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)

   - Categorial variables can help to analyze data on different basis, like period of year when bikes have used the most with temperature and season and months.
   - Above information helps to understand exactly when people are comfortable to ride a bike or at which temperature as well.
   - Season or temp variable affects a lot on temp, hum & windspeed variables and by default then affects the overall count as well.

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

   - Drop_first=True helps to reduce the extra columns created while creating the dummy variables, so it reduces the correlation.

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

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

   - If line passes from X & Y bar then it's correct and that's how we validate it.

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

   1. Demand of bike is high in summer or rainy season.
   2. Demand increases on the holidays so supply can be added in that time.
   3. Demand increases on weekends as well

General Subjective Questions

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

   - Linear Regreassion is Machine Learning Algorithm based on Superwised Learning which performs the regression tasks.
   - Use to find the relationship between Independent & Dependent variables.
   - If input is only one then it's Simple Linear Regression but if inputs are more than one then it's Multiple Linear Regression.
   - If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.

- If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.

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 analyzing it and the effect of outliers on statistical properties.

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

- Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

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

- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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 lead 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 (which show an infinite VIF as well).

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.