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)

Ans: A categorical variable has levels which rarely occur. Many of these levels have minimal chance of making a real impact on model fit

For example for the categorical variable 'weathersit' one of the level "Heavy Rain" not present in data.

the main effect of the categorical variable is comparable to the difference in the y-intercepts.

For example, The main effect of the "yr" variable is comparable to the difference in the y-intercepts.

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

Ans: 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)

Ans.

The 'temp' variable among the numerical variables 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)

Ans.

The linear regression has five key assumptions and the following steps were performed to validate each of them

- **Linear relationship**

  Plot of actual vs predicted values is linear
- **Multivariate normality**

  The distplot indicates that the residuals are normally distributed.
- **No or little multicollinearity**

  Using VIF on the predictor variables showed very little multicollinearity
- **No auto-correlation**

  - Using durbin_watson method on the residuals showed no auto correlation.

- **Homoscedasticity**
  The residual plot showed 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)

Ans.

The top 3 features are “yr”, “weathersit” and “temp”. This was determined from the t values. A higher t values indicate variable is significant

General Subjective Questions

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

Ans.

Linear regression is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable

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

Ans.

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician ***Francis Anscombe*** in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

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

Ans. 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.

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

Ans.

Feature Scaling also known as data normalization is a data preprocessing step. It is a method used to normalize the range of independent variables or features of data.

It is performed so that model does not confuse a feature with a larger magnitude as a better one. Feature scaling in Machine Learning would help all the independent variables to be in the same range, for example- centered around a particular number(0) or in the range (0,1), depending on the scaling technique.

Standardization is a technique is to re-scale features value with the distribution value between 0 and 1. The result of standardization (or Z-score normalization) is that the features will be rescaled to ensure the mean and the standard deviation to be 0 and 1, respectively.
Normalization technique is to re-scales features with a distribution value between 0 and 1. For every feature, the minimum value of that feature gets transformed into 0, and the maximum value gets transformed into 1.

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

Ans. In regression analysis, the variance inflation factor (VIF) is a measure of the degree of multicollinearity of one regressor with the other regressors. The greater the VIF, the higher the degree of multicollinearity.

In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.

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

(3 marks)

Ans.   A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

A Q-Q plot tells us whether a data set is normally distributed