Capstone Project - 2

Supervised machine learning(regression)- Bike sharing demand prediction

Mohammed arifuddin atif

All about this presentation:

1. Defining problem statement.
2. Overview of data.
3. Performing exploratory data analysis.
4. Model preparation.
5. Building different models.
6. Evaluation of all models.
7. Extracting the best model.

Problem statement

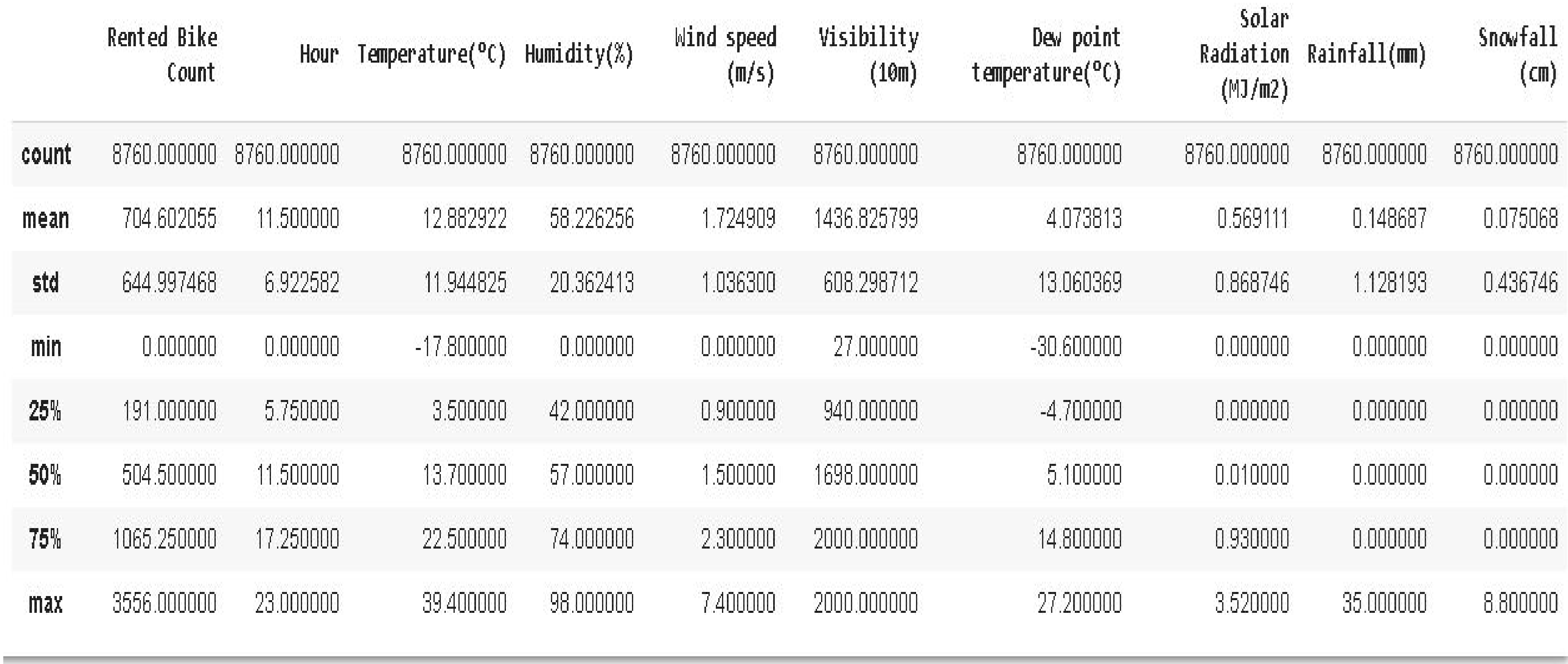
We are tasked with predicting the number of bikes rented each hour so as to make an approximate estimation of the number of bikes to be made available to the public given a particular hour of the day.

Overview of given data

We are given the following columns in our data:

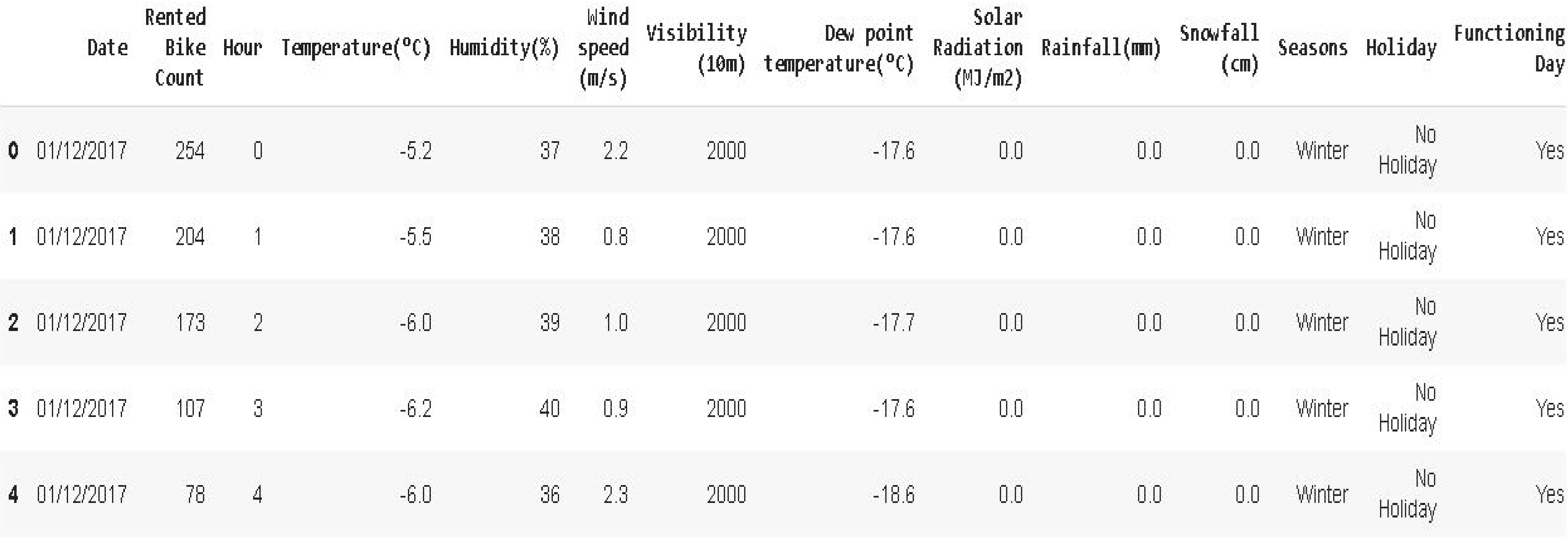
1. **Date : year-month-day**
2. **Rented Bike count - Count of bikes rented at each hour**
3. **Hour - Hour of he day**
4. **Temperature-Temperature in Celsius**
5. **Humidity - %**
6. **Wind Speed - m/s**
7. **Visibility - 10m**
8. **Dew point temperature - Celsius**
9. **Solar radiation - MJ/m2**
10. **Rainfall - mm**
11. **Snowfall - cm**
12. **Seasons - Winter, Spring, Summer, Autumn**
13. **Holiday - Holiday/No holiday**
14. **Functional Day - No(Non Functional Hours), Yes(Functional hours)**

Description of data



	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)
count	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000
mean	704.602055	11.500000	12.882922	58.226256	1.724909	1436.825799	4.073813	0.569111	0.148687	0.075068
std	644.997468	6.922582	11.944825	20.362413	1.036300	608.298712	13.060369	0.868746	1.128193	0.436746
min	0.000000	0.000000	-17.800000	0.000000	0.000000	27.000000	-30.600000	0.000000	0.000000	0.000000
25%	191.000000	5.750000	3.500000	42.000000	0.900000	940.000000	-4.700000	0.000000	0.000000	0.000000
50%	504.500000	11.500000	13.700000	57.000000	1.500000	1698.000000	5.100000	0.010000	0.000000	0.000000
75%	1065.250000	17.250000	22.500000	74.000000	2.300000	2000.000000	14.800000	0.930000	0.000000	0.000000
max	3556.000000	23.000000	39.400000	98.000000	7.400000	2000.000000	27.200000	3.520000	35.000000	8.800000

Sample of data

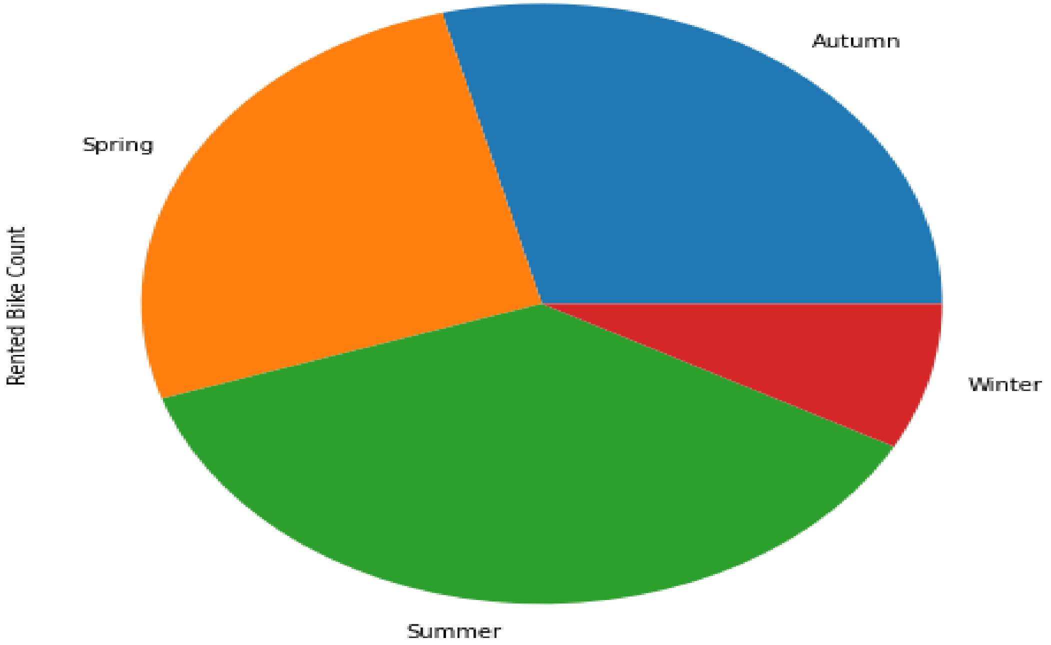


	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

Exploratory data analysis

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.
- EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

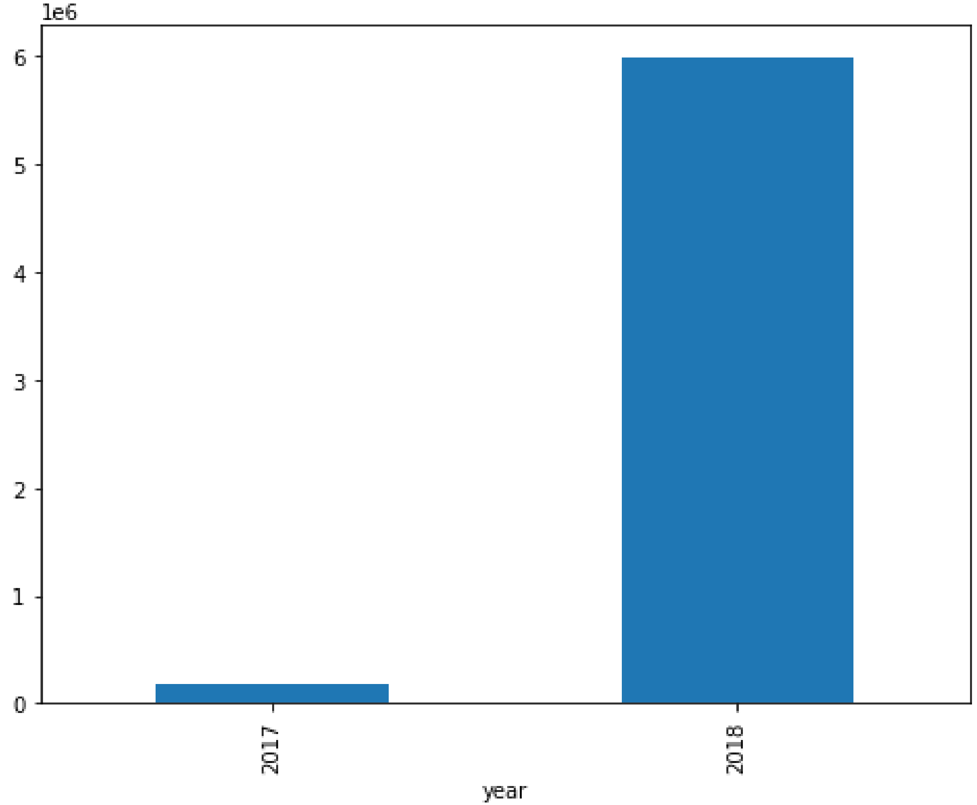
Comparison of bikes rented seasonally



conclusions from above pie chart:

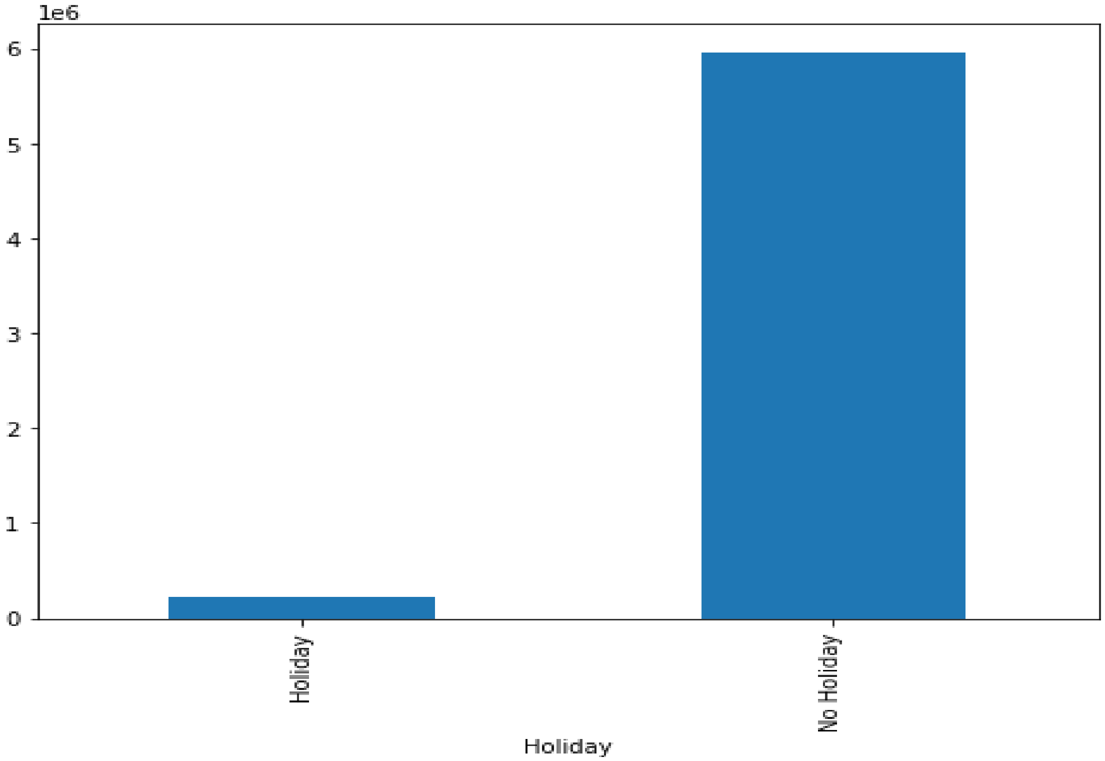
1. most bikes have been rented in the summer season.
2. least bike rent count is in winter season.
3. autumn and spring seasons have almost equal amounts of bike rent count.

Comparison of number of bikes rented (year)



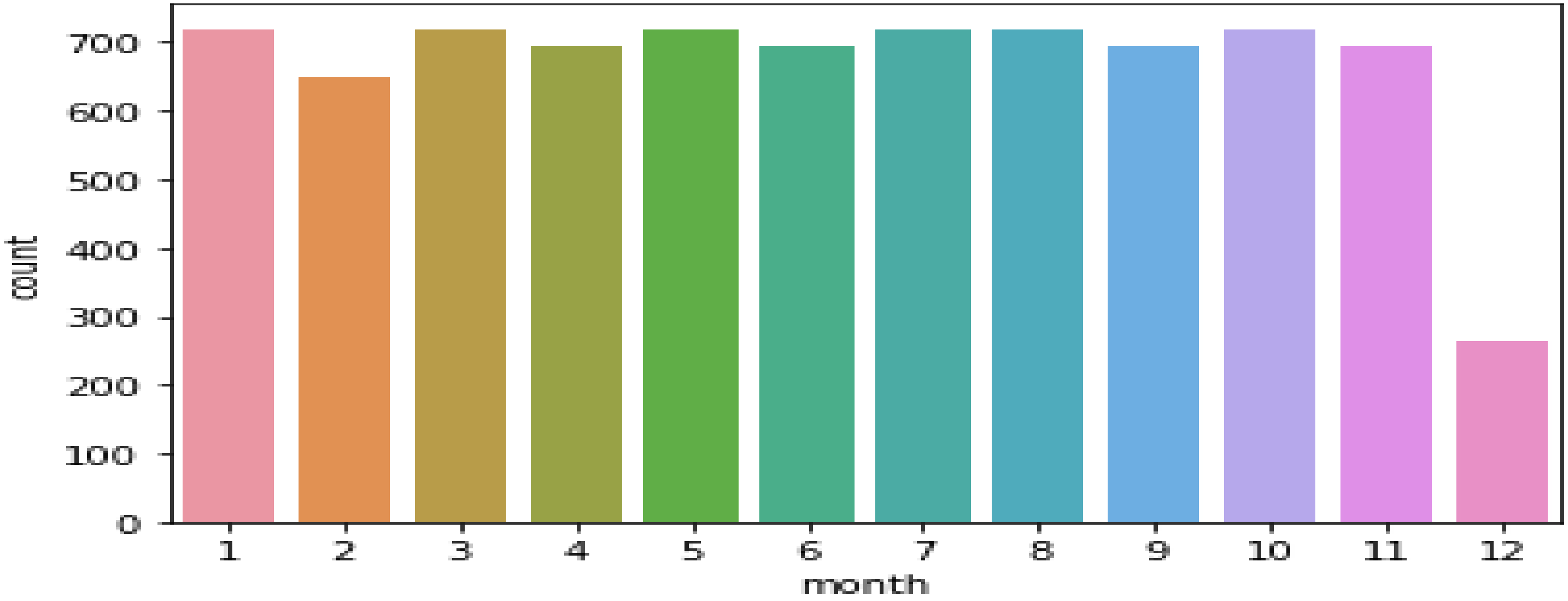
Above plot shows that most of the bikes have been rented in the year 2018.

Comparison of number of bikes rented (type of day)



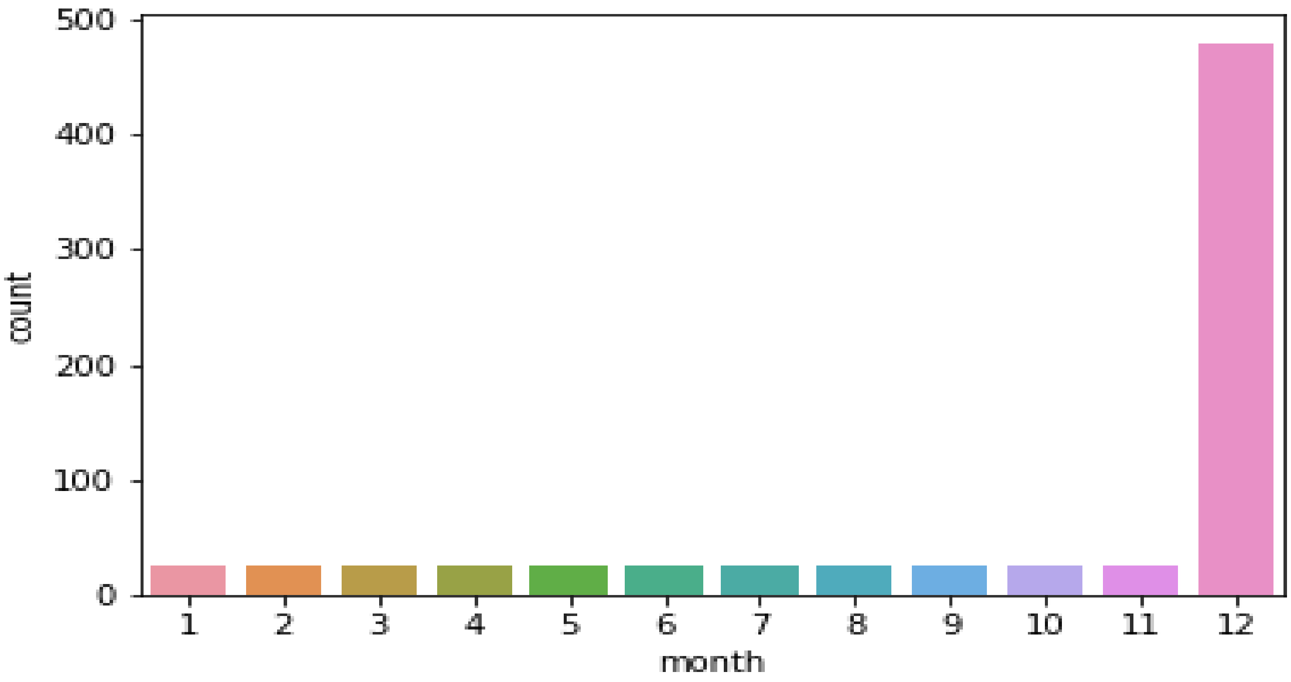
Above plot shows that most of the bikes have been rented on working days.

Comparison of number of bikes rented in year 2018



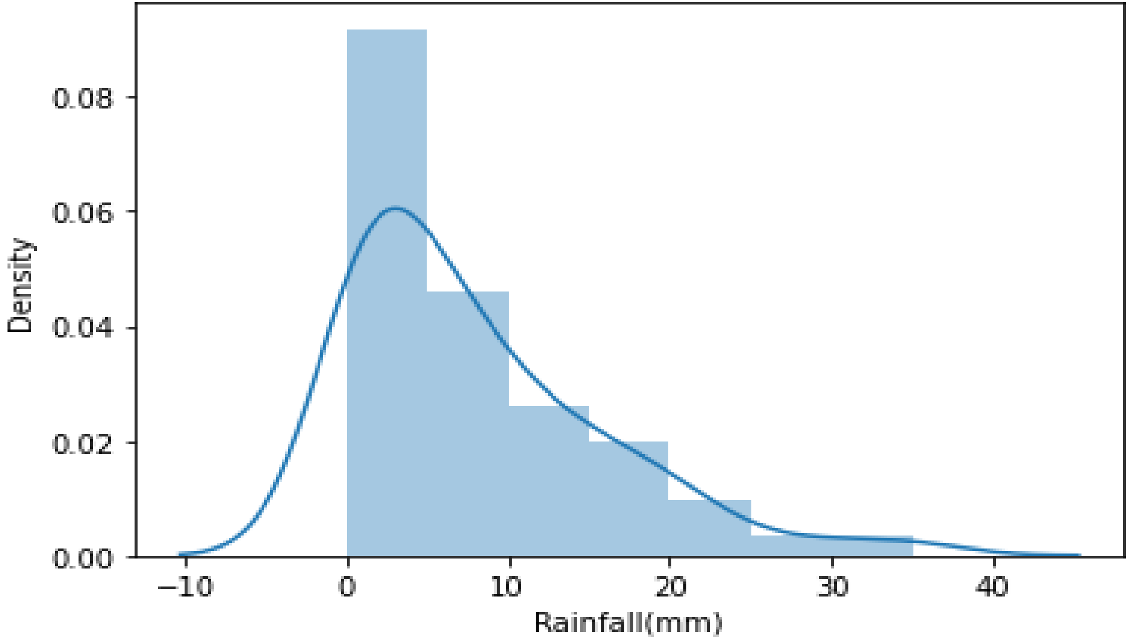
Above plot shows that most of the bikes have been rented in december (winter).

Comparison of number of bikes rented in year 2017



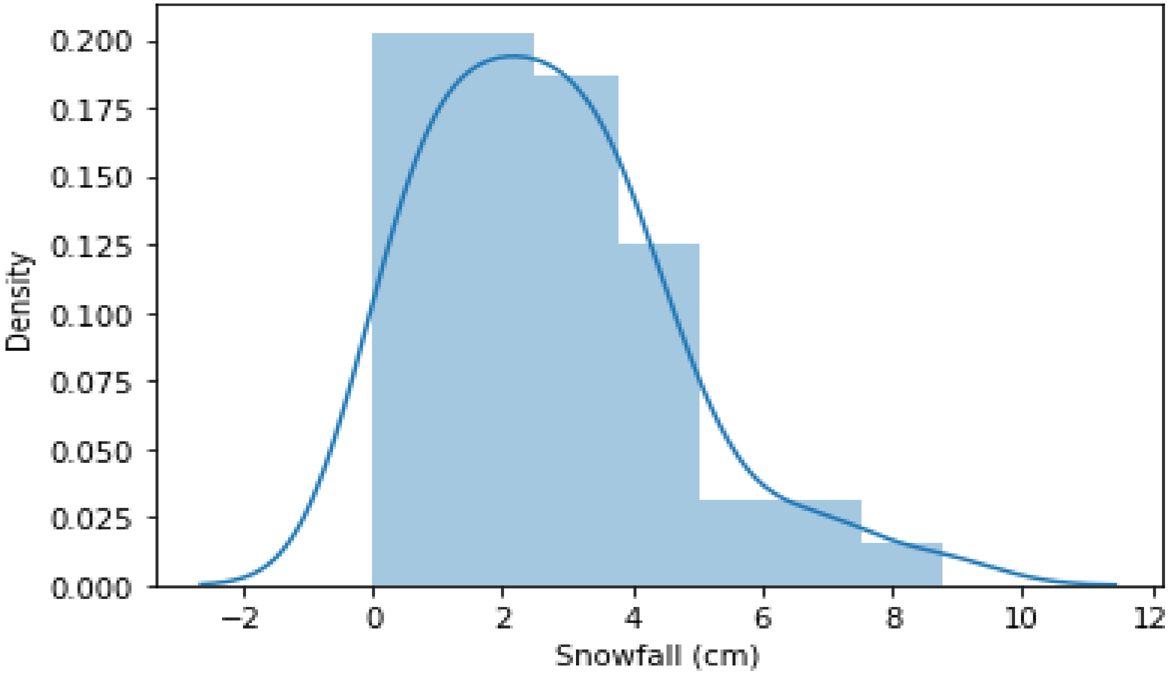
Above plot shows that most of the bikes have been rented in december in the year 2017.

Distribution of bike rentals according to rainfall intensity



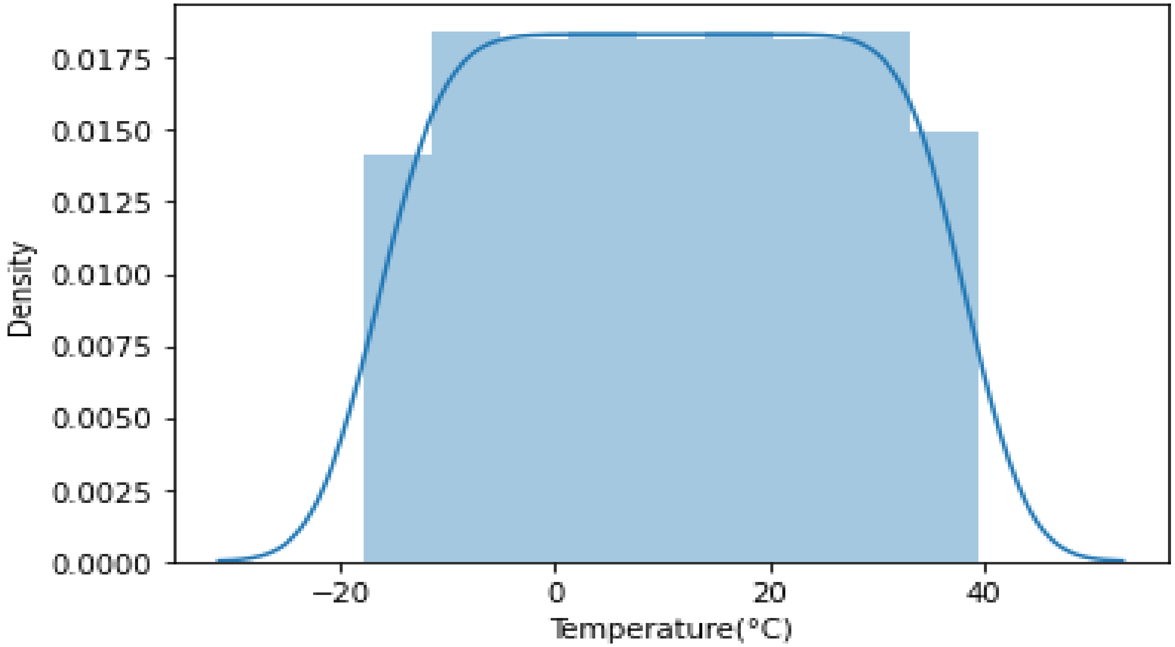
Above plot shows that most people tend to rent bikes when there is no or less rainfall.

Distribution of bike rentals according to snowfall intensity



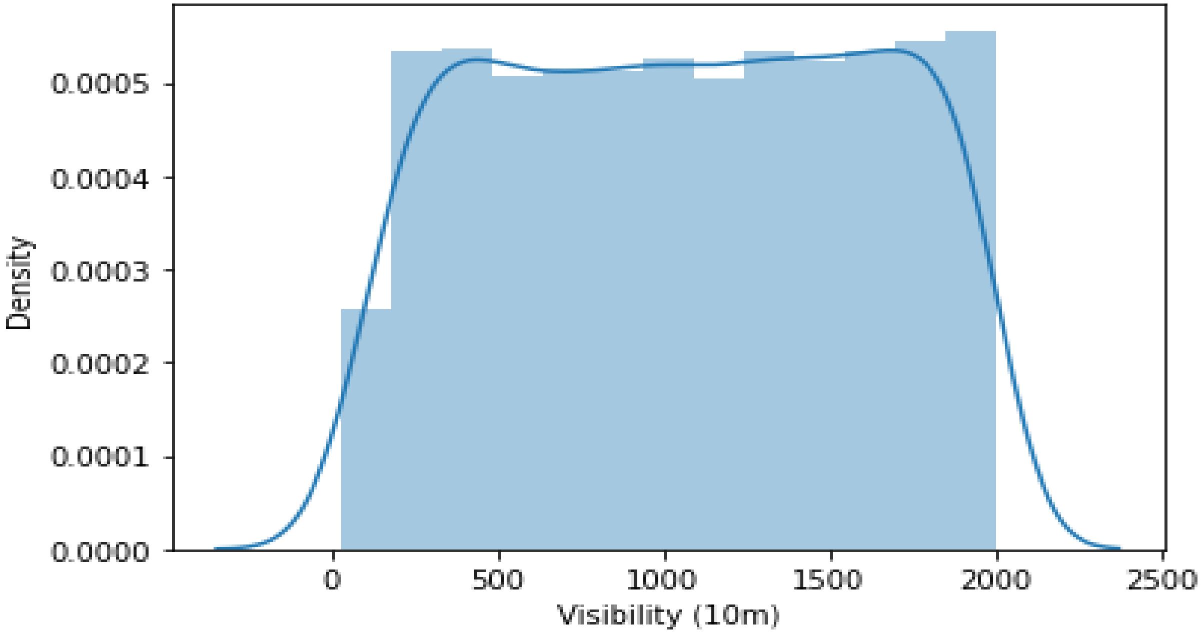
Above plot shows that most people tend to rent bikes when there is no or less snowfall.

Distribution of bike rentals according to temperature intensity



Above plot shows that most people tend to rent bikes when temperature is between -5 t0 25 degrees.

Distribution of bike rentals according to visibility



Above plot shows that most people tend to rent bikes when visibility is between 300 t0 1700.

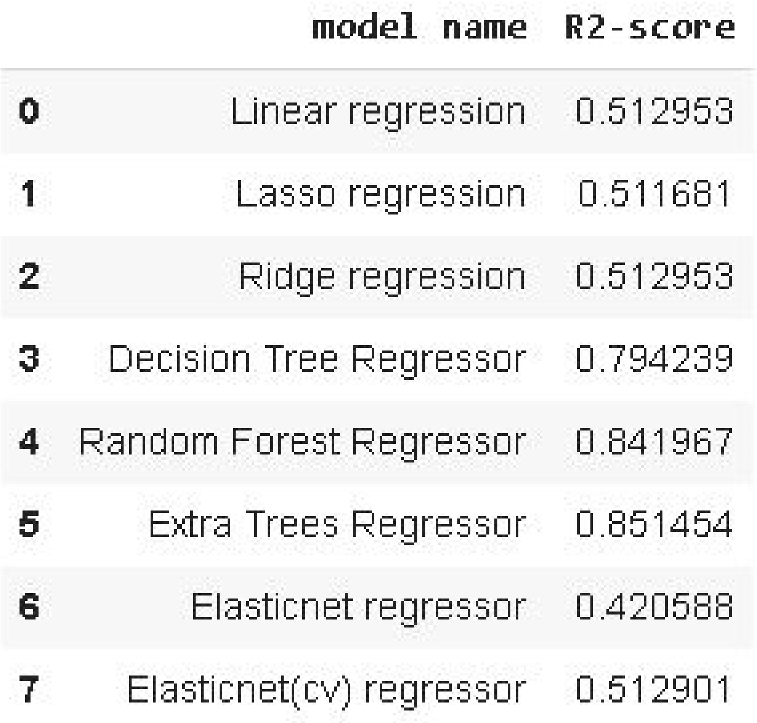
Model preparation

1. Plotting the correlation heatmap and removing variables which are highly correlated.
2. Calculating multicollinearity through VIF and filtering our data.
3. Converting data types of variables into relevant data types.
4. Filling the null values in our data with mean of particular values.

Models used

- Linear regression model
- Lasso regression model
- Ridge regression model
- Decision tree regression model
- Random-forest regression model
- Extra-trees regression model
- Elastic net regression model

Evaluation of models



	model name	R2-score
0	Linear regression	0.512953
1	Lasso regression	0.511681
2	Ridge regression	0.512953
3	Decision Tree Regressor	0.794239
4	Random Forest Regressor	0.841967
5	Extra Trees Regressor	0.851454
6	Elasticnet regressor	0.420588
7	Elasticnet(cv) regressor	0.512901

From above it is clear that extra-trees regression model has done very well with our dataset

Challenges faced

1. Pre-processing the data was one of the challenges we faced which includes removing highly correlated variables from the data so as to not hinder the performance of our regression model.
2. Exploring all the columns and calculating VIF for multicollinearity was challenging because it might decrease the models performance.
3. Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.

Conclusion

We are finally at the conclusion of our project!

Coming from the beginning we did EDA on the dataset and also cleaned the data according to our needs.After that we were able to draw relevant conclusions from the given data and then we trained our model on linear regression and other models .

Out of all models used , with extra-trees regression model we were able to get the r2-score of 0.85.The model which performed poorly was elastic net regularization with r2-score of 0.42.

Given the size of data and the amount of irrelevance in the data , the above score is good.