Capstone Project

Bike Sharing Demand Prediction

By Nayana Pradeep

PROBLEM STATEMENT

- Given is a rich dataset of the Seoul Bike Demand Analysis which includes the weather condition details with number of bikes rent every hour for an entire year.

- Explore and analyze the dataset to have an understanding on how the bike rental demand change with the weather and time.

- Predict the count of bikes being rented every hour.

OVERVIEW

Bike riding has become quite popular in last few years owing to the ecofriendly and economical reasons.

South Korea is actually one of the best countries in the world to explore by bicycle, and almost no one knows, not even cyclists!

So it is no surprise that the bike rental services keep in check with the supply and demand of bikes.

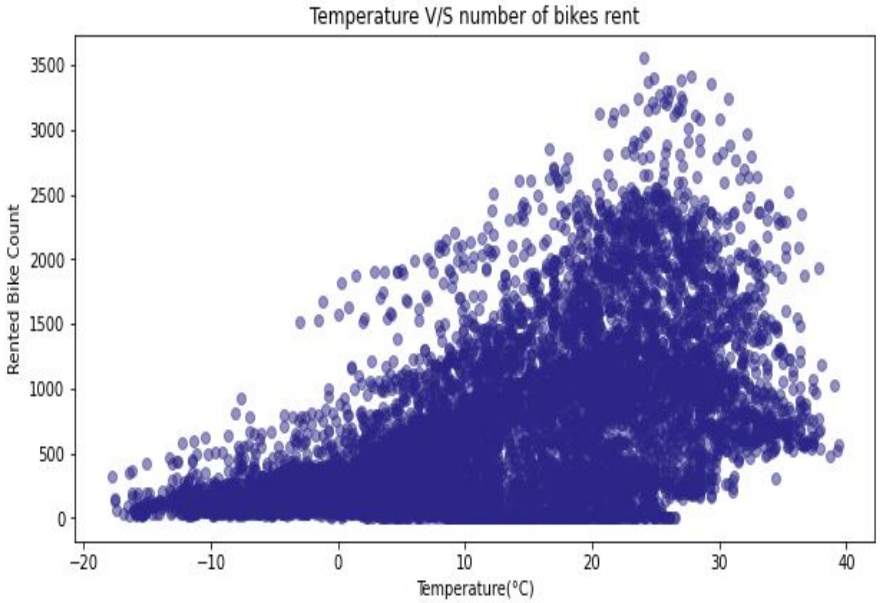
DATA SUMMARY

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday – Holiday and No holiday
- Functional Day - Non Functional Days and Functional Days

EXPLORATORY DATA ANALYSIS

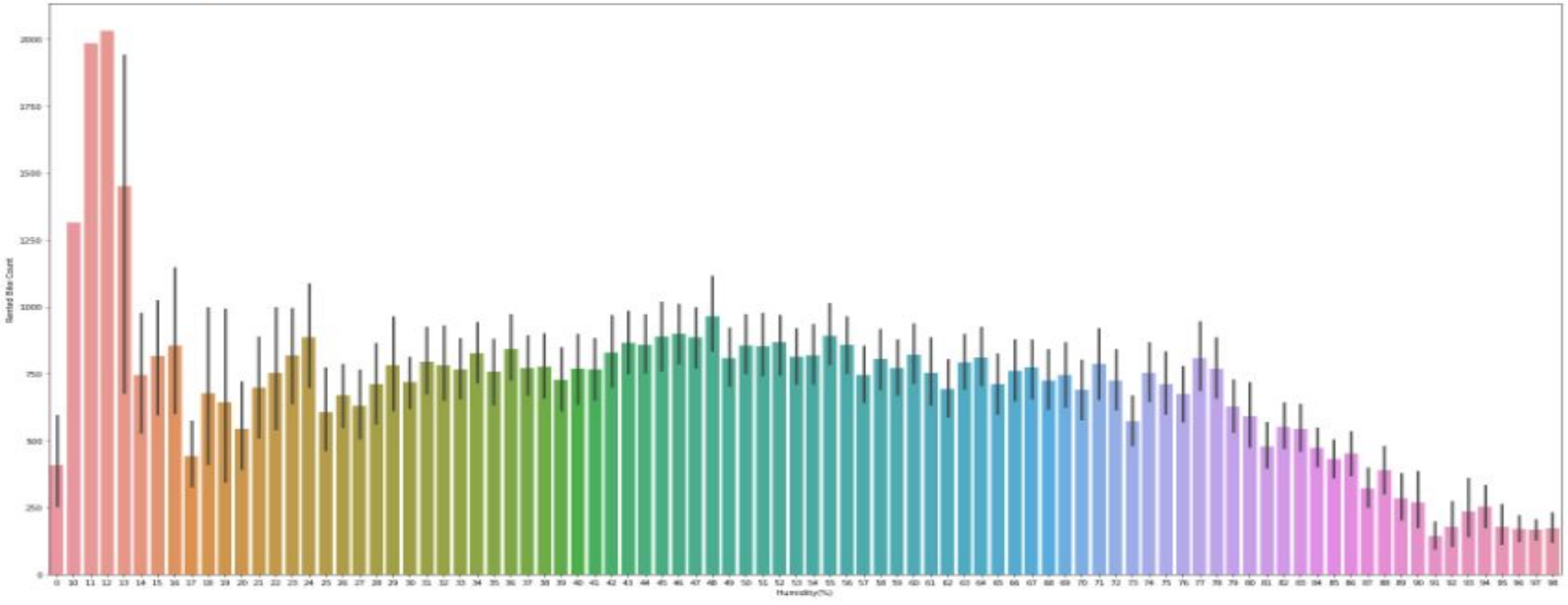
- 8760 rows and 14 columns
- No null values or duplicates
- Date, Hour , Seasons, Functioning Day and Holiday are the categorical features
- People rent most number of bikes during the summer season and least during the winter season.
- Bike rental demand is less on holidays. This indicates that people prefer to use these bikes as mode of transportation to work.

TEMPERATURE AND BIKE COUNT

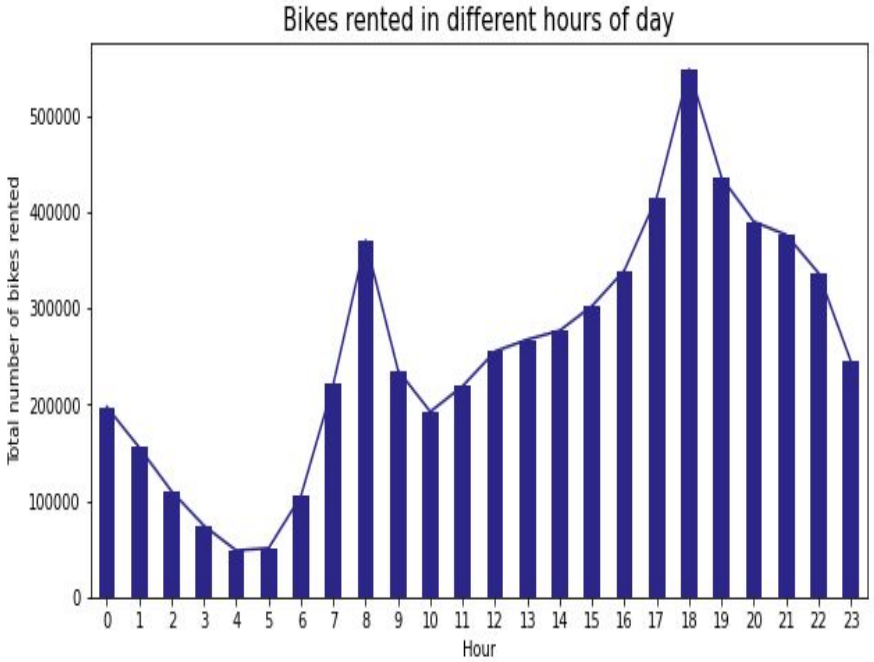


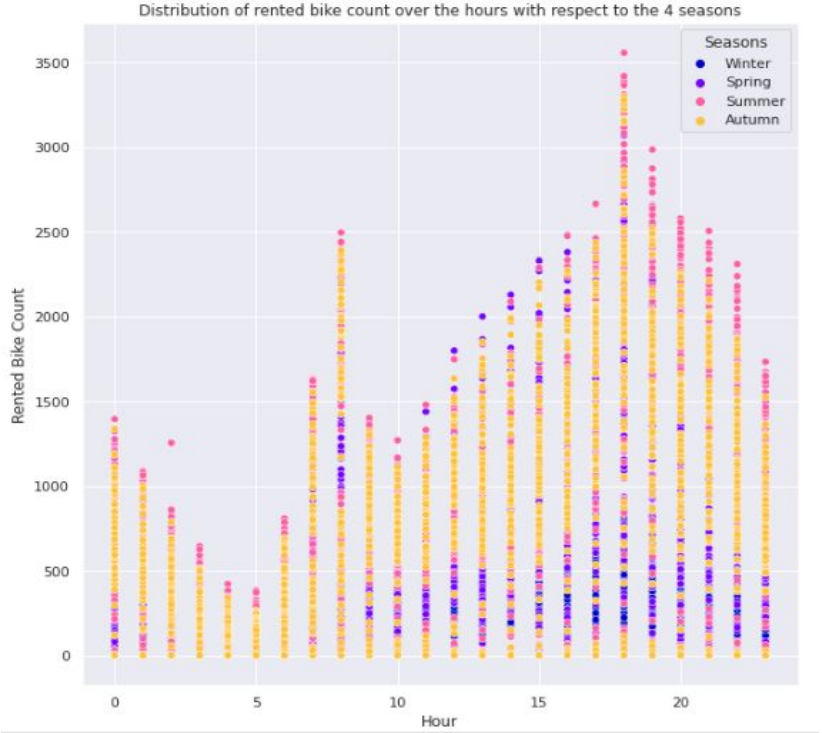
- Rent Bike Count has an almost linear relationship with the temperature.
- The conclusion of increasing demand during the summer is clear.

HUMIDITY AND BIKE COUNT



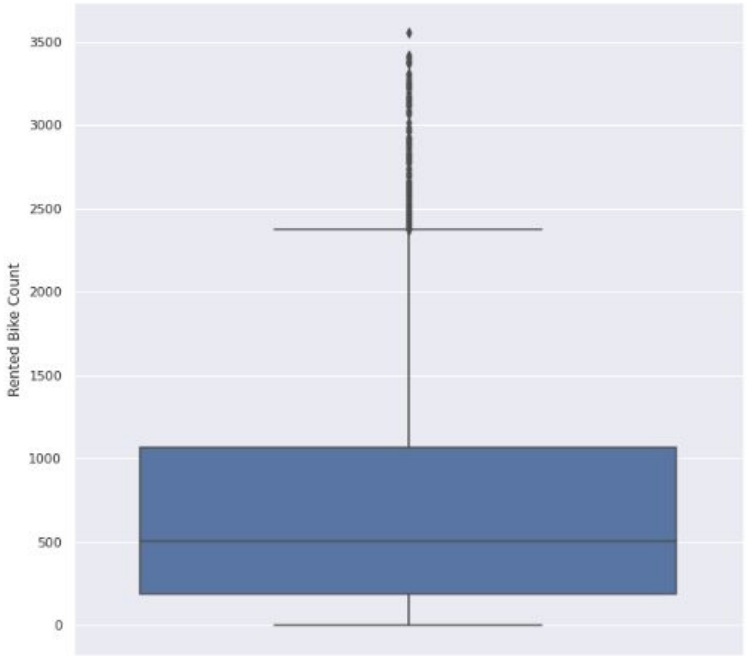
SEASONS AND HOURS WITH BIKE COUNT





FEATURE EXTRACTION AND OUTLIERS

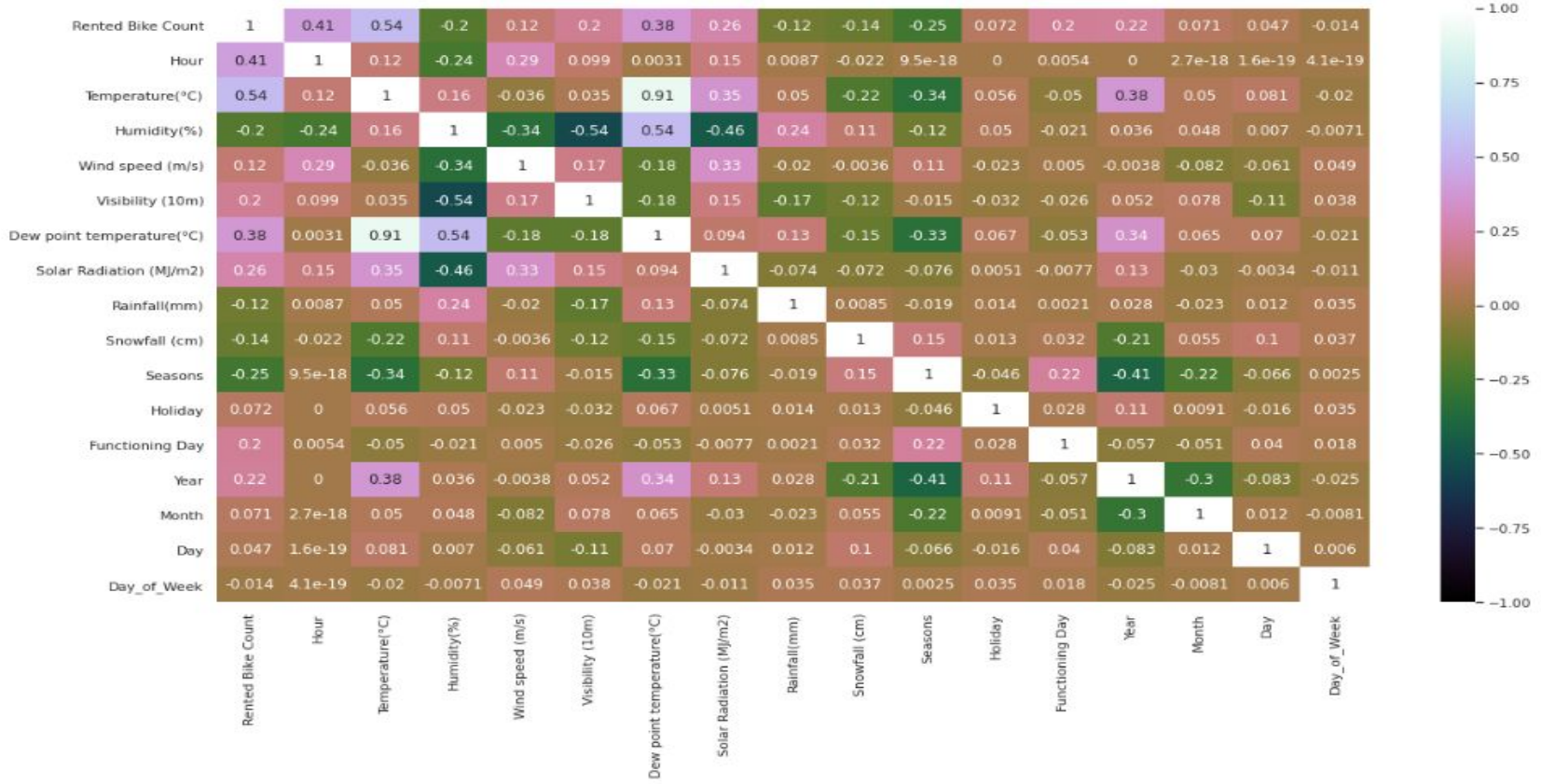
- Day, month , year and weekday extracted from date column and made new columns out of it
- Enabled better analysis
- Median of the Rent Bike Count is 500
- Less number of outliers



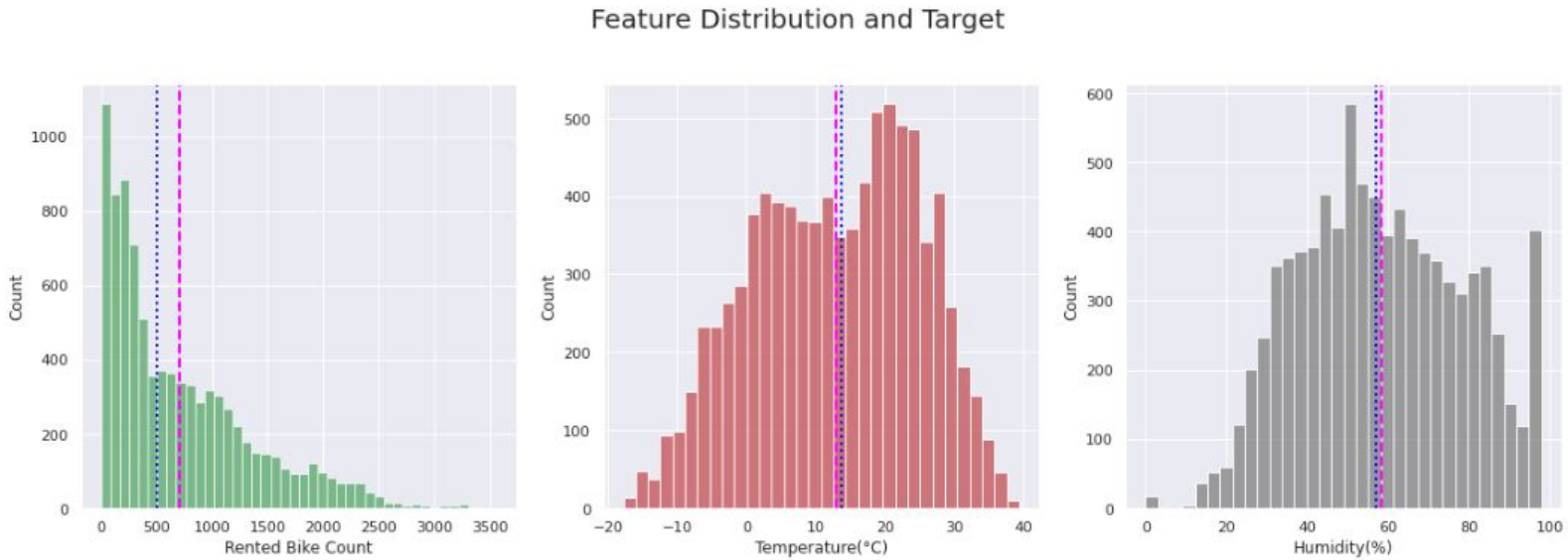
LABEL ENCODING

- Holiday
- Functioning Day
- Seasons
- Day of the week

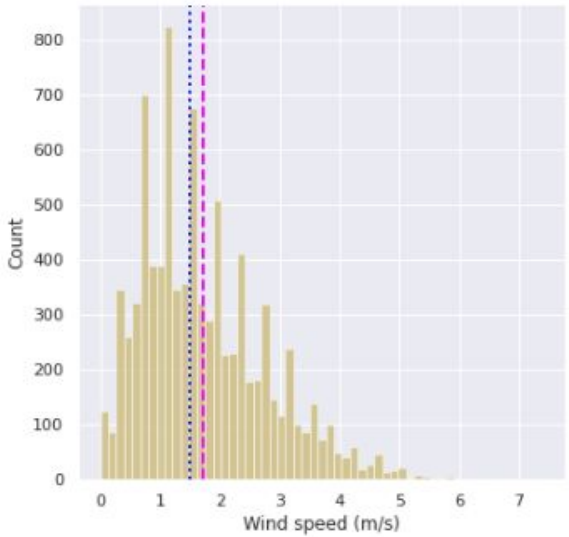
CORRELATION HEATMAP

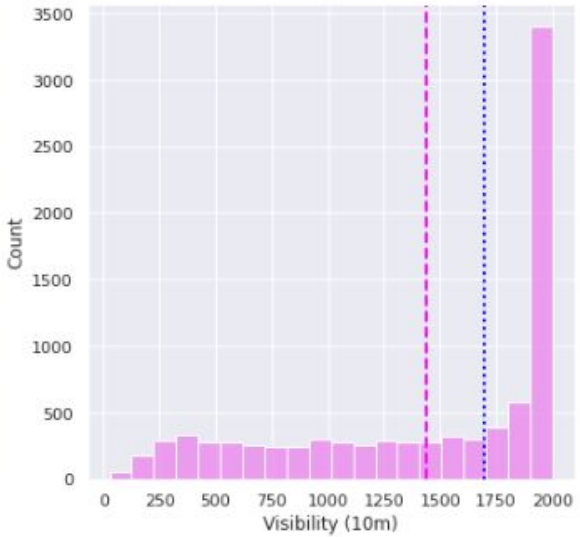


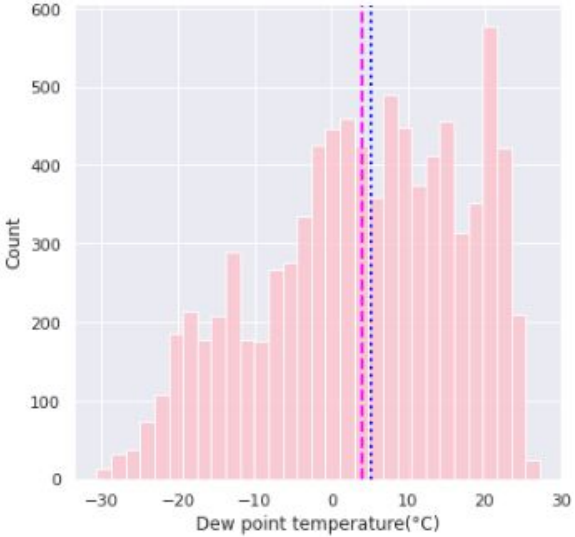
FEATURE DISTRIBUTION



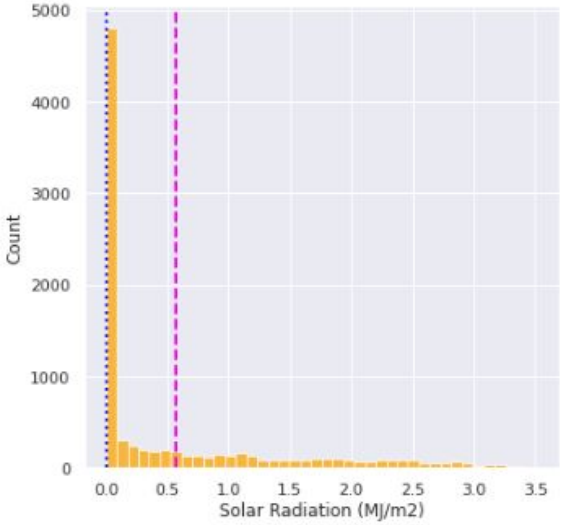
FEATURE DISTRIBUTION

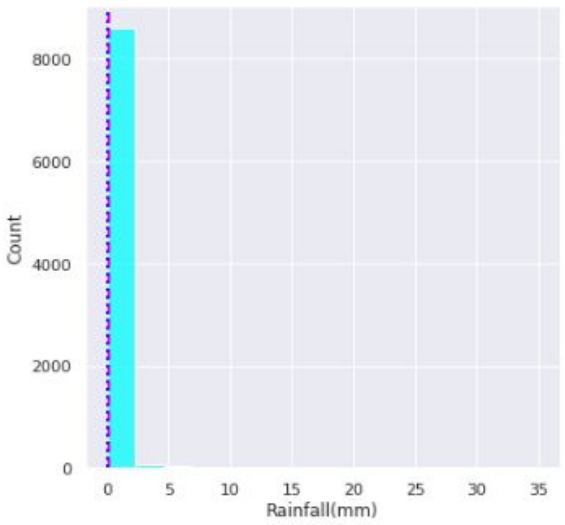


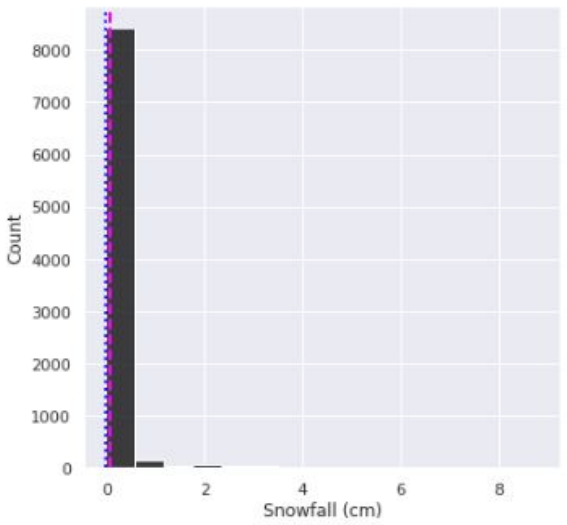




FEATURE DISTRIBUTION







CORRELATION WITH BIKE RENT COUNT

	Correlation with Rented Bike Count
Rented Bike Count	1.000000
Temperature(°C)	0.538558
Hour	0.410257
Dew point temperature(°C)	0.379788
Solar Radiation (MJ/m2)	0.261837
Year	0.215162
Functioning Day	0.203943
Visibility (10m)	0.199280
Wind speed (m/s)	0.121108
Holiday	0.072338
Month	0.070861
Day	0.046849
Day_of_Week	-0.014004
Rainfall(mm)	-0.123074
Snowfall (cm)	-0.141804
Humidity(%)	-0.199780
Seasons	-0.253058

LINEAR REGRESSION MODEL

- MAE : 5.660417893430779
- MSE : 54.05488568978613
- RMSE : 7.352202778064961
- R2 : 0.6578077075559197

LASSO REGRESSION MODEL

- MAE : 5.660956563638153
- MSE : 54.0561990479736
- RMSE : 7.352292094848626
- R2 : 0.6577993933944293

RIDGE REGRESSION MODEL

- MAE : 5.703254454109305
- MSE : 54.60897217620785
- RMSE : 7.389788371544062
- R2 : 0.6543000852090879

GRADIENT BOOSTING

- MAE : 3.1410591115406703
- MSE : 19.487435690184608
- RMSE : 4.4144575759865
- R2 : 0.8766355675794004

XGBOOST MODEL

- MAE : 3.1139086468938473
- MSE : 19.202553766527622
- RMSE : 4.382071857754916
- R2 : 0.8784390012059464

RANDOM FOREST MODEL

- MAE : 2.4127930867482856
- MSE : 13.38508831250127
- RMSE : 3.658563695290991
- R2 : 0.9152662336480208

HYPERPARAMETER TUNING

XGBoost – GridSearchCV	max_depth	8
	min_samples_leaf	30
	min_samples_split	10
	n_estimators	100

Random Forest – GridSearchCV	max_depth	12
	min_samples_leaf	30
	min_samples_split	10
	n_estimators	100

Random Forest – RandomizedSearchCV	bootstrap	TRUE
	max_depth	None
	max_features	Auto
	n_estimators	11

XGBOOST GRIDSEARCHCV

- MAE : 2.0239312402965672
- MSE : 10.00143590368713
- RMSE : 3.1625046883265058
- R2 : 0.9366863099247676

RANDOM FOREST GRIDSEARCHCV

- MAE : 3.240885412547648
- MSE : 21.01221206348057
- RMSE : 4.58390794666304
- R2 : 0.8669830317172941

RANDOM FOREST RANDOMIZEDSEARCHCV

- MAE : 2.671382889859473
- MSE : 15.449442001863549
- RMSE : 3.930577820354604
- R2 : 0.9021979251618604

MODEL COMPARISON

	MODEL NAME	MAE	MSE	RMSE	R2
0	Linear Regression	5.660418	54.054886	7.352203	0.657808
1	Lasso Regression	5.660957	54.056199	7.352292	0.657799
2	Ridge Regression	5.703254	54.608972	7.389788	0.654300
3	Gradient Boosting	3.140610	19.483598	4.414023	0.876660
4	XGBoost	3.113909	19.202554	4.382072	0.878439
5	Random Forest	2.404020	13.401257	3.660773	0.915164
6	XGBoost- GridSearchCV	2.023931	10.001436	3.162505	0.936686
7	Random Forest- GridSearchCV	3.252191	21.248756	4.609637	0.865486
8	Random Forest- RandomizedSearchCV	2.554616	15.153617	3.892765	0.904071

CONCLUSION

- **XGBoost model with hyperparameter tuning** promises the best accuracy among the experimented models with R2 score of **0.9366** and the least error rate**.**
- This is followed by Random Forest with R2 score of 0.9151.
- Least accurate model is linear regression model with an accuracy of 65.7%.

CHALLENGES

- Large dataset
- Lot of analyzing and visualization required to unfold the relationship between the features

THANK YOU