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 the 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. Cloudy: Has negative coefficient, hence the count decreases on a cloudy day.
5. Thunderstorm: Has negative coefficient, hence the count decreases on thunderstorm much more than in cloudy day.
6. 2019: Has positive coefficient, hence the count increases than in 2018.
7. Aug: Has positive coefficient, hence the count increases in August.
8. Sept: Has positive coefficient, hence the count increases in September more than in August.

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

Ans: Using drop_first=True during dummy variable creation is crucial to prevent multicollinearity. When creating dummy variables for categorical features, one extra column is generated than the number of categories. This leads to perfect correlation between these columns, causing multicollinearity issues in models like linear regression. Dropping the first column removes this redundancy, ensuring model stability and accurate coefficient interpretation.

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

Ans: Temperature (temp) is the numerical variable most strongly positively correlated with bike rental count (cnt), as indicated by the pair plot.

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: Temperature, thunderstorm condition, and year 2019 are the most influential features in the model due to their large, significant coefficients.

General Subjective Questions:

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

1. 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.
2. 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).

3. Objective Function: The objective in linear regression is to minimize the difference between the actual values and the predicted values.
4. 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.

5. 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.
6. 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 set of four small datasets that have nearly identical statistical properties. Despite this similarity in numerical summaries (mean, variance, correlation, regression line), when visualized as scatter plots, they exhibit vastly different patterns.

- This quartet serves as a powerful illustration of the importance of data visualization. It highlights the limitations of relying solely on numerical summaries, emphasizing that visual exploration is crucial for understanding the underlying structure of data. By revealing hidden patterns, outliers, and relationships, visualization complements statistical measures and provides a more comprehensive understanding of the data.

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:

1. Scaling is a preprocessing technique used in data analysis and machine learning to adjust the range and distribution of features (Variables) in a dataset.
2. Scaling can improve the performance and convergence speed of many algorithms, especially those that involve distance calculations or gradient-based optimization.
3. Difference:
   - Normalization scales the data to a fixed range, making it suitable for algorithms sensitive to the magnitude of data.
   - Standardization centres the data on 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:

1. A VIF value of infinity indicates perfect multicollinearity. This means at least two independent variables in your model are perfectly correlated. In simpler terms, one variable can be perfectly predicted from another. This causes issues in regression analysis as the model becomes unstable and the coefficients become unreliable.

2. Perfect multicollinearity often arises due to data entry errors, redundant variables, or derived variables. To address this, you typically need to remove one of the perfectly correlated variables from the model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 mark)
Ans: Q-Q plot is a graphical method to compare two probability distributions by plotting their quantiles against each other. In simpler terms, it plots the quantiles of one dataset against the quantiles of another dataset (or a theoretical distribution).

In linear regression, Q-Q plots are essential for:

- Checking normality of residuals: One of the key assumptions of linear regression is that the residuals (errors) are normally distributed. By plotting the residuals against the quantiles of a standard normal distribution, you can visually assess if this assumption holds. If the points on the plot roughly form a straight line, it suggests normality. Deviations from the line indicate departures from normality.
- Comparing distributions: Q-Q plots can be used to compare the distributions of different variables or datasets, helping to identify potential outliers, differences in shape, or scale.
- Identifying transformations: If the residuals are not normally distributed, transforming the data (e.g., log transformation) might help. Q-Q plots can be used to assess the impact of different transformations.