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)**

   → Categorical variables also affect the dependent variable. Before using categorical variable, the same needs to be encoded, using methods like one hot encoding. In case of regression analysis, categorical variables cannot be used for identifying the trend however these variables help in understanding clusters of the data.

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

   → get_dummies generates n variables for n categories. However, first variable can be identified by all others. Therefore, to avoid duplicate information drop_first=True is important to be used.

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

   → Based on the scatter plot the temperature variables had highest linear 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)**

   i. After predicting values with the trained model, residue was calculated. The histogram of the same showed normal distribution

   ii. R-squared value was checked to ensure that it has a higher value.

   iii. P-values of all the predictor variables was checked to be less than 0.5

   iv. VIF (Variance inflation factor) was checked to ensure there is minimum collinearity amongst the predictor variables.

   v. After finalizing model, the model was used on test data set to predict the values. R-squared was computed again using actual dependent variable from test data set and predicted data set.

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

   i. Actual temperature (atemp)

   ii. Windspeed (windspeed)

   iii. Season (season)

General Subjective Questions

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

   Linear regression is a method of training model based on data and using the model to predict new values. In case of linear regression two or more variables are considered out of which one is dependent variable and others are predictor variables. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

   Mathematically, linear regression can be written as,

$$y = mx + c$$

   Where,

m is slope
c is y - intercept
x is predictor variable
y is dependent variable
Typically following steps are followed for Multivariate linear regression analysis:

i. Data preparation: Data cleaning methods are applied to update / remove variables with insignificant values.

ii. Model building: Once data is ready it is divided in to training and test data set. The training dataset is then applied with scaling. The model is built on training data set. Post which it is applied on the test dataset to check validity.

iii. Residue analysis: Once predicted values are ready, the difference between actual and predicted values is calculated.

iv. Prediction: if the residue analysis is satisfactory the model is used to predict future or upcoming values.

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

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed (scatter plot).

The quartet was constructed to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties.

Currently it is not known how Anscombe creates datasets.

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

The 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.

Pearson r Formula

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Here,

r	=correlation coefficient
x_i	=values of the x-variable in a sample
$\bar{x}$	=mean of the values of the x-variable
y_i	=values of the y-variable in a sample
$\bar{y}$	=mean of the values of the y-variable

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

In machine learning, feature scaling refers to putting the feature values into the same range.

Scaling brings all the variables on same scale, thus making it easy for the model to understand the variables. This also reduces bias, due to variables having very high values.

There are two major methods to scale the variables, i.e. standardisation and MinMax scaling. Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one. MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1. The formulae in the background used for each of these methods are as given below:

- Standardization: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - \min(x)}{\max(x) - \min(x)}$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

   In this case, there were multiple variables derived by using encoding, due to which there was a very high correlation between encoded variables. Once one of the encoded variable was removed the VIF value used to come down.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
   The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

   The Q-Q plot is used to help assess if a sample comes from a known distribution such as a normal distribution. For regression, when checking if the data in this sample is normally distributed, we can use a Normal Q-Q plot to test that assumption.