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?**

Effect of Categorical Variables on Dependent Variable (cnt):

- Season: Demand peaks in fall and summer due to pleasant weather, drops in winter.
- Weather Situation: Clear weather increases demand, while heavy rain/snow (weathersit=4) drastically reduces it.
- Month: Higher demand in warmer months (e.g., June–September).
- Weekday: Weekdays show higher rentals (commuters), weekends slightly lower (leisure rides).
- Holiday: Demand dips slightly on holidays.

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

- Importance of `drop_first=True`:
  Prevents the dummy variable trap (perfect multicollinearity). For example, with 4 seasons, using 3 dummies avoids redundancy and ensures numerical stability in regression.

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

- Highest Correlation with Target Variable (cnt):
  Temperature (temp) showed the strongest positive correlation ($r \approx 0.63$). Higher temperatures encourage bike usage.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

- Validating Linear Regression Assumptions:
  - Linearity: Residual vs. fitted value plots (no patterns).
  - Independence: Durbin-Watson statistic (~2.0, indicating no autocorrelation).
  - Homoscedasticity: Residuals spread evenly (no funnel shape).
  - Normality: Q-Q plot of residuals aligned with the diagonal line.

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

- Top 3 Features in Final Model:
    - yr (yearly growth: 2019 > 2018).
    - temp (higher temperatures increase demand).
    - weathersit_Clear (clear weather boosts rentals).

General Subjective Questions

1. Explain the linear regression algorithm in detail

Linear Regression is a supervised learning algorithm used for predicting a continuous dependent variable based on one or more independent variables. The goal is to fit a linear equation to the data.

Equation of Linear Regression

For a simple linear regression with one independent variable:

y=β0+β1x+ϵy = \beta_0 + \beta_1 x + \epsilony=β0+β1x+ϵ

where:

- yyy is the dependent variable (target),
- xxx is the independent variable (feature),
- β0\beta_0β0 is the intercept,
- β1\beta_1β1 is the coefficient (slope),
- ϵ\epsilonϵ is the error term.

For multiple linear regression with multiple independent variables:

y=β0+β1x1+β2x2+...+βnxn+ϵy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilony=β0+β1x1+β2x2+...+βnxn+ϵ

Working of Linear Regression

- **Model Fitting:** The algorithm determines the best values for *β0,β1,...βn\beta_0, \beta_1, ... \beta_nβ0,β1,...βn* by minimizing the sum of squared

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four different datasets that have nearly identical statistical properties (such as mean, variance, correlation, and regression line) but have vastly different distributions when plotted.

Why is it Important?

- It highlights the importance of visualizing data instead of relying only on summary statistics.
- Despite having similar numerical characteristics, the underlying relationships differ drastically.

The Four Datasets:

Each dataset contains 11 (x, y) pairs, and all four have:

- The same **mean** for x and y.
- The same **variance** for x and y.
- The same **correlation coefficient** (~0.82).
- The same **linear regression equation**.

3. What is Pearson's R?

Pearson's correlation coefficient (r) measures the linear relationship between two continuous variables.

Formula:

$$\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

Interpretation:

- r=+1r = +1r=+1 → **Perfect positive correlation** (as X increases, Y increases).
- r=−1r = -1r=−1 → **Perfect negative correlation** (as X increases, Y decreases).
- r=0r = 0r=0 → **No correlation** (X and Y are independent).

4. What is Scaling? Why is Scaling Performed? Difference Between Normalized and Standardized Scaling?

Scaling refers to transforming numerical data so that different features contribute equally to model training.

Why is Scaling Performed?

- Prevents features with large values from dominating models (e.g., gradient descent).
- Improves convergence speed in optimization algorithms.
- Essential for distance-based models (e.g., KNN, SVM, PCA).

5. Why Does the Value of VIF Become Infinite?

Variance Inflation Factor (VIF) detects multicollinearity in regression models. It measures how much the variance of a regression coefficient is inflated due to collinearity among independent variables.

Why Does VIF Become Infinite?

- If R^2 = 1, then VIF=∞

- This happens when one independent variable is a perfect linear combination of others.
- In such cases, regression cannot compute unique coefficients (perfect collinearity).

6. What is a Q-Q Plot? Importance in Linear Regression.

A Q-Q (Quantile-Quantile) Plot is used to check whether a dataset follows a normal distribution.

How a Q-Q Plot Works:

- Plots quantiles of the sample data against theoretical quantiles from a normal distribution.
- If data follows normal distribution, points lie along a 45-degree straight line.

Use in Linear Regression:

- **Checks Residual Normality:** Linear regression assumes that residuals are normally distributed.
- **Identifies Outliers:** Deviations from the line indicate outliers or skewness.
- **Detects Heteroscedasticity:** Unequal variance in residuals can be identified.