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)

    - Bike demand doesn't change whether day is working day or not
    - The demand of bike is almost similar throughout the weekdays
    - Bike demand in the fall is the highest
    - Bike demand takes a dip in spring
    - Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow
    - Bike demand is high in the months from May to October
    - Bike demand in year 2019 is higher as compared to 2018

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

It helps in reducing the extra column created during dummy variable creation, this in turn will reduce the no of independent variables. The individual effect of the dummy variables on the prediction model cannot be interpreted well because of multicollinearity.So it helps in dropping redundant dummy variables.

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

Based on heat map and pair plot we can come to a conclusion that atemp and temp variables have highest correlation of 0.63 against the target variable

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

    - Less multicollinearity amongst the variables using VIF
    - Normal distribution of error terms by using distribution plot.(y_train - y_train_pred)
    - Constant variance of errors
    - Linear relationship between independent and predictive variables

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

    - Light snow – Highest Negative co-efficient meaning affects the demand negatively.
    - Temp – Affects the demand positively.

- Winter – Affects the demand positively.

General Subjective Questions

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

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting linear relationship (line) that describes the relationship between the variables. In simple linear regression, there is only one independent variable, while in multiple linear regression, there are two or more independent variables.

Assumptions:

Linearity: The relationship between the independent and dependent variables is assumed to be linear.

Independence: Observations are assumed to be independent of each other.

Homoscedasticity: Residuals (the differences between observed and predicted values) should have constant variance across all levels of the independent variables.

Normality: Residuals are assumed to be normally distributed.

Objective Function:

The goal is to minimize the sum of squared differences between the observed and predicted values.

This is known as the least squares method.

Parameter Estimation:

The coefficients are estimated using various methods, such as the least squares method.

The coefficients are determined in such a way that they minimize the objective function.

Model Evaluation:

Common metrics for evaluating the performance of the model include:

R-squared : Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

Mean Squared Error (MSE): Measures the average squared difference between observed and predicted values.

Making Predictions:

Once the model is trained, it can be used to make predictions on new data by substituting the values of the independent variables into the regression equation.

Interpretation:

The coefficients represent the strength and direction of the relationship between the independent

and dependent variables.

Assumptions Checking:

It is essential to check the assumptions of linear regression, including linearity, independence, homoscedasticity, and normality of residuals

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

Anscombe's quartet is a famous example in statistics that consists of four distinct datasets, each with 11 data points. It was created by the British statistician Francis Anscombe in 1973 to emphasize the importance of graphical exploration and visualization in understanding data. The quartet is remarkable because, despite having very different statistical properties, the datasets have nearly identical simple descriptive statistics when examined on their own.

Here are the four datasets within Anscombe's quartet:

Dataset I:

X values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
Y values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
Dataset II:

X values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
Y values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
Dataset III:

X values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
Y values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
Dataset IV:

X values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
Y values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89
When you calculate basic statistics for each of these datasets (mean, variance, correlation coefficient, regression line, etc.), you will find that they are very similar or even identical. However, when you plot these datasets, you will notice significant differences in their distributions and relationships between the variables.

The key takeaway from Anscombe's quartet is that relying solely on summary statistics can be misleading, as it may not capture the true nature of the data. Visualizing data through scatter plots and other graphical representations is crucial for gaining a more comprehensive understanding of the data and identifying patterns, outliers, and relationships that might not be apparent from summary statistics alone. This highlights the importance of data exploration and visualization as an essential step in statistical analysis.

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

Pearson's r, also known as the Pearson correlation coefficient or Pearson's correlation, is a statistic used to measure the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson in the early 20th century and is widely used in statistics and data analysis to quantify how closely two variables are related to each other.

The Pearson correlation coefficient is a value between -1 and 1, where:

1 indicates a perfect positive linear relationship: As one variable increases, the other also increases in a linear fashion.
-1 indicates a perfect negative linear relationship: As one variable increases, the other decreases in a linear fashion.
0 indicates no linear relationship: There is no systematic linear relationship between the variables.

Pearson's r is a measure of linear correlation, and it assumes that the relationship between the variables is linear. It is sensitive to outliers and can be influenced by extreme values in the data. If r is close to 1 or -1, it suggests a strong linear relationship, while r close to 0 indicates a weak or no linear relationship. When r is positive, it suggests a positive linear relationship, and when r is negative, it indicates a negative linear relationship.

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

Scaling is a preprocessing step in data analysis and machine learning that involves transforming the features of a dataset to a standard range. The purpose of scaling is to ensure that all variables contribute equally to the analysis and that no variable dominates due to its scale. In many machine learning algorithms, the scale of the features can affect the performance and convergence of the algorithm.
Scaling is performed for several reasons:

Algorithm Sensitivity: Some machine learning algorithms are sensitive to the scale of the input features. For example, distance-based algorithms like k-nearest neighbors (KNN) or clustering algorithms can be influenced by the scale of the features.

Convergence: Gradient-based optimization algorithms, such as those used in linear regression or neural networks, may converge faster when features are on similar scales.

Interpretability: Scaling ensures that coefficients or feature importances in models are comparable and can be interpreted in a meaningful way.

Regularization: Regularization terms in models, like in Ridge or Lasso regression, treat coefficients uniformly when features are on the same scale.

Difference Between Normalized Scaling and Standardized Scaling:

Normalized Scaling:

Normalization (or Min-Max scaling) scales the features to a specific range, usually between 0 and 1.

Normalization is suitable when the distribution of the data does not follow a Gaussian (normal) distribution, and the algorithm you are using (e.g., neural networks) requires input features to be within a specific range

Standardized Scaling (Z-score normalization):

Standardization transforms the features to have a mean of 0 and a standard deviation of 1.
Standardization assumes that the data follows a Gaussian distribution.
Standardized scaling is often preferred when the algorithm relies on measures of variability, such as in principal component analysis (PCA) or support vector machines (SVM).

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

   The formula for the VIF of an independent variable is given by:
   VIF=1/1-R2
   Now, if the VIF for a particular variable is infinite, it means that the R2 value is equal to 1. This situation occurs when one or more independent variables in the model can be perfectly predicted by a linear combination of the other independent variables.

   Here are a few common reasons why the VIF might be infinite:

Perfect Collinearity:

   If there is perfect collinearity among the independent variables, meaning that one or more variables can be expressed as a perfect linear combination of the others, the VIF becomes infinite.
   In this case, the correlation matrix of the independent variables will have a determinant of zero, indicating a singular matrix.

Redundant Variables:

   If one or more variables are redundant because they can be predicted exactly using a combination of other variables, the VIF for those redundant variables will be infinite.

Dummies Trap:

   In the case of dummy variables, if you include a dummy variable for each category of a categorical variable without excluding one category as the reference, it can lead to perfect multicollinearity, and VIF values may become infinite.

Data Issues:

   Data errors or extreme values can sometimes lead to perfect multicollinearity and, consequently, infinite VIF.

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

   A Q-Q plot, short for "quantile-quantile plot," is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical probability distribution. It is a way to visually

compare the quantiles (percentiles) of the data against the quantiles of the expected theoretical distribution, typically the normal distribution. Q-Q plots are especially useful for checking the assumption of normality in a dataset, which is important in various statistical analyses, including linear regression.

Here's how to interpret and use a Q-Q plot in the context of linear regression:

Generating a Q-Q plot:

Start by sorting the data in ascending order.

Calculate the expected quantiles for a theoretical distribution (e.g., the normal distribution) based on the sample size and the specific quantile (percentile) of interest. The most common theoretical distribution used for Q-Q plots is the standard normal distribution (mean = 0, standard deviation = 1).

Plot the observed quantiles of your dataset against the expected quantiles of the theoretical distribution. This typically results in a scatterplot.

Interpretation:

If the data points in the Q-Q plot roughly follow a straight line, it suggests that the dataset is approximately normally distributed. Deviations from a straight line may indicate departures from normality.

If the Q-Q plot shows a pronounced curvature, heavy tails, or systematic deviations from a straight line, it may suggest non-normality or the presence of outliers in the data.

Use in linear regression:

Normality assumption: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot of the residuals can help you assess this assumption. If the Q-Q plot of residuals shows a relatively straight line, it suggests that the residuals are approximately normally distributed, which is important for valid hypothesis tests and confidence intervals in regression analysis.

Detecting outliers: Q-Q plots can also help identify outliers or data points that do not conform to the expected distribution. Outliers can have a significant impact on regression results, and a Q-Q plot may reveal their presence.

The importance of Q-Q plots in linear regression lies in their ability to visually and quantitatively assess the normality assumption of the residuals. Departures from normality can affect the validity of statistical tests, confidence intervals, and the reliability of regression results. If the Q-Q plot reveals substantial deviations from the expected straight line, you may need to consider data transformations or alternative modeling approaches to address the non-normality of the residuals.