Assignment-based Subjective Questions and Answers

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   Based on the analysis of categorical variables, it's clear that seasonal factors and weather conditions significantly influence bike rental demand. Spring tends to have lower rentals, while fall see higher usage. Favourable weather, such as clear or partly cloudy conditions, also positively impacts bike rentals. While day of the week or working days seem to have minimal effect, there's a noticeable upward trend in bike sharing popularity over time. These findings suggest that bike-sharing services can benefit from tailoring their operations to seasonal patterns and weather conditions to optimize demand.

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

   Using drop_first = True during dummy variable creation is important for the following reasons:

   Avoiding Multicollinearity: When creating dummy variables, each category of a categorical variable is converted into a separate binary column. If all categories are included, it can lead to multicollinearity, where one variable can be perfectly predicted from the others. By dropping the first category, you avoid this issue.

   Reducing Redundancy: Dropping the first dummy variable reduces redundancy in the dataset. Since the dropped category can be inferred from the remaining categories, it simplifies the model without losing any information.

   Improving Model Performance: By reducing multicollinearity and redundancy, the model can perform better and provide more reliable coefficients. This leads to more accurate and interpretable results.

Efficient Computation: Fewer variables mean less computational complexity, which can speed up the training process and reduce the risk of overfitting.

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

   Based on the pair plot analysis of numerical variables, atemp 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)**

   To validate the linear regression model, a scatter plot was created to visualize the relationship between predicted and actual values on the test data. The regression line should fit the data well. Additionally, multicollinearity among features was checked using VIF values, ensuring they were within an acceptable range (≤5). Finally, the normality of residuals was confirmed through a distribution plot, which should show a normal distribution with a peak at 0.

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

   The top 3 features contributing significantly to the demand of shared bikes are:
   Temperature: Higher temperatures lead to increased demand.
   Weather: Light rain or snow reduces demand.
   Year: Demand increases year over year.

General Subjective Questions

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

Linear regression is a statistical method used to model the relationship between a target variable and one or more predictor variables. The goal is to establish a linear equation that can predict the value of the target variable based on the values of the predictor variables.

The equation of the line is represented as: y = mx + c, where:

- **y:** is the target variable
- **x:** is the predictor variable
- **m:** is the slope of the line (how much y changes for every unit increase in x)
- **c:** is the y-intercept (the value of y when x is 0)

Linear regression aims to establish a linear relationship between a dependent variable and one or more independent variables. To achieve this, the model calculates optimal coefficients (intercept and slope) that minimize the mean squared error (MSE). This metric measures the average squared difference between predicted and actual values.

Once trained, the model can be used to predict the dependent variable for new data points. A common optimization technique for finding these optimal coefficients is gradient descent.

To evaluate the model's performance, metrics like R-squared and adjusted R-squared are employed. These metrics assess how well the model fits the data. To prevent overfitting, regularization techniques like Ridge and Lasso can be used. These techniques introduce penalties to the cost function, discouraging the model from becoming too complex and relying heavily on any particular feature.

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

Anscombe's Quartet is a set of four datasets that, despite having identical means, standard deviations, correlation coefficients, and regression lines, exhibit vastly different patterns when visualized. This illustrates the importance of visualizing data before relying solely on numerical summaries. Each of the four datasets contains 11 data points, and while their statistical properties are identical, their underlying distributions and relationships are quite distinct.

Anscombe's quartet serves as a powerful reminder that visual inspection is crucial for understanding the nuances of data and avoiding misleading conclusions based on numerical statistics alone.

3. **What is Pearson’s R? (3 marks)**

Pearson's correlation coefficient (r) is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1.

r = 1: Indicates a perfect positive correlation, meaning the two variables increase or decrease together perfectly.
r = -1: Indicates a perfect negative correlation, meaning one variable increases as the other decreases perfectly.
r = 0: Indicates no correlation between the variables.

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

Scaling is a data preprocessing technique that transforms numerical data into a common range. It's essential for ensuring fair variable comparison, improving algorithm performance, and preventing numerical instability.
Normalized scaling (min-max scaling) rescales data to a specific range (e.g., 0-1), preserving relative differences. Standardized scaling (z-score standardization) transforms data to have a mean of 0 and a standard deviation of 1, making values comparable in terms of standard deviations.

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

An infinite VIF indicates perfect multicollinearity, meaning a predictor is perfectly correlated with another. This can occur due to below reasons.
Exact linear relationships: If two or more predictors are perfectly linearly related, their columns in the design matrix will be linearly dependent, resulting in a singular matrix.
Dummy variable trap: When creating dummy variables for categorical predictors, including all levels can lead to perfect multicollinearity. To avoid this, one level is typically omitted.
Data entry errors: Incorrect data entry or inconsistencies can introduce spurious correlations between variables, leading to infinite VIF.
Numerical precision limitations: In some cases, due to numerical precision limitations, a near-perfect correlation can be interpreted as a perfect correlation, resulting in an infinite VIF.

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

Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of two datasets. It plots the quantiles of one dataset against the quantiles of another dataset. If the two datasets have the same distribution, the Q-Q plot will be a straight line.

In linear regression, Q-Q plots are used to assess the normality of the residuals. Residuals are the differences between the actual values and the predicted values from the regression model. If the residuals are normally distributed, it is a key assumption of linear regression.

Q-Q plots are important for assessing residual normality, identifying outliers, and evaluating the appropriateness of linear regression models.