# 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)

A – Rental was more in the year 2019, rental was significantly more on holidays, rental was significantly similar on weekdays, rental is significantly more from June to October and the highest in the month of September, rental is significantly higher when the weather is clear.

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

A – It is done for 2 reasons, first is to make sure that no extra column is created during the creation of dummy variable. Secondly, to make sure that there is minimum correlation between these variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark)

A – ‘registered’ variable has the highest correlation with our target var ‘cnt’.

4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

A – After building the model the 2 things that we are to check from the summary and calculations are P value wiz the significance of the variable and the VIF of the variable that gives the multicollinearity with other variables. Ideally P value has to be less than 0.05 and the VIF has to be less than 5.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

A – Temperature, Season (Summer, Winter, Spring) and Weather (Mist + Cloudy, Light Snow).

General Subjective Questions

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

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

A – Anscombe’s quartet states the importance of visualizing the data before analysing it and building model. It states caution that different sets of data can have peculiarities that can fool a linear regression model or can depict nearly identical simple descriptive statistics.

3. What is Pearson’s R? (3 marks)

A – Pearson’s R is the most common way of measuring linear correlation, it’s measure lies between -1 and 1 that measures the strength and the type of correlation between the target and the other variables.

-1 means strong -ve correlation, 1 means a strong +ve correlation.

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

and standardized scaling? (3 marks)

A – Scaling is a way by which we normalize the data within a particular range. Each time data is collected the unit of measure and the magnitude of the data varies vastly. In order to uniform it we scale it so that the whole data is confined within a very small range making it easier for calculations and plotting.

Standardised scaling rings all data by their respective Z score. The data scaled by this method has its mean 0 and standard deviation 1.

Normalised Scaling brings all data in the range of 0 and 1. This is also known as min max scaling.

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

(3 marks)

1. It is a sign of very high correlation between the 2 variable. To proceed further with model building we have to dop this variable that is causing this very high correlation.

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

(3 marks)

A – It is a quantile quantile plot. The main purpose of a quantile plot is to find out if the distribution of the 2 different set of data is common or not. A 45 degree line will be plotted if the 2 data sets have common distribution.