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)

- **Season:** Different seasons significantly impact rentals, with certain seasons having higher rentals, likely due to weather changes.
- **Holidays & Working days:** Rentals vary slightly, suggesting that bike usage changes on holidays versus regular working days.
- **Weather Situation:** Clearly influences rentals, with worse weather leading to lower rentals.

This suggests categorical variables significantly affect bike rental demand.

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

- Although I have not explicitly used it, typically, using drop_first=True is important to avoid the dummy variable trap, preventing multicollinearity. This ensures the independence of dummy variables in regression models.

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

- temp (temperature) has the highest correlation with bike rentals (cnt).

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

- **Linearity:** Visualized by plotting actual versus predicted values.
- **Homoscedasticity:** The visualized scatter plots implicitly showed consistency in variance across predictions.

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

- I have explicitly analysed only one feature (temp), which significantly contributed to predicting the demand.

General Subjective Questions

1. **Explain the Linear Regression Algorithm in Detail.** (4 marks)

Linear Regression is a **supervised learning algorithm** used for **predicting a continuous dependent variable** based on one or more independent variables. The core idea is to **fit a straight line** (in simple linear regression) or a hyperplane (in multiple linear regression) that best explains the relationship.

Equation:

y=β0+β1x1+β2x2+⋯+βnxn+εy

Steps:

1. Estimate the best-fit line by **minimizing the sum of squared errors** (ordinary least squares - OLS).
2. Evaluate model performance using metrics like **R^2, RMSE**.
3. Validate assumptions: linearity, normality of residuals, homoscedasticity, independence, and absence of multicollinearity.

2. **Explain Anscombe's Quartet in Detail.** (3 marks)

Anscombe's Quartet is a famous set of four datasets created by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data rather than relying solely on summary statistics.

This quartet elegantly demonstrates that:

1. Statistical summaries alone can be misleading
2. Exploratory data visualization is essential before drawing conclusions
3. Outliers can dramatically influence statistical calculations
4. Different data distributions can produce identical summary statistics

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

Pearson's correlation coefficient, commonly denoted as "r" or "Pearson's r," is a statistical measure that quantifies the linear relationship between two continuous variables. It ranges from -1 to +1, where:

1. +1 indicates a perfect positive linear correlation (as one variable increases, the other increases proportionally)
2. 0 indicates no linear correlation (variables appear unrelated)

3. -1 indicates a perfect negative linear correlation (as one variable increases, the other decreases proportionally)

- It measures only linear relationships (may miss non-linear patterns)
- It is sensitive to outliers
- It is dimensionless (unaffected by changes in scale)
- It is symmetric (correlation of X with Y equals correlation of Y with X)

4. **What is Scaling? Why is Scaling Performed? What is the Difference Between Normalized and Standardized Scaling?** (3 marks)

Scaling is a preprocessing technique that transforms numerical features to a similar range of values. It involves adjusting the range or distribution of data variables without changing the underlying relationships between data points.

Many machine learning algorithms (especially distance-based ones like k-means, kNN, and SVMs) are sensitive to the scale of input features. Features with larger ranges can dominate those with smaller ranges.

Normalization (Min-Max Scaling) is used to transform data to a fixed range, typically [0,1]. Normalization preserves the shape of the original distribution and is sensitive to outliers.

5. **You Might Have Observed That Sometimes the Value of VIF is Infinite. Why Does This Happen?** (3 marks)

Variance Inflation Factor (VIF) becomes infinite when perfect multicollinearity exists in a dataset. VIF for a predictor variable is calculated as:

VIF = $1 / (1 - R^2)$

Where R^2 is the coefficient of determination obtained when that variable is regressed against all other predictor variables.

When perfect multicollinearity exists, R^2 equals exactly 1, meaning the variable can be perfectly predicted from other variables. This makes the denominator $(1 - R^2)$ equal to zero which makes it infinite.

6. **What is a Q-Q Plot? Explain Its Use and Importance in Linear Regression.** (3 marks)

A Quantile-Quantile plot (Q-Q plot) is a graphical technique used to assess whether a dataset follows a particular theoretical distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution (most commonly the normal distribution).

In linear regression, Q-Q plots are primarily used to verify the normality assumption of residuals.

1. **Assess Normality**: Check if the residuals (differences between observed and predicted values) follow a normal distribution, which is a key assumption of linear regression
2. **Identify Outliers**: Points that deviate significantly from the diagonal line indicate outliers
3. **Detect Skewness**: If points curve away from the diagonal line, it suggests skewness in the data

Q-Q plots are especially valuable because they provide a more complete picture of distributional characteristics than summary statistics or histograms alone.