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. We can see that the demand for bikes is more in fall and summer compared to other seasons.
   B. We can see that the demand for bikes is more when the weather is Clear, Few clouds, Partly cloudy, Partly cloudy.
   C. We can see that the demand for bikes is more between May and October.
   D. We can see that the demand for bikes is more during the year 2019 as compared to 2018.

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

   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

   Variable registered has the highest correlation with the target variable.

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

   The simple way to determine if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

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

As per the final model below are the top 3 features contributing significantly towards explaining the demand of shared bikes

Temperature (0.049)

Year (0.0082)

Weekday_Tuesday (0.0039)

General Subjective Questions

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

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear, that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as: y = a + bx

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

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

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit

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

It is a step of data Pre-Processing which is applied to independent variables to normalise the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then the algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Normalisation typically rescales the values into a range of [0,1]. Standardisation typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.