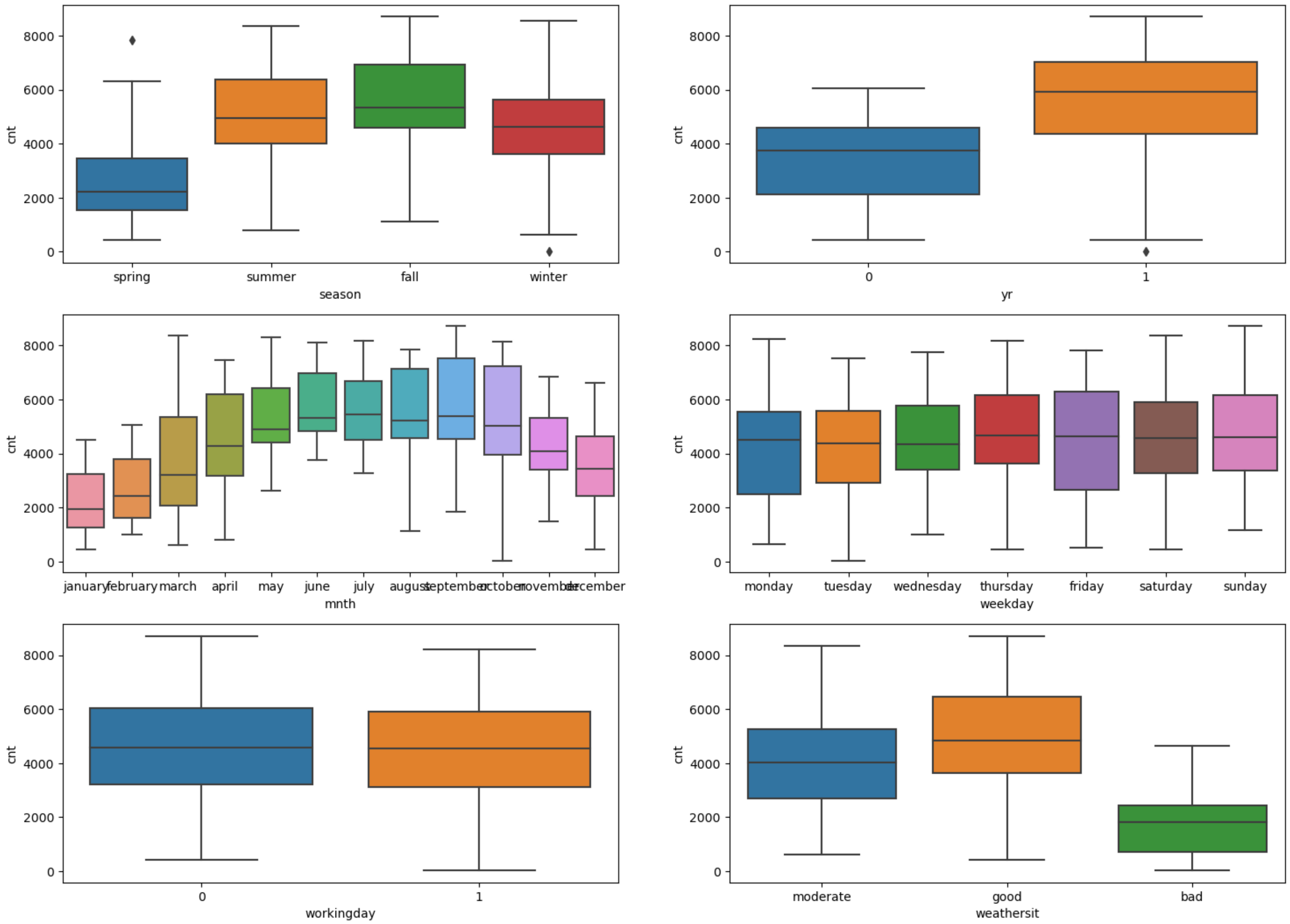
Assignment based - Subjective Questions

Q1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

The categorical variables from the dataset were Season,Holiday,Month,Weather Situation,Working Day,Week Day .For these columns I have plotted boxplot to visualize . These variables have significant effect on dependent variable ie. CNT.

1) Season: The season of Fall has highest median then summer then winter. Lowest median is of spring.
2) Year: Median have increased in year 2019 as compared to 2018.
3) Month: The months from June to October have higher median value than other months.
4) Week day: The bike rentals are more on non-holidays as compared to holidays. This could be due to people reach their workplace by using rental bikes.
5) Working Day:Median for both working day and non working day seems to be same.
6) Weather Situation: Good weather indicates here clear weather situation which has the highest median value.Whereas Bad weather represents snow which has lowest median.

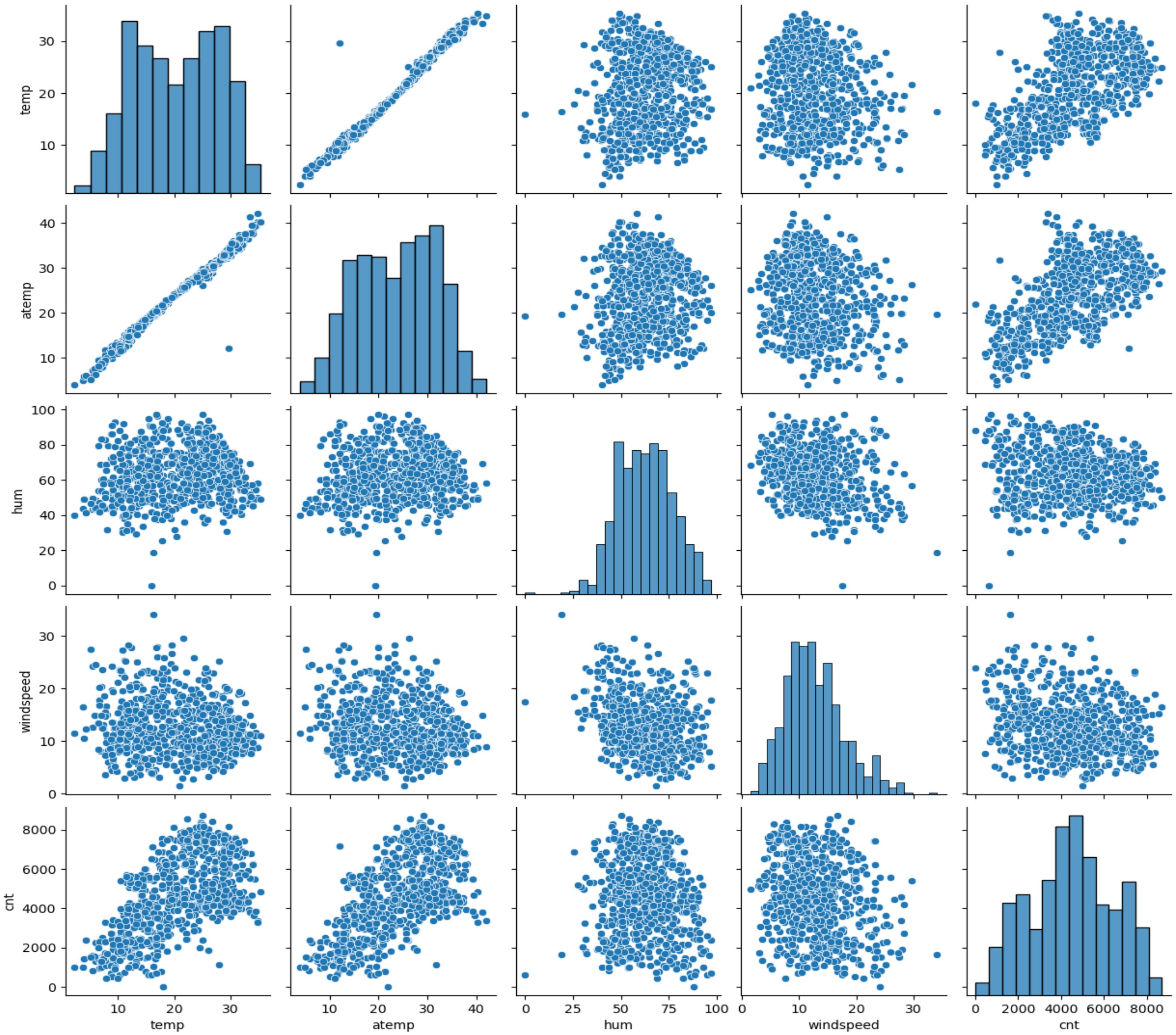

cnt
0
2000
4000
6000
8000
spring
summer
fall
winter
season
cnt
0
2000
4000
6000
8000
0
1
yr
cnt
0
2000
4000
6000
8000
january
february
march
april
may
june
july
august
september
october
november
december
mnth
cnt
0
2000
4000
6000
8000
monday
tuesday
wednesday
thursday
friday
saturday
sunday
weekday
cnt
0
2000
4000
6000
8000
0
1
workingday
cnt
0
2000
4000
6000
8000
moderate
good
bad
weathersit

Q2) **Why is it important to use drop_first=True during dummy variable creation?**

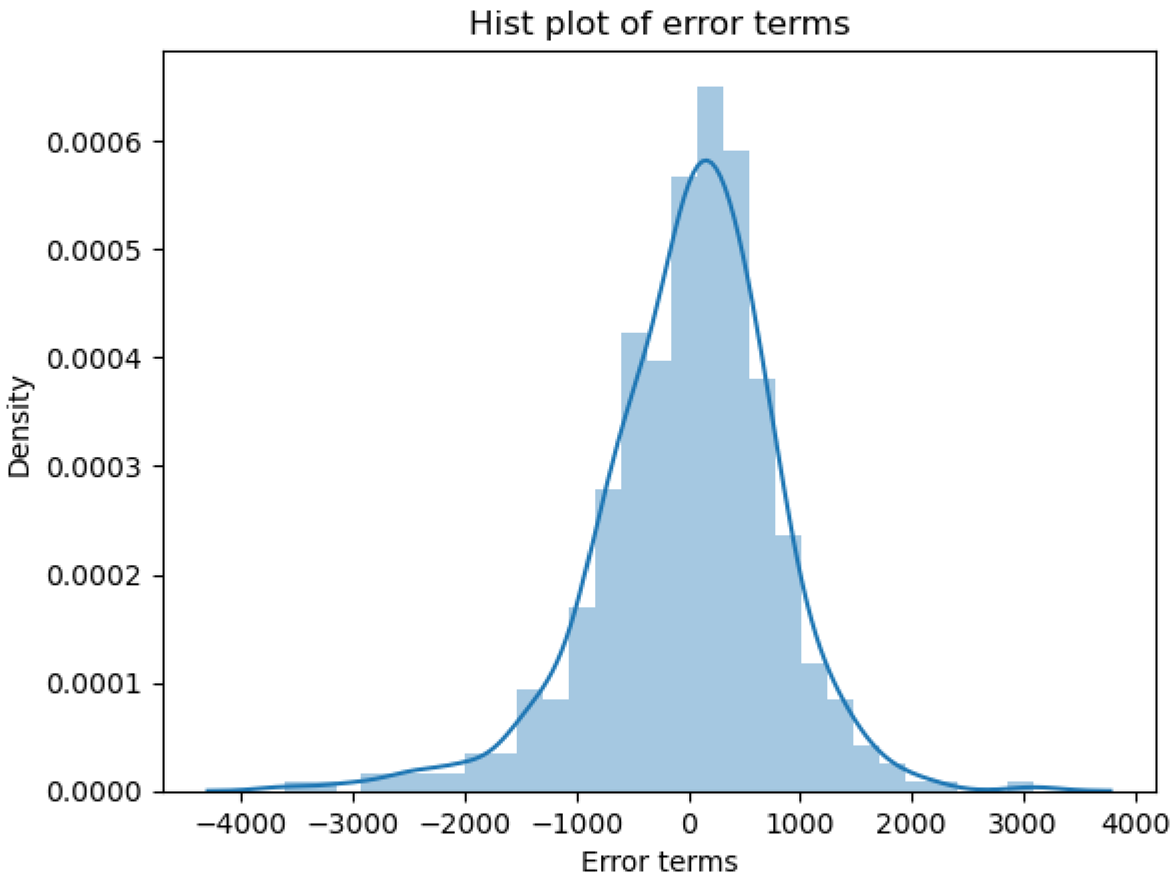
It is important to use drop_first=True during dummy variable creation because while creating dummy variables extra column is created which is of no use. For avoiding Multicollinearity we drop first column .For example If we have 3 types of colour categorical variables such as green , violet, purple when you create dummy variables with drop_first=true so when value with 0 both in violet and purple it indicates that it is of green colour.

Q3)Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

By seeing the below pair plot we can easily see ‘Temp’ and ‘Atemp’ has highest co-relation with target variable ‘CNT’.


temp
atemp
hum
windspeed
cnt
0
10
20
30
40
0
20
40
60
80
100
0
10
20
30
0
2000
4000
6000
8000
10
20
30
10
20
30
40
0
25
50
75
100
0
10
20
30
0
2000
4000
6000
8000
temp
atemp
hum
windspeed
cnt

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?



The distribution of Actual value - Predicted value ie.Residual values are normally distributed as the mean is near zero. I have plotted Histogram for proving the same .

Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The top 3 predictor variables that influence demand of shared bikes are as follows:

1) Temperature(temp): With coefficient of 4274 , it indicates that a unit increase in temp variable the number of bike rentals will increase by 4274

2) Weather Situation(weathersit_bad): With coefficient of '-2476' it indicates that a unit increase in weathersit_bad variable reduces number of bike rentals by 2476 . Whereas weathersit_bad = Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

3) Year(yr):With coefficient of '2034' it indicates that a unit increase in year variable the number of bike rentals will increase by 2034.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear Regression is a type of Machine Learning algorithm that is used for predicting numerical features. Linear Regression is a primary type of regression. With the help of regression we predict values which will be use for future purpose. Linear regression has equation of **‘y = mx + c ’**.It assumes there is a linear relationship between dependent variable and independent variables. Regression is performed when target variable that is dependent variable is of numerical or continuous data type and independent variable is of numerical or categorical data type. Regression module tries to find best fit line for dependent variable and independent variable using square of mean error with minimizing the error terms. Error terms are nothing but difference between actual values and predicted values. Regression is divided into 2 types which are as follows:

1) Simple Linear Regression: When there is only one dependent variable and one independent variable we call Simple Linear Regression.

2) Multiple Linear Regression: When there is one dependent variable and more than one independent variable we call Multiple Linear Regression.

Equation for MLR is as follows:

Y=c+m1X1+m2X2+m3X3+……+mnXn
Where c= intercept
m1=coefficient of X1 variable

m2=coefficient of X2 variable
m3=coefficient of X3 variable and so on

Q2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartel was developed by statistician Francis Anscombe . It was developed to explain the importance of data visualization and limitations based only on summary statistics to understand data. Sometimes the statistical data does not find any outliers in data but there are some in data. Anscombe’s quartel has four data sets that have similar statistical features such as mean,variance,correlation. It proves that sometimes relying on stats features can be deceiving.It has given the importance of data visualization to understand the data and explore through it. It avoids concluding erroneous results.

Q3 What is Pearson’s R?

Pearson’s R or r is a statistical numerical feature which represents a linear relationship between two numerical variables. It’s coefficient ranges from -1 to 1 . It is developed by British statistician Karl Pearson. A value close to -1 indicates that there is a negative slope, value close to 1 indicates that there is a positive slope while value=0 indicates that there is no linear relationship between them. It assumes that variables are normally distributed which means there mean is near zero. One of the feature is it does not imply causation.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a part of pre processing stage where all the variables are scaled into a similar set of range values to avoid any misleading conclusion. It also handles any outliers if present in data. Sometimes data can have very big numbers and very small numbers so the machine learning algorithm will process in that manner which is not appropriate that’s why we do scaling , so that all numbers will be scaled into a similar set of range values. There are various types of scaling some of which are normalized scaling and standardized scaling. Normalized scaling is also known as min-max scaling converts all data between 0 to 1 . On the other hand standardized scaling is also known as z-scale normalization which converts data where mean =0 and standard deviation=1 It has formula as follows: x-mean(x)/std(x).
Difference between normalized scaling and standardized scaling are as follows: In normalized scaling all values are between 0 to 1 whereas in standardized scaling mean will be 0 and standard deviation will be 1.In standardized scaling outlier has been take care of but in min max scaling it is affected by outlier. Normalized can be done when values all variables needs to be in certain range. Standardization helps when algorithm assumes that data is normally distributed.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF stands for Variable Inflation Factor. It is a measure that indicates how two or more independent variable are highly co-related with each other. It can cause problem for interpreting individual variables with respect to dependent variable. VIF formula is given by 1/(1-Rsqured).VIF when is infinity states that there is perfect correlation. Here 1-Rsquared represents how well it is explained by other independent variables . If independent variable has value =1 it will result into VIF=1/(1-1)=1/0 which leads to ‘Infinity’.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plot means quantile-quantile plot, which is a tool for comparing the shapes of different distributions. It helps us to determine whether the data is following any specific pattern or not. In Linear regression it is used to assess the distribution of residuals which is nothing but error terms.It is important because it assumes that residuals i.e. difference between actual values and predicted values, are normally distributed. If the plot is deviated from line it represents that residuals are not normally distributed. If residuals does not follow a specific pattern it represents that there are some anomalies which needs to be take care of.