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)

- Fall season has the highest demand followed by Summer
- With respect to year 2019 has more demand than 2018

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

- To avoid unnecessary extra column

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

- 'registered' has highest correlation followed by 'casual'
- essentially because these two sum up to result in the target variable i.e, 'cnt'

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

- Scatter plot proves the point
- The relationship between X and Y is linear

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

- temperature
- season
- weather

General Subjective Questions

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

- Linear regression algorithm is a type of machine learning algorithm used for predictive analysis.

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

- A set of four datasets that are nearly similar in simple descriptive statistics is called Anscombe's quartet, when these datasets also have some peculiarities that would trick regression models at build time.

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

- Pearson's R is a measure used to measure the connections such as numerical between two continuous variable.

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

and standardized scaling? (3 marks)

- Scaling is part of data preparation, used to normalize independent variables

- In order to Normalize the data between measurable values, usually to speedup calculations

- Normalized is used to bring down values between 0 and 1, whereas Standardized brings all of the data into a standard normal distribution which has mean, zero and standard deviation 1.

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

(3 marks)

- Basically means perfect correlation between 2 Variables (independent)

- Simple, Rsq becomes 1, 1/(1-R2) becomes infinity

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

(3 marks)

- Q-Q plot also known as Quantile-Quantile plots, is a technique used to plot the quantiles of a sample distribution against quantiles of a theoretical distribution.

- Q-Q plots are commonly used to compare a data set to a theoretical model.