**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)*

--The demad of bike is less in the month of spring when compared with other seasons. The demand bike increased in the year 2019 when compared with year 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)*

--The numerical variable 'registered' has the highest correlation with the target variable 'cnt' , if we consider all the features.

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

--Linear functional form: The response variable y should be a linearly related to the explanatory variables *X*.

Residual errors should be i.i.d.: After fitting the model on the training data set, the residual errors of the model should be independent and identically distributed random variables.
Residual errors should be normally distributed: The residual errors should be normally distributed.
Residual errors should be homoscedastic: The residual errors should have constant variance.

*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 (temp)
Weather Situation Moderate
Year (yr)

**General Subjective Questions**

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

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

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

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation Coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope ( i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope ( i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

*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.

Standardization brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

*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)*

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.