Capstone Project – 2

Bike Sharing Demand Prediction

By – Nitin Kumar

Content

- Introduction
- Problem definition
- Hypothesis
- EDA on given data
- Statistical Data Analysis
- Model implementation
- Model validation and selection
- Conclusion
- References

Introduction

- According to recent studies, it is expected that more than 60% of the population in the world tends to dwell in cities.
- Urban mobility usually fills 64% of the entire kilometres travelled in the world.
- Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort.
- As a convenient, economical, and eco-friendly travel mode, bike-sharing greatly improved urban mobility.

Introduction

- It is eco-friendly and comfortable compared to driving.
- Ggenerally travelling by bike takes less time than walking.
- Usually it is difficult to maintain the balance in between demand of bikes in cities.
- Aim of the project is to put a model on given data for prediction of bikes.

Problem definition

- The main goal of the project is to Finding factors and cause those influence shortages of bike and time delay of availing bike on rent.
- Maximize: The availability of bikes to the customer.
- Minimize: Minimise the time of waiting to get a bike on rent.

Hypothesis

- From data and problem statement we would like to put hypothesis on 5 features as.
- In season summer bike demand will be more.
- If there is less visibility bike demand will be less.
- In the hours 9 am and 7 pm demands will be more.
- On Sunday bike demands will be less.
- With rainfall and snow fall demands of bike will be reduced.
- At end we will see either we reject the hypothesis or fail to reject it.

EDA

Digging into data we understand that

- Dataset contains 13 features such as
  - Date : year-month-day
  - 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

EDA

- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - No Func (Non Functional Hours), Fun (Functional hours)

- Response variable is :
- Rented Bike count - Count of bikes rented at each hour
- Graphical representation according to various columns and with manipulation of columns.

EDA

- We can see there is no null values in the given data set.

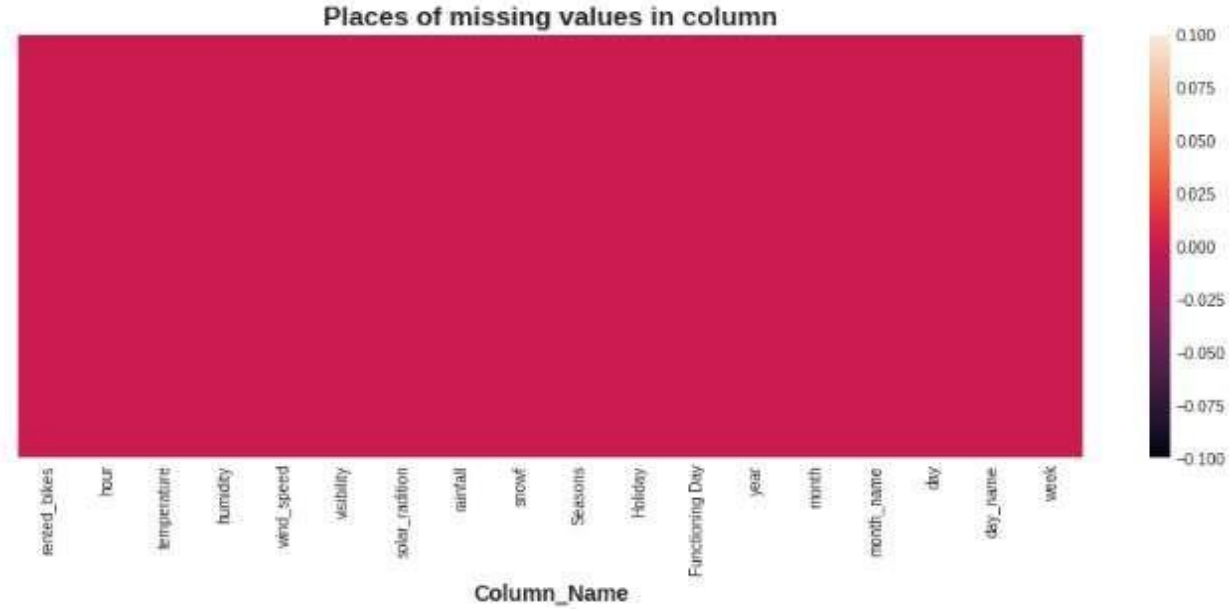

Fig 1:Missing values in data set.

EDA

- Rented bike counts are positively skewed and temperature is normally distributed.

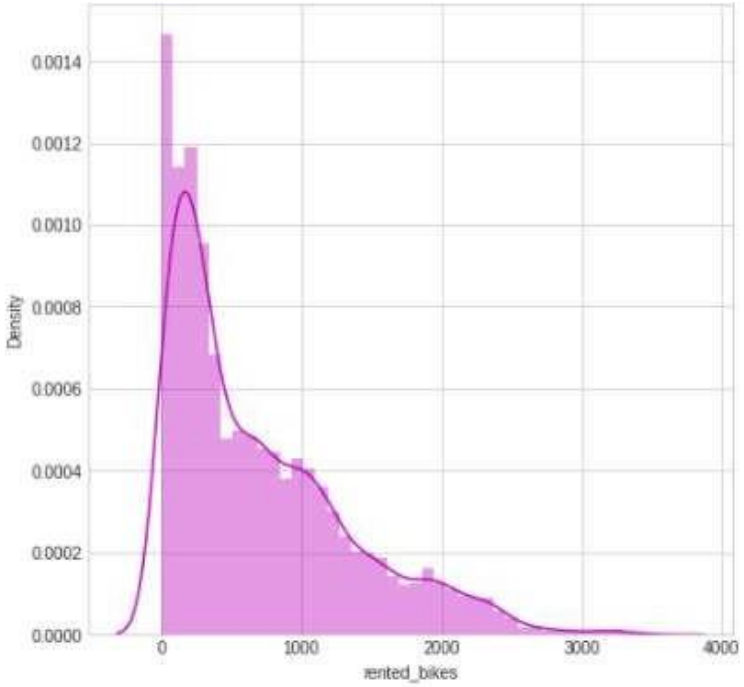

Fig 2: Distribution of rented bike counts.

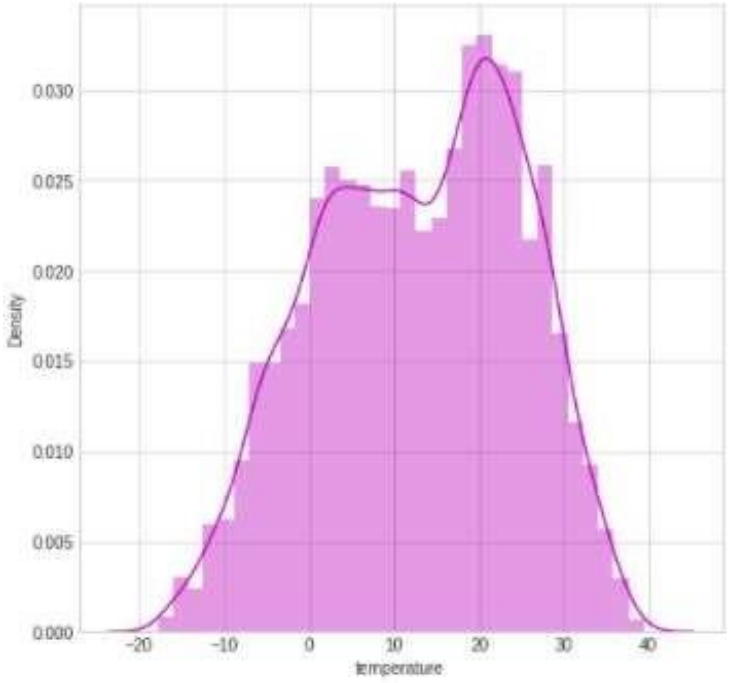

Fig 3: Distribution of temperature.

EDA

- Humidity is normally distributed and wind speed is positively skewed.

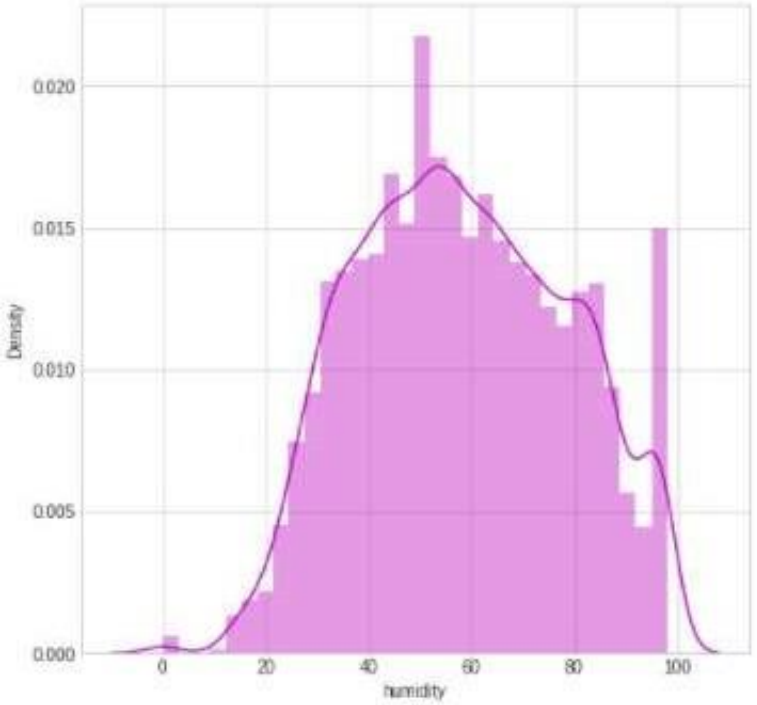

Fig 4: Distribution of Humidity.

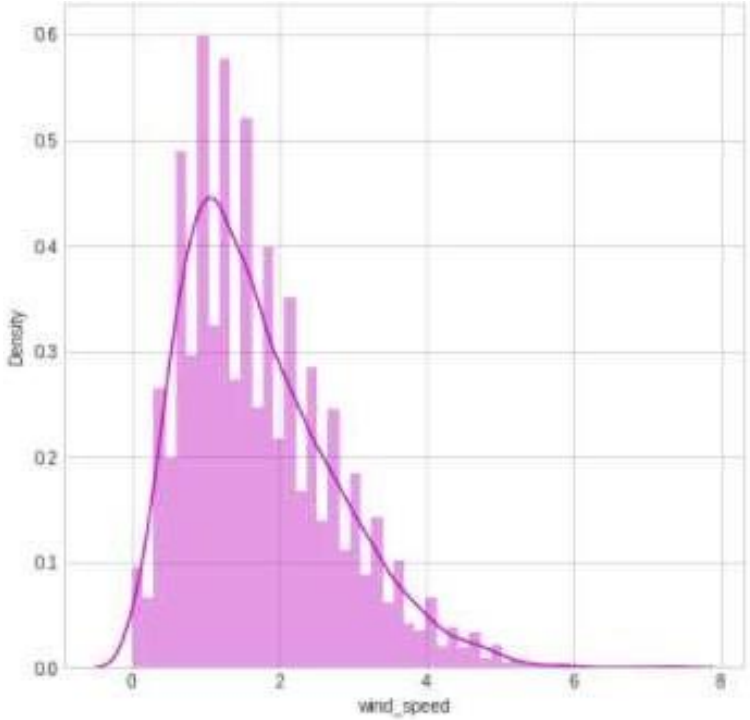

Fig 5: Distribution of Wind speed.

EDA

- Visibility is negatively skewed and solar radiation is positively skewed.

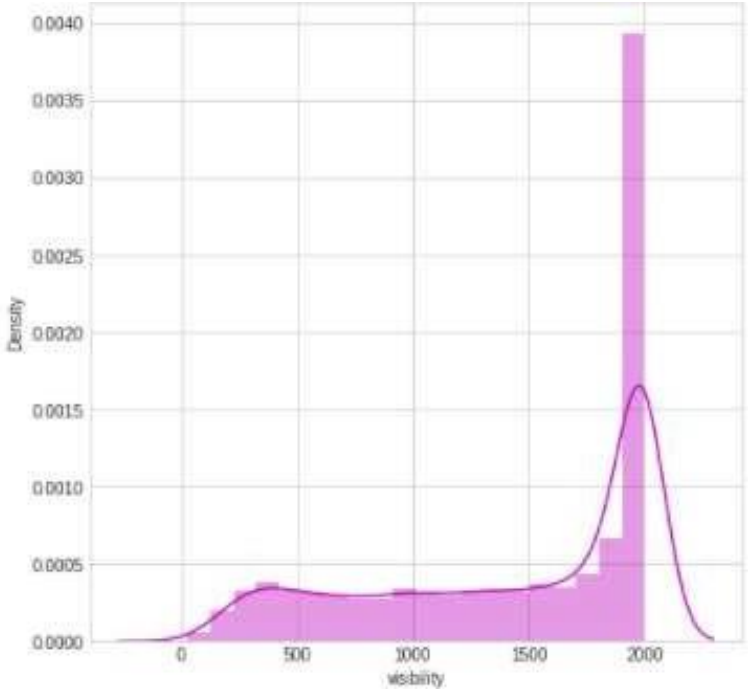

Fig 6:Distribution of Visibility.

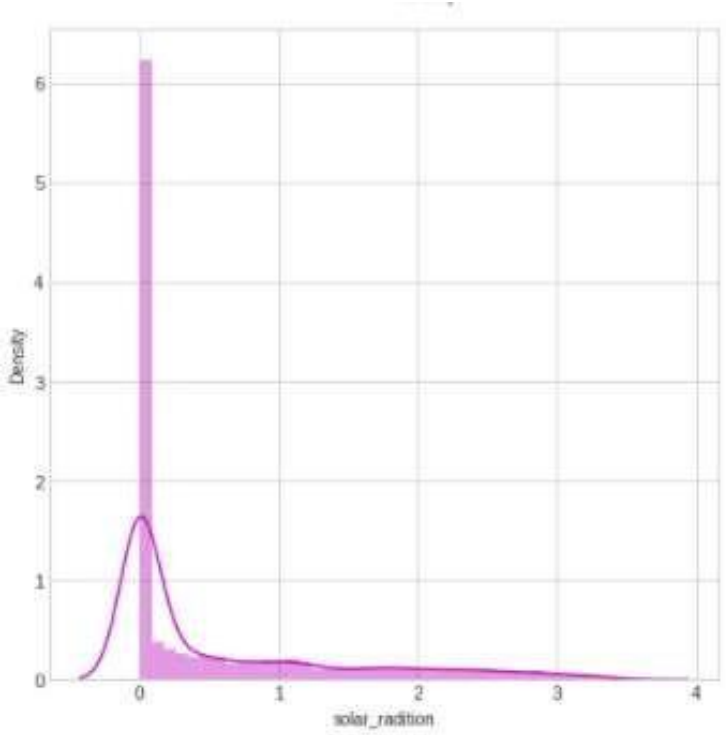

Fig 7:Distribution of Solar radiation.

EDA

- Rainfall and Snowfall both are positively skewed.

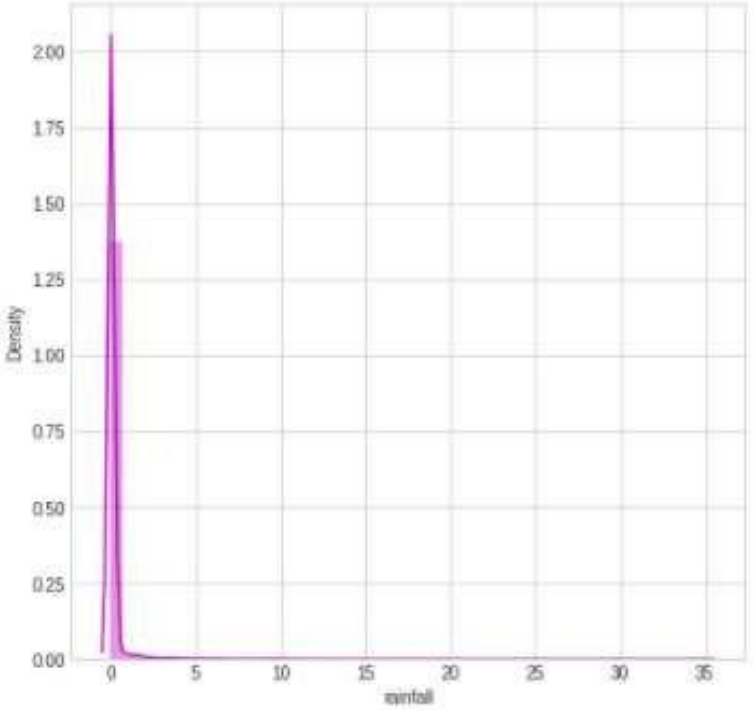

Fig 8: Distribution of Rainfall.

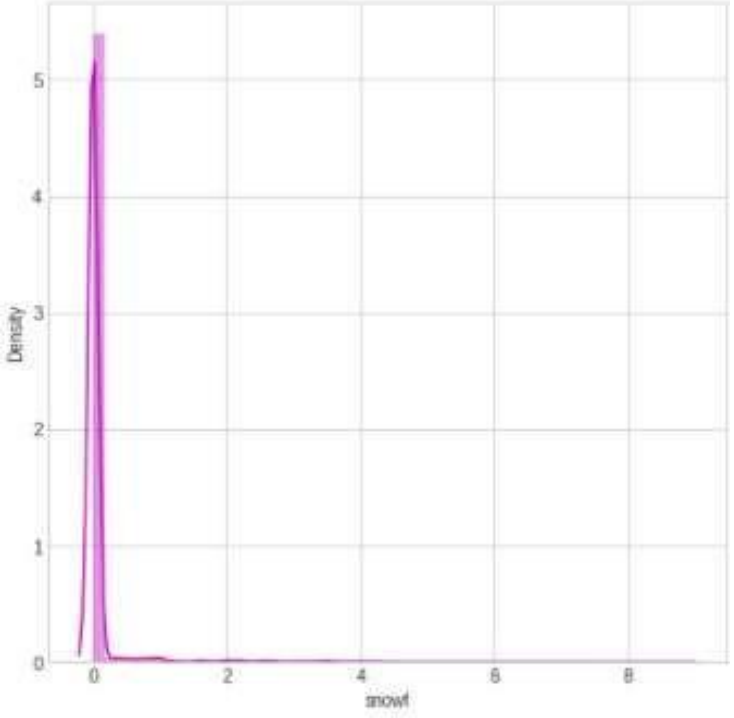

Fig 9: Distribution of Snowfall.

EDA

- From graph we can say less bikes are rented on Sunday as compared to other days.

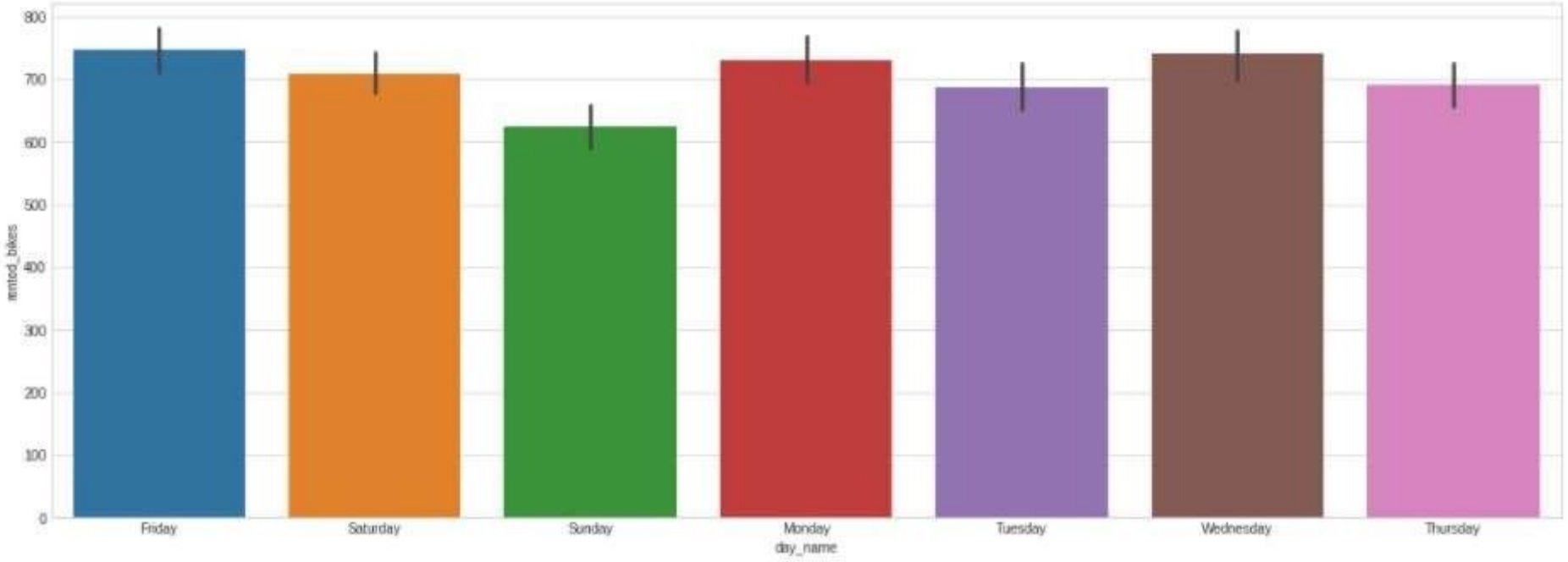

Fig 10 Bar plot in between Rented bikes and Days

EDA

- Maximum bikes are rented on 8 am and 6 pm.

Fig 11Count plot in between Rented bikes and Hours

EDA

- In June month (in summer) more bikes are rented as compared to other.

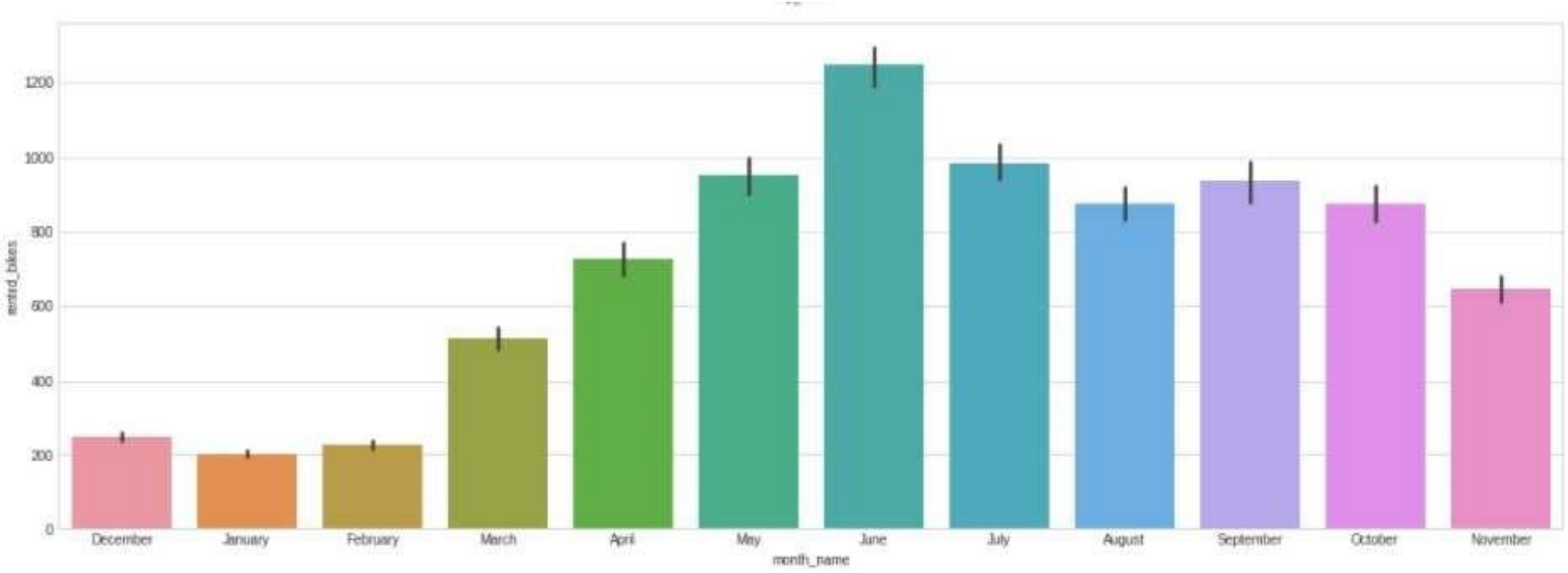

Fig 12 Count plot in between Rented bikes and Months

EDA

- From graph we can say more bikes are rented in summer and less bikes are rented in winter.

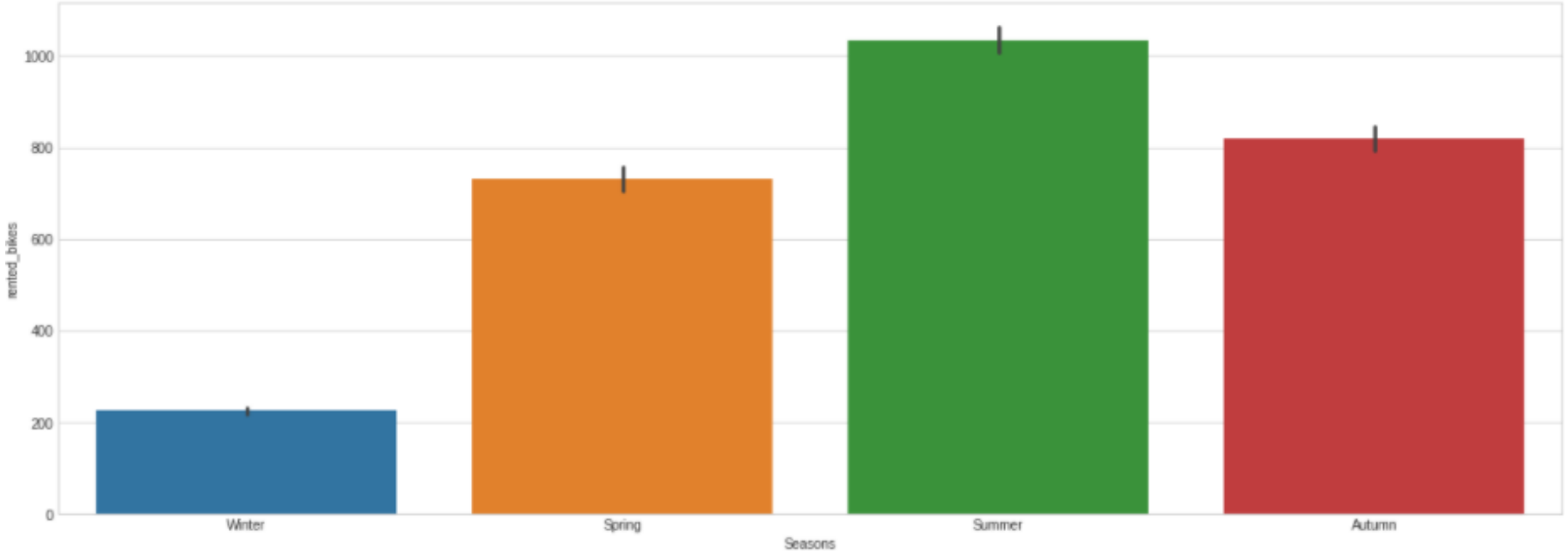

Fig 13Count plot in between Rented bikes and Seasons

EDA

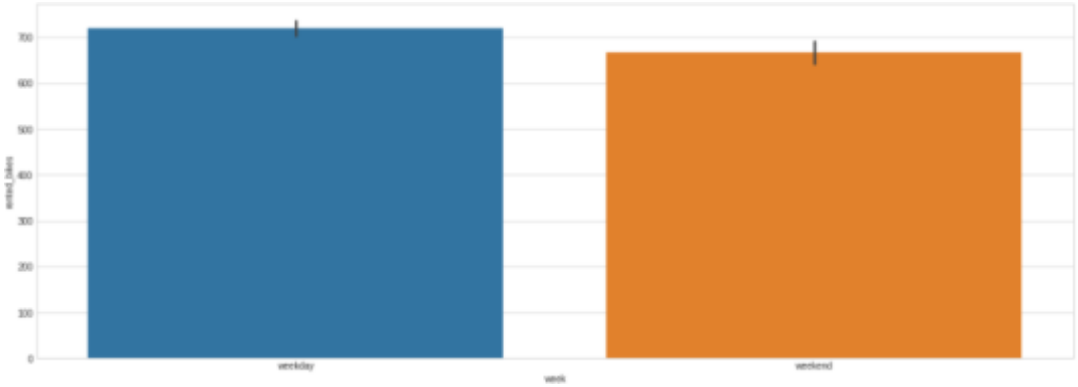

Fig 14 Count plot of Week day and Week end.

- From fig 14 we can say there not more difference in demand of bikes between weekday and weekend
- From fig 15 we see there is a difference in demands of bike in No holiday and Holiday.

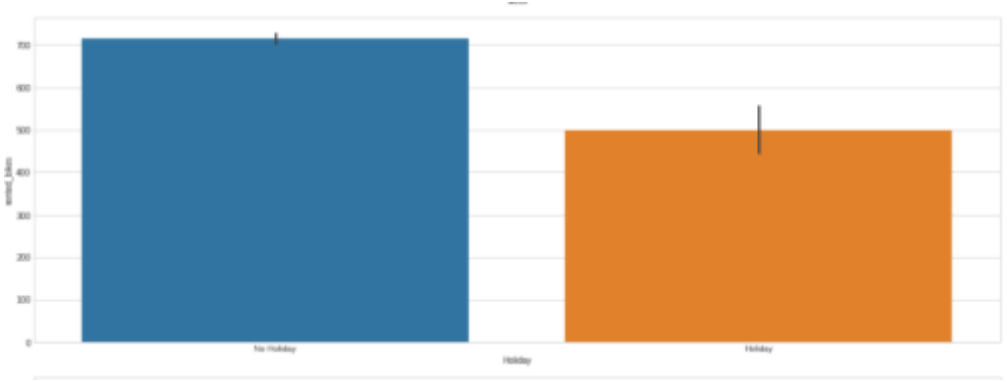

Fig 15 Count plot of No holiday and Holiday.

EDA

- In data there are very less No function day as compared to Function day.

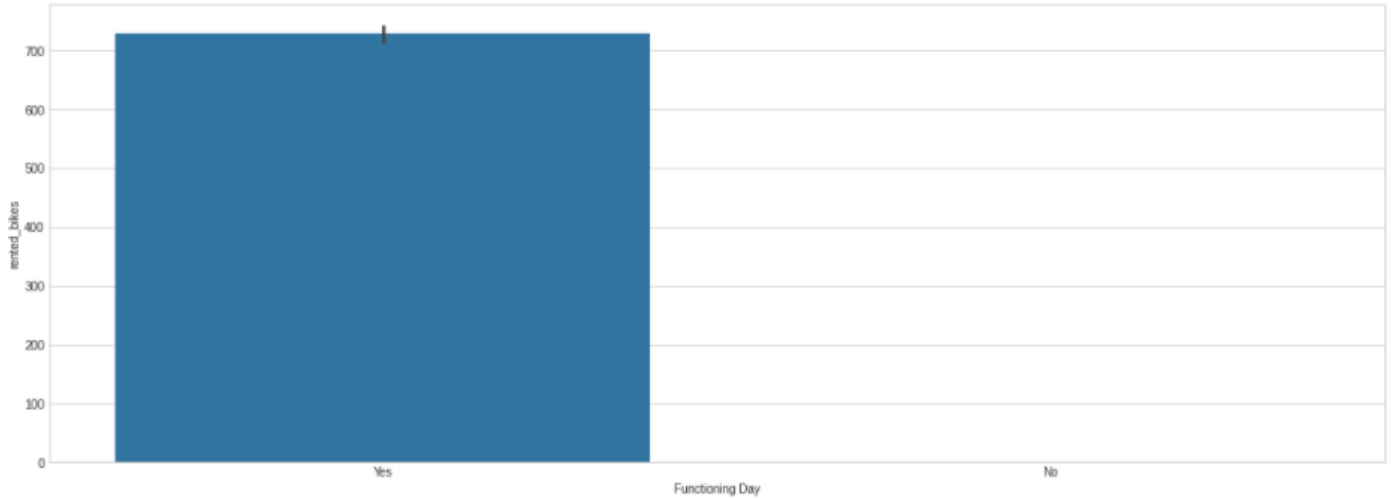

Fig 16 Count plot of Function day and No Function day.

EDA

Seasons	rented_bikes
Summer	2283234
Autumn	1790002
Spring	1611909
Winter	487169

Fig 17 Count plot of Rented bike and Seasons

Seasons	solar_radition
Summer	1680.850000
Spring	1520.840000
Autumn	1139.650000
Winter	644.070000

Fig 18 Count plot of Solar radiation and Seasons

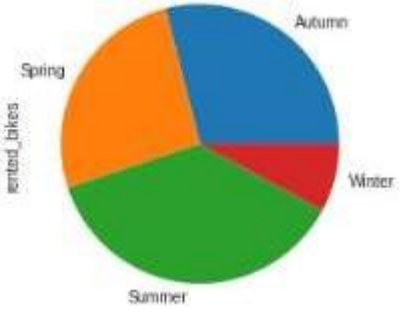

Fig 19 Pi plot of Seasons

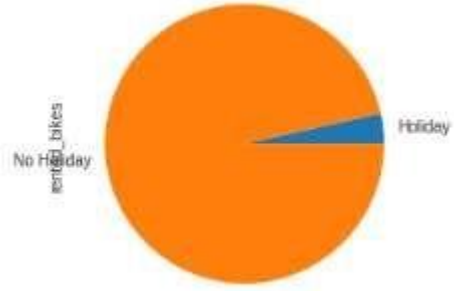

Fig 20 Pi plot of Holiday No holiday

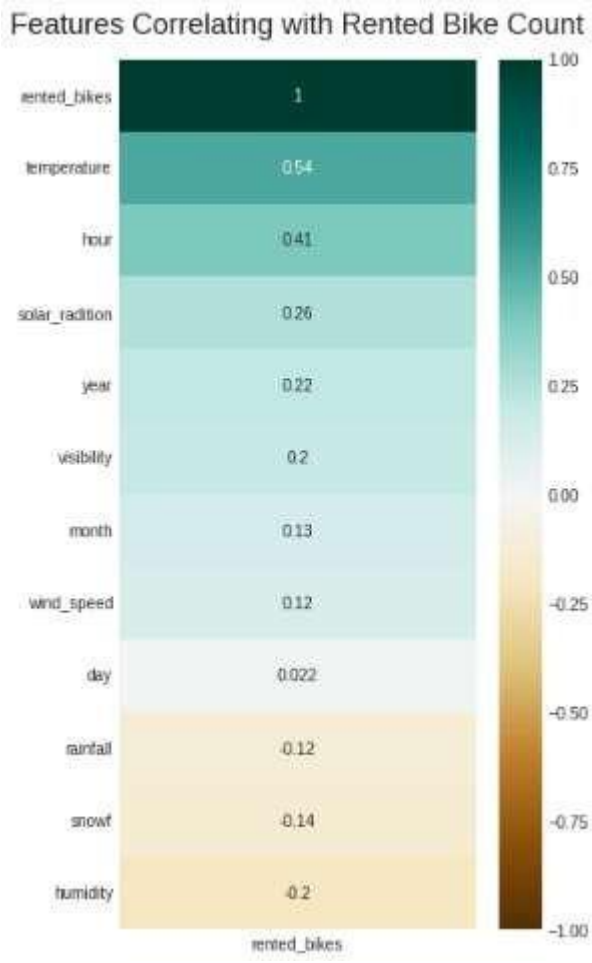

Fig 21 Heat map Rented bike and other factors

EDA

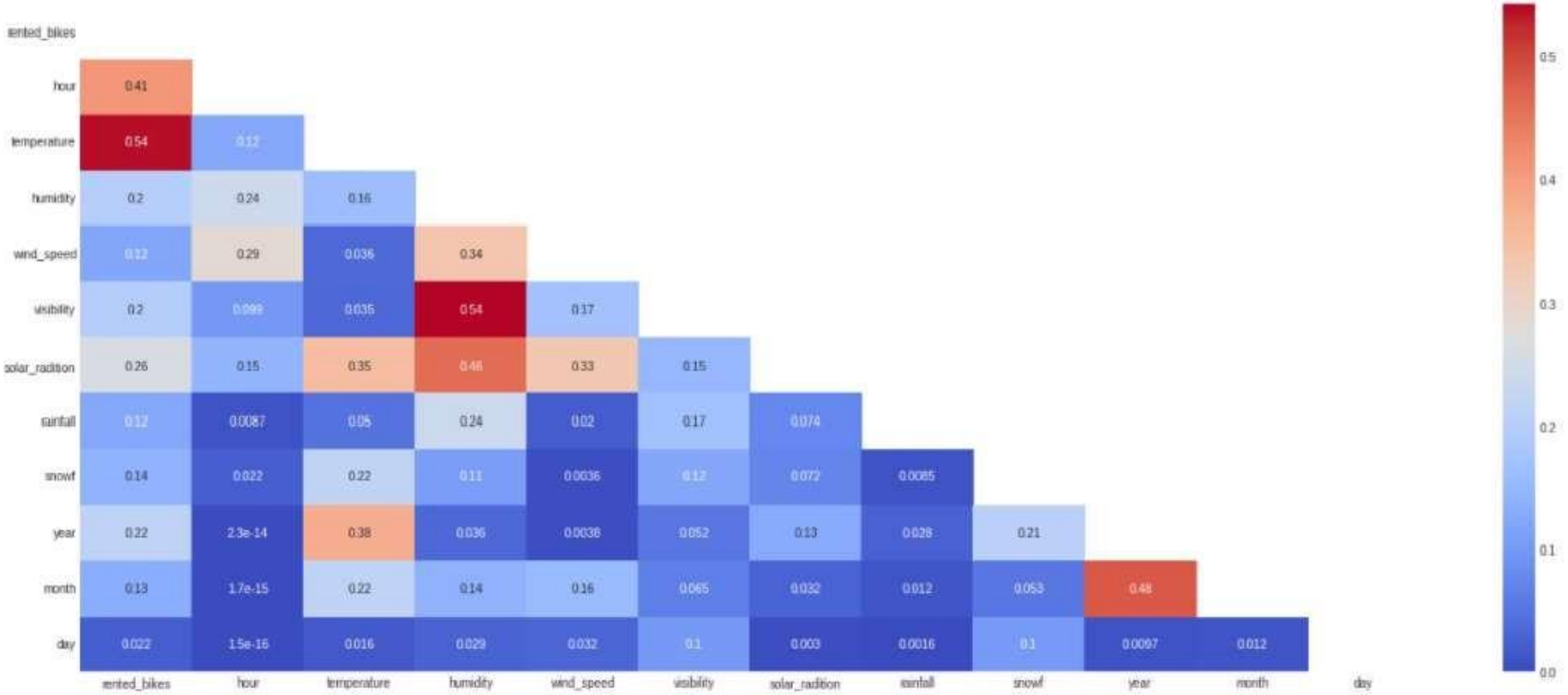

Fig 22 Heat map of variables

EDA

Fig 23 Feature importance bar plot.

Statistical Data Analysis

- From fig 24 temperature and rented bikes are positively co related and from fig 25 humidity and rented biked are negatively co related.

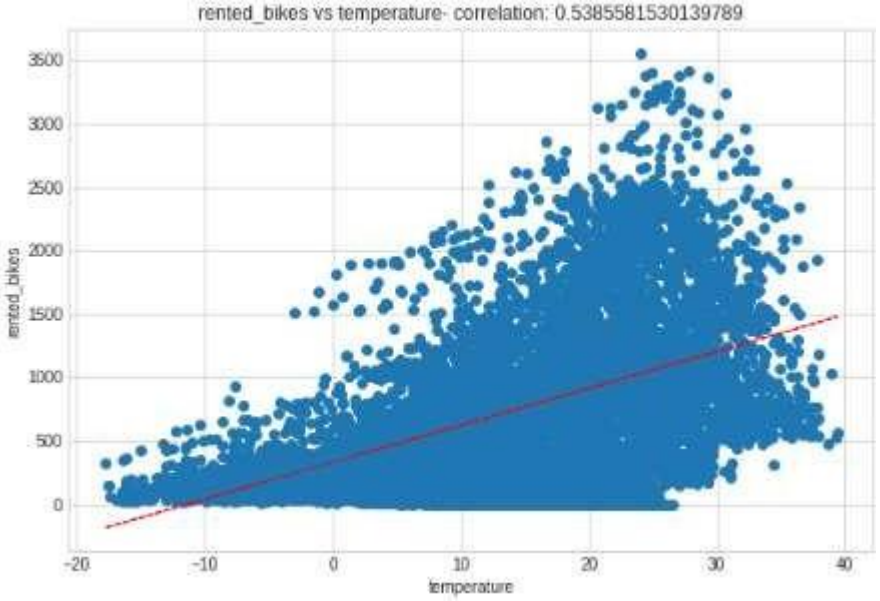

Fig 24 Co-relation plot in Rented bike and Temperature.

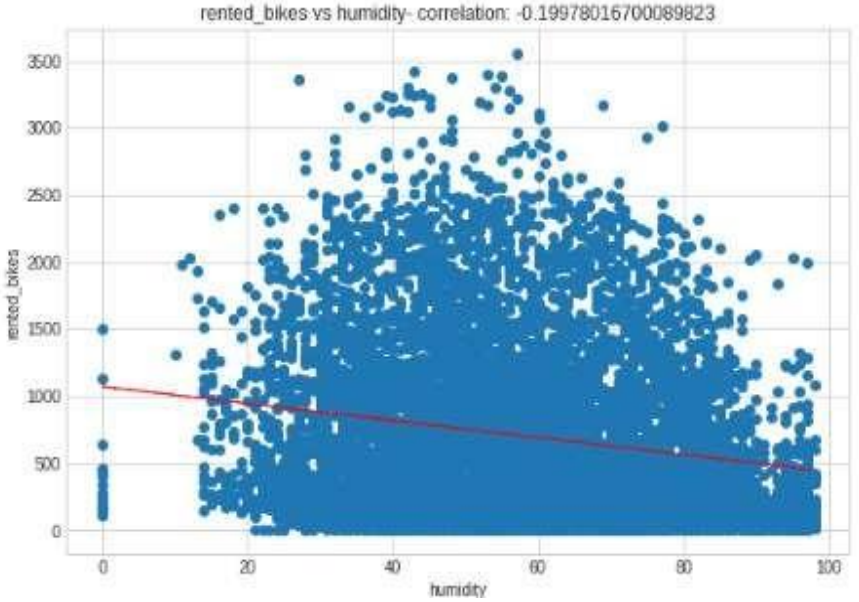

Fig 25 Co-relation plot in Rented bike and Humidity.

Statistical Data Analysis

- Fig 26 and fig 27 Wind speed and visibility both are positively co related with rented bikes.

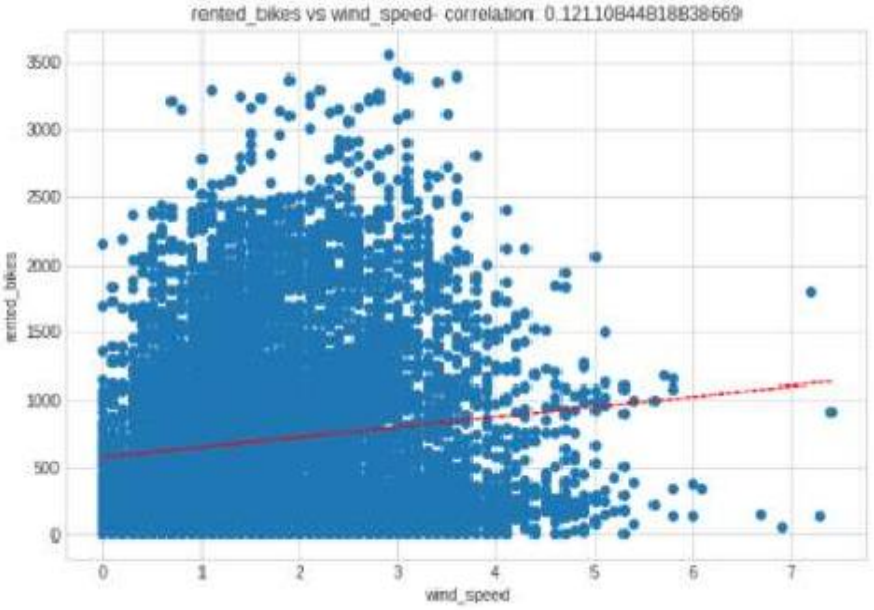

Fig 26 Co-relation plot in Rented bike and Wind speed.

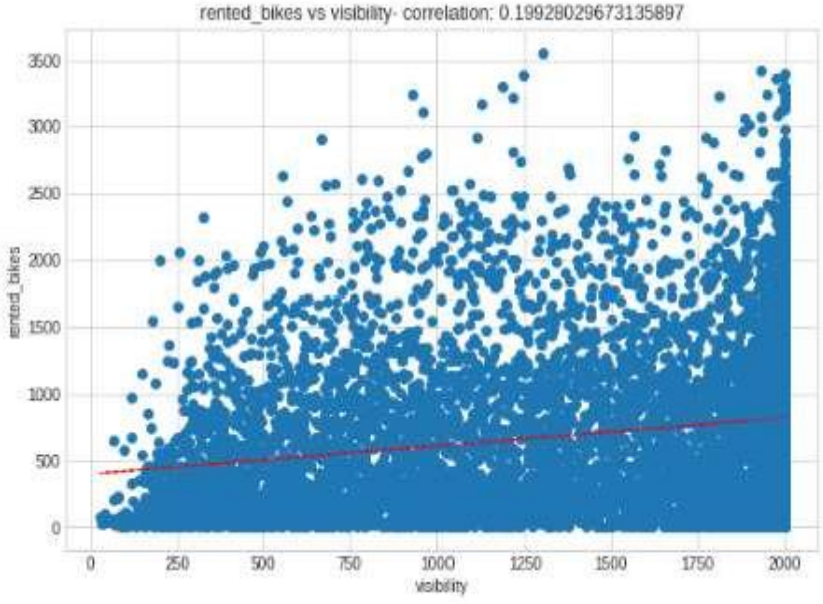

Fig 27 Co-relation plot in Rented bike and Visibility.

Statistical Data Analysis

- Solar radiation and rented bike counts are positively co related.

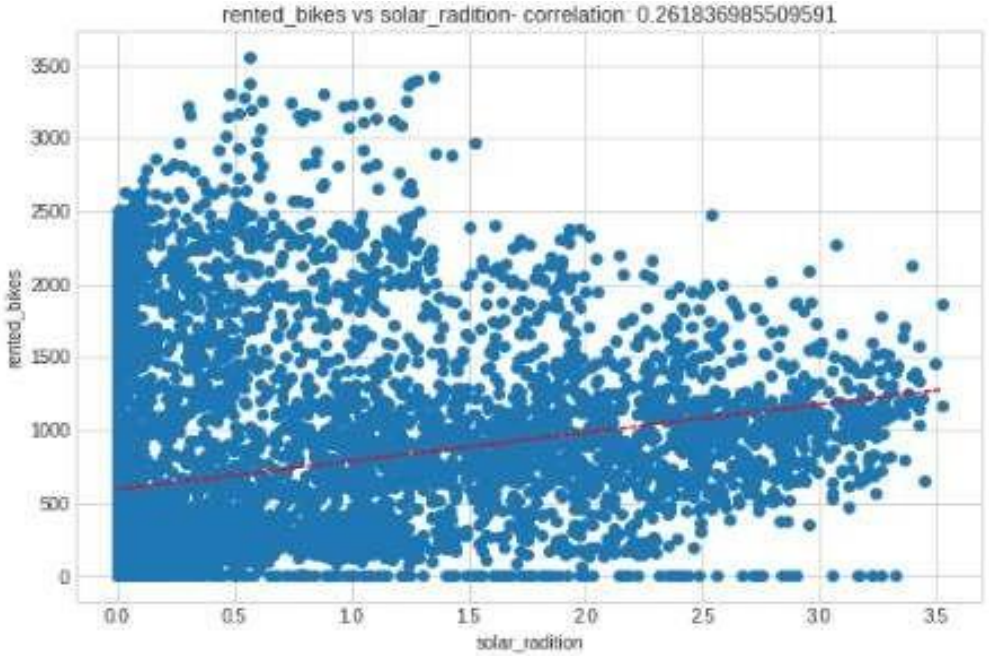

Fig 28 Co-relation plot in Rented bike and Solar radiation

Statistical Data Analysis

- Rain fall and Snow fall both are negatively co related with rented bike count.

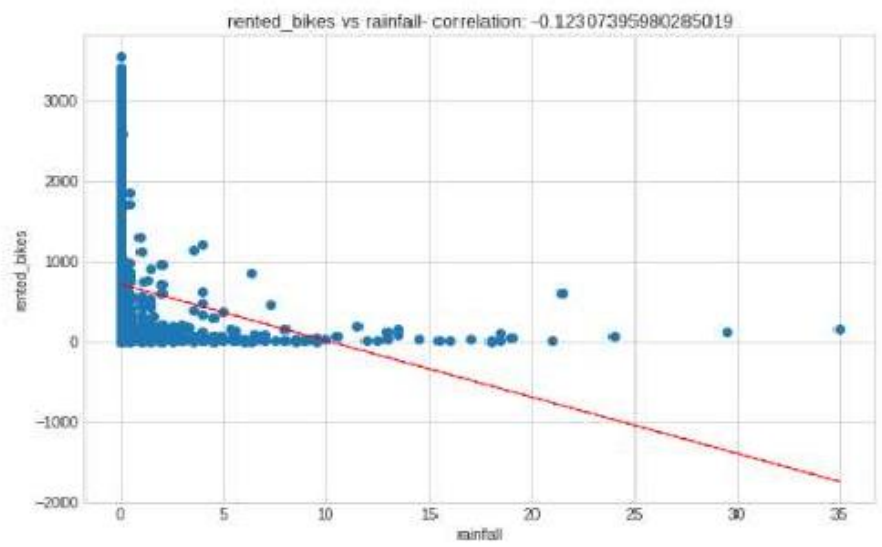

Fig 29 Co-relation plot in Rented bike and Rainfall.

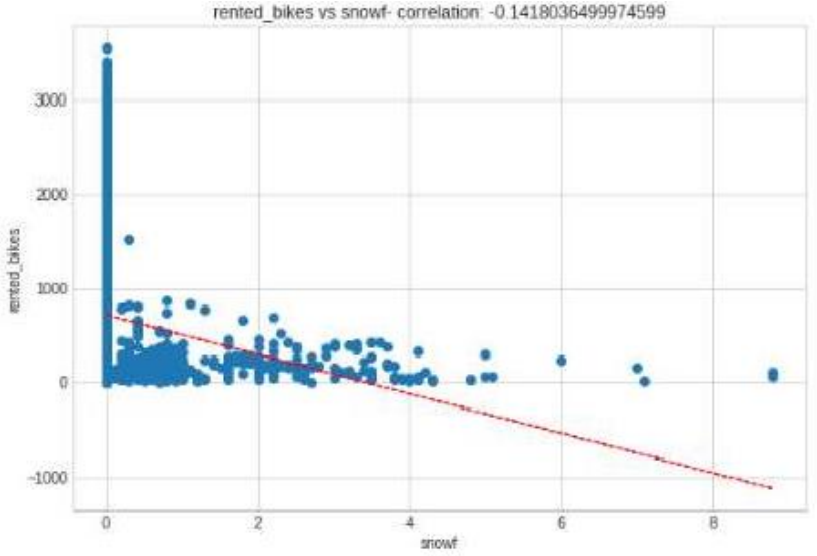

Fig 30 Co-relation plot in Rented bike and Snowfall.

Hypothesis

- With EDA we can justify our hypothesis.
- In season summer bike demand will be more.(Fail to reject)
- If there is less visibility bike demand will be less.(Reject)
- In the hours 9 am and 7 pm demands will be more.(Reject)
- On Sunday bike demands will be less.(Fail to reject)
- With rainfall and snow fall demands of bike will be reduced.(Fail to reject)

Model implementation

- Various types of linear model are implemented on data such as Linear regression, Ridge regression, Lasso regression, Elastic net regression and Polynomial feature regression. As we can see from table the result accomplished are good but not satisfying.

Model	MSE (Train)	RMSE(Train)	R2(Train)	MSE (Test)	RMSE(Test)	R2(Test)
Linear Regression	115071.02	339.22	0.72	112112.97	339.22	0.72
Ridge	115073.23	339.22	0.73	112092.51	334.80	0.72
Lasso	115071.14	339.22	0.72	112102.68	334.81	0.72
Elastic Net	115080.73	339.23	0.72	112074.51	334.77	0.72
Polynomial Feature	29566.78	171.94	0.92	40725.48	201.80	0.90

Table 1Result table of linear models.

Model implementation

- So we have to move towards some complex models such as Decision tree regressor, Random forest regressor, XG boost regressor and Cat boost regressor

Model	MSE (Train)	RMSE(Train)	R2(Train)	MSE (Test)	RMSE(Test)	R2(Test)
Decision tree regressor	0.0	0.0	1.0	63029.50	251.05	0.84
Random forest regressor	4811.62	69.36	0.98	33822.95	183.91	0.91
XG boost	10073.27	100.36	0.97	36204.99	190.27	0.97
Cat boost	101028.40	317.84	0.75	104605.09	323.42	0.75

Table 2 Result table of Complex models.

Model validation and selection

- Best results on model is obtained from Random forest regressor and XG boost regressor.
- After implementation of Grid search CV on both models final accuracy (r2) for XGB was more.

Model	MSE (Train)	RMSE(Train)	R2(Train)	MSE (Test)	RMSE(Test)	R2(Test)
Random forest regressor	4811.62	69.36	0.98	33822.95	183.91	0.91
XG boost	10073.27	100.36	0.97	36204.99	190.27	0.97

Table 3 Result table after implementation of Grid search CV on Complex models.

Conclusion

- Bicycle sharing systems can be the new boom in India, with use of various prediction models the ease of operations will be increased.
- During implementation of linear models accuracy obtained was not more so we moved towards more complex models.
- With application of Decision tree regressor, Random forest regressor, XG boost regressor and Cat boost regressor we got higher accuracy.
- Out of which Random forest regressor, XG boost regressor have more accuracy.
- With the application of Grid search CV on both finally we got more accuracy for XG boost regressor. As r2 =0.97 and RMSE =190.27.

References

1"Short-Term Prediction of Bike-Sharing Demand Using Multi-Source Data: A Spatial-Temporal Graph Attentional LSTM Approach" by Xinwei Ma 1, Yurui Yin 1, Yuchuan Jin 2, Mingjia He 3 and Minqing Zhu 4

2 https://www.computerscijournal.org/vol10no1/prediction-of-bike-sharing-demand/

[3]https://link.springer.com/chapter/10.1007/978-3-030-94751-4_25

4Foell, S., Phithakkitnukoon, S., Kortuem, G., Veloso, M., Bento, C., 2015. Predictability of public transport usage: A study of bus rides in lisbon, portugal. IEEE Transactions on Intelligent Transportation Systems 16, 2955–2960.

5 Gers, F.A., Schmidhuber, J., Cummins, F., 1999. Learning to forget: Continual prediction with lstm .

Thank You…