Applied Data Science and Artifical Intelligence
Assignment 1-b

Ayush Raina

August 31, 2024

Data Preprocessing

1. There are no missing values in the dataset.
2. I have removed the Non Functioning Days, and after removing the non functioning days, I have removed that column also because all values in that column are same.
3. Then converted the date column to datetime format in pandas.
4. Added a column which shows the day of the week.
5. Added a column which shows whether it is a weekend or not.
6. Added a columnn which shows month of the year.
7. All integer or float data types are of numerical type.
8. 4 of them have data type object, which are categorical data type, 1 is of pandas datetime type and remaining are of numerical type.

I will be doing more feature engineering in further parts.

Data Visualization

1. We need to visualize how rented bike count varies hourly for different categorical features

We have 3 categorical features, which are:

1. Season
2. Holiday
3. Day of the Week

Visualizing with respect to Season

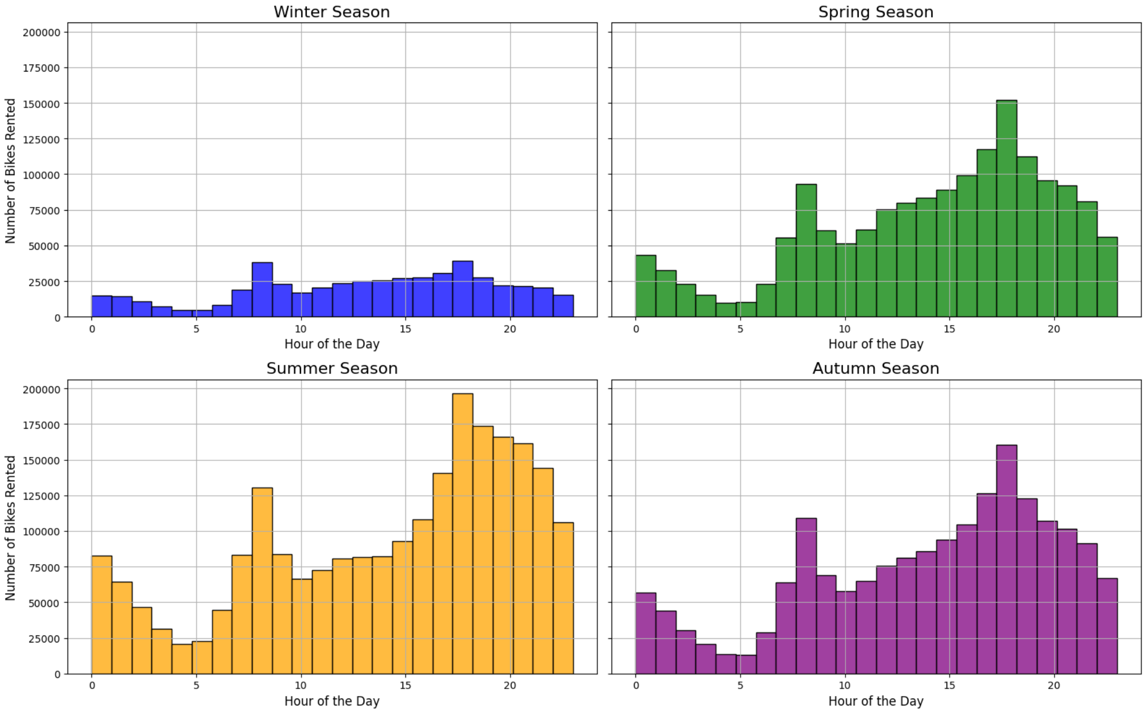


Figure 1: Rented Bike Count vs Hourly for different Seasons

Observations

1. Bike rentals are lowest in winter season than other 3 seasons.
2. From above plots we can observe that highest bike rentals are between 3pm to 8pm in all seasons.
3. Summer season has highest number of bike rentals followed by autumn, spring and winter.

Visualizing with respect to Holiday

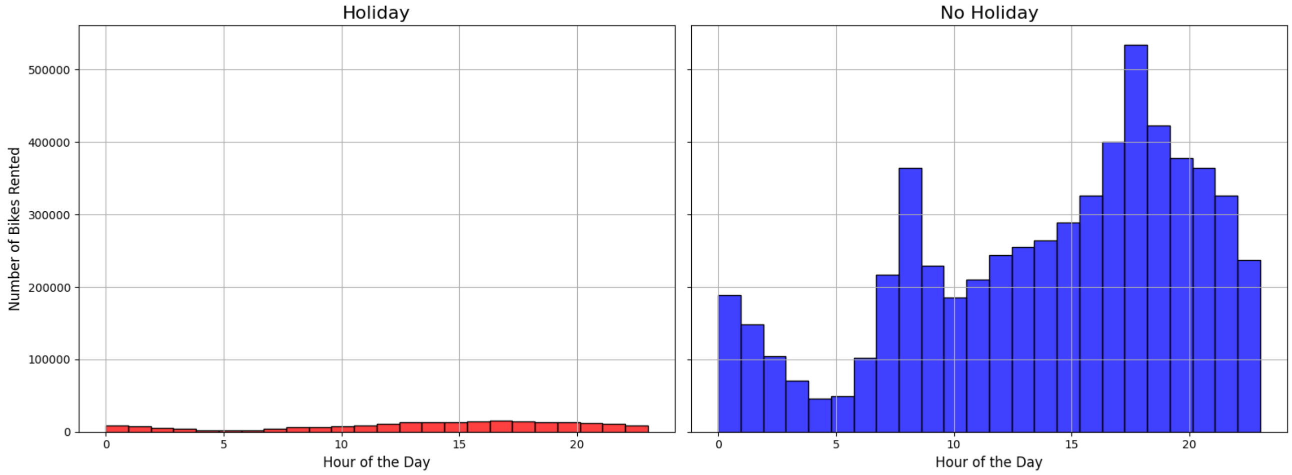


Figure 2: Rented Bike Count vs Hourly for different Holidays

Observations

1. We can clearly observe that there are a lot more bike rentals on non-holidays than on holidays.

2. Again highest number of bike rentals are between 3pm to 9pm.

Visualizing with respect to Day of the Week

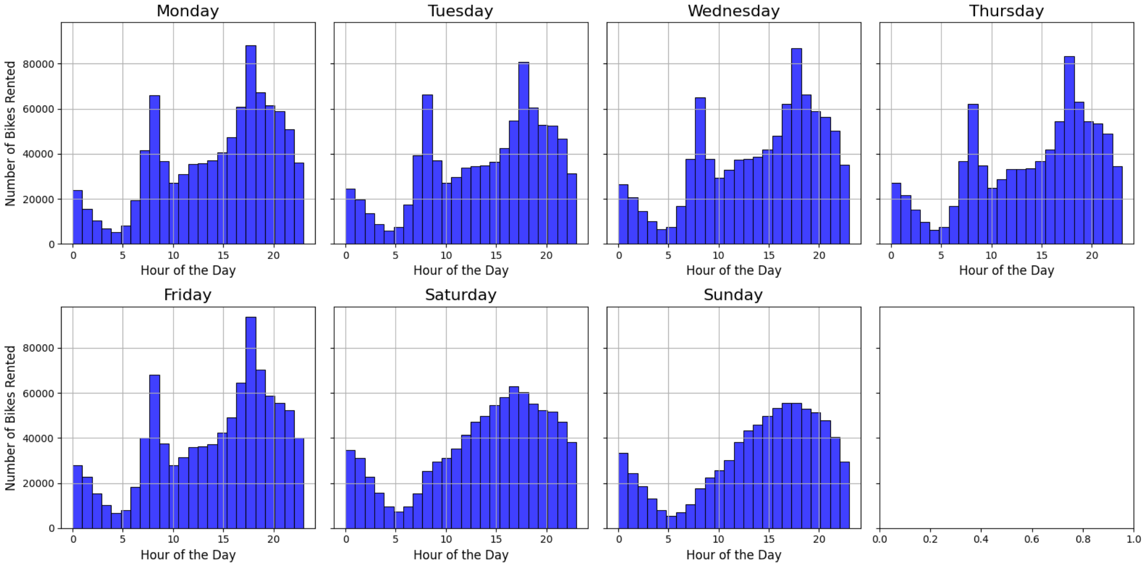


Figure 3: Rented Bike Count vs Hourly for different Days of the Week

Observations

1. From above hourly distribution of bike rentals are almost same for weekdays but different for weekends.

2. We can also observe the peak in the plot when highest numbers of bikes are rented on weekdays between 3pm to 8pm is missing on weekends.

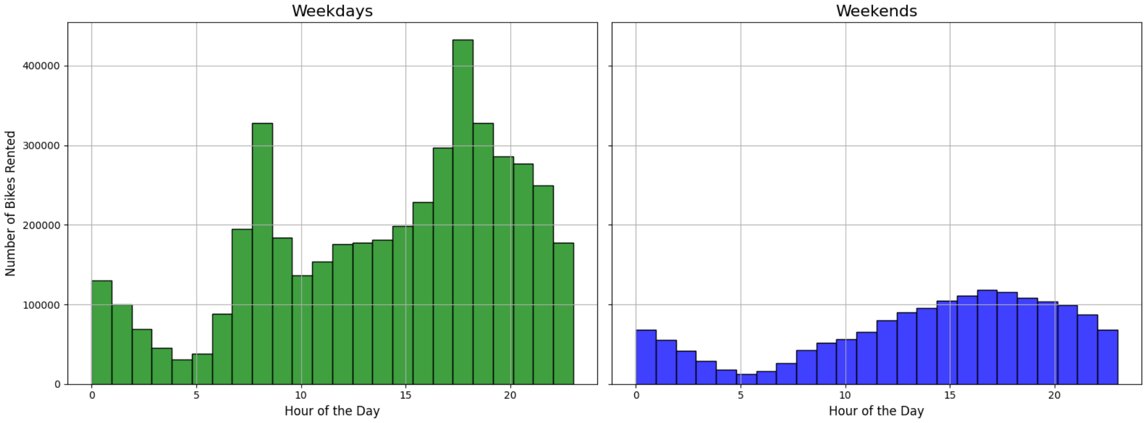


Figure 4: Rented Bike Count vs Hourly weekend vs weekdays

2. We need to visualize the outliers in rented bike count for different categorical features

Visualizing outliers with respect to Season

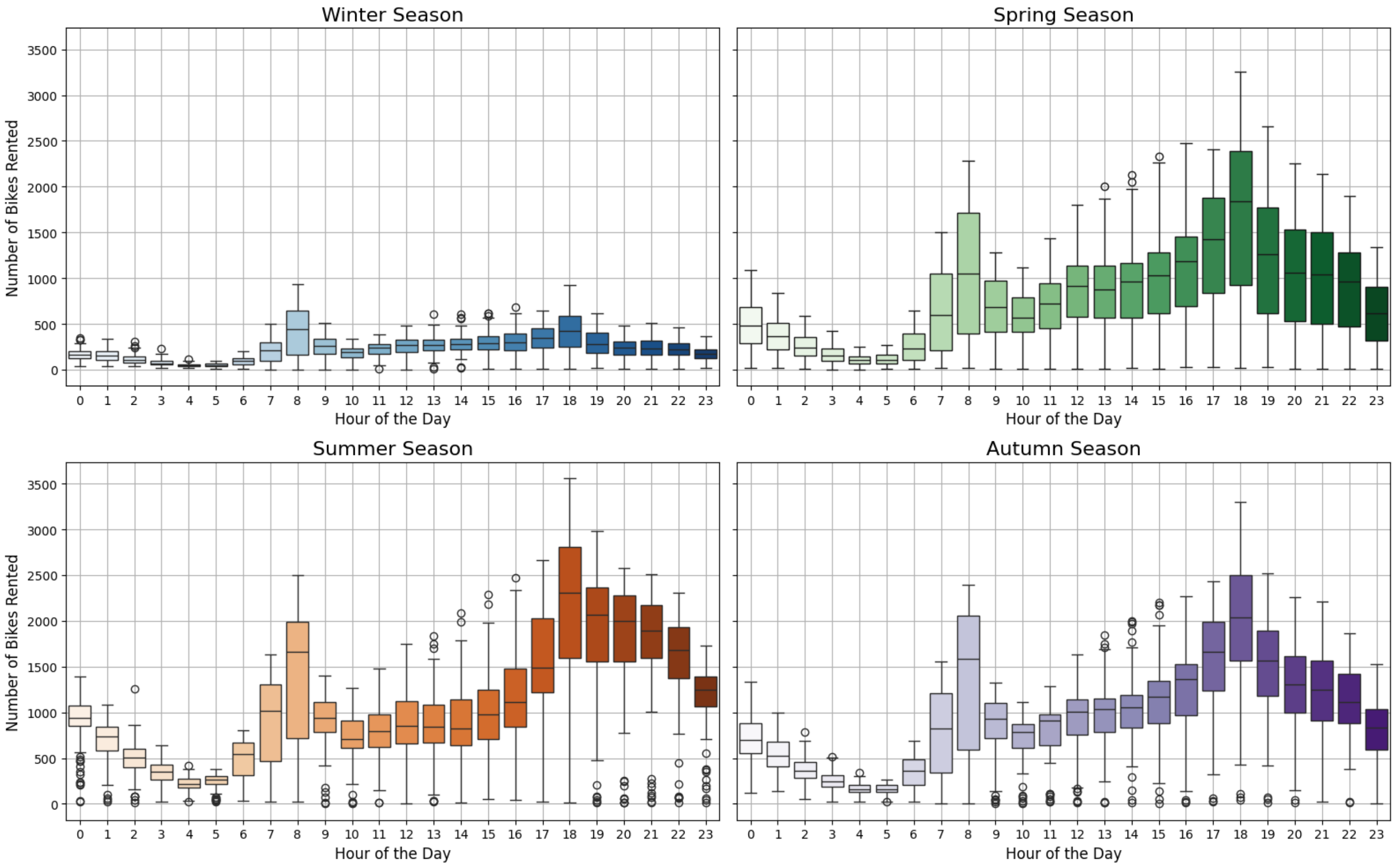


Figure 5: Outliers in Rented Bike Count for different Seasons

Observations

1. We can clearly observe that there are a lot of outliers in the summer season and there are very few outliers in the spring season.

Visualizing outliers with respect to Holiday

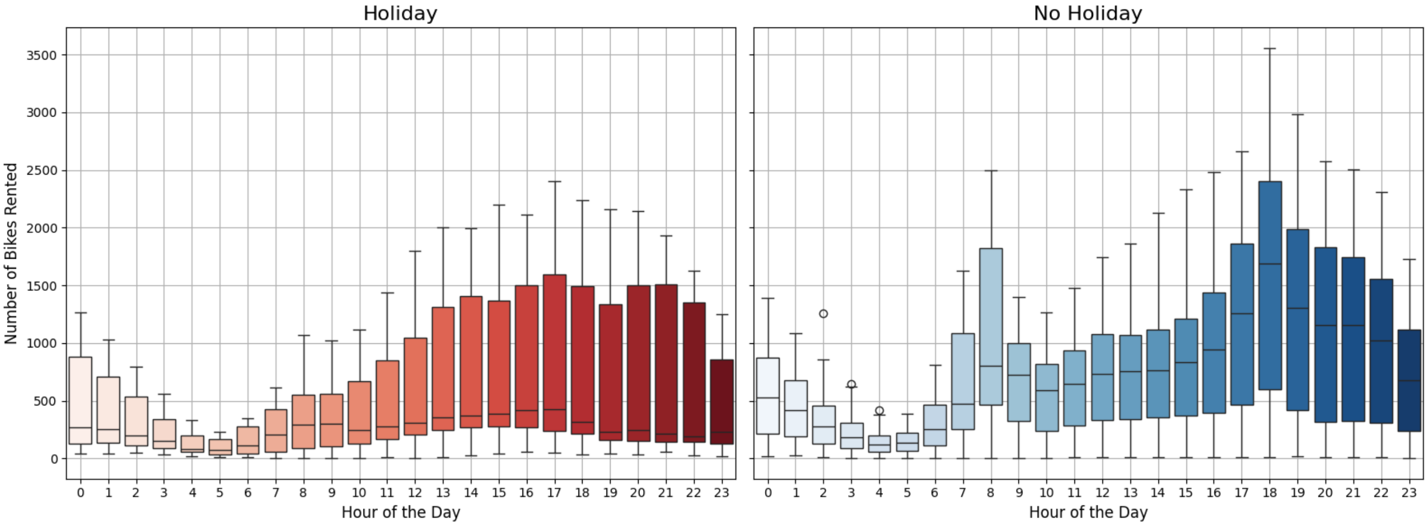


Figure 6: Outliers in Rented Bike Count for different Holidays

Observations

1. There are some outliers when there is no holiday but there almost no outliers when there is a holiday.
2. People are really enjoying their holidays.

Visualizing outliers with respect to Day of the Week

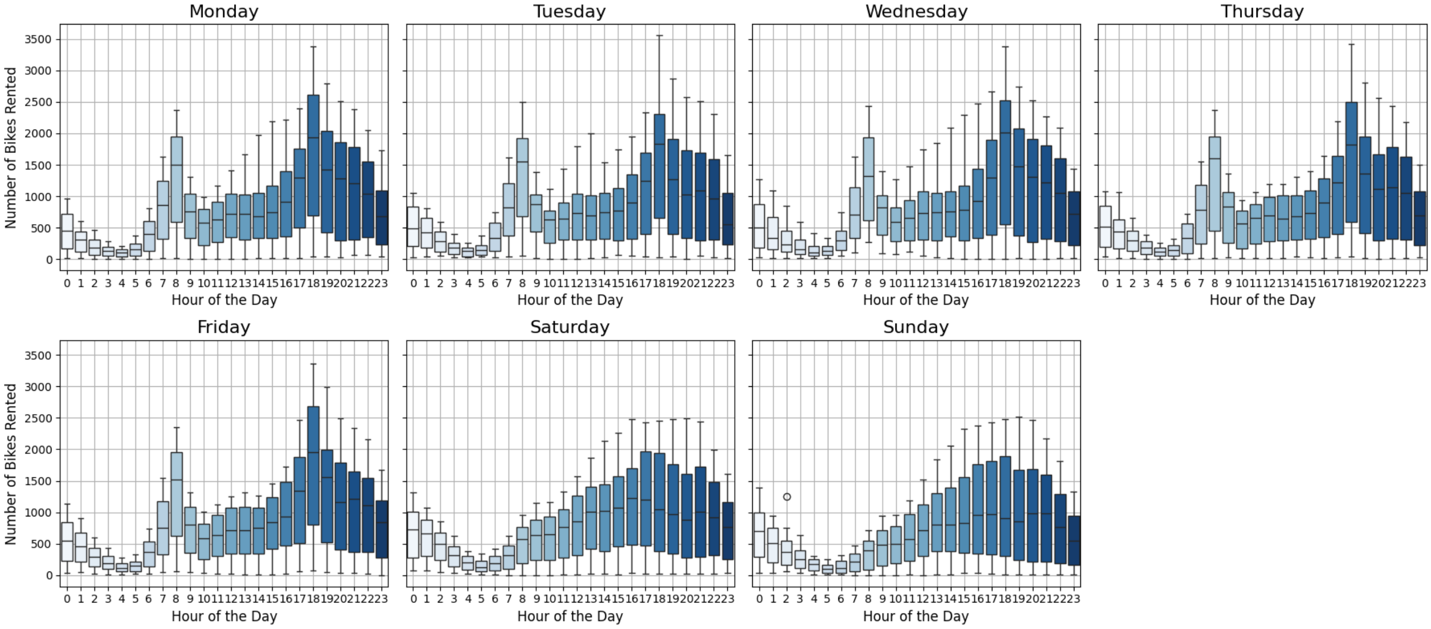


Figure 7: Outliers in Rented Bike Count for different Days of the Week

Observations

1. There are very few outliers that we can observe in above plots.

3. We need to visualize the data distribution for each numerical feature. We also need to show mean and median of the data distribution.

We have to plot the data distribution for Rented Bike Count, Humidity, Wind speed, visibility, dew point temperature, solar radiation, rainfall, snowfall.

Visualizing data distribution for Rented Bike Count

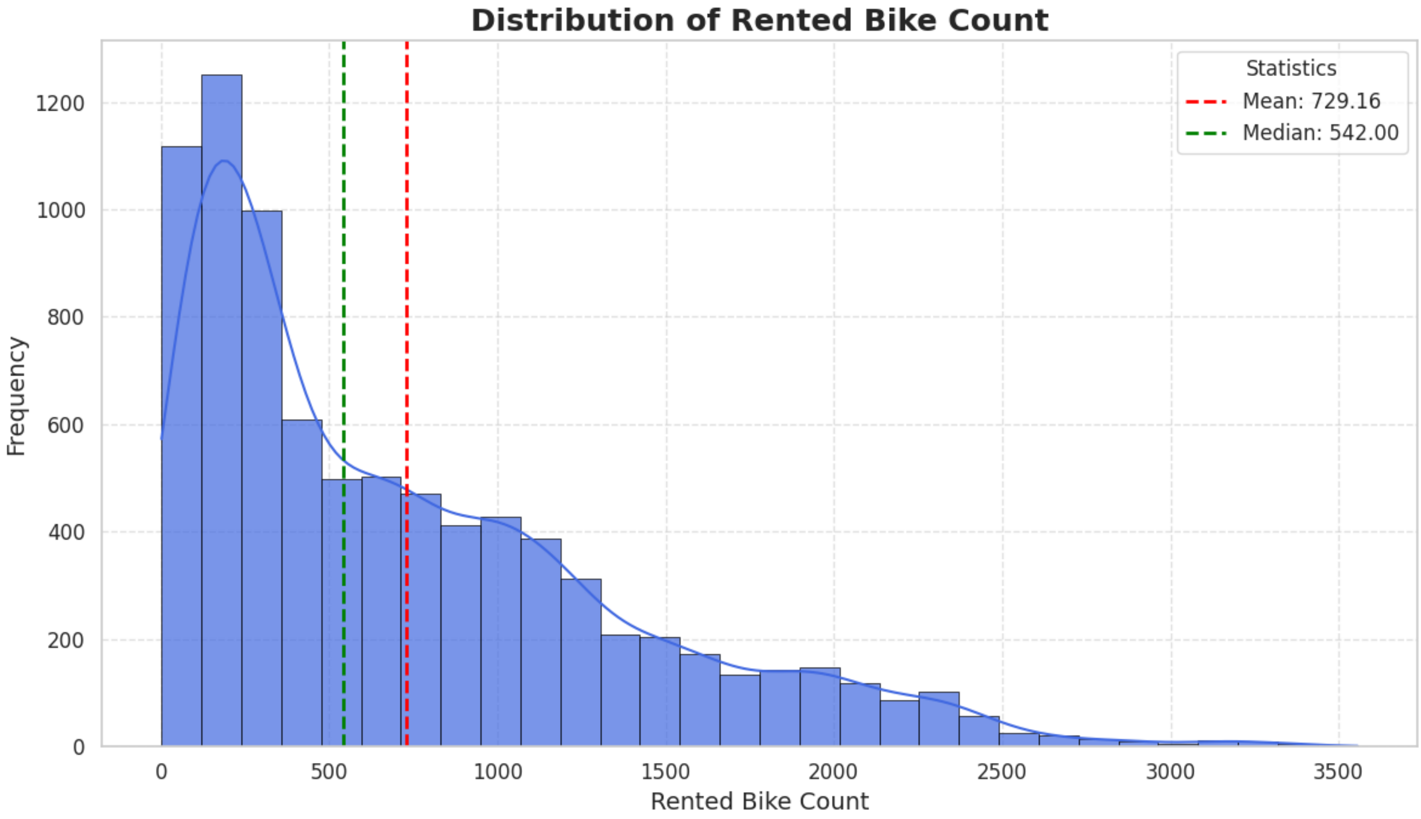


Figure 8: Data Distribution for Rented Bike Count

Observations

1. We can observe that the data distribution is right skewed.
2. Frequency of bike rentals is highest between 0 to 500.

Visualizing data distribution for Temperature

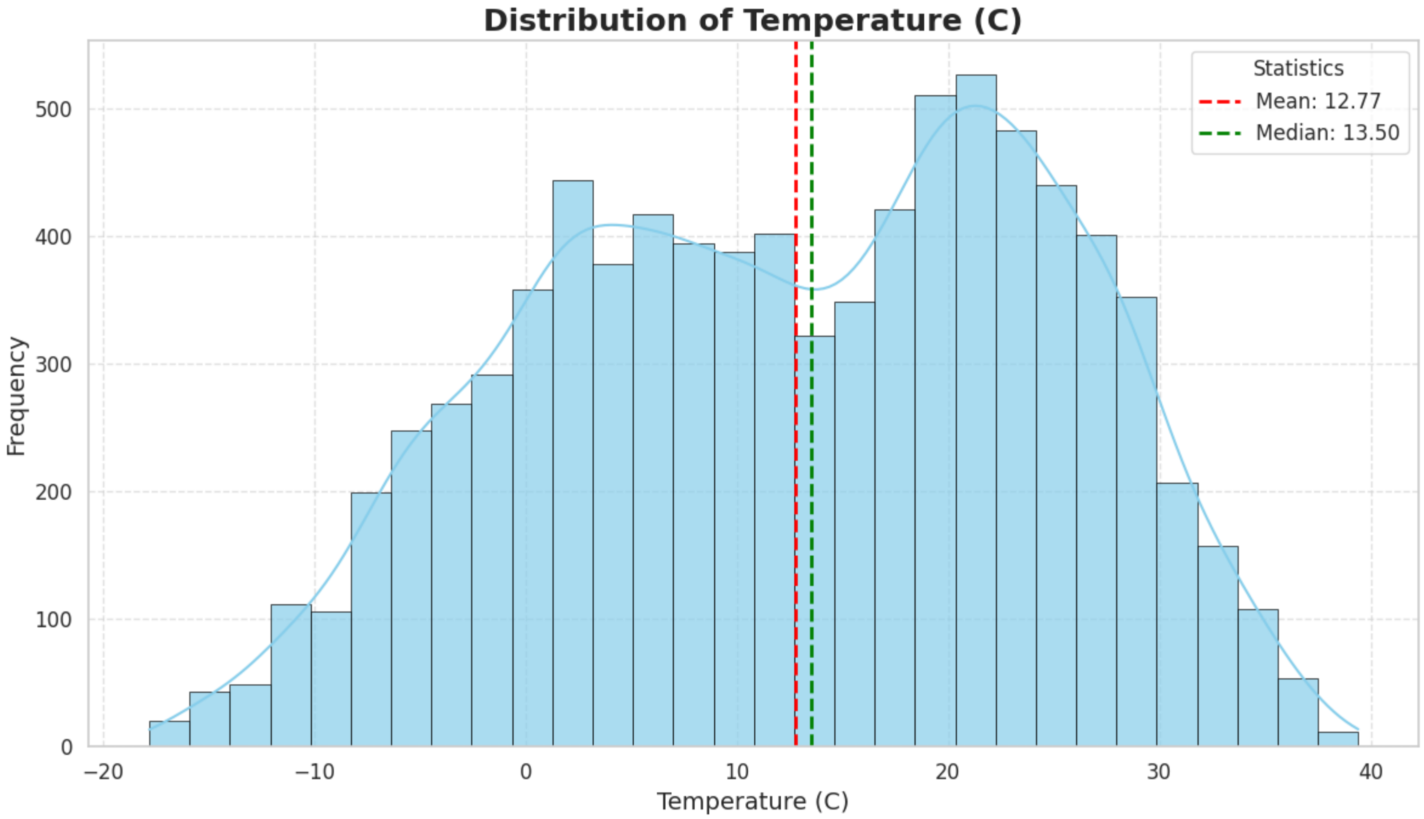


Figure 9: Data Distribution for Temperature

Observations

1. From the plot, it looks like balanced distribution.

Visualizing data distribution for Humidity

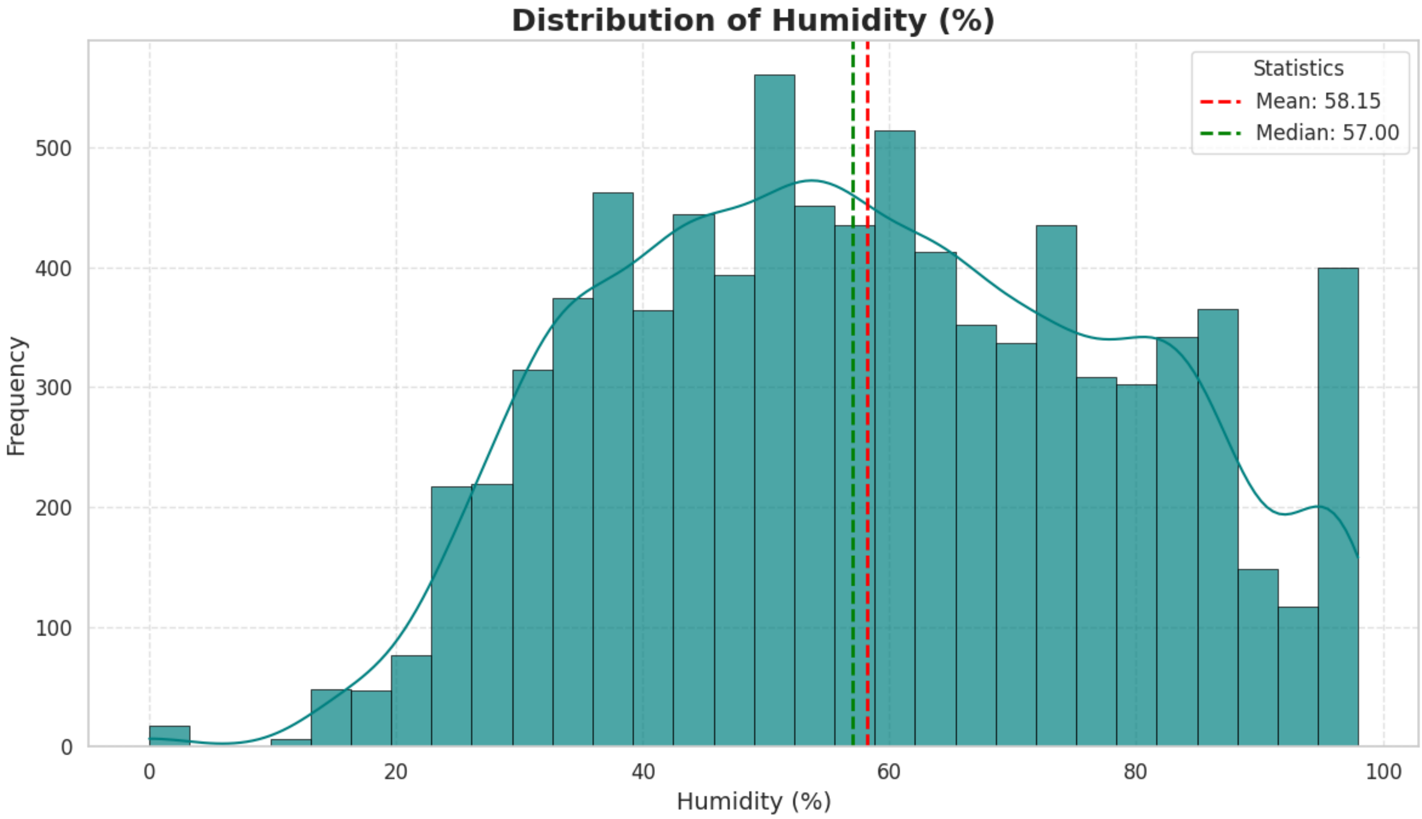


Figure 10: Data Distribution for Humidity

Observations

1. From the plot, it looks like humidity follows gaussian distribution.

Visualizing data distribution for Wind Speed

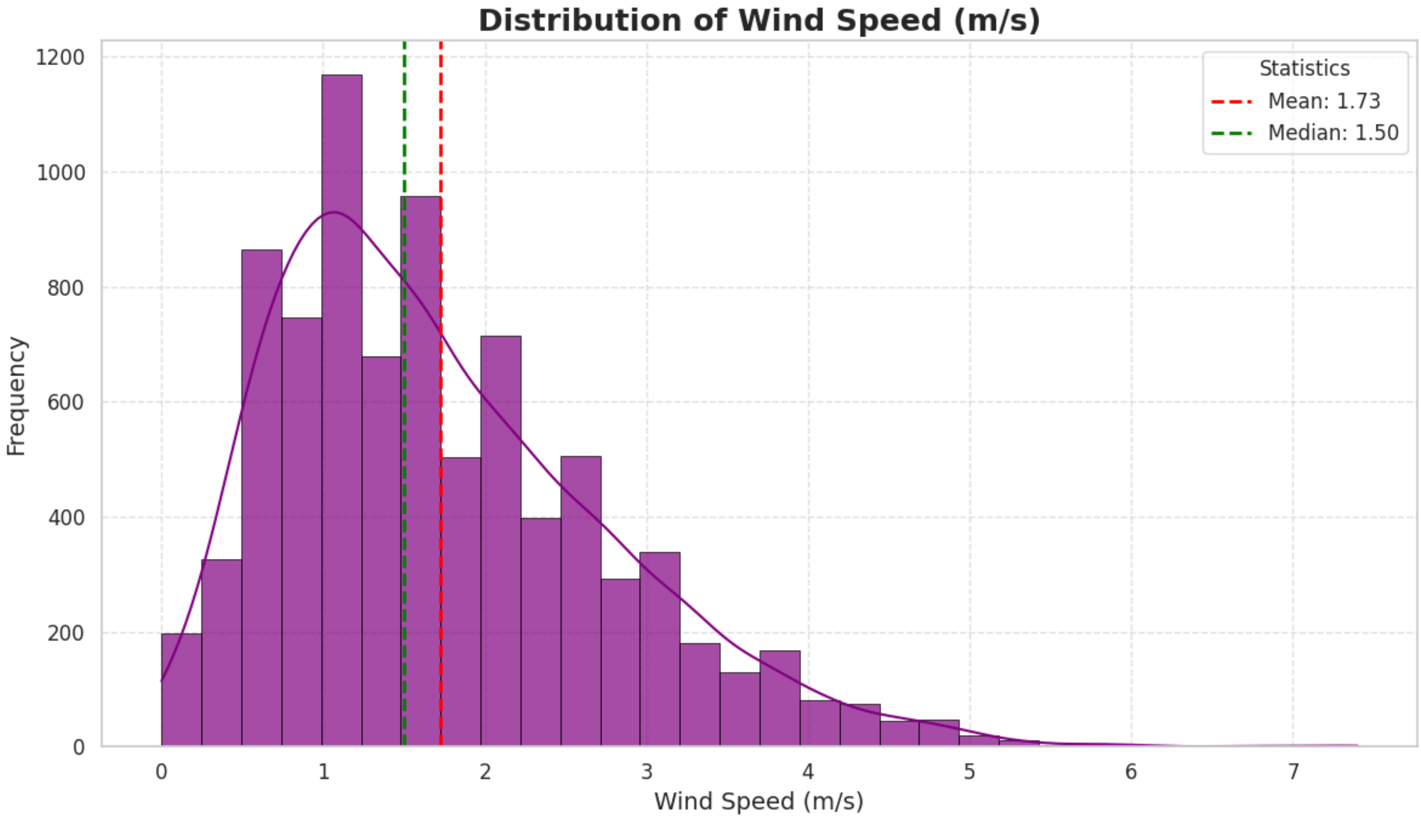


Figure 11: Data Distribution for Wind Speed

Observations

1. From the plot, it also looks like wind speed follows gaussian distribution.

Visualizing data distribution for Visibility

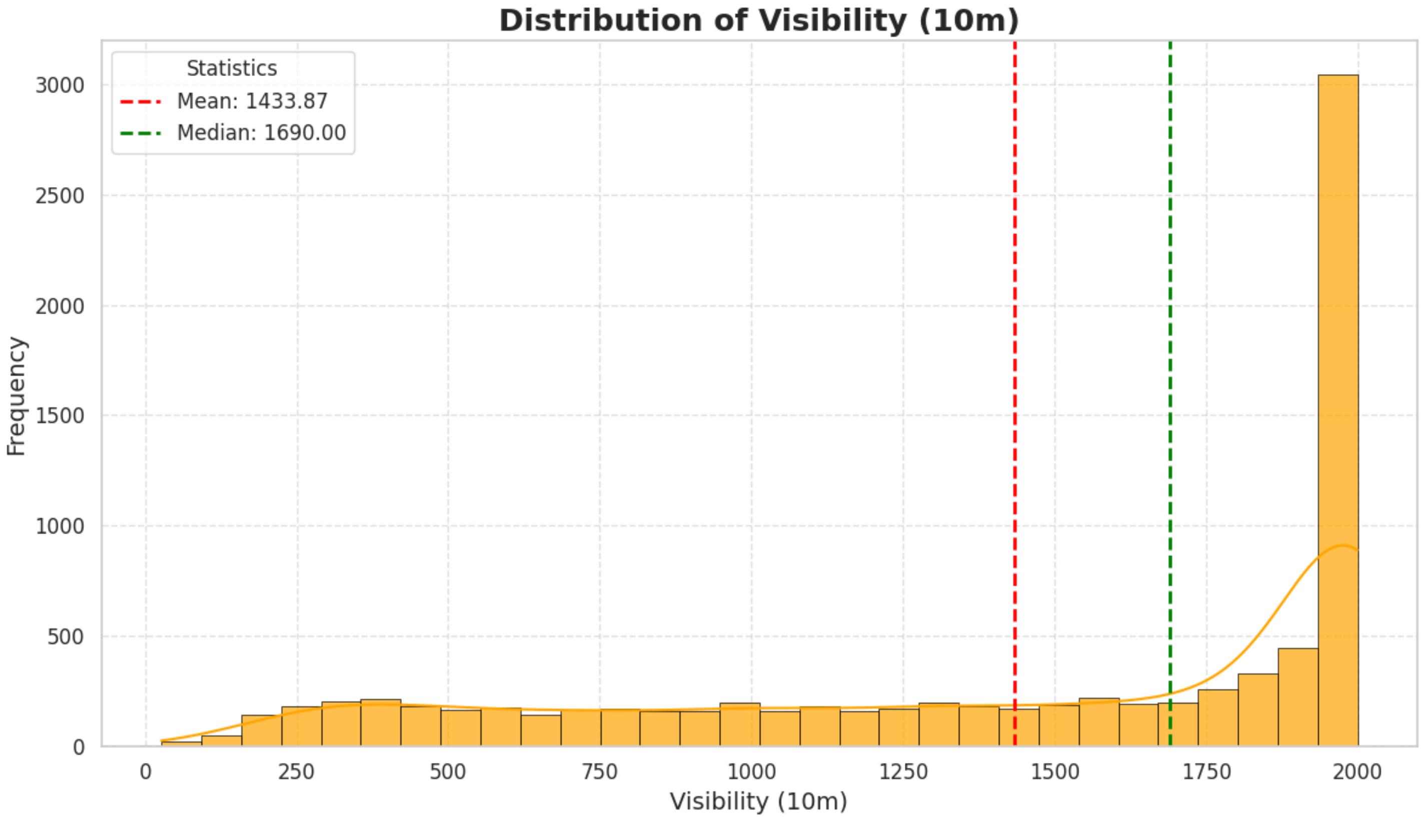


Figure 12: Data Distribution for Visibility

Observations

1. From the plot it can be seen that this distribution is highly left skewed.

Visualizing data distribution for Dew Point Temperature

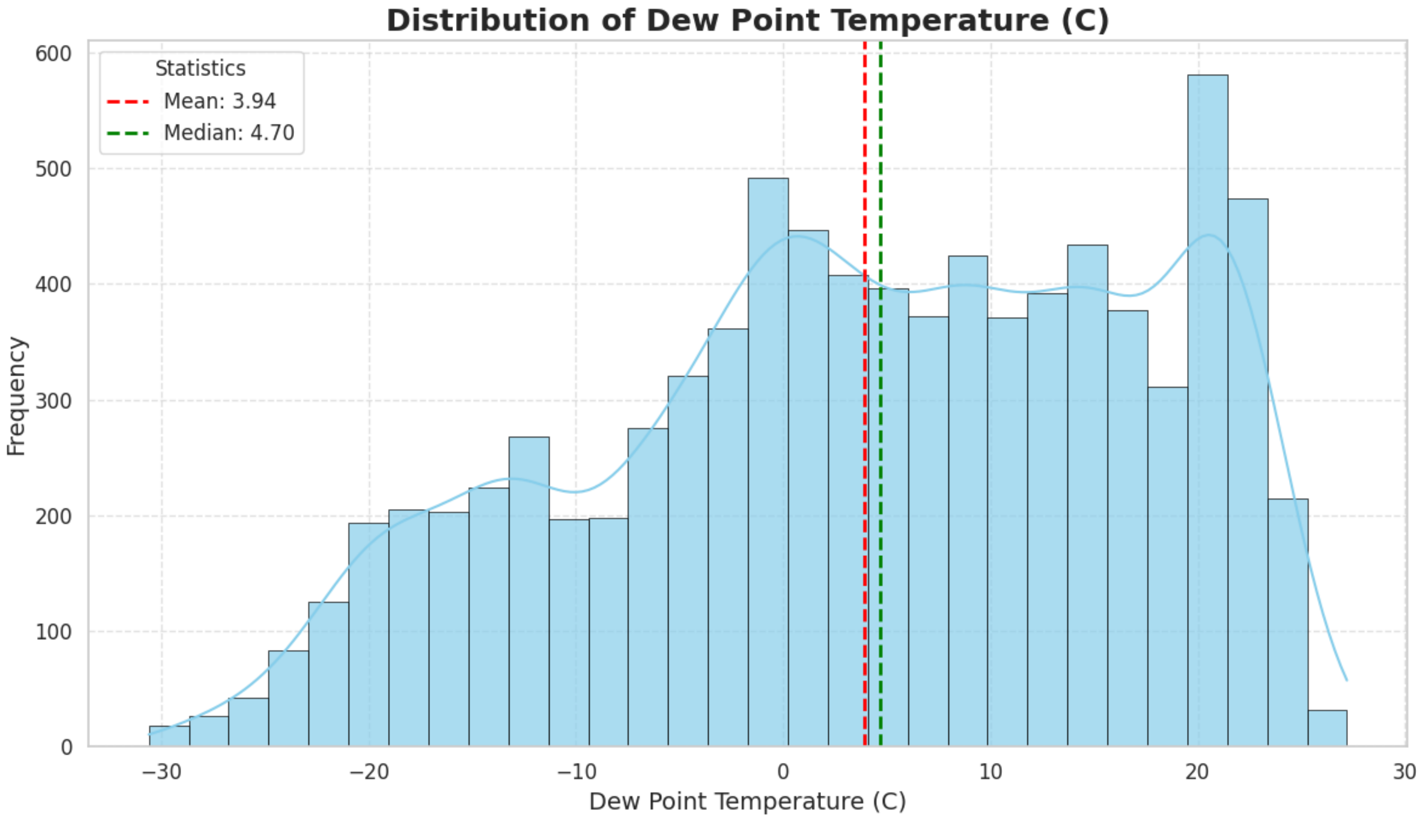


Figure 13: Data Distribution for Dew Point Temperature

Observations

1. This is a balanced distribution, mean-median are also close to each other.

Visualizing data distribution for Solar Radiation

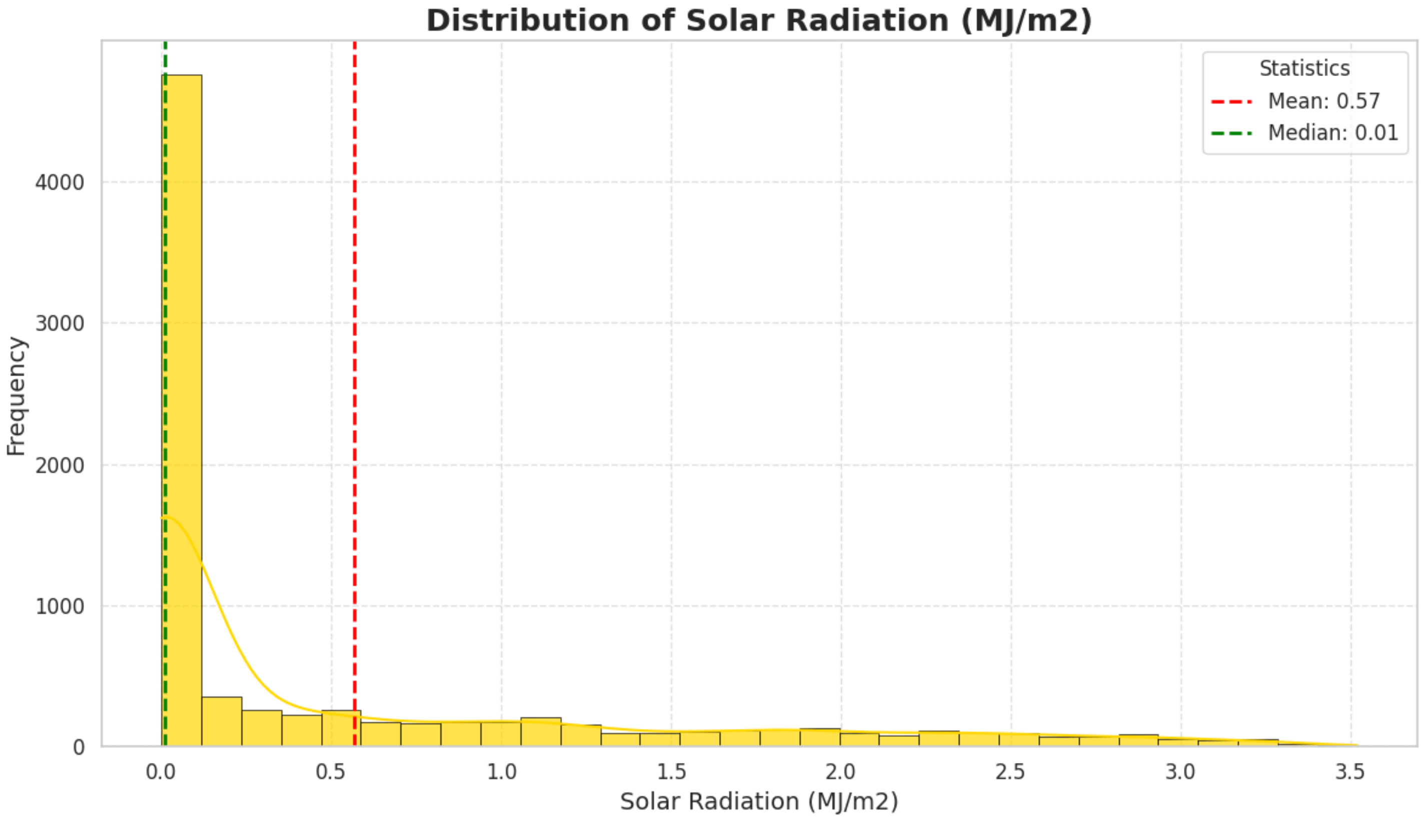


Figure 14: Data Distribution for Solar Radiation

Observations

1. This is a highly right skewed distribution.
2. Mean is greater than median.
3. Most of the values are between 0 to 1.

Visualizing data distribution for Rainfall

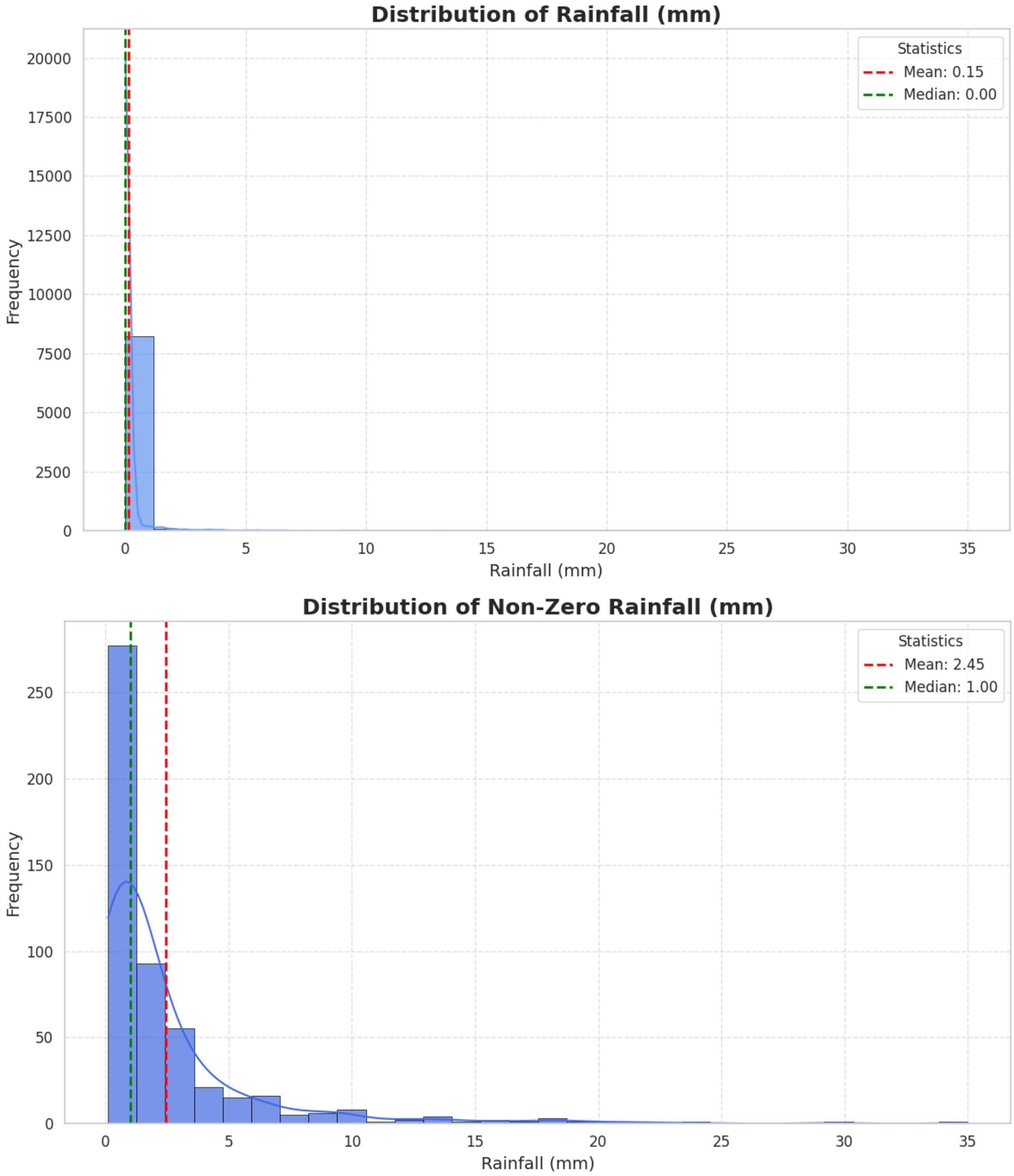


Figure 15: Data Distribution for Rainfall

Observations

1. Both zero-rainfall and non-zero rainfall data distribution is right skewed.

Visualizing data distribution for Snowfall

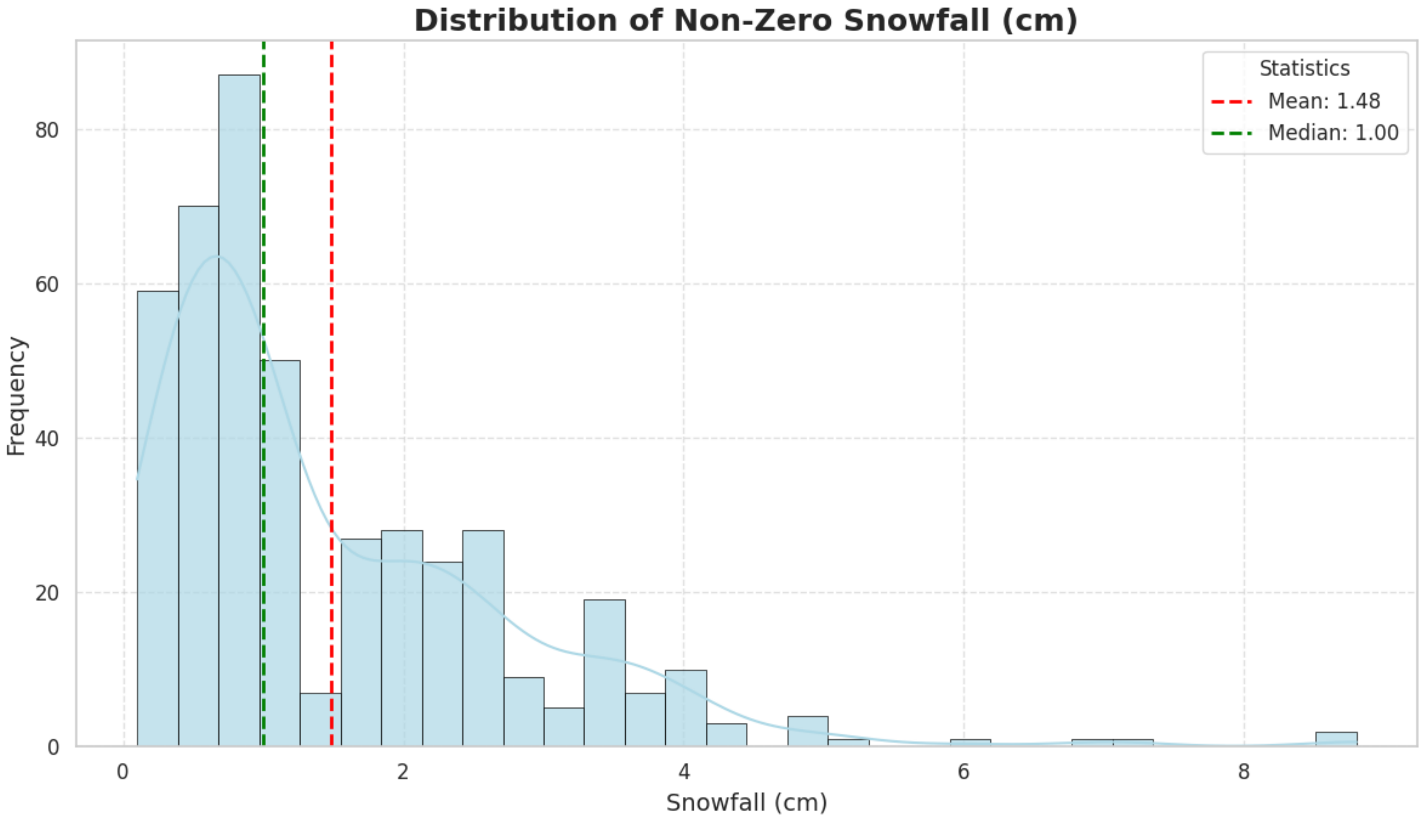


Figure 16: Data Distribution for Snowfall

Observations

1. This is a highly right skewed distribution.
2. Most of the values are between 0 to 1.

5. We want to visualize the outliers in numerical features.

Visualizing outliers for Rented Bike Count

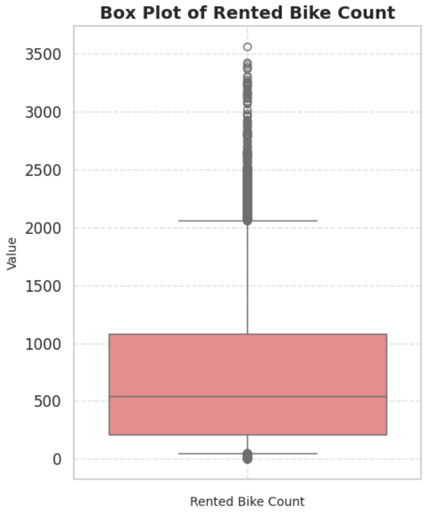


Figure 17: Outliers in Rented Bike Count

Visualizing outliers for Humidity

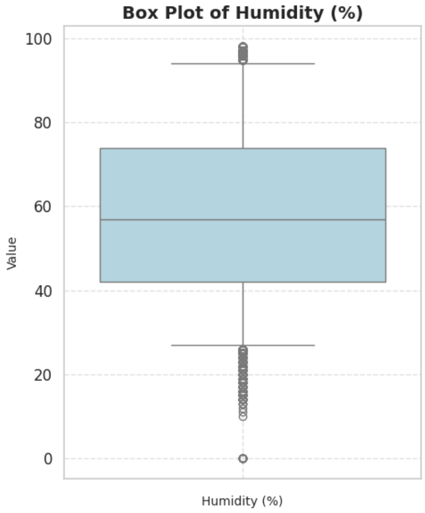


Figure 18: Outliers in Humidity

Visualizing outliers for Wind Speed

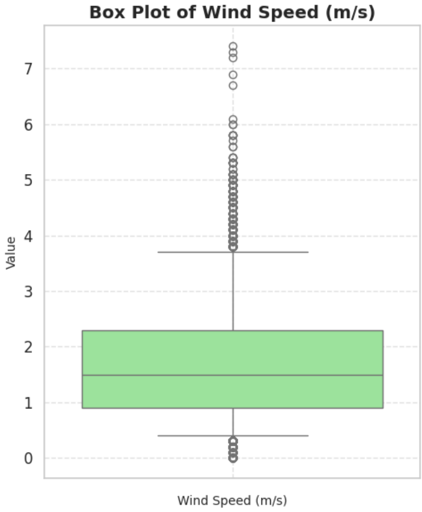


Figure 19: Outliers in Wind Speed

Visualizing outliers for Visibility

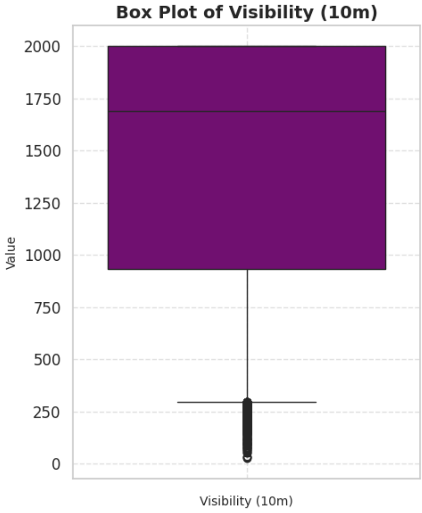


Figure 20: Outliers in Visibility

Visualizing outliers for Dew Point Temperature

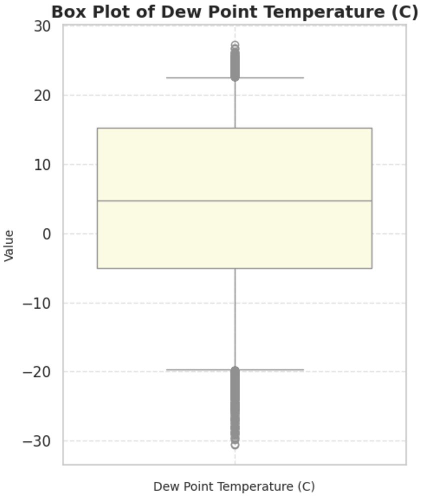


Figure 21: Outliers in Dew Point Temperature

Visualizing outliers for Solar Radiation

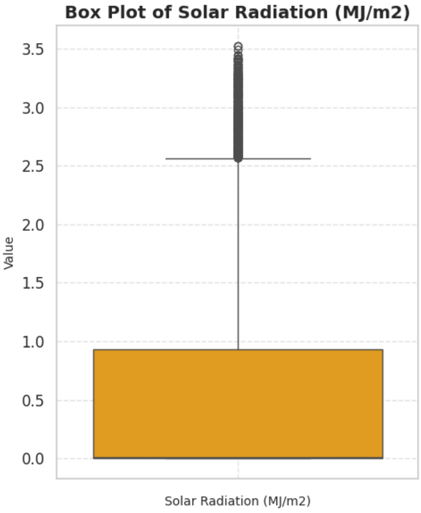


Figure 22: Outliers in Solar Radiation

Visualizing outliers for Rainfall

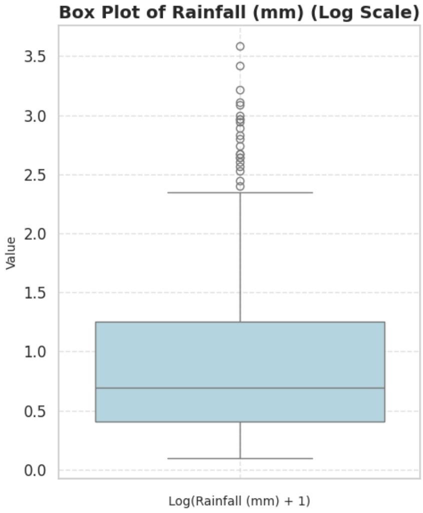


Figure 23: Outliers in Rainfall (Used log scale otherwise plot is not visible)

Visualizing outliers for Snowfall

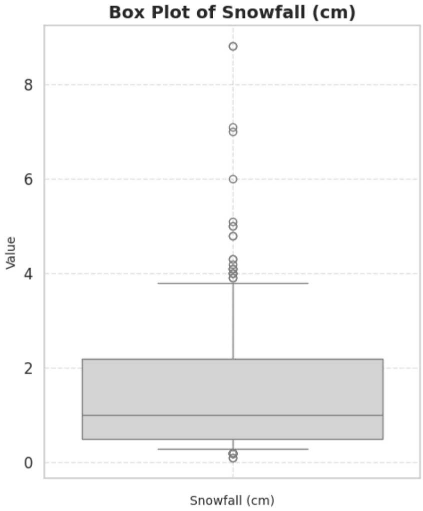


Figure 24: Outliers in Snowfall

5. We need to create regression plot between Rented Bike Count and other numerical features.

Regression plot between Rented Bike Count and Temperature

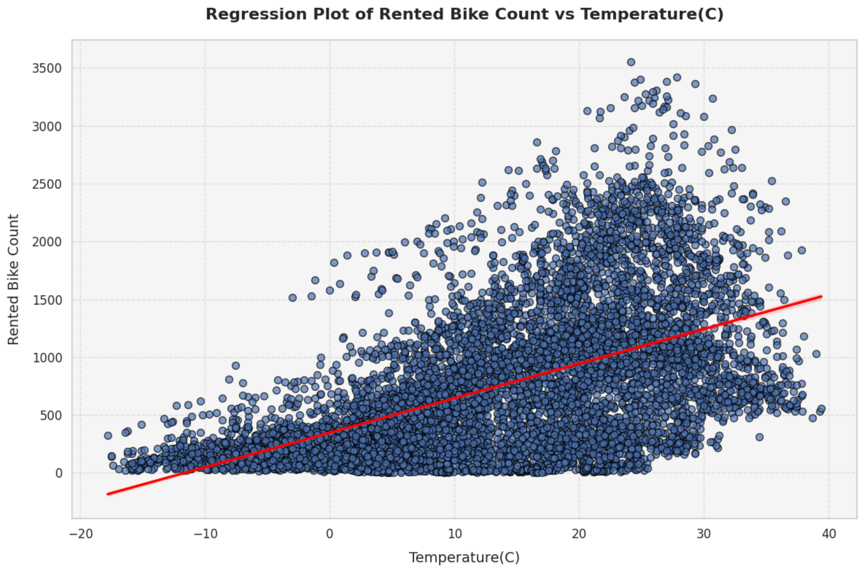


Figure 25: Regression plot between Rented Bike Count and Temperature

Regression plot between Rented Bike Count and Humidity

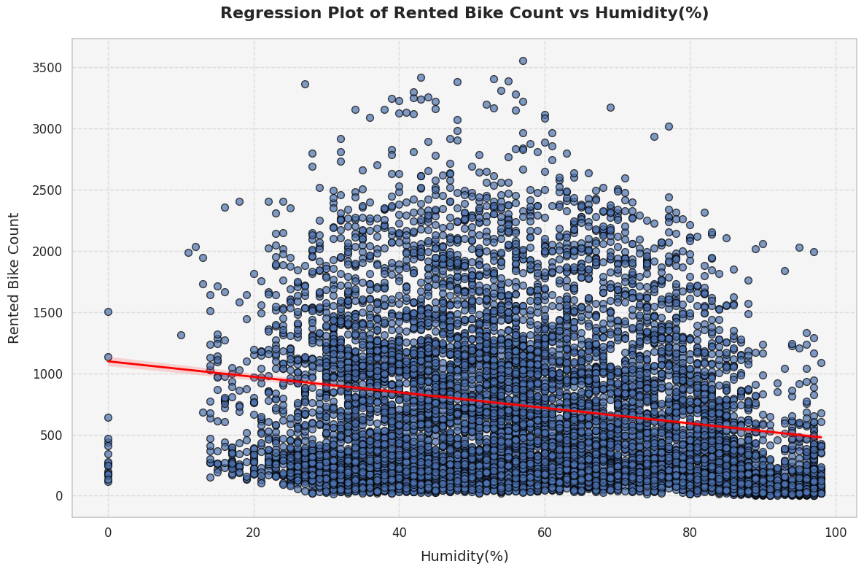


Figure 26: Regression plot between Rented Bike Count and Humidity

Regression plot between Rented Bike Count and Wind Speed

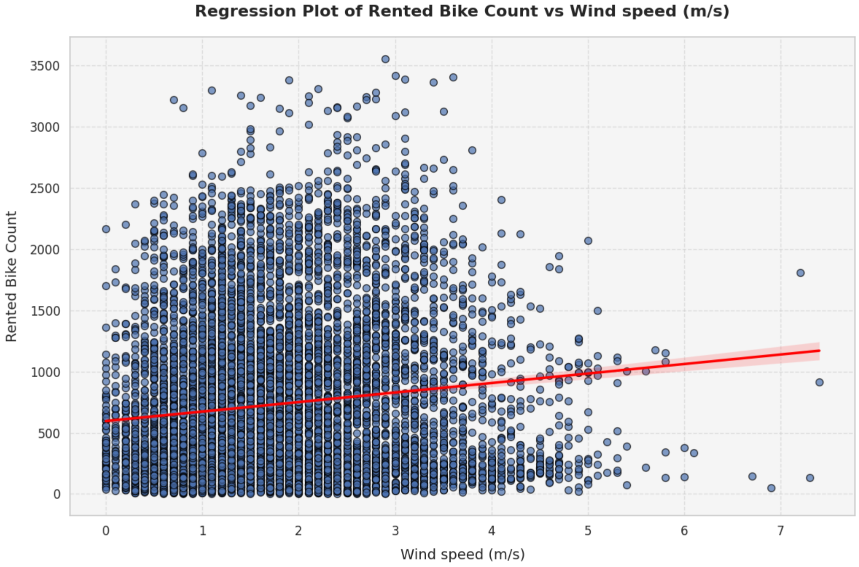


Figure 27: Regression plot between Rented Bike Count and Wind Speed

Regression plot between Rented Bike Count and Visibility

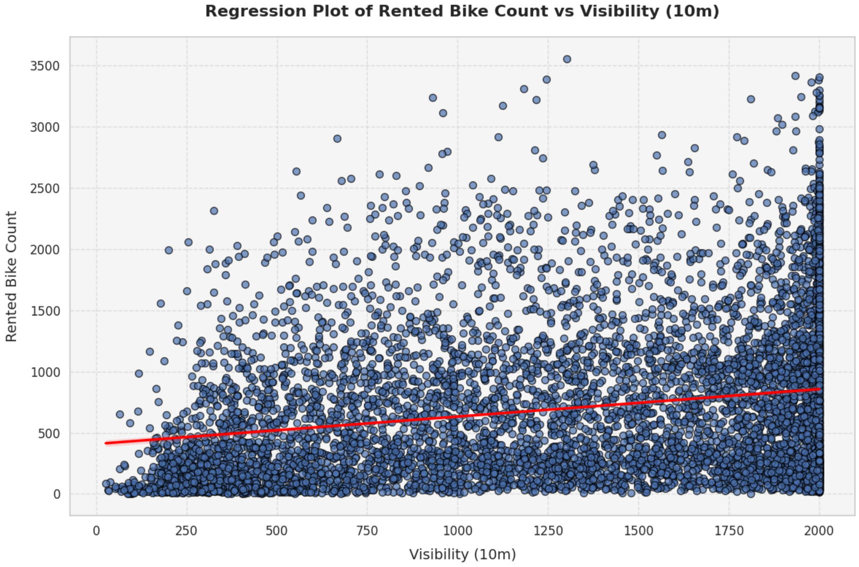


Figure 28: Regression plot between Rented Bike Count and Visibility

Regression plot between Rented Bike Count and Dew Point Temperature

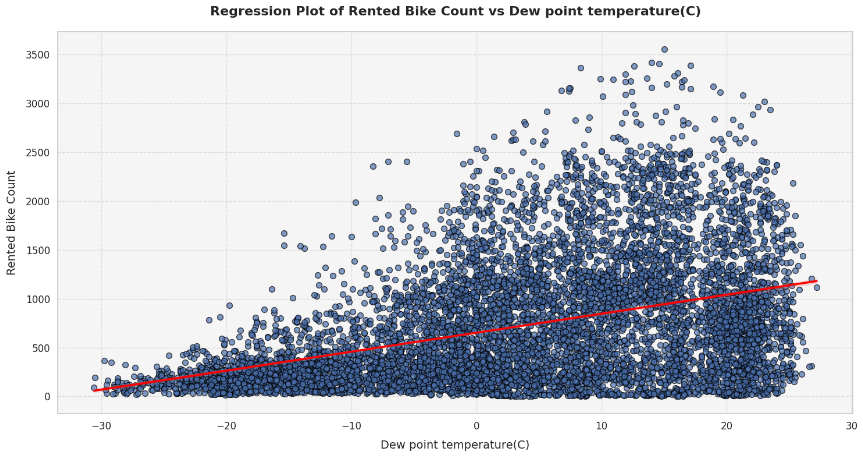


Figure 29: Regression plot between Rented Bike Count and Dew Point Temperature

Regression plot between Rented Bike Count and Solar Radiation

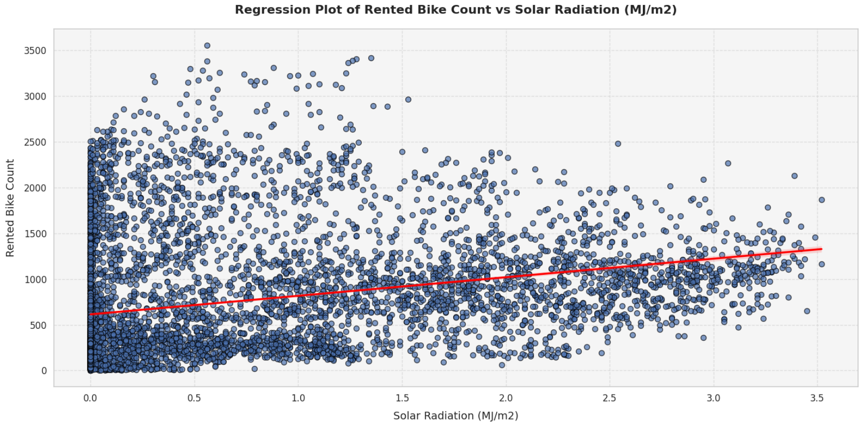


Figure 30: Regression plot between Rented Bike Count and Solar Radiation

Regression plot between Rented Bike Count and Rainfall

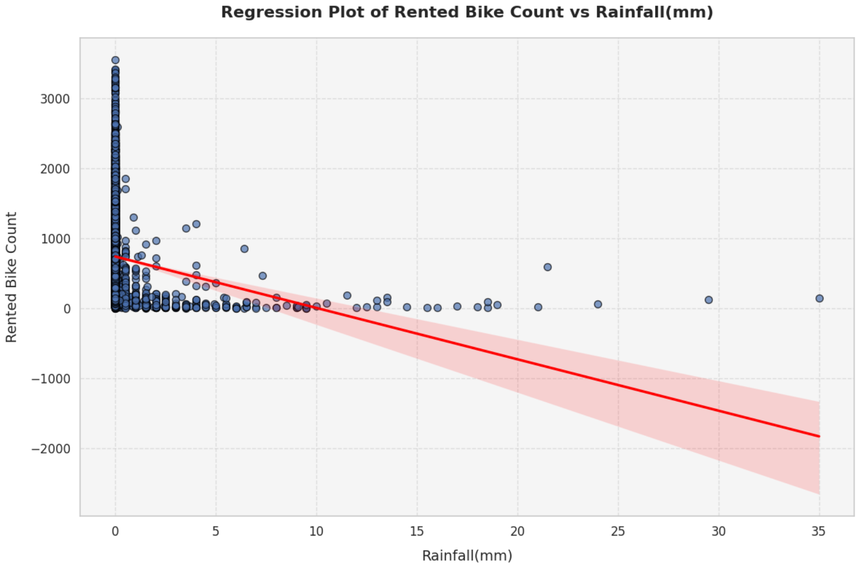


Figure 31: Regression plot between Rented Bike Count and Rainfall

Regression plot between Rented Bike Count and Snowfall

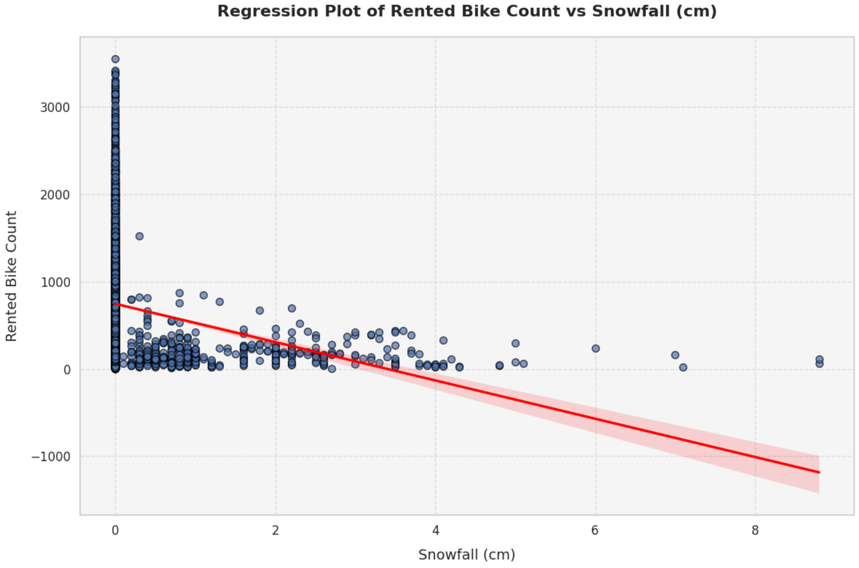


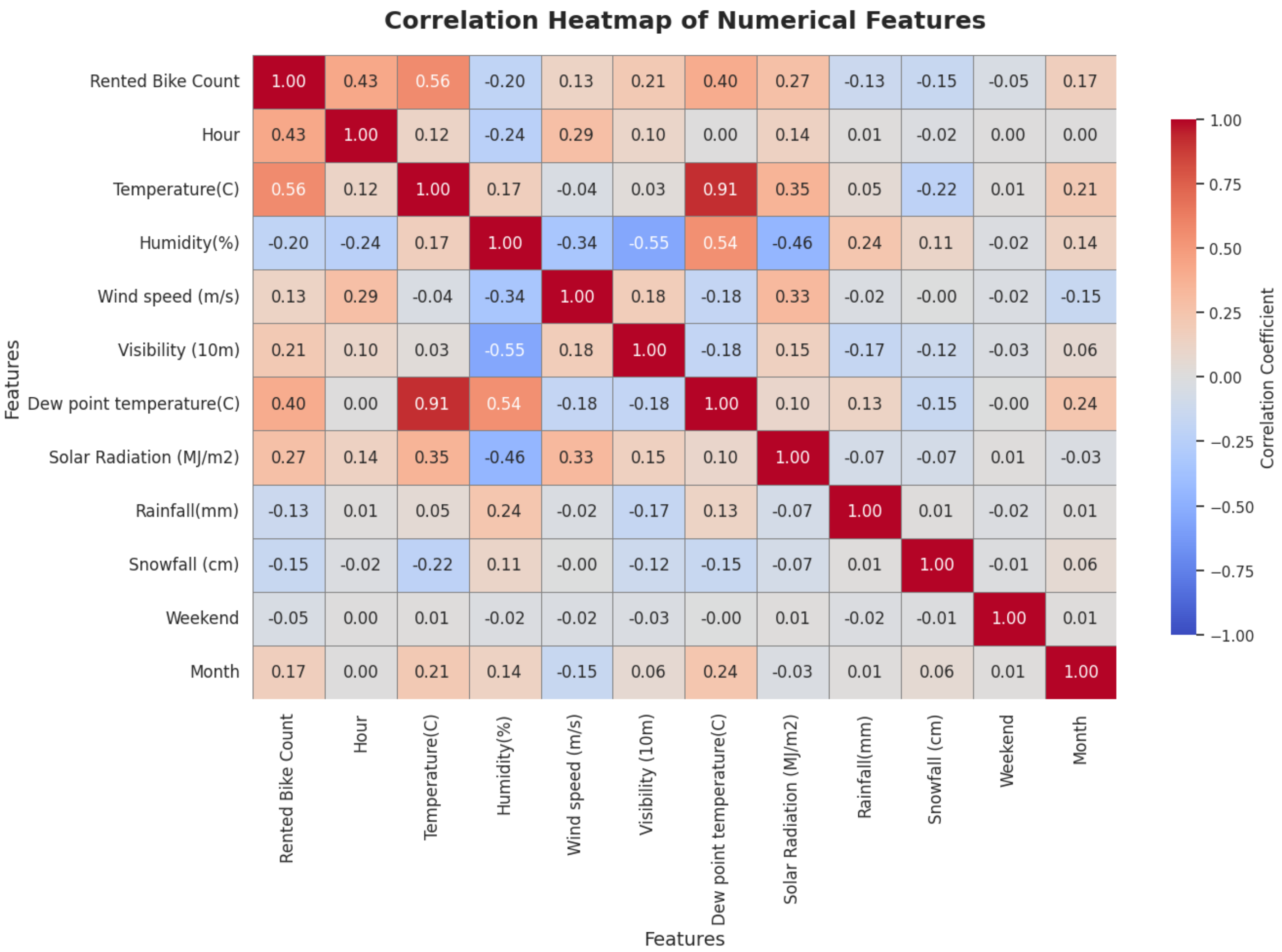
Figure 32: Regression plot between Rented Bike Count and Snowfall

Observations from regression plots

1. Most of the bike rentals are when rainfall and snowfall are zero, as a result we are getting such plots.
2. Similar case is there in case of solar radiation.
3. In case of visibility, we can see that bike rentals are more when visibility is high.
4. From plot it seems that wind speed is not affecting the bike rentals.
5. Humidity,Temperature, Dew Point Temperatureare the ones which are affecting the bike rentals the most.

6. Now we need to visualize the correlation heatmap between all numerical features.

Correlation Heatmap between all numerical features



Features	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)	Weekend	Month
Rented Bike Count	1.00	0.43	0.56	-0.20	0.13	0.21	0.40	0.27	-0.13	-0.15	-0.05	0.17
Hour	0.43	1.00	0.12	-0.24	0.29	0.10	0.00	0.14	0.01	-0.02	0.00	0.00
Temperature(C)	0.56	0.12	1.00	0.17	-0.04	0.03	0.91	0.35	0.05	-0.22	0.01	0.21
Humidity(%)	-0.20	-0.24	0.17	1.00	-0.34	-0.55	0.54	-0.46	0.24	0.11	-0.02	0.14
Wind speed (m/s)	0.13	0.29	-0.04	-0.34	1.00	0.18	-0.18	0.33	-0.02	-0.00	-0.02	-0.15
Visibility (10m)	0.21	0.10	0.03	-0.55	0.18	1.00	-0.18	0.15	-0.17	-0.12	-0.03	0.06
Dew point temperature(C)	0.40	0.00	0.91	0.54	-0.18	-0.18	1.00	0.10	0.13	-0.15	-0.00	0.24
Solar Radiation (MJ/m2)	0.27	0.14	0.35	-0.46	0.33	0.15	0.10	1.00	-0.07	-0.07	0.01	-0.03
Rainfall(mm)	-0.13	0.01	0.05	0.24	-0.02	-0.17	0.13	-0.07	1.00	0.01	-0.02	0.01
Snowfall (cm)	-0.15	-0.02	-0.22	0.11	-0.00	-0.12	-0.15	-0.07	0.01	1.00	-0.01	0.06
Weekend	-0.05	0.00	0.01	-0.02	-0.02	-0.03	-0.00	0.01	-0.02	-0.01	1.00	0.01
Month	0.17	0.00	0.21	0.14	-0.15	0.06	0.24	-0.03	0.01	0.06	0.01	1.00

Figure 33: Correlation Heatmap between all numerical features

Observations

1. From the heatmap, we can see that there is a strong positive correlation between temperature and Dew Point Temperature.
2. There is a good negative correlation between visibility and humidity.
3. We will remove dew point temperature because it is highly correlated with temperature.

7. Now we need to encoding of categorical features.

1. I have done one hot encoding for seasons and removed the season feature, however 0123 encoding would have been better.

2. I have done 0-1 encoding for holiday and removed the holiday feature.

3. My months were already marked as 1-12 so no need to encode that.

8. I need to delete non relevant features from the dataframe and comment on it.