# Capstone project 2

on

# Bike sharing demand prediction

By

**Shubhangi Dharmik**

(Individual)

# Procedure

1. Introduction
2. Data summary
3. Data cleaning
4. Exploratory data analysis
5. One hot encoding
6. Train test split
7. Modeling
8. Conclusions

# Introduction

- The goal of this project is to combine the historical bike usage patterns with the weather data to forecast bike rental demand.
- The main objective is to build a predictive model, which could help to train a model to predict the number of bike rentals of the year given the weather conditions.  This would in turn help to predicting quickly and efficiently.

# Data summary

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

# Data cleaning

- **Null values Treatment and Outliers :** Dataset contains a no null values to disturb the accuracy but outliers are present which can disturb the accuracy. So Again, we use z-score to remove outliers.

# Exploratory data analysis

After loading and reading the dataset in notebook, we performed EDA. Comparing target variable which is bike rentals counts with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables and also we observed the distribution of variables. It gave us a better idea that how feature behaves with the target variable.

# One hot encoding

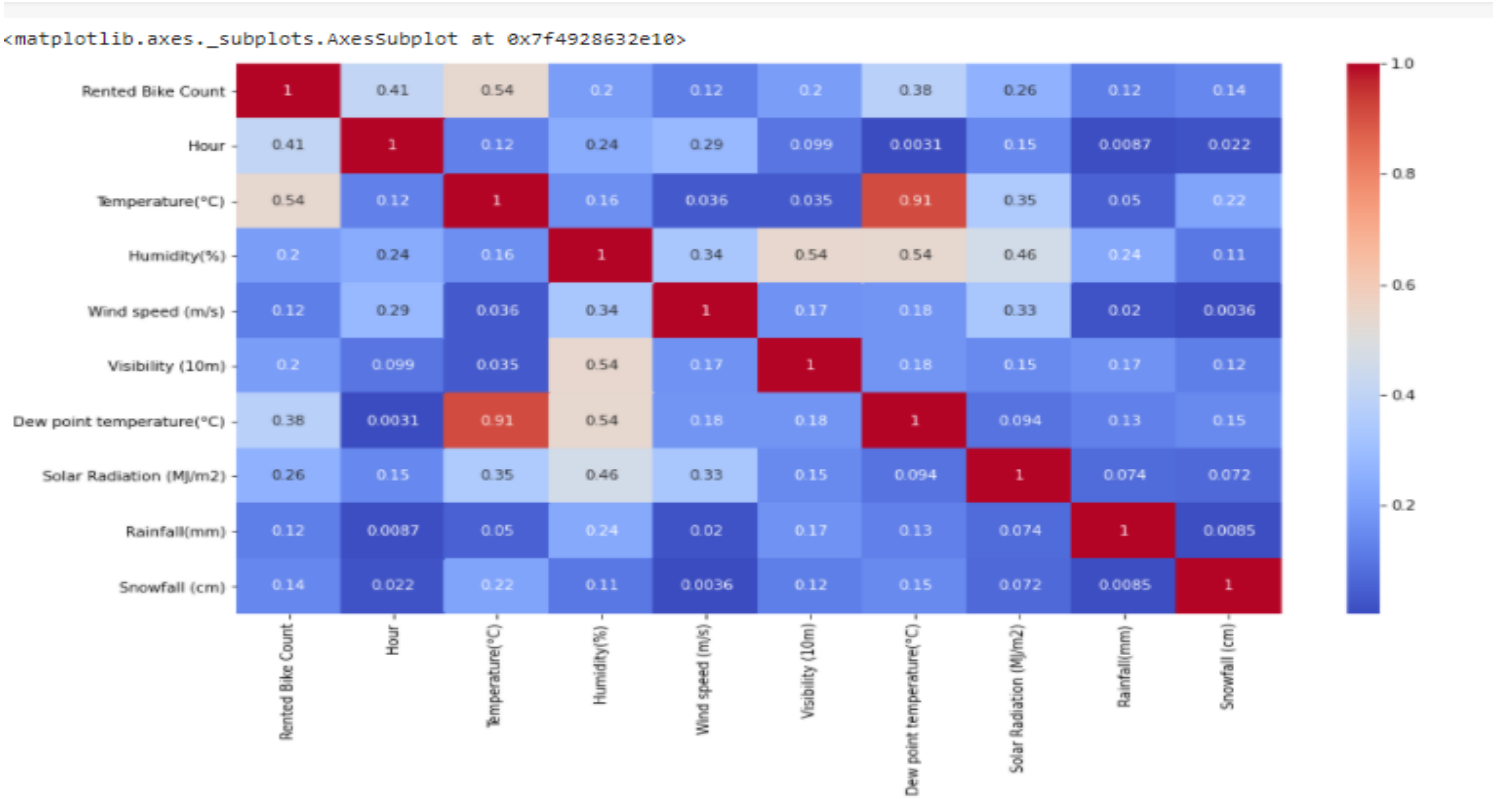
In this dataset some categorical variables

1. **Seasons**
2. **Holiday**
3. **Function day**

we change it with numerical database.

# Correlation Analysis :

- We plot the heatmap to find  the correlation between both dependent variable and independent variables.

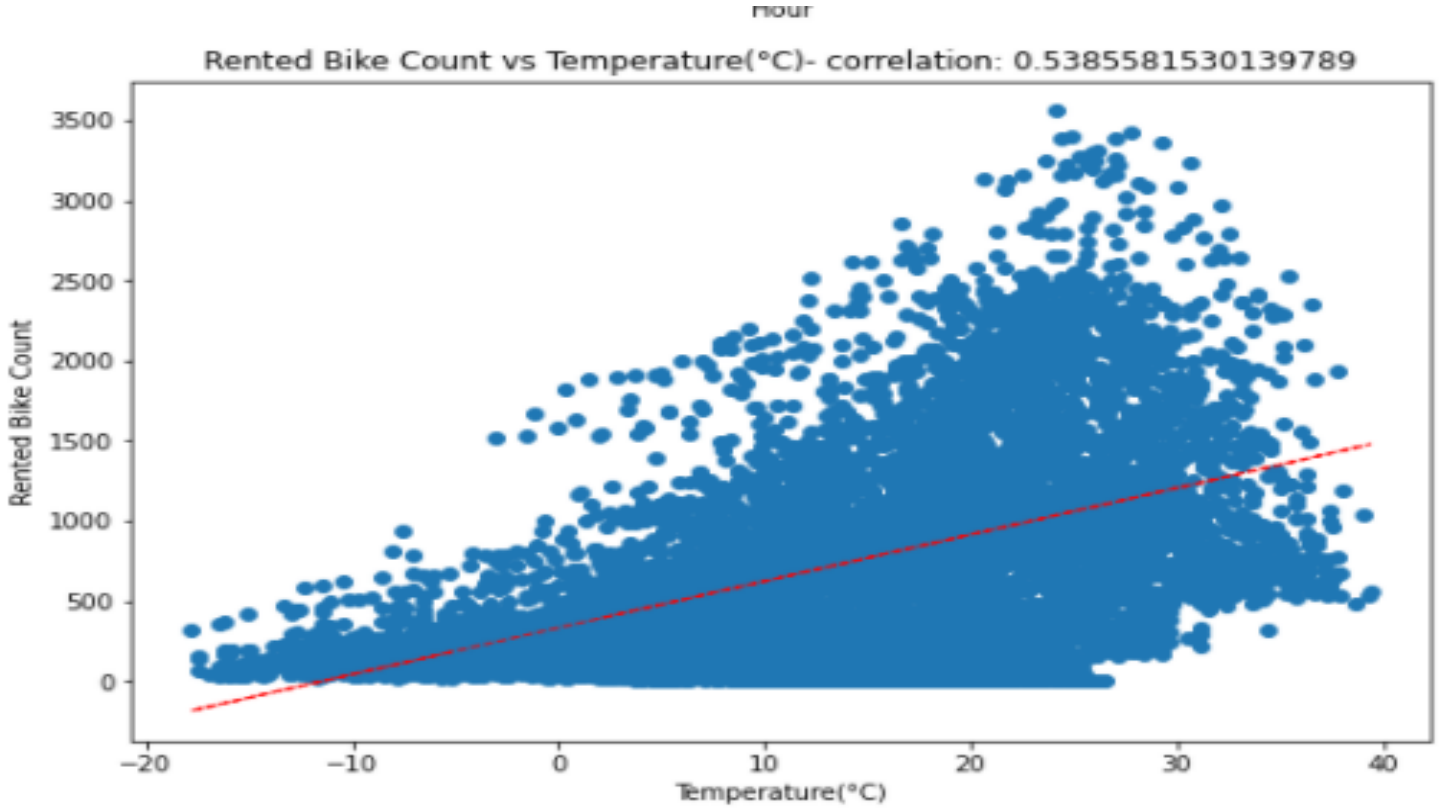


## From the heatmap we observed that :

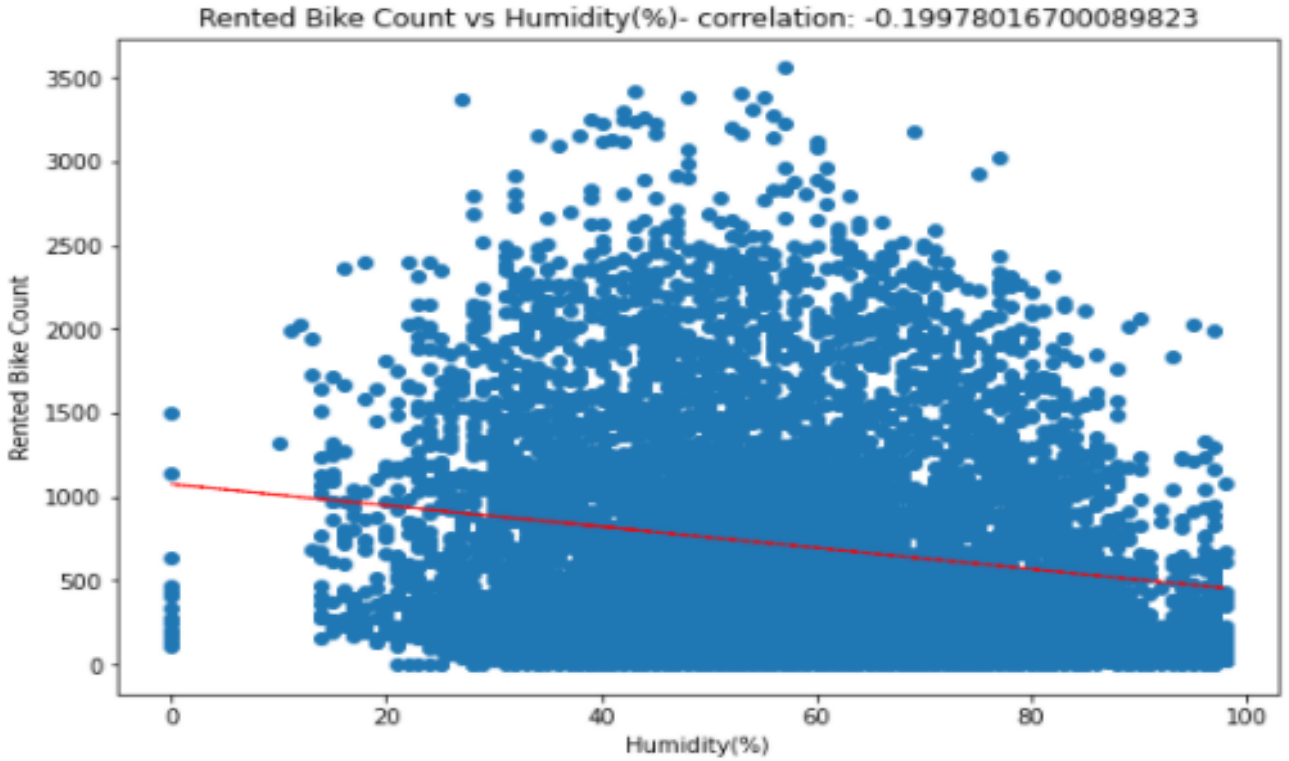
- Temperatures are highly correlated.
- There is a positive correlation between bike rentals counts and temperature.
- We observed a correlation between bike rentals counts and humidity. The more the humidity, the less people prefer to rental bikes.
- Bike rentals counts has a weak dependence on wind speed.

# Regression Plot

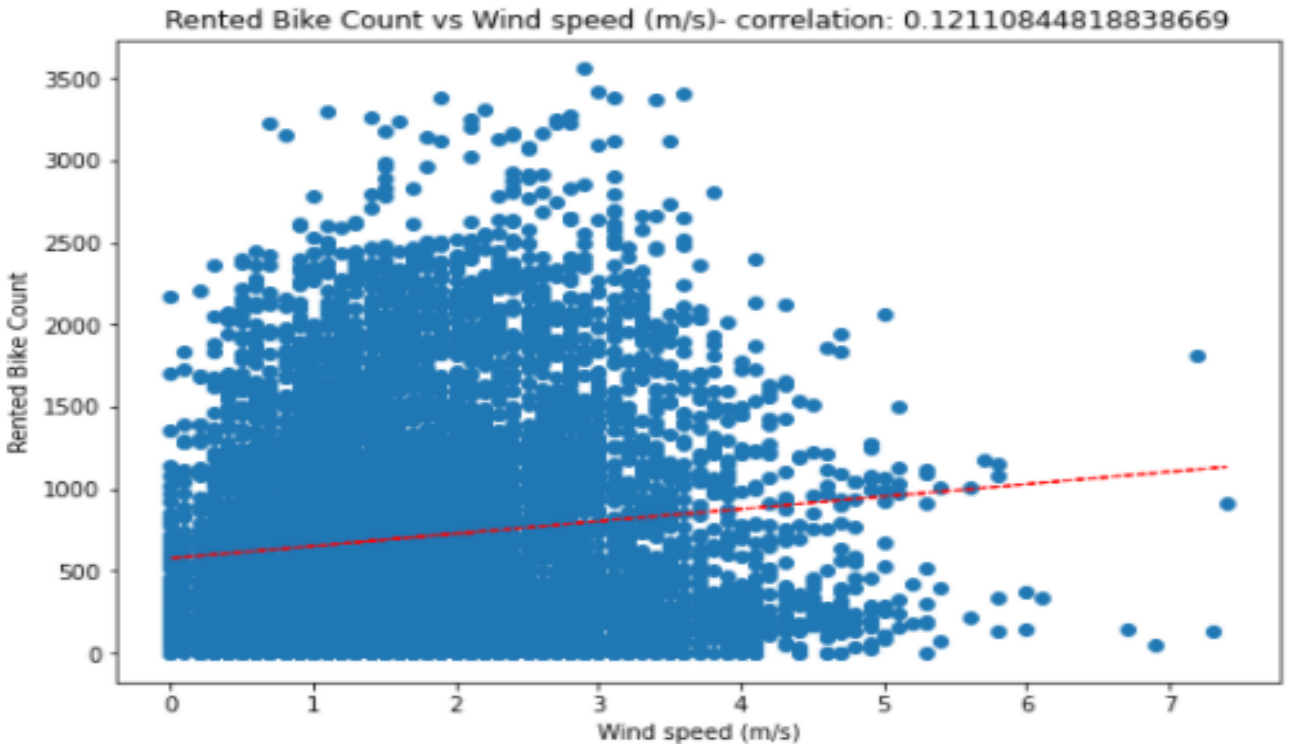
Below are regression plots of the bike rental count with Temperature, Humidity and Wind speed, respectively.




There is a positive correlation between bike rentals counts and temp.




We observed a correlation between bike rental counts and humidity. The more the humidity, the less people prefer to rental bikes.




Bike rentals counts has a weak dependence on wind speed. As we see in heatmap.

# Train test split

In the train test split we take two variables ie X and Y where X contain all the independent variables and Y contain dependent variable. Here the independent variable is bike rentals counts and dependent variables is affecting the bike rentals counts like temperature, weather, seasons etc.

# Modeling

## 1. Linear Regression :

We train model by linear regression and we get results as follows:

- R Squared for Training Data: 0.55
- R Squared for Testing Data: 0.54
- RMS for Training Data: 431.48
- RMS for Testing Data: 436.59
- MAE for Training Data: 322.21
- MAE for Testing Data: 326.49

# 2. Lasso regression

By performing lasso regression we get the results are as follows :

- R Squared for Training Data: 0.55
- R Squared for Testing Data: 0.54
- RMS for Training Data: 431.48
- RMS for Testing Data: 436.6

# 3. Ridge regression

By performing ridge regression we get the results are as follows :

- R Squared for Training Data: 0.55
- R Squared for Testing Data: 0.54
- RMS for Training Data: 431.49
- RMS for Testing Data: 436.63

# 4. Elastic Net

By performing ridge regression we get the results are as follows:

- R Squared for Training Data: 0.55
- R Squared for Testing Data: 0.54
- RMS for Training Data: 431.52
- RMS for Testing Data: 436.74

# Conclusions

- Bike rental count is mostly correlated with the time of the day as it is peak at 10 am morning and 8 pm at evening.
- We observed that bike rental count is high during working days than non working day.
- We see that people generally prefer to bike at moderate to high temperatures. We observed highest rental counts between 32 to 36 degrees Celsius.
- Hour of the day holds most importance among all the features for prediction of dataset.

- It is observed that highest number bike rentals counts in Autumn Summer seasons & the lowest in Spring season.
- We observed that the highest number of bike rentals on a clear day and the lowest on a snowy or rainy day.
- We observed that with increasing humidity, the number of bike rental counts decreases.
- We get quite less accuracy because of outliers are present in dataset.

# Thank You