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)

Categorical variables such as **season** and **weathersit** showed a significant impact on bike demand. For example:

- **Season:** Bike demand was highest during **summer** and **fall** seasons, indicating favorable weather conditions.
- **Weathersit:** Clear weather conditions positively influenced bike demand, while heavy rain or snow reduced it significantly.
- **Year (yr):** Demand increased in 2019 compared to 2018, showing a year-over-year growth trend.

These insights suggest that both seasonal and weather-based patterns are critical factors in predicting bike-sharing demand.

---

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

When creating dummy variables, **drop_first=True** ensures that **multicollinearity** is avoided by dropping one of the dummy categories. In a regression model, if all dummy variables for a categorical column are included, they will introduce **perfect multicollinearity** because one variable can be predicted from the others. Dropping the first dummy variable serves as a **reference category** and prevents this issue.

---

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

From the pair plot analysis, **temperature (temp)** had the **highest positive correlation** with the target variable (cnt). This indicates that warmer temperatures are directly associated with increased bike demand.

---

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

The assumptions of **Linear Regression** were validated as follows:

1. **Linearity:** Scatter plots between predicted and actual values showed a linear trend.
2. **Homoscedasticity:** Residuals vs Fitted Values plot displayed random distribution without visible patterns.
3. **Independence of Residuals:** Durbin-Watson test was checked for residual independence.
4. **Normality of Residuals:** A histogram and Q-Q plot of residuals indicated a normal distribution.
5. **No Multicollinearity:** Variance Inflation Factor (VIF) values were checked for all predictors.

These validations confirmed that the model met all the necessary assumptions.

---

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 significant features identified from the model coefficients were:

1. **Temperature (temp)** – Higher temperatures increased bike demand.
2. **Year (yr)** – Demand increased in 2019 compared to 2018.
3. **Season_Fall (season_fall)** – Bike demand peaked in the fall season.

These features had the highest positive coefficients, indicating their strong contribution to predicting bike demand.

General Subjective Questions

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

Linear Regression is a statistical technique used to model the relationship between a dependent variable (Y) and one or more independent variables (X).

- **Equation:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$
- **Objective:** Minimize the **Residual Sum of Squares (RSS)** to determine the best-fit line.
- **Assumptions:**
    - Linearity
    - Independence of residuals
    - Homoscedasticity
    - Normality of residuals
    - No multicollinearity
- **Training:** Model coefficients (β_0, β_1...) are estimated using **Ordinary Least Squares (OLS)**.
- **Evaluation:** Metrics such as **R^2 score**, **RMSE**, and **Residual Analysis** are used to evaluate model performance.

****

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

Anscombe's Quartet is a set of four datasets that have nearly identical statistical properties (mean, variance, correlation, regression line) but appear very different when plotted graphically.

- It highlights the importance of **visualizing data** instead of relying solely on summary statistics.
- Each dataset demonstrates different data patterns:
    1. Linear Relationship
    2. Non-linear Relationship

3. Outlier Effect
4. High-leverage Point Effect

****

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

Pearson's Correlation Coefficient (R) measures the **strength and direction** of a linear relationship between two variables.

- **Range:** -1 to 1
- **+1:** Perfect positive correlation
- **-1:** Perfect negative correlation
- **0:** No correlation
- **Formula:** $R = Cov(X, Y) / (\sigma_X * \sigma_Y)$
  It is commonly used in feature selection to identify highly correlated predictors.

****

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

Scaling adjusts the range or distribution of data features:

- **Normalization:** Scales data between 0 and 1.
  - Formula: (X - Xmin) / (Xmax - Xmin)
- **Standardization:** Scales data to have a mean of 0 and a standard deviation of 1.
  - Formula: (X - Mean) / Std Dev
    Why: Scaling ensures features with different ranges do not dominate model training.

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

VIF (Variance Inflation Factor) becomes **infinite** when there is **perfect multicollinearity** between predictor variables. This happens when:

- One feature is an exact linear combination of others.
- Duplicate or highly correlated variables exist in the dataset.
  Solution:
- Remove redundant features.
- Use dimensionality reduction techniques like PCA.

****

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

(3 marks)

What is a Q-Q Plot?

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (commonly a normal distribution). It plots the quantiles of the sample data against the quantiles of the theoretical distribution.

- If the data follows the assumed theoretical distribution, the points on the Q-Q plot will align along a straight diagonal line.
- Deviations from this line indicate departures from the assumed distribution.

Use of Q-Q Plot in Linear Regression

In Linear Regression, one of the key assumptions is that the residuals (errors) are normally distributed. The Q-Q plot is used to visually check this assumption:

Normality Check:

- If residuals follow a normal distribution, the points on the Q-Q plot will closely align with the diagonal reference line.

Outlier Detection:

- Points that deviate significantly from the line indicate outliers or heavy tails in the data.

Identifying Skewness:

- An upward or downward curve in the Q-Q plot suggests skewness in the residuals.

Importance of Q-Q Plot in Linear Regression

1. Validation of Assumptions: Ensures that residuals meet the assumption of normality, which is critical for valid hypothesis testing and confidence intervals.
2. Improved Model Reliability: If residuals are normally distributed, the model's predictions and statistical inferences (e.g., p-values, confidence intervals) are more reliable.
3. Diagnostic Tool: Helps in diagnosing non-normal residuals, indicating the need for transformations or alternative modeling techniques.