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? (3 marks)

Ans. From the model, it can be inferred, that,

- Const has a negative effect as its value is negative
- 2019 (yr) has a negative effect as its value is negative
- Oct in month with value 8 has a negative effect as its value is negative
- Hum has a positive effect as its value is positive
- Casual has a positive effect as its value is positive
- registered has a positive effect as its value is positive
- Dec in month with value 12 has a positive effect as its value is positive

```
lm.params
```

```
const         -6.314393e-16
hum            8.326673e-16
casual         3.774735e-01
registered     7.968247e-01
2019          -7.285839e-17
Oct           -5.898060e-17
Dec            1.179612e-16
dtype: float64
```

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

Ans. drop_first=True is used when we convert the categorical data into dummies and to drop the first column of the dummies created as for n type of catefories we need only n-1 of columns as otherwise it would add to the processing power and redundancy of the calculation which could be limited in large datasets

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans. registered column has the highest correlation of 0.95

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans. After building the model, it is to be checked if the features probability t-score is less that 0.05 to have it significant and also in VIF if it is less than 3. Then if all the conditions are satisfied we check he residual curve, it is follows normal distribution around mean 0

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans. hum, registered and yr

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans. Linear Regression is used to predict the relationship between the dependent variable and independent variables (that is independent among themselves) on whom the dependent variable depends. It tries to fit a line on whose y axis is the dependent variable. Along with the independent variables or X values, there is an intercept value or constant value and thus they together are similar to line equation of y=mx+c with m as the slope, and linear regression equation sets to $Y=\beta0 + \beta1X1 + \beta2X2 + ….. + \beta nXn$ with β as the slope.

Linear regression can only be used in range specified and it is not good to predict extrapolation

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans. Anscombe's Quartet shows the importance of plotting or visually analysing the dataset before proceeding with the model building. This can be understand as for example, take four datasets, with similar statistical features that is similar mean, variance and thus it can fool the engineer that the data is similar, but what can happen is that we have an outlier or other values which shows that they are same. So, visualizing data helps us understand how the data actually represents and if indeed they are similar or very different but with same statistical features.

3. What is Pearson's R? (3 marks)

Ans. Pearson's R value is the representation of the correlation between two features in the range of -1 to 1 where, values at the both extremes of the range represents high correlation let it be negative or positive and values near 0 represents no to less correlation that is presence of these features doesn't affect the other feature in any way. The higher values means that the both features can be represented with straight line with either positive or negative slopes.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans. Scaling is to convert the qualitative values to a scale or space. It is used as when the values of one of the feature is too large, it makes other smaller values look insignificant for the regression model and so it behaves as only changes to this value leads to changes to dependent value which might not be the real case.

In normalized scaling, the values are scaled between 0 and 1, thus the range of the values are kept same for all the features.

In standardized scaling, the quantitative features are scaled such that there standard deviation is 1 and they are cantered on the mean which is 0. Thus all features have same standard deviation of 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans. VIF is the relation in which we try to find how much the feature is correlated with other features. We do VIF and try to keep it low and remove features with high VIF as we want to have the

features as much independent of each other to find a dependent variable which is the gist of Linear Regression.

$$VIF = \frac{1}{1 - R_i^2}$$

Here, as we can see, VIF is dependent upon R-squared of the feature with respect to other features. When we have a R-squared value as 1, that is the feature is perfectly correlated with others, we get VIF as infinity. Thus VIF infinity refers to a feature which can be perfectly represented by other features, and thus it can be removed.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans. Q-Q plot or Quantile-Quantile plot is used to represent the similarity between the two populations of the two data sets, the plot represents the relationship between the two datasets with respect to 45 degree of the x-y line, where if the visualization s over the line shows, the normal distribution between the two datasets, or if they are far from straight line shows skewness or if completely away, no such distribution exists.