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 1. For the Categorical Dataset like Season (Spring, Winter & Summer) and Light_snowrain, Misty has negative coefficient. Which means when the value of these independent variables increases the value for dependent variable increase.

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

Ans 2. "drop_first=True" drops the first dummy variable , As the dummy variables are themselves correlated , it leads to multicollinearity .

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

Ans 3. "Atemp" variable has the highest correlation with target variable.

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

Ans 4.

- **RFE is normally distributed.**
- **There is no multicollinearity between the independent variables. VIF is less than 5 for each independent variable and eliminated the once which have VIF higher than 5**
- **Error terms are independent of each other/**

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 5. Based on the final model the top 3 feature contributing significantly toward the demand of shared bikes are , Spring season, light_snowrain weather situation and year.

General Subjective Questions

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

Ans 1.

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features.

Types of Linear Regression

There are two main types of linear regression:

- **Simple Linear Regression – Involves one independent variable**
- **Multiple Linear regression – Involves more than one independent variable**

  - ✓ **The equation for simple linear regression is:**

    $y = \beta 0 + \beta 1X$

    where:

    y is the dependent variable & X is the independent variable

    β0 is the intercept

    β1 is the slope

  - ✓ **The equation for simple linear regression is:**

    $y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + ... \beta nXn$

    where:

    y is the dependent variable & X1, X2, X3 …Xn are independent variable

    β0 is the intercept

    β1, β2, … βn are slopes

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

Ans 2. Anscombe's quartet is a set of four datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same

summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

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

Ans 3. Pearson's Correlation Coefficient, often denoted as r, measures the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

The formula is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

where X_i and Y_i are individual data points, and $\bar{X}$ and $\bar{Y}$ are the means of the variables.

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

Ans 4. 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 must do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling: It brings all the data in the range of 0 and 1.

Standardization Scaling: 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)**

Ans 5. A large value of VIF of a variable indicates that the variable has high collinear relationship with other variables. When there is perfect correlation of a variable with other variable then VIF becomes Infinite.

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

Ans 6.

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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Key advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution.

ii. have common location and scale

iii. have similar distributional shapes.

iv. have similar tail behavior.