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: In my final analysis, the following are the categorical variables and their analysis:

1. holiday: Has negative coefficient, hence the count decreases on holiday
2. summer: Has positive coefficient, hence the count increases in summer
3. winter: Has positive coefficient, hence the count increases in winter, more than in summer.
4. 2019: Has positive coefficient, hence the count increases than in 2018.
5. Aug: Has positive coefficient, hence the count increases in August.
6. Sept: Has positive coefficient, hence the count increases in September more than in August.
7. cloudy: Has negative coefficient, hence the count decreases on a cloudy day.
8. thunderstorm: Has negative coefficient, hence the count decreases on thunderstorm much more than in cloudy day.

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

Ans: When creating dummy variables for a categorical variable with K categories, K
dummy variables are typically created, one for each category. However, this introduces perfect multicollinearity because the sum of these dummy variables will always equal one. Including all K dummy variables along with the intercept in a regression model results in a situation where one variable is a perfect linear combination of the others, causing multicollinearity issues.

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

Ans: temp

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

Ans: Histogram of Residuals: The histogram of residuals resemble a bell curve.

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: These three features (temperature, thunderstorm condition, and year 2019) have the largest absolute coefficients and are highly statistically significant, making them the top contributors to the model.

General Subjective Questions:

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

Algorithm:

Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the input variables and the output variable. This algorithm is widely used in predictive modeling and data analysis.

Mathematical Foundation:

The goal of linear regression is to find the best-fitting line through the data points. The linear regression model can be represented by the equation:

$y = B0 + B1*x + E$

where:

- y is the dependent variable (target).
- B0 is the intercept of the line.
- B1 is the slope of the line.
- x is the independent variable (feature).
- E is the error term (residual).

Objective Function:

The objective in linear regression is to minimize the difference between the actual values and the predicted values.

Training the Model:

- Gradient Descent: An iterative optimization algorithm used to find the optimal coefficients by minimizing the cost function. The algorithm updates the coefficients in the direction that reduces the error.
- Normal Equation: An analytical method that provides a closed-form solution to find the optimal coefficients without iterative optimization.

Model Evaluation:

After training, the performance of the linear regression model is evaluated using metrics such as:

- R-squared: Represents the proportion of variance explained by the model.
- Mean Squared Error (MSE): Measures the average of the squared errors between predicted and actual values.

Implementation Details: To implement linear regression, follow these steps:

1. Prepare the data (features and target variables).
2. Split the data into training and testing sets.
3. Train the model using either gradient descent or the normal equation.
4. Evaluate the model using appropriate metrics.
5. Make predictions on new data.

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

Anscombe's Quartet is a collection of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed. It was introduced by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data before analyzing it.

Explanation:

Anscombe's Quartet consists of four datasets, each with 11 data points. Despite having identical means, variances, and correlation coefficients, the datasets reveal different underlying patterns when plotted. This demonstrates that relying solely on statistical measures can be misleading and highlights the importance of visualizing data.

Datasets Overview:

Dataset I: Linear relationship with a small amount of scatter.
Dataset II: Linear relationship but with a significant outlier.
Dataset III: Non-linear relationship (quadratic curve).
Dataset IV: A single outlier influencing the dataset.

Anscombe's Quartet highlights the necessity of graphical data analysis in conjunction with statistical summaries. It underscores how different datasets with identical statistics can tell vastly different stories, emphasizing the importance of not relying solely on numerical summaries for data interpretation.

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

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is used to understand how strongly two variables are related.

Formula:
r = Cov(X, Y) / (Ïƒ_X * Ïƒ_Y)

Where:
- Cov(X, Y) is the covariance of variables X and Y.
- Ïƒ_X and Ïƒ_Y are the standard deviations of X and Y, respectively.

Pearson's R assumes:
1. Linearity: The relationship between the variables is linear.
2. Homogeneity of Variance: The variance around the regression line is constant.
3. Normality: The data for both variables should be approximately normally distributed.

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

Scaling is a preprocessing technique used in data analysis and machine learning to adjust the range and distribution of features (variables) in a dataset.

Scaling can improve the performance and convergence speed of many algorithms, especially those that involve distance calculations or gradient-based optimization.

Difference:

Normalization scales the data to a fixed range, making it suitable for algorithms sensitive to the magnitude of data.

Standardization centers the data around zero and scales it based on standard deviation, making it suitable for algorithms that assume data is normally distributed or when variance is important.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?(3 marks)

Ans: This occurs when one predictor variable is a perfect linear combination of one or more other predictor variables. In other words, if there is an exact linear relationship among the predictors, the R2 value becomes 1, leading to infinite value.

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

Ans:

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a specified theoretical distribution, such as the normal distribution. The Q-Q plot compares the quantiles of the sample data against the quantiles of a theoretical distribution. If the data follows the theoretical distribution, the points in the Q-Q plot will approximately lie on a straight line.

Uses of a Q-Q Plot:

1. Assessing Normality: In linear regression, a common assumption is that the residuals (errors) are normally distributed. A Q-Q plot can visually check if the residuals conform to this normal distribution.

2. Identifying Deviations from Normality: It helps in identifying if the data follows a distribution or if there are deviations, such as heavy tails or skewness.

3. Checking Fit for Other Distributions: Beyond normality, Q-Q plots can be used to check if data follows other theoretical distributions like exponential, uniform, or t-distributions.

Importance:

Linear regression assumes that the residuals (errors) of the model are normally distributed. This assumption is important for accurate hypothesis testing and confidence intervals. The Q-Q plot helps to check this assumption.