing set. We have a randomly selected 20% subset of the data for testing. Also, we have used just the numeric and encoded columns.

Checked various models and applied hyperparamter tuning :-

We have used around 12 models and have applied hyperparamter tuning to get us the best accuracy with least error

Graphical Representation

We started with Univariate Analysis then bivariate Analysis and concluded with various prediction models driving the Demand for bikes

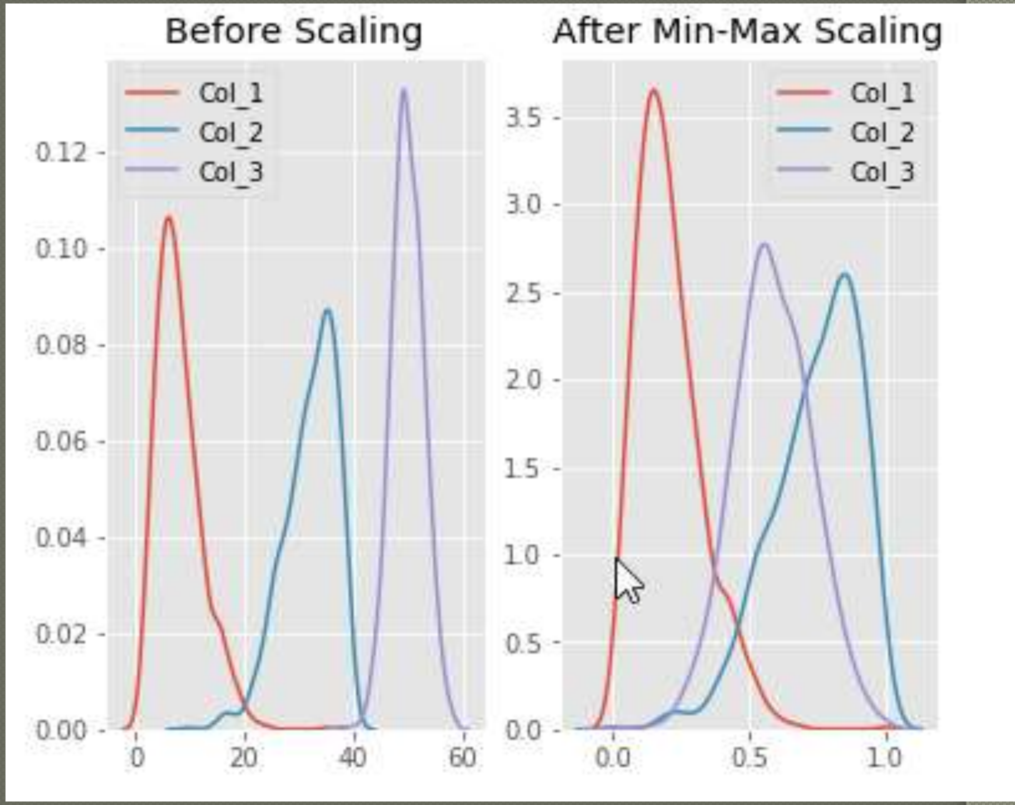
Scaling

Types of scaling :

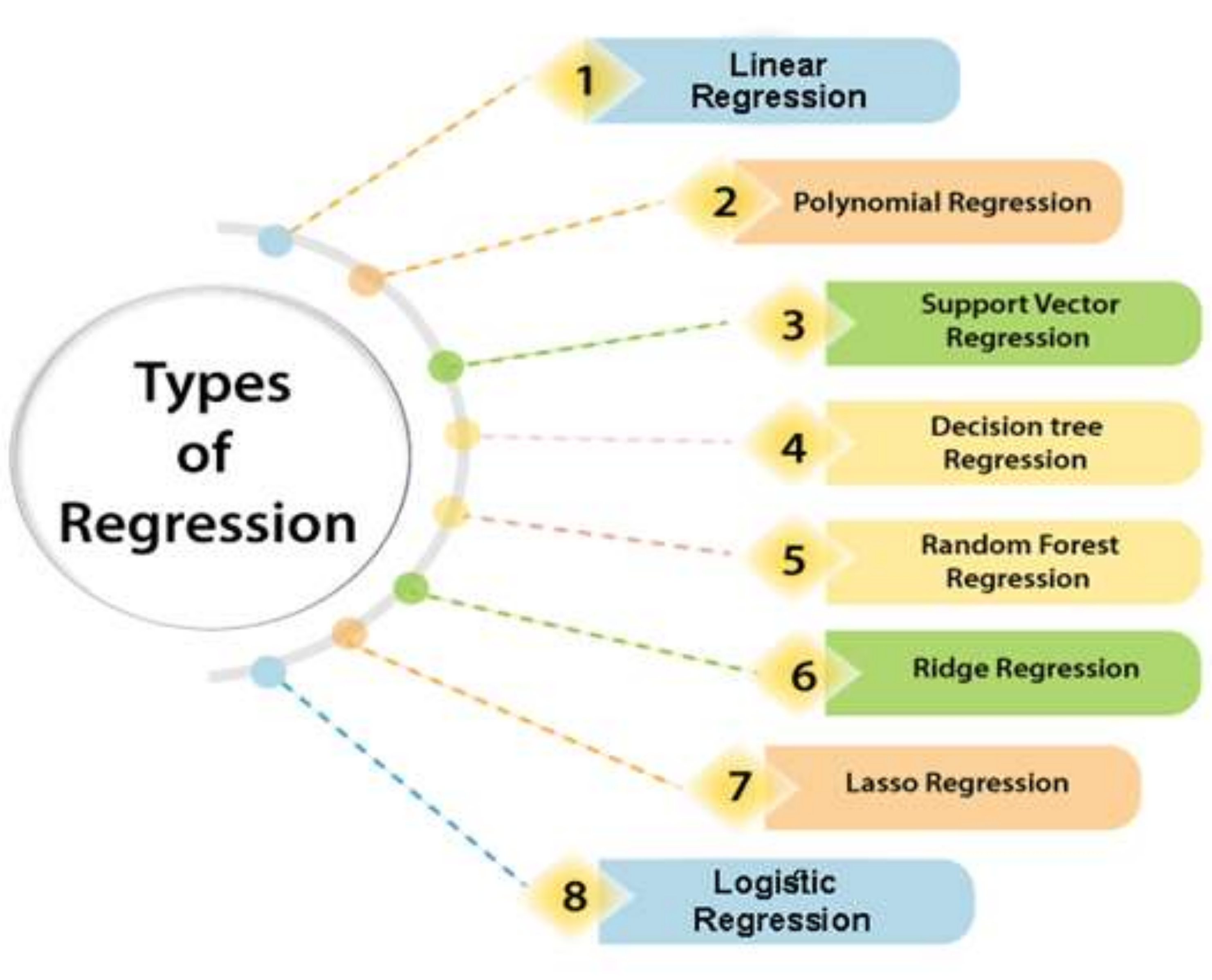
MinMaxScalar- scales all the data features in the range [0, 1] or else in the range [-1, 1] if there are negative values. It scales the values to a specific value range Without changing the shape of the original distribution.

StandardScalar-In Machine Learning, StandardScaler is used to resize the distribution of values so that the mean of the observed values is 0 and the standard deviation is 1.

RobustScalar-This Scaler removes the median and scales the data according to the quantile range (defaults to IQR: Interquartile Range).



Scaling Data and Model Building

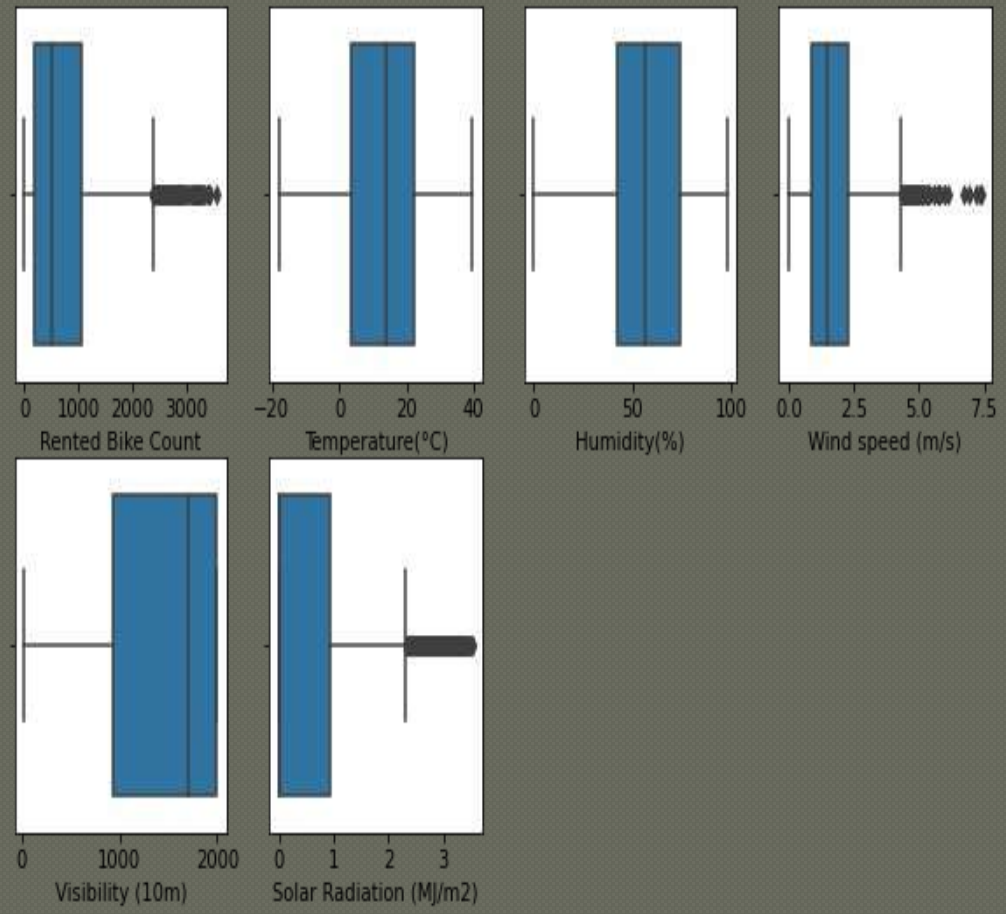


We checked the accuracy of our model using different scaling methods & different Regression’s also.

We apply all 3 different scaler and check accuracy difference between scalers.

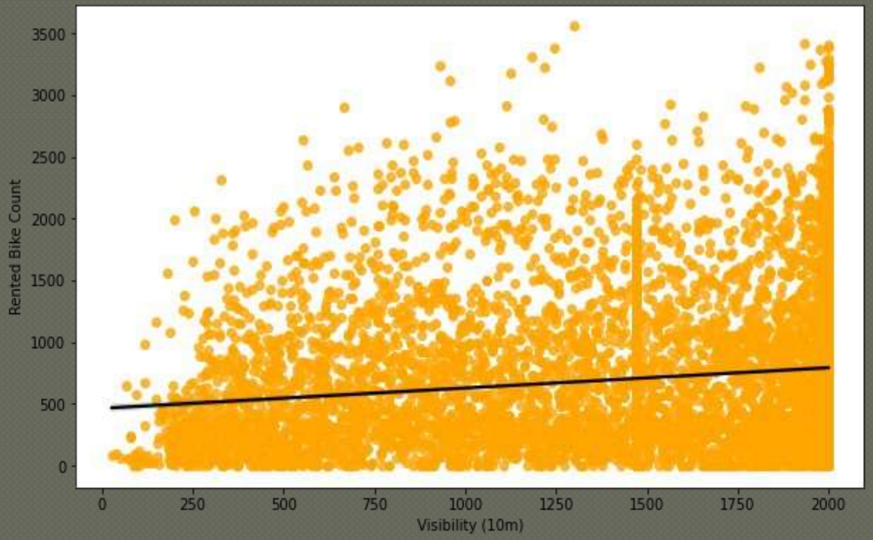
Checking difference between Actual test value and Predicted value

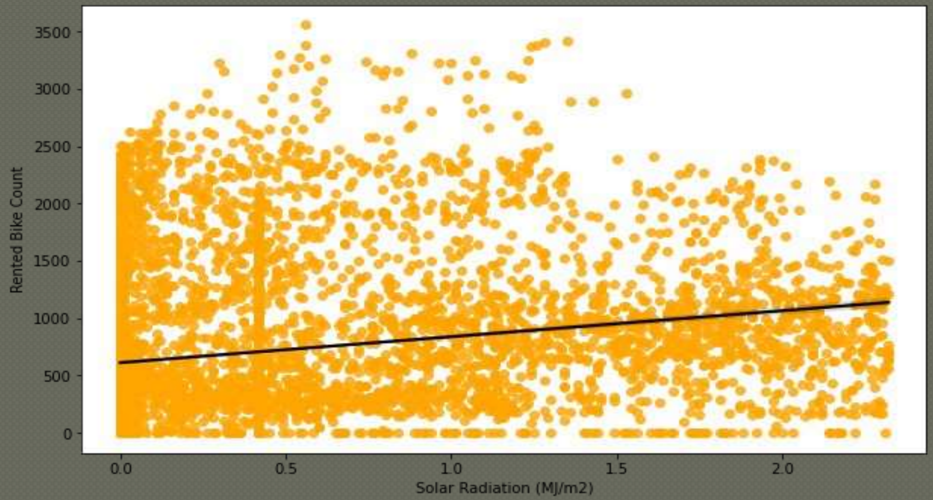
Handling outliers

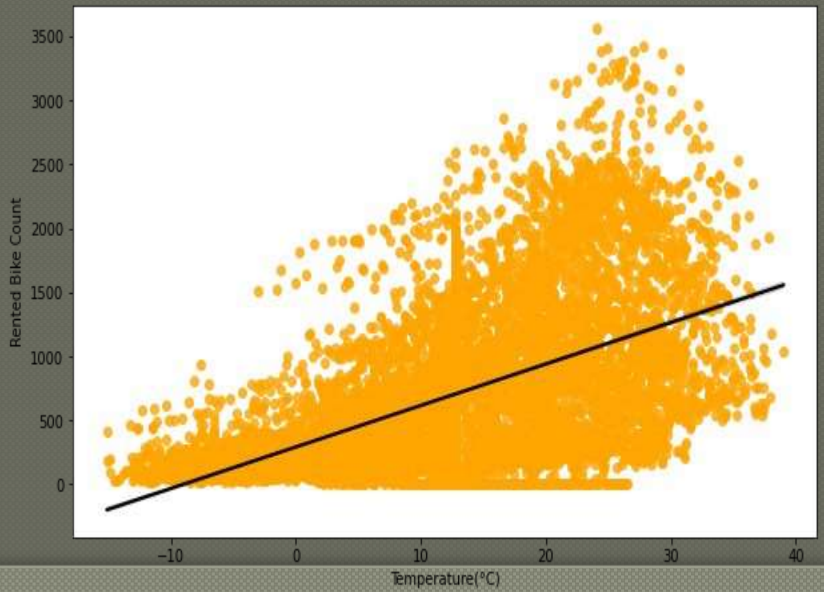


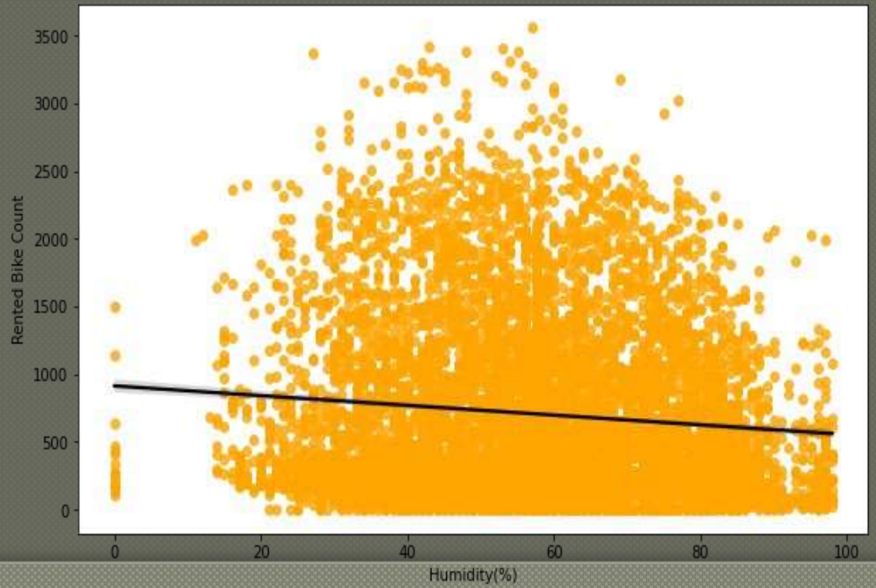
An Outlier is a data-item/object that deviates significantly from the rest of the (so-called normal)objects,The interquartile range (IQR) is the difference between the 75th and 25th percentile of the data. It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outlier

Regression plot









Machine learning algorithms

1. Linear regression
2. Ridge regression
3. Elastic net
4. Decision tree
5. Random Forest regressor
6. SVR 7. Gradient boosting

Conclusion

- Most numbers of Bikes were rented in **Summer**, followed by **Autumn**, **Spring**, and **Winter**. **May-July** is the peak Bike renting Season, and **Dec-Feb** is the least preferred month for bike renting.

- Majority of the client in the bike rental sector belongs to the Working class. This is evident from EDA analysis where bike demand is more on weekdays, working days in Seoul.

- **Temperature** of **20-30 Degrees**, evening time **4 pm- 8 pm, Humidity** between **40%-60%** are the most favorable parameters where the Bike demand is at its peak.

- **Temperature, Hour** of the day, **Solar radiation**, and **Humidity** are major driving factors for the Bike rent demand.

- Feature and Labels had a weak linear relationship, hence the prediction from the linear model was very low. Best predictions are obtained with GradientBoosting Regressor with applied hyperparameter tuning with r2 score of **0.917** and RMSE of **3.2018**