CAPSTONE PROJECT - II

Bike sharing demand prediction

by

Nakshith D N

Contents

- Introduction
- Problem statement
- Dataset
- Feature engineering
- Data visualization
- Plots between dependent and independent variables
- Model building- Linear regression, Decision tree, Random forest.
- Conclusion

Introduction

Currently rental bikes are increasing throughout the world and especially in fast growing cities, it helps individuals to Move easily without waiting for any public transport and it helps reduce dependency on public transport and reduces pollution.

Problem Statement

The objective of the project is to perform an exploratory data analysis, data pre-processing, data cleaning & imputation, and in the end, apply different Data Visualization techniques to get meaningful insights from the given data and predict the results.

Also to explore and analyze the data to discover the following key understandings.

Here the crucial part is to predict the bike count which is required at each hour for stable supply of rental bikes.

Dataset

This dataset has 14 columns and it is a mix of categorical and numeric values.

The features of the dataset are

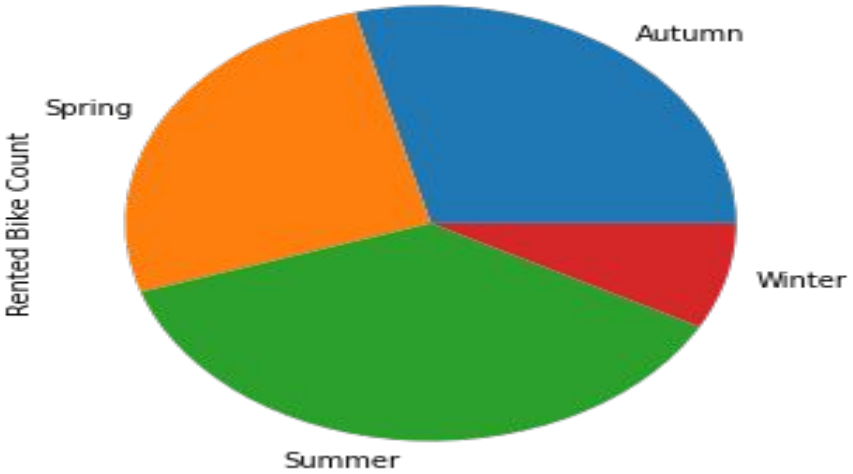
‘Date' - Indicates the date of specific year
'Rented Bike Count' - Total rented bike counts
'Hour' - Total hours
'Temperature(°C)'
'Humidity(%)'
'Wind speed (m/s)'
'Visibility (10m)'
'Dew point temperature(°C)'

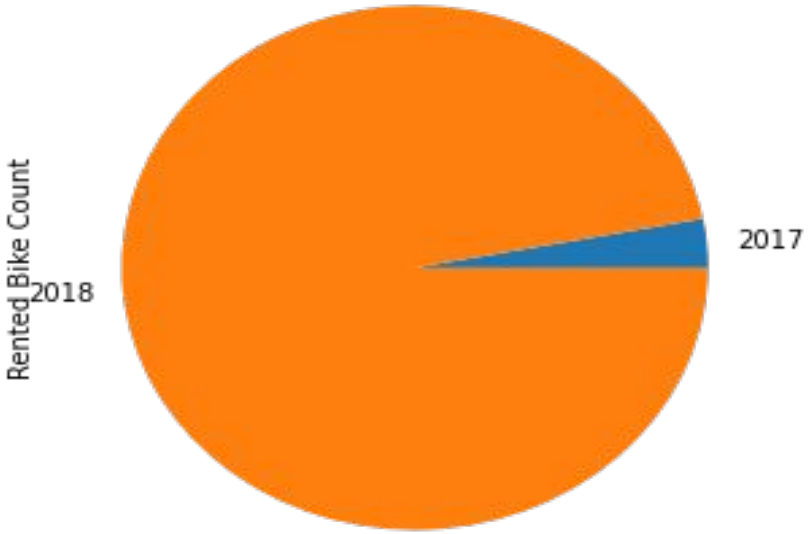
'Solar Radiation (MJ/m2)'
'Rainfall(mm)'
'Snowfall (cm)'
'Seasons'
'Holiday'
'Functioning Day',

Feature engineering and Data visualization

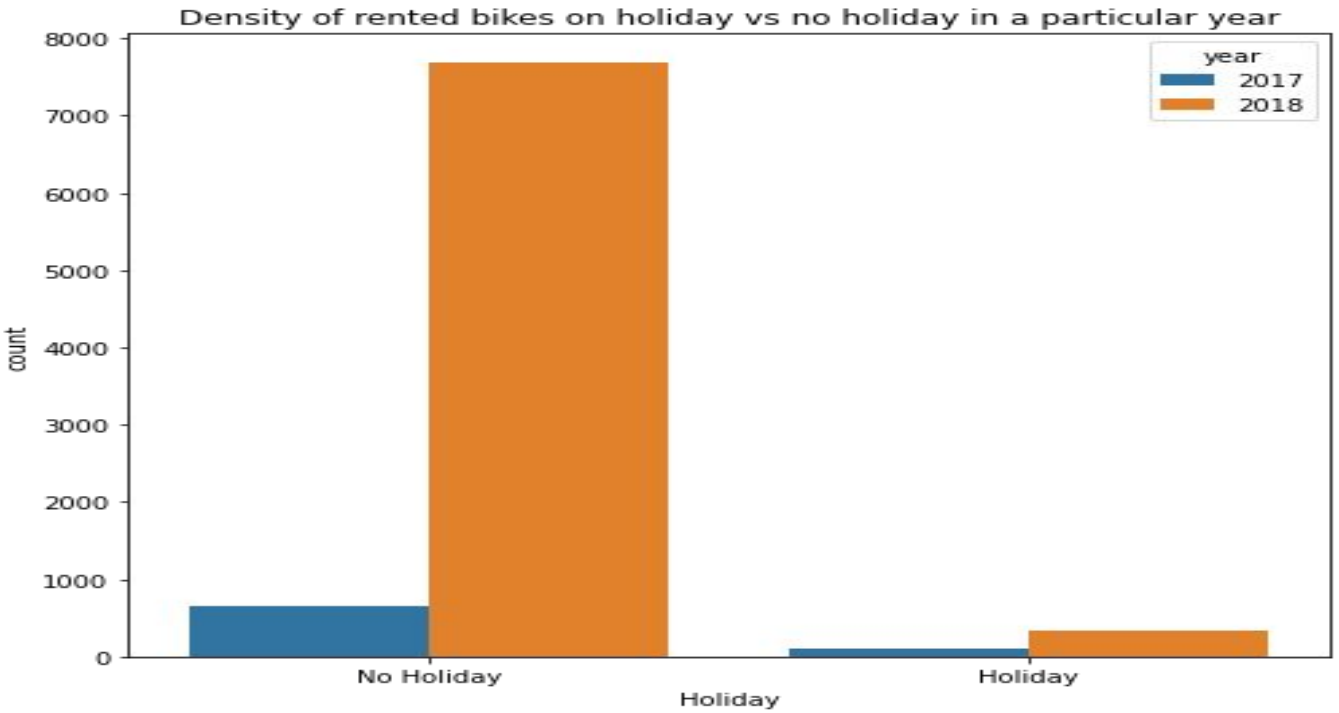
This shows the total count of rented bikes throughout all seasons in a pie chart.
Most bikes have been rented in summer,
Whereas during winter least number of bikes have been rented.

Around 1790002 in autumn
1611909 in spring
2283243 in summer
487169 in winter.

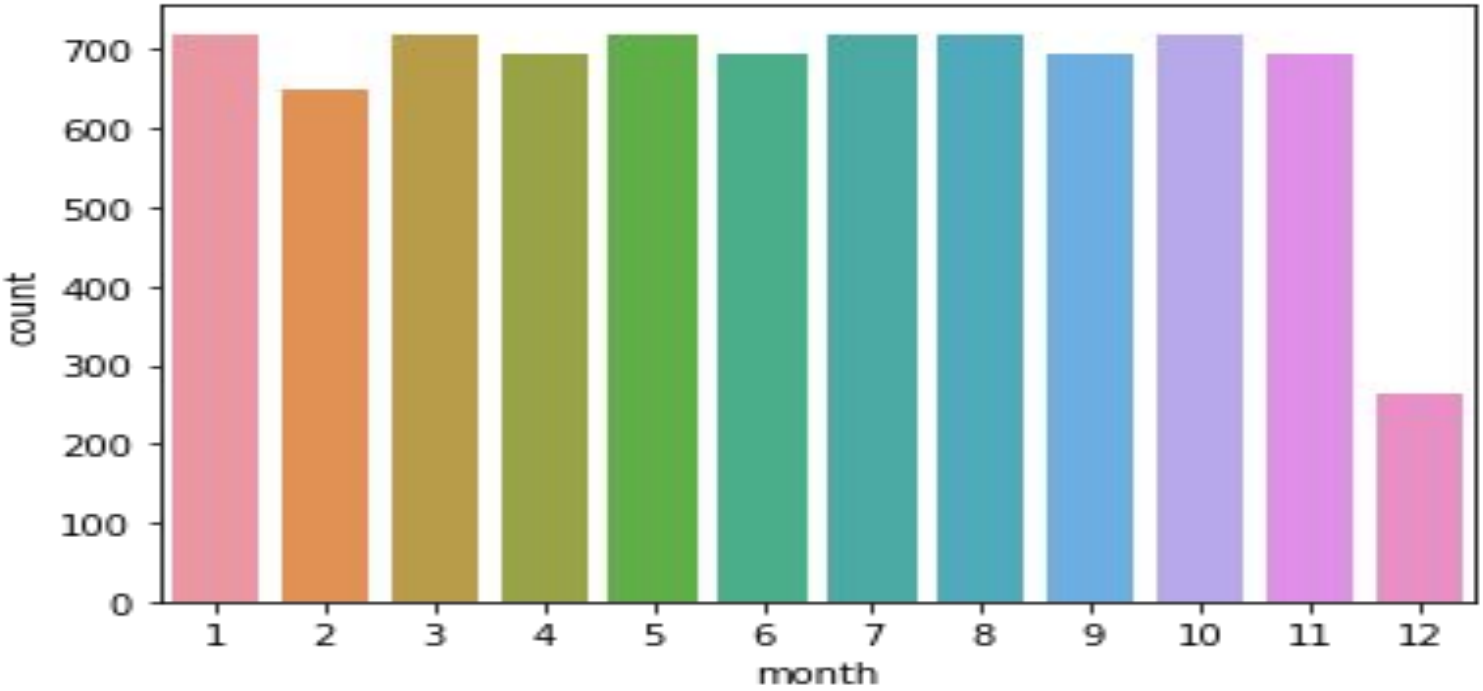




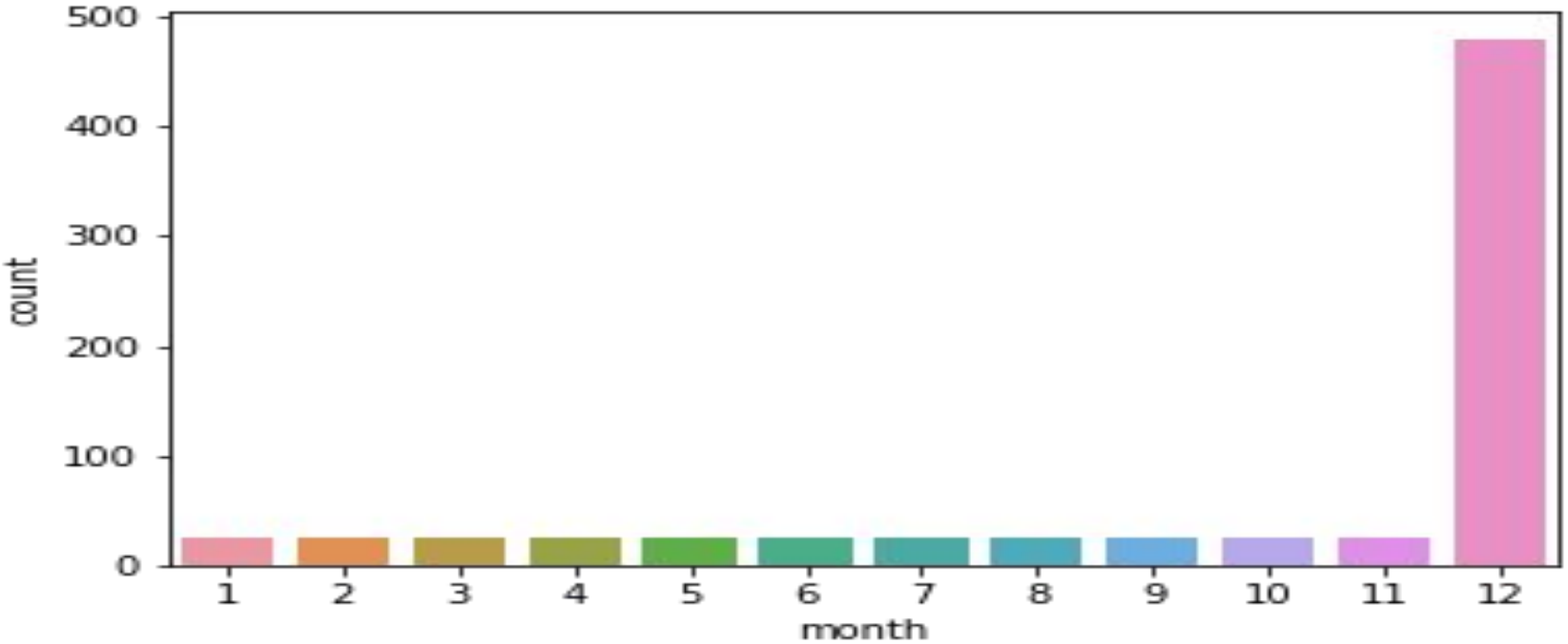
Total number of bikes rented each year. In 2017 around 185330 and in 2018 around 5986984.



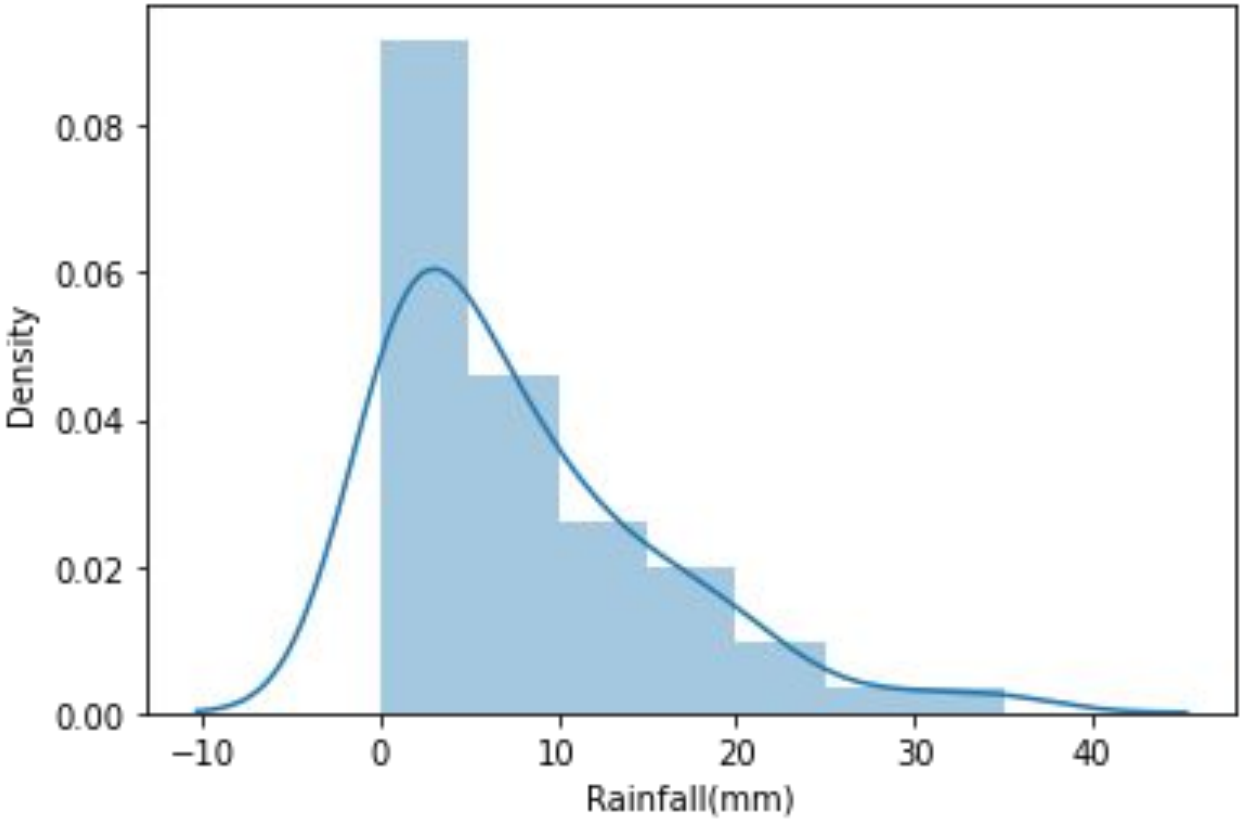
This shows the density of bikes rented on holidays and no holidays in particular year



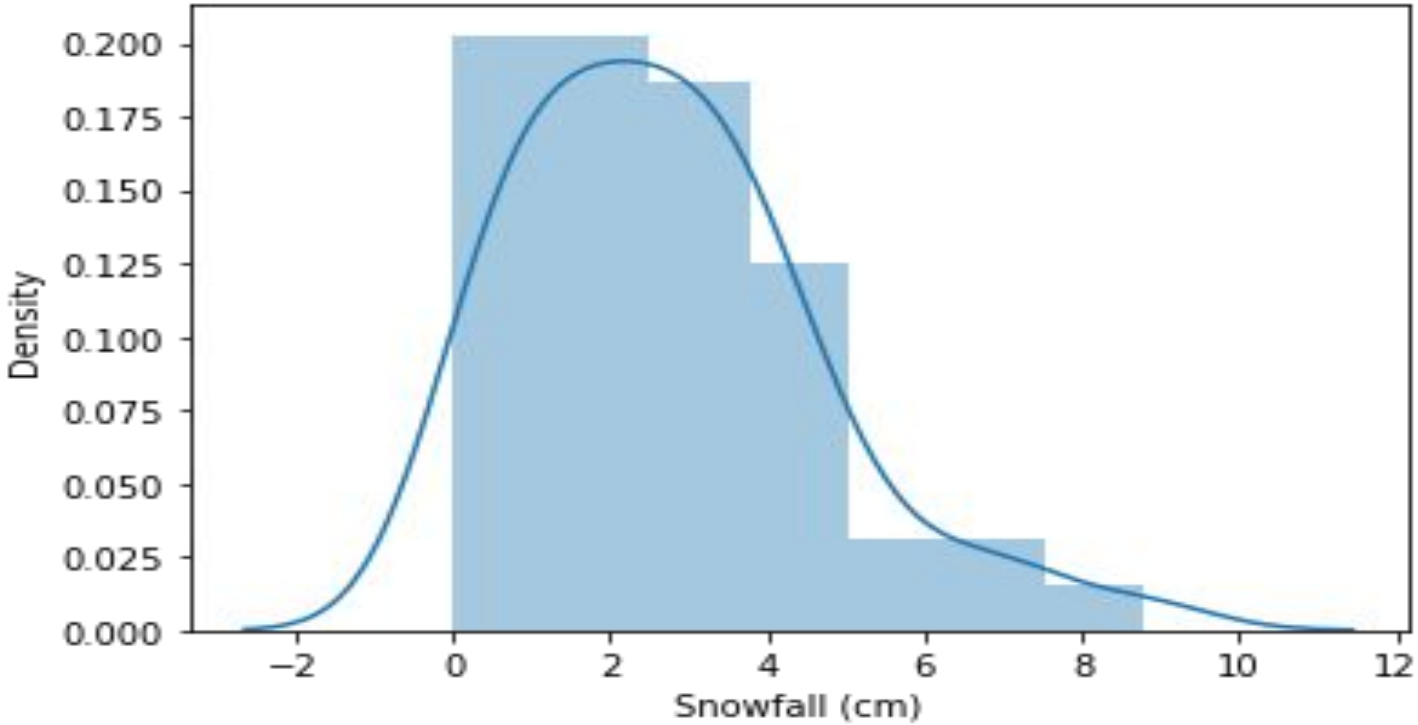
This shows the bikes rented in different months in the year 2018



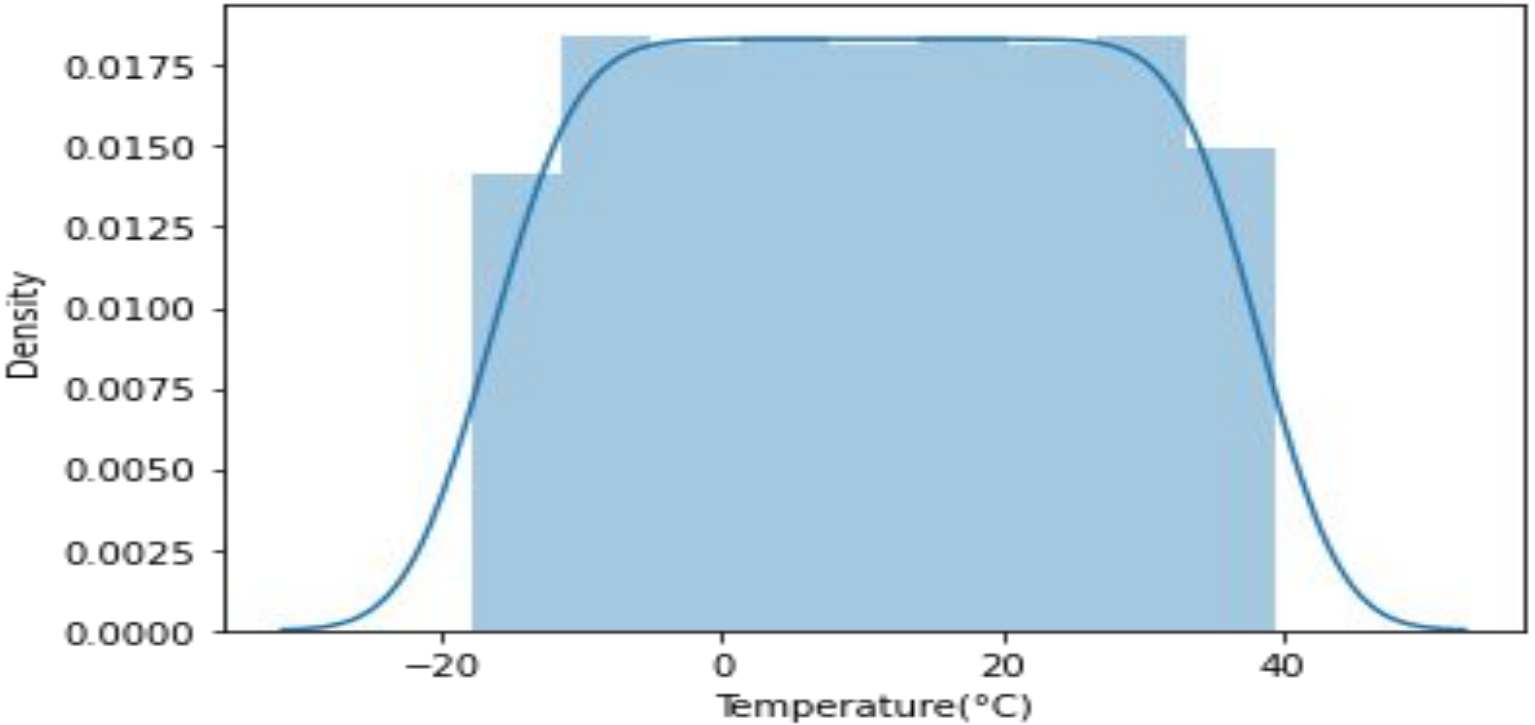
This shows the bikes rented in different months in the year 2017.



This plot shows the bike rentals in accordance to rainfall intensity.



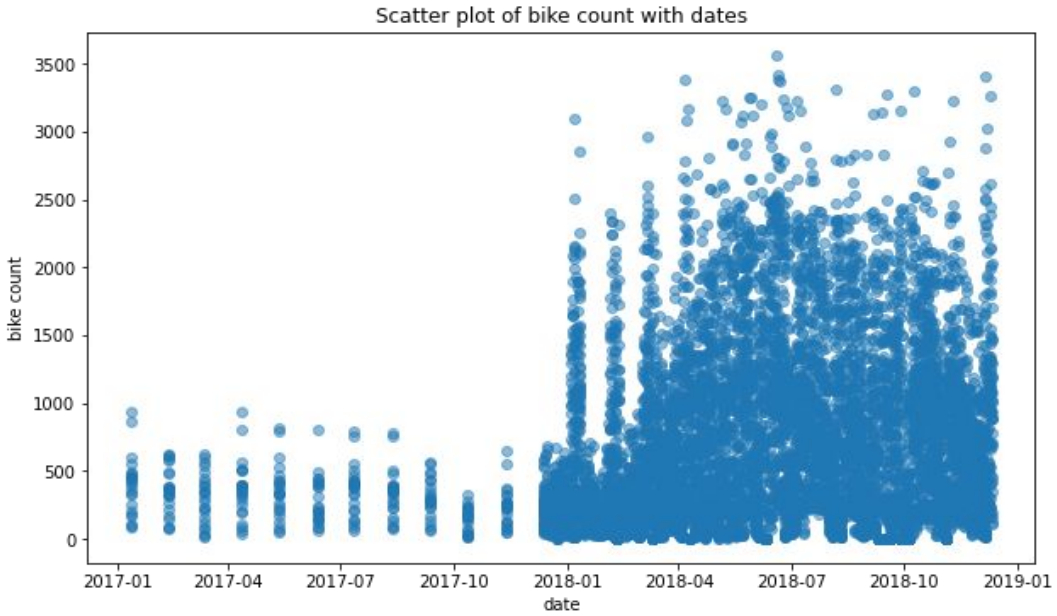
This plot shows the bike rentals in accordance to snowfall intensity.

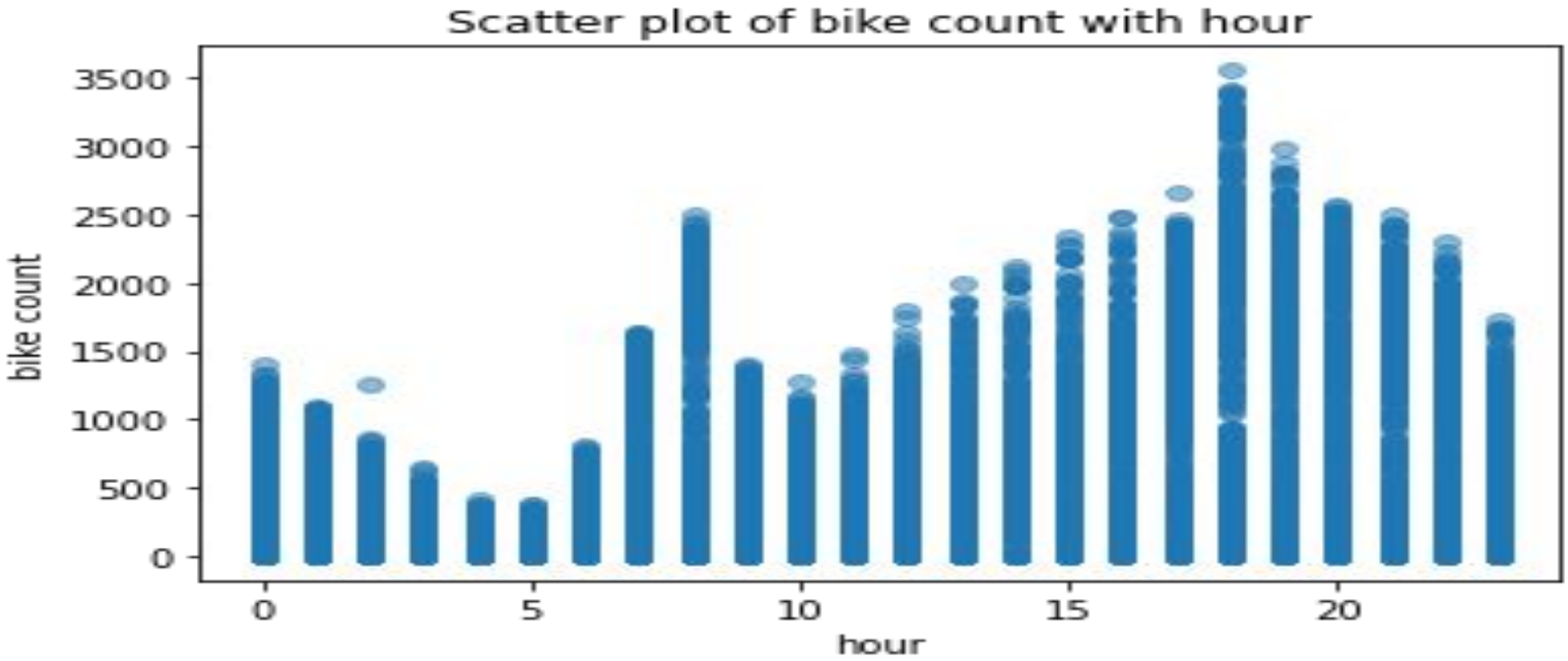


This plot shows the temperature variations in accordance to bike rentals.

Scatter plots.

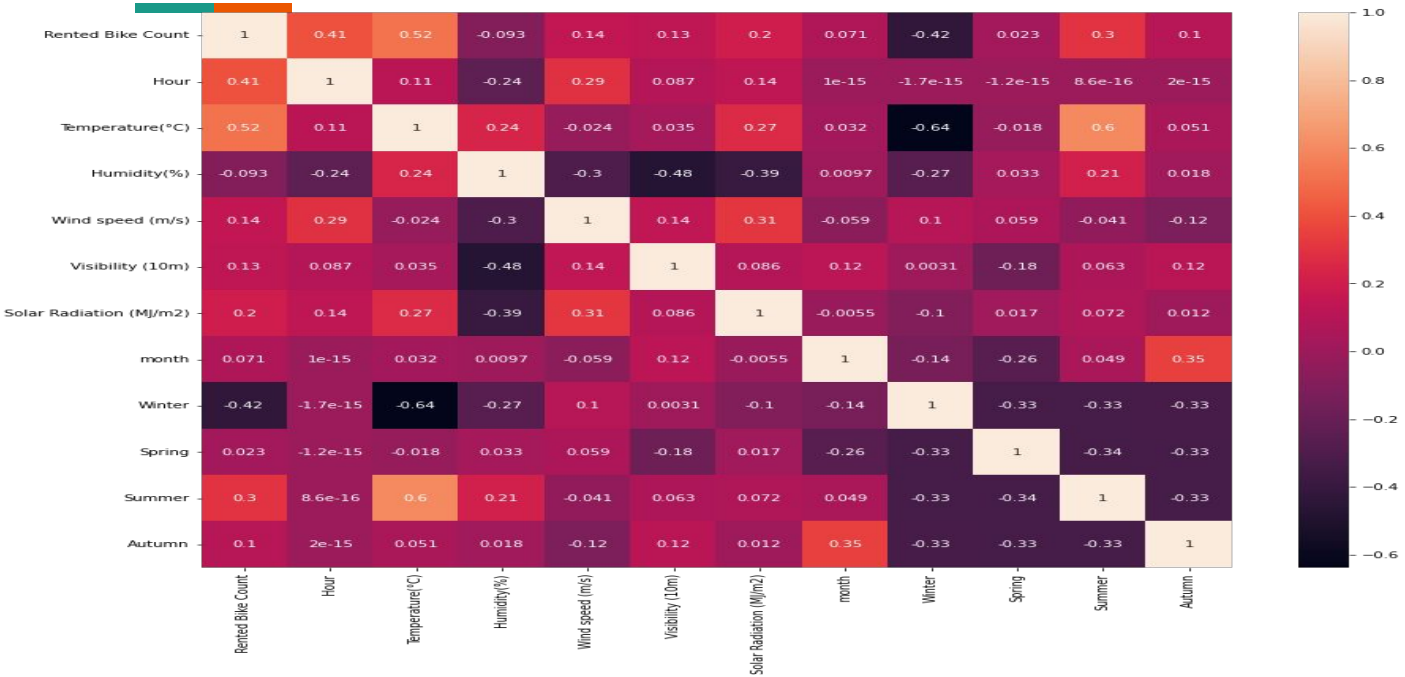
Scatter plot of bike counts on different dates.





From this plot we can see that rentals were more in morning and evening, this is specifically due to People who do not have own vehicles rent bikes to go to work or to go home.

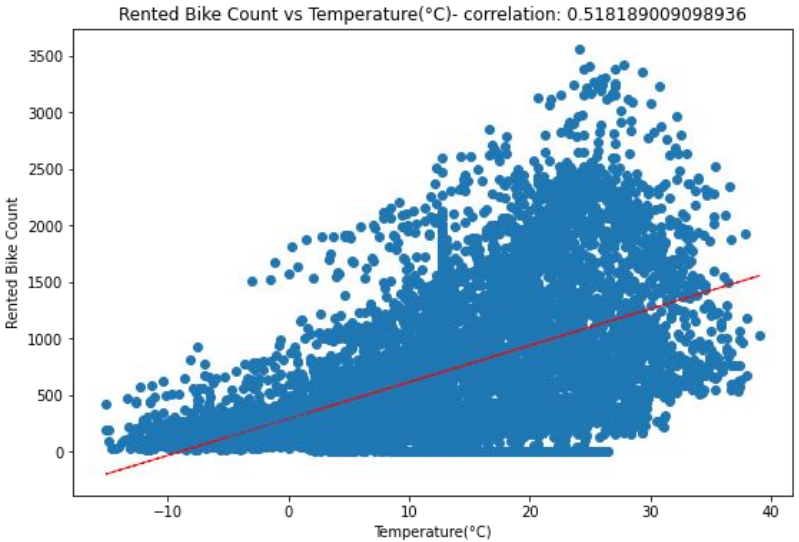
Correlation Analysis

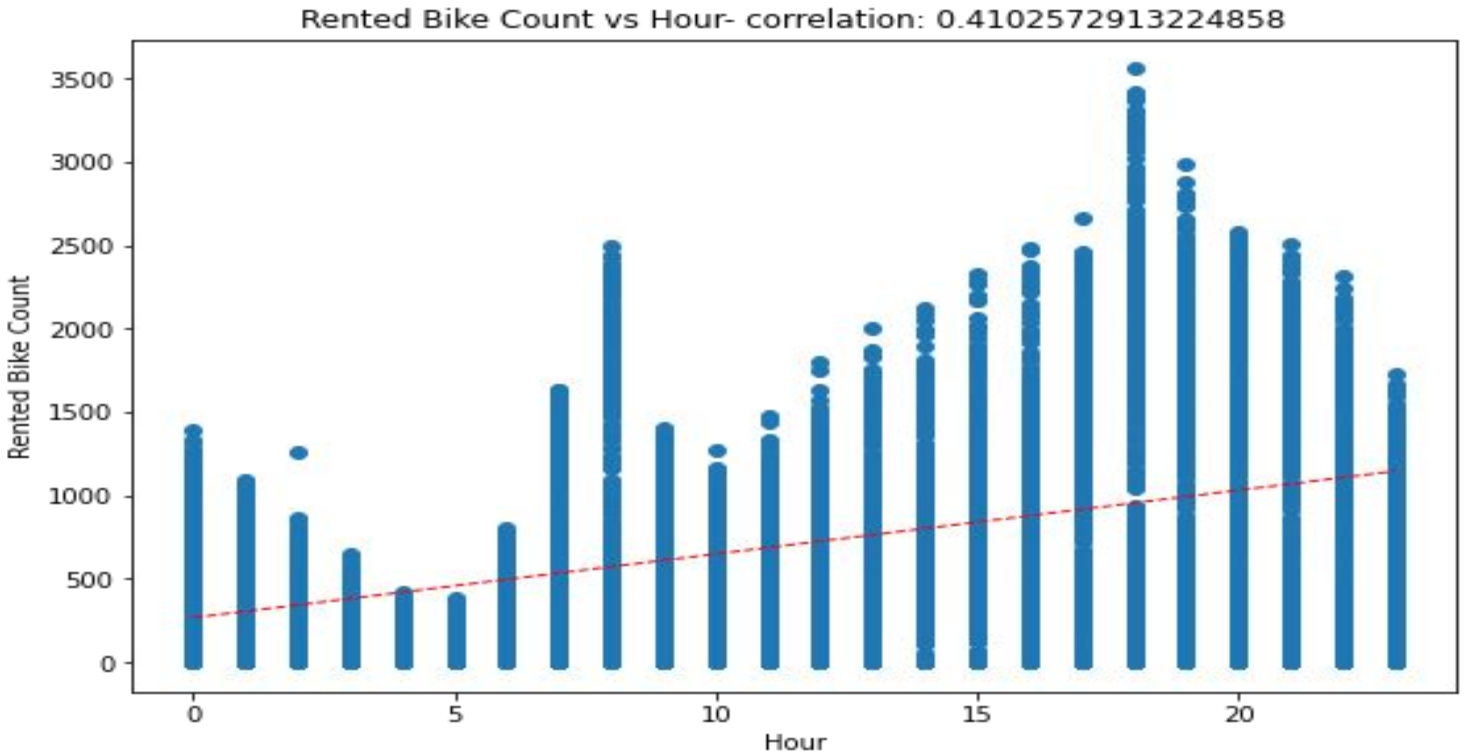


VIF has been used to detect multicollinearity and summer feature is highly collinear.so it is dropped.

Correlation plots between dependent and independent variables

Rented bike count V/S temperature.



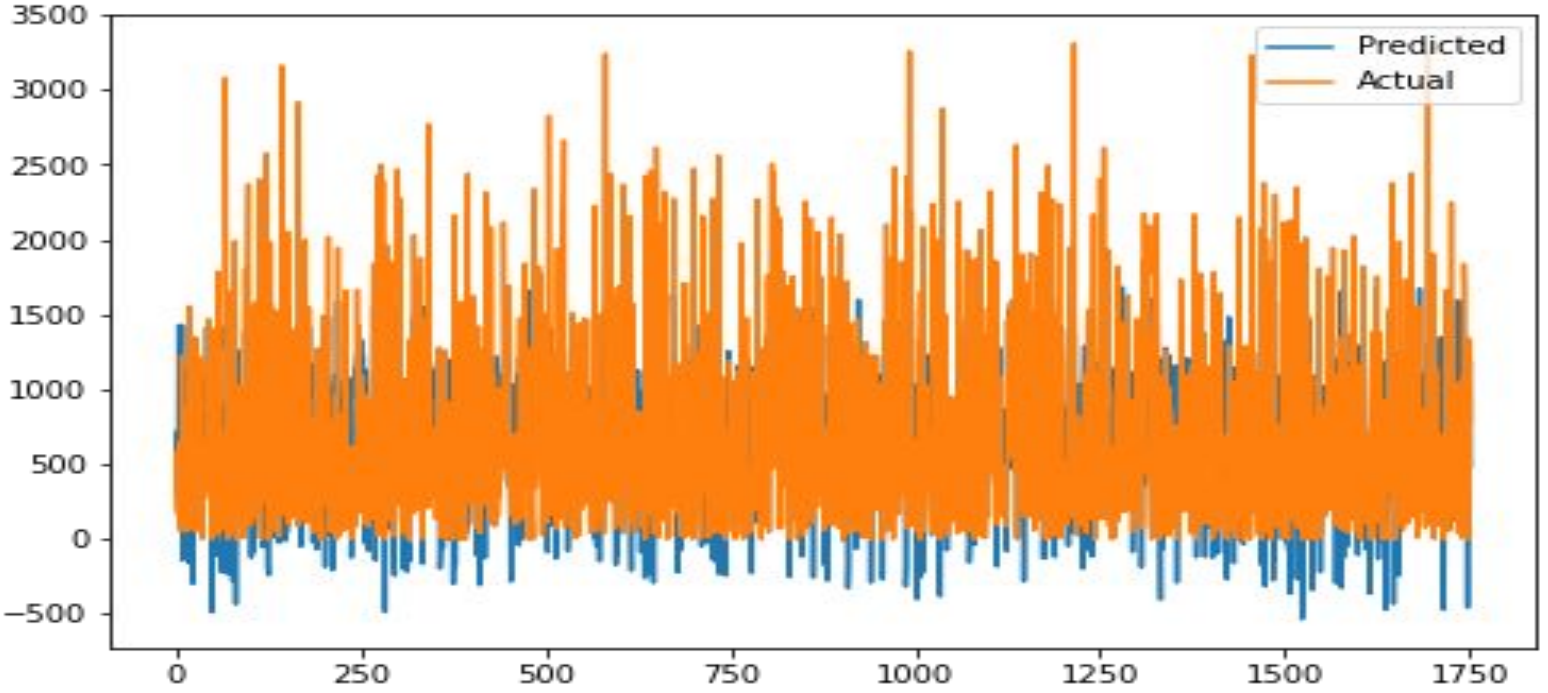


This shows the rented bike count vs hour.

Model Building - Linear regression model

After selecting rented bikes feature as the independent feature and the rest other columns as dependent feature, we split the data into train set and test set , later we transformed the data using minmax scalar and we fitted LINEAR REGRESSION MODEL to the dataset.

The following results were obtained,
MSE : 202937.66869256634
RMSE : 450.4860360683407
R2 : 0.5151094008043042
Adjusted R2 : 0.5117634047201476



Results obtained from linear regression model.

Decision tree model

After applying the decision tree model we obtained,

The best Decision Tree R2 score is 0.7753673960792731 with max depth 10

The best R2 test score is : 0.7948642204947705 with max depth = 10

Random forest regression model

The best Random Forest R2 train score is : 0.8278994903973604 with n estimators = 20, max depth : 15, min samples split : 4 and min samples leaf : 1

The best Random Forest R2 test score is : 0.8516768166059918 with n estimators = 20, max depth : 15, min samples split : 4 and min samples leaf : 1

Conclusion

The exploratory data analysis and modelling for Bike sharing demand prediction dataset has been successfully done and the following inferences have been made from the obtained visualizations and also from the dataset,

For the linear regression model the obtained results are,

- MSE : 202937.66869256634
- RMSE : 450.4860360683407
- R2 : 0.5151094008043042
- Adjusted R2 : 0.5117634047201476

Conclusion

For the decision tree model the obtained results are

The best Decision Tree R2 score is 0.7753673960792731 with max depth 10
The best R2 test score is : 0.7948642204947705 with max depth = 10

For the random forest model the obtained results are
The best Random Forest R2 train score is : 0.8278994903973604 with n estimators = 20, max depth : 15, min samples split : 4 and min samples leaf : 1

The best Random Forest R2 test score is : 0.8516768166059918 with n estimators = 20, max depth : 15, min samples split : 4 and min samples leaf : 1

The results are obtained and compared thoroughly and the best predictions are made.

Thank you