MACHINE LEARNING

LINEAR REGRESSION

MULTIPLE LINEAR REGRESSION

PROBLEM STATEMENT

- A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free.
- You are required to model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations. Further, the model will be a good way for management to understand the demand dynamics of a new market.

APPROACH

- IMPORTING LIBRARIES
- READING AND UNDERSTANDING DATASET
- DATA CLEANING
- EXPLORATORY DATA ANALYSIS
- DATA PREPARATION FOR LINEAR REGRESSION
- MODEL BUILDING

IMPORTING LIBRARIES

- NUMPY
- PANDAS
- MATPLOTLIB
- SEABORN
- SKLEARN
- STATSMODEL

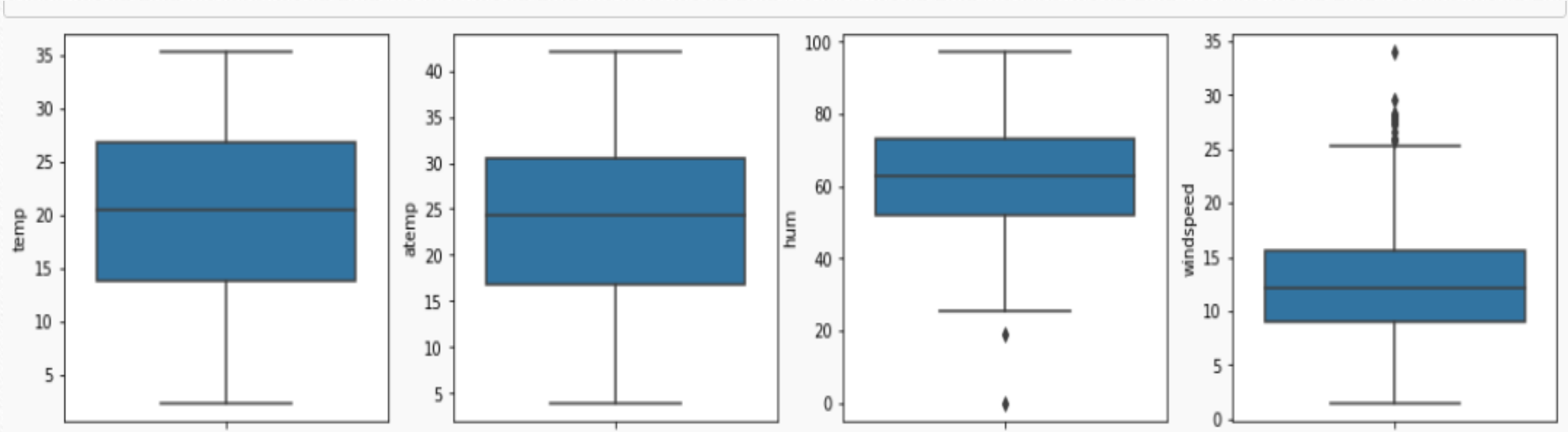
READING DATASET AND UNDERSTANDING

- First we have read the csv file using pandas.
- Then we use certain functions to know more about the data like .shape, .dtype(), .describe(), .head(), .columns() these function help us to understand data.
- Then we check for null values using .info() and we found that there is no null values in csv file.
- After analyzing data we found that there are 730 rows and 16 columns in the dataset.

DATA CLEANING

- Dropping the columns that are not useful for us like instant, dteday, casual and registered.
- Then we handle the missing values and we found that there are no missing values in the data.
- We were used boxplot to figure out the outliers and we found that there are no outliers in the data.

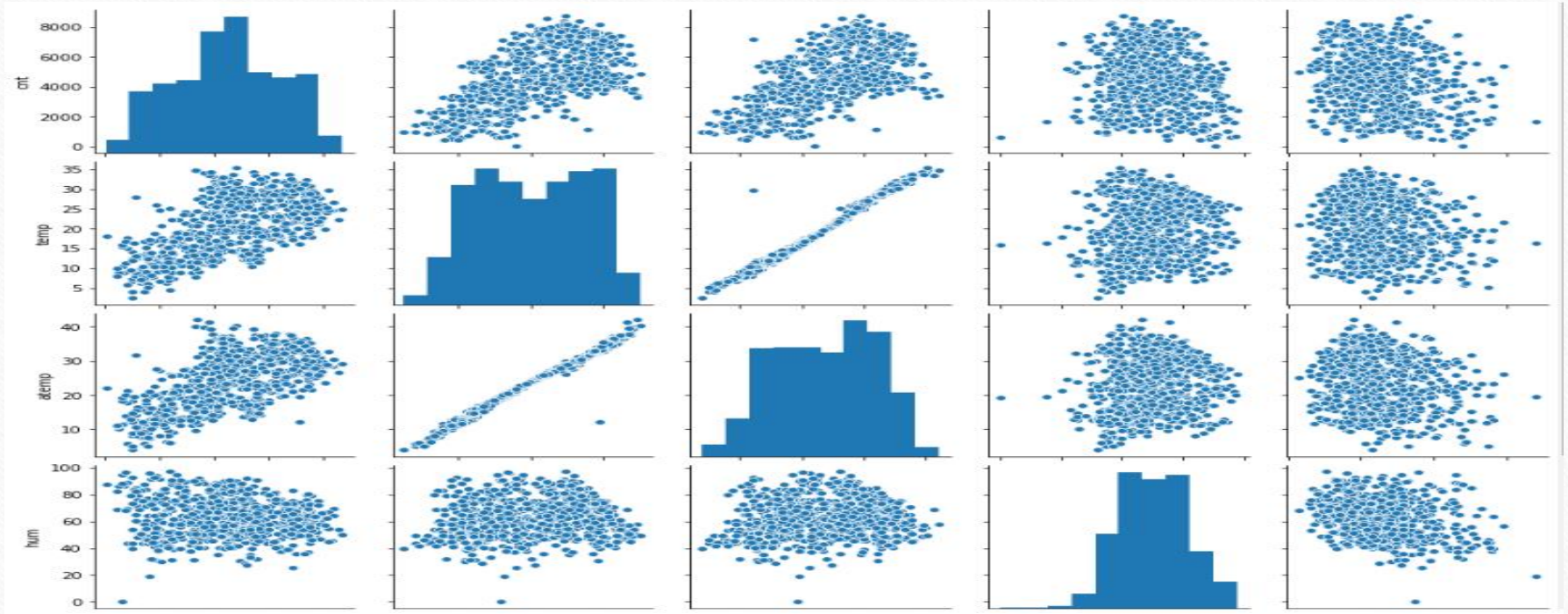
OUTLIERS


temp
35
30
25
20
15
10
5
atemp
40
35
30
25
20
15
10
5
hum
100
80
60
40
20
0
windspeed
35
30
25
20
15
10
5
0

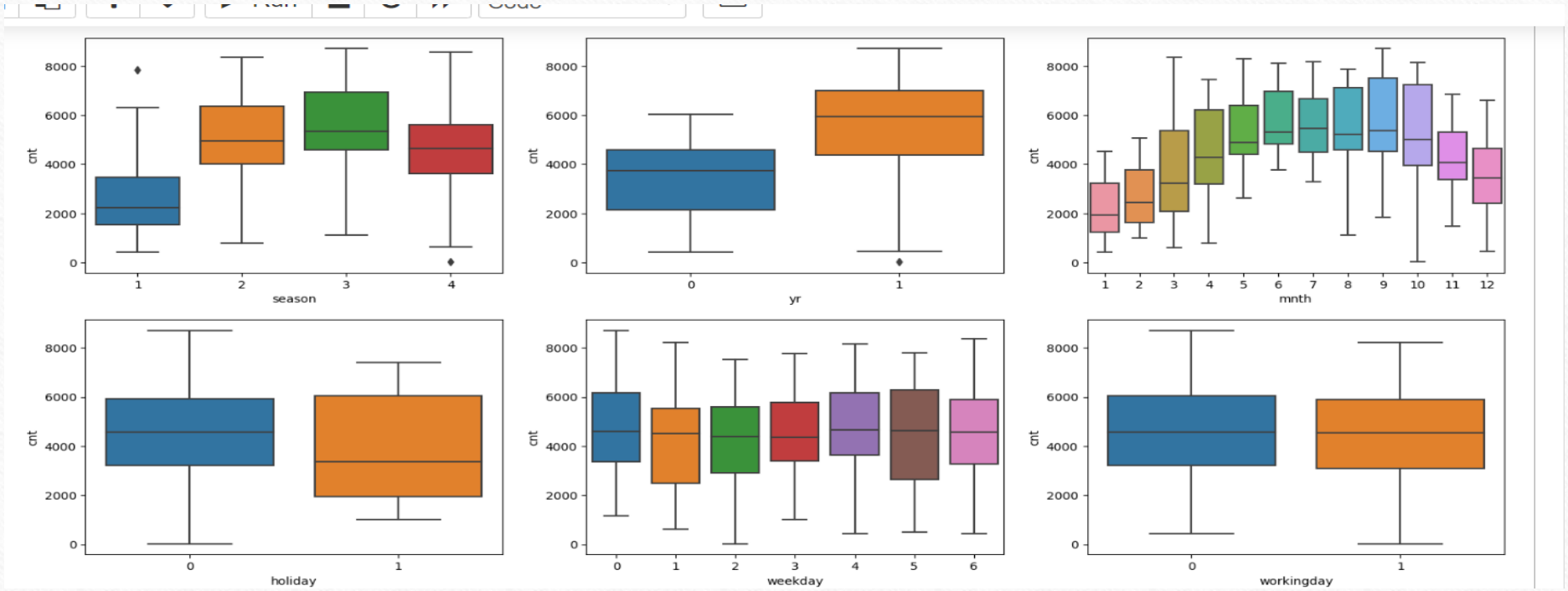
EXPLORATORY DATA ANALYSIS

- First we draw the pairplots to check linear relationship.
- After analyzing the pair plots we found that ‘temp’ and ‘atemp’ has highest correlation with the ‘cnt’ target variable.
- Then we are Using heatmap to know the continuous variable relationship with each other and we found that the “temp” and “atemp” has correlation more than 0.99 means almost 1

EXPLORATORY DATA ANAYSIS



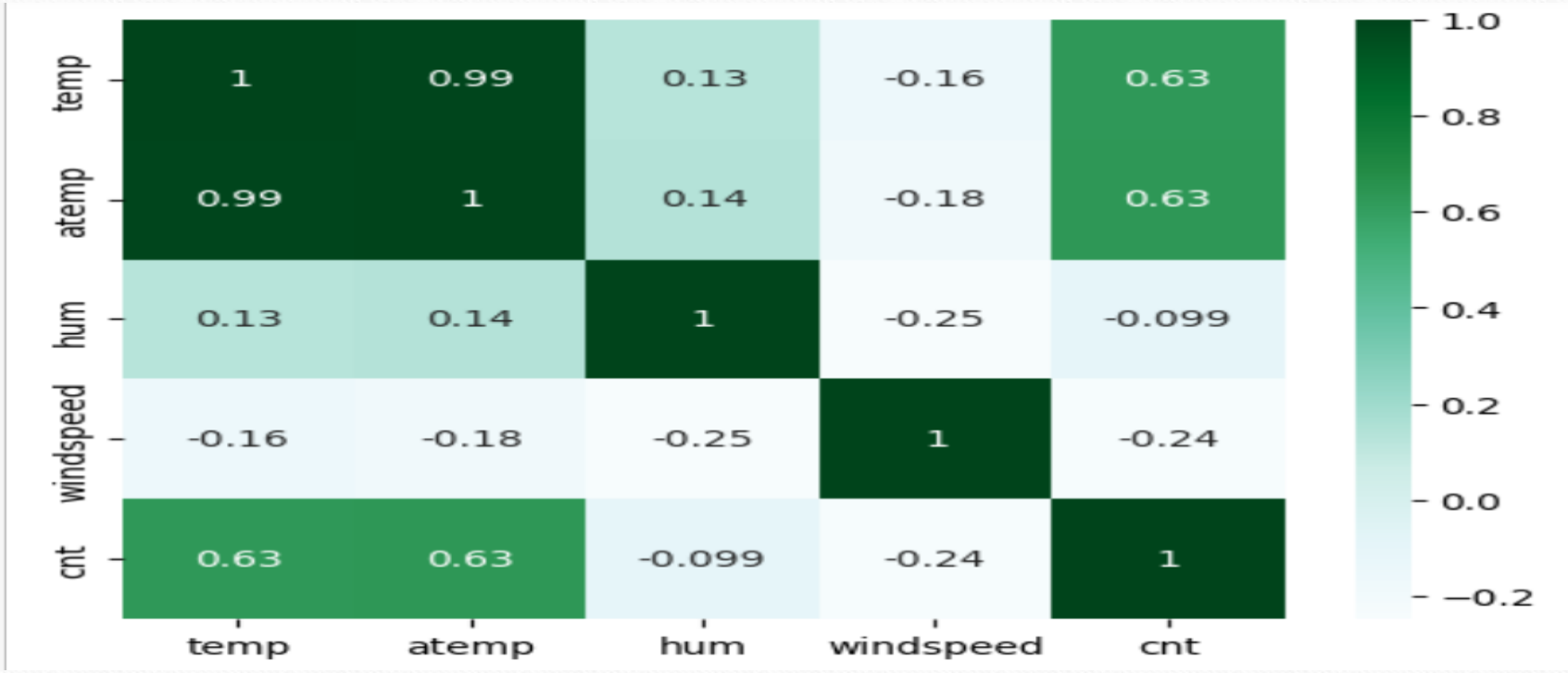
CATEGORICAL ANALYSIS



CATEGORICAL ANALYSIS

- Season 3 has highest demand for rental bikes.
- When there is holiday demand is decreased.
- Weekday is not giving clear picture about demand.
- Demand is continuously growing each month, September month has highest demand .

HEATMAP


temp
atemp
hum
windspeed
cnt
1
0.99
0.13
-0.16
0.63
0.99
1
0.14
-0.18
0.63
0.13
0.14
1
-0.25
-0.099
-0.16
-0.18
-0.25
1
-0.24
0.63
0.63
-0.099
-0.24
1
temp
atemp
hum
windspeed
cnt
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

DATA PREPARATION FOR LINEAR RIGRESSION

- In this section we have create dummies variable.
- Dropping the first columns as (p-1) dummies can explain p categories.
- In weathersit first column was not dropped so as to not lose the info about severe weather situation.

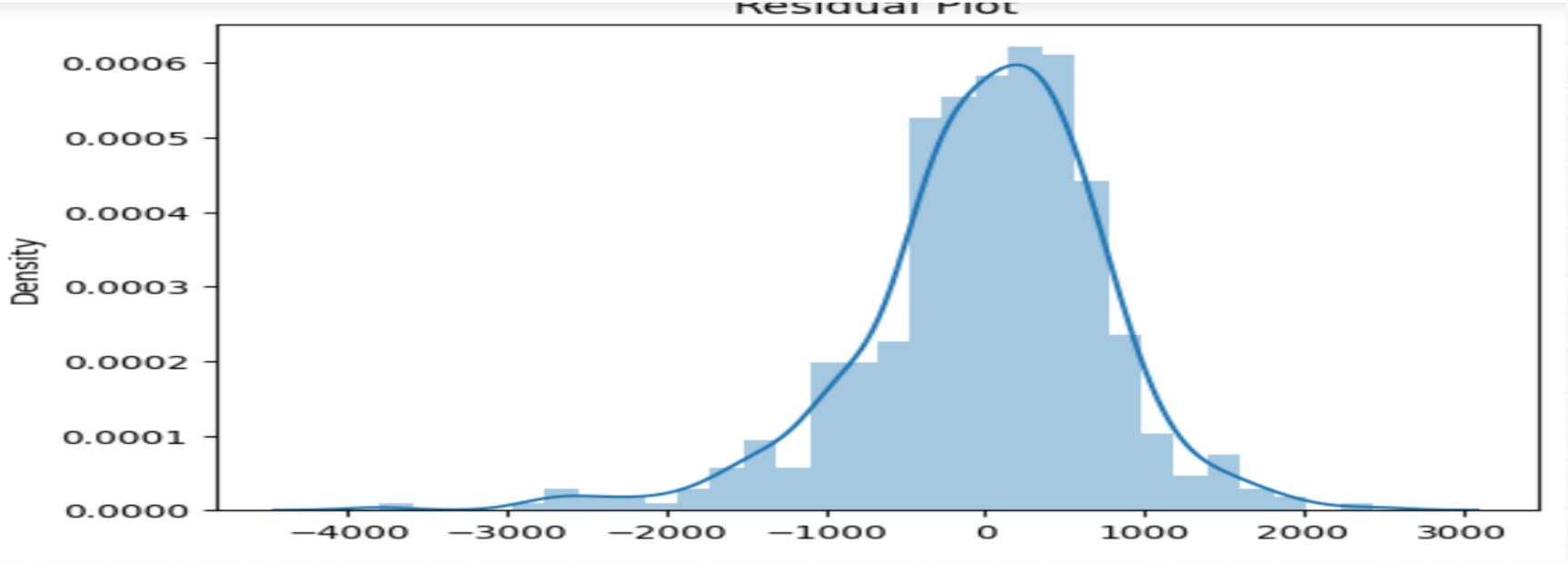
MODELS

- First we build the functions for Model and Variance Inflation factor(VIF).
- Here we have two methods to do it first one is either we can add all the columns and then start dropping one by one .
- Second one is that we can add columns one by one, but I prefer 1st one because it is easy to do.
- Then we drop all the insignificant variables.

MODELS

- After analyzing on heat map we clearly figure out some of the variables have negative values that means insignificant variable we can drop them.
- Then we drop the high Variance Inflation Factor (vif) value variables as its improves the accuracy and the stability of the linear Regression model.
- Then we perform the Residual analysis for the evaluation of the goodness of the model.

RESIDUAL ANALYSIS



CONCLUSION

- Company should expand their business both winter and spring.
- Company should focus on expanding their business during july.
- There would be less bookings during Light Snow or Rain, they could probably use this time to serive the bikes without having business impact.
- Test predicted r2 = 0.78 , Train predicted r2 = 0.847 as per the data we can say our model is good.
- Error are normally distributed around mean = 0 we can clearly seen in the figure.

THANK YOU

DHAMRNEDRA KUMAR YADAV

BATCH – DS53