Assignment-based Subjective Questions

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

The final equation of line I get after the building the model is provided below:

cnt = 0.2309yr- 0.0699holiday+ 0.4665atemp -0.0844windspeed - 0.2911Thunderstom -0.0806Cloudy - 0.1299Spring +0.0350Winter +0.0436March -0.0543July + 0.0685September - 0.0415*Sunday + 0.2478

Interference Obtained

- The demand of the bike increases by 23.09% in 2019.
- Feeling temperature plays a key role in the bike count and it increase the demand by 46.65%
- weather situations which includes Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds decreases the bike counts by 29.11%. Similarly, cloudy weather like Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist decreases the count by 8.0 6%.
- Windspeed also decreases the demand by 8.44%.
- Season like Spring decrease the demand by 12.99% whereas winter season increase the bike count by 3.5%.
- Month also depend on bike demand, it is evident from the equation that March and September the count increases by 4.36% and 6.85% whereas the demand decreases in July.
- The demand of the bike decrease approximately by 4% on Sunday.
- The demand decreases on Holidays by around 7%.

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

The answer to this is best explained with an example. Suppose, we have a variable which give the direction name (east, west, north, south).

Then after creating dummies for the above it will create 4 variable one for each direction name. Suppose in the dataset if a record has direction name as north then it will values will be 1 for north and 0 for other 3.

If we do drop first then it will create only 3 variables and drop any one direction name let say it dropped east then if in dataset the record has direction east then the coding for variable will be 000.

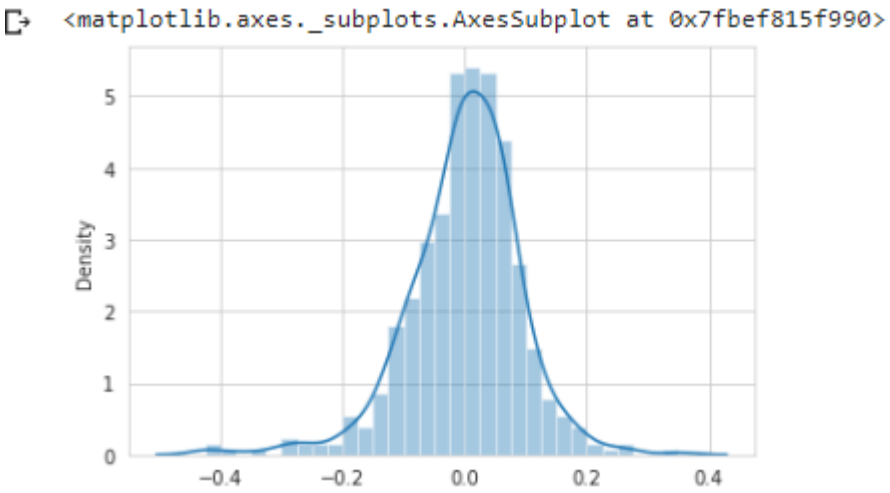
This also help to reduce the number of variables.

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

Registered

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

The distribution of error term are normally distributed.



The target and the independent variable are linearly dependent.

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

- Year(yr) 23.61% positive relation
- Feeling temperature (atemp) 44.25% positively correlated.
- Weather situation(weathersit)with Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 28.41 negatively correlated.

General Subjective Questions

1.) **Explain the linear regression algorithm in detail.**

Linear Regression is a machine learning algorithm which works on regression tasks(numerical dependent variable).

Regression models a target prediction value based on independent variables.

General Linear regression line when there are 2 feature

Y = b1x1 +b2x2 +b0

Where b0 = intercept, b1 and b2 are the coefficients of x1 and x2 which also shows the importance of each variable on Y.

2.) **Explain the Anscombe's quartet in detail.**

Anscombe's quartet demonstrate the importance for visualizing the data before applying the dataset to build the model. It help to visualize the abnormalities present in the dataset like outliers and wheather there is a linear relationship exists.

Anscombe's quartet is a group of 4 dataset which are very similar in their statistical properties but different when plotted on graph.

3.) What is Pearson's R?

Pearson Rank Coffecient is used to measure between the 2 random variable. The coefficient lies between -1 to +1.

Properties:

If the value is 0, there is no relationship.

Limitation:

a) Not good for non linear relationship.
b) The slope in graph doesn't matter.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

Scaling is a process of getting rid of the scale. Suppose if we have 3 independent numerical variable like age, salary, height with their range as (18-100 years), (25000-75000 USD) and (1-2 meters) respectively. Then feature scaling will be helpful to get them all in same range. Normalization scaling will scale all the feature between 0 to 1 and standardization scaling will scale the feature with mean 0 and 1 std deviation.

Use Normalization when your dataset has outliers and standardization when you care about the distribution.

Normalization: X_new = (X - X_min)/(X_max - X_min

Standardization: X_new = (X - mean)/Std

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

The Value of VIF is infinite only when there is perfect correlation between 2 independent variable.

If the variables are perfectly correlated their R2 is 1 therefore, 1/1-R2 = infinite.

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

Q-Q plot is used to check whether the random variable X follow a certain distribution(normal, powerlaw etc). Here we check whether the provided random variable is normally distributed or not. It is done by taking percentile of the provided random variable X and comparing with other Normally distributed random variable. Suppose, X is a random variable of sample size 500 then 500/100 = 5 .

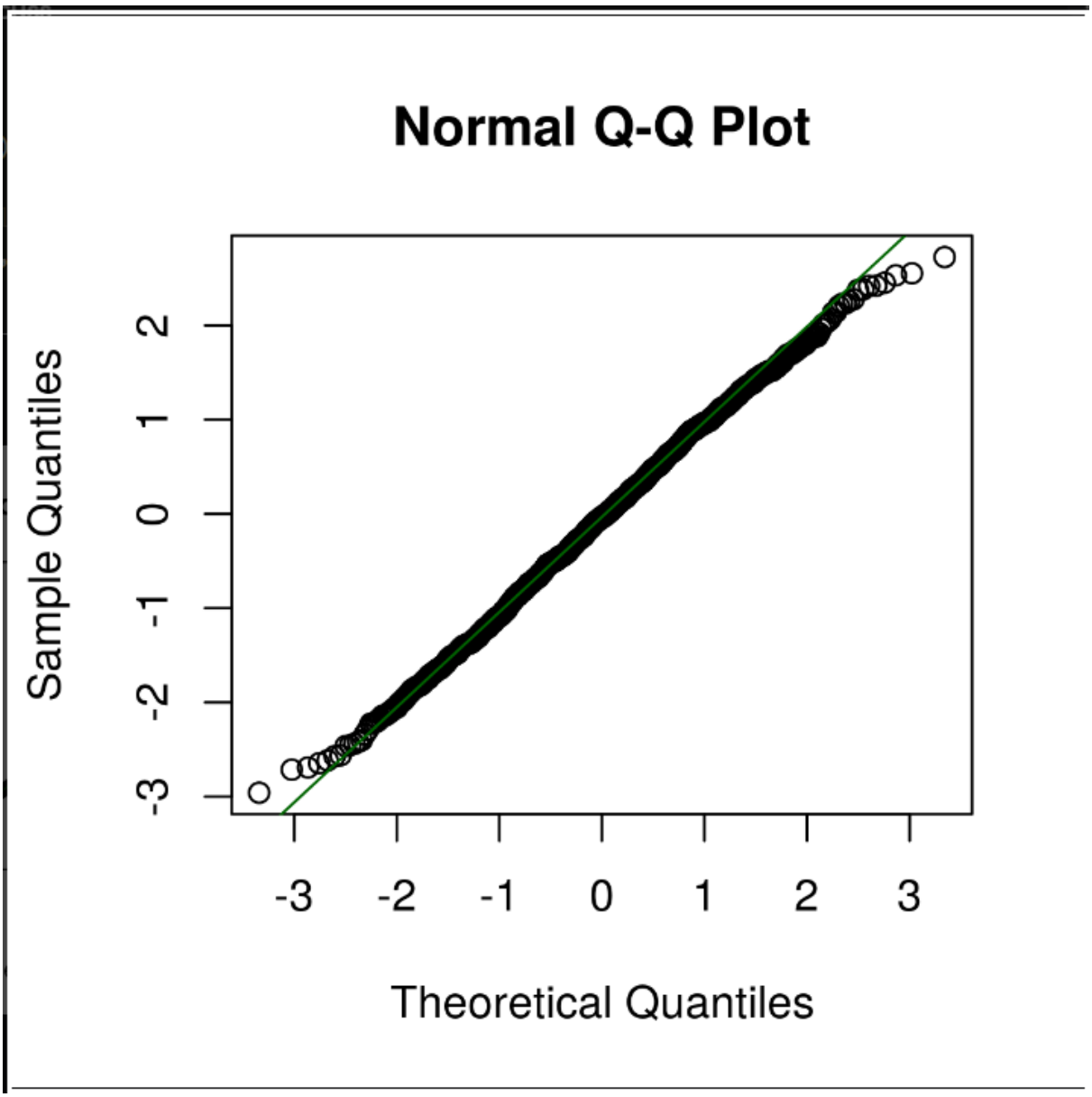
Then x1 will be the 5^{th} percentile of X, x2 will be 10^{th} percentile and so on till the 100^{th} percentile.

Then we take/generate a random variable Y of same sample size which is normally distributed N(0,1). Take percentile the same way we took for X.

Then plot x1, x2, x3….x100 on x-axis and y1,y2,y3… y100 on y-axis.

If they both lies roughly on the same line then the randon variable X is Normally distributed.

Similarly we can check for other distribution as well.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-3
-2
-1
0
1
2
3