# Assignment-based Subjective Questions

# Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

# Total Marks: 3 marks (Do not edit)

# Answer: <Your answer for Question 1 goes below this line> (Do not edit)

# Categorical variables like 'season', 'weathersit', 'holiday', and 'weekday' significantly affect the dependent variable 'cnt'. For example:

# - 'season': Demand is higher during summer and fall due to favorable weather conditions, and lower in winter.

# - 'weathersit': Clear weather leads to higher bike usage compared to adverse weather.

# - 'holiday' and 'weekday': Demand patterns vary based on holidays or weekdays.""",

# 

# 

**Question 2.** Why is it important to use **drop\_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop\_first=True avoids the dummy variable trap, which occurs due to multicollinearity among dummy variables. This ensures only \( n-1 \) categories are represented, allowing regression coefficients to be interpretable.

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

# Answer: <Your answer for Question 3 goes below this line> (Do not edit)

# The variable 'temp' (temperature) has the highest positive correlation with 'cnt'. Warmer temperatures increase bike usage, making it a key factor.

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

# Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Validating the assumptions of Linear Regression is a crucial step to ensure the model is reliable and interpretable. After building the model on the training set, the following steps were performed to validate its assumptions:

* **Linearity**: Residuals showed no discernible patterns, confirming linearity.
* **Homoscedasticity**: Residuals had a uniform spread, confirming homoscedasticity.
* **Normality**: Residuals were approximately normally distributed based on histograms and Q-Q plots.
* **Multicollinearity**: VIF values were checked to ensure no significant multicollinearity.
* **Independence**: Durbin-Watson statistic indicated no significant autocorrelation.

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 5 goes below this line> (Do not edit)

* **temp (Temperature)**:

High positive correlation with cnt. Warmer temperatures encourage outdoor activities, leading to increased bike demand.

* **year (Year)**:

Indicates a growing trend in bike-sharing popularity over time (e.g., more rentals in 2019 compared to 2018).

* **season\_summer or season\_fall**:

Seasons like summer and fall tend to see more rentals due to favorable weather conditions.

# General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)

**Total Marks:** 4 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 6 goes here>

Linear Regression is a statistical and machine learning algorithm used for predicting a continuous target variable y based on one or more independent variables X. The algorithm assumes a linear relationship between the independent and dependent variables.

Key Components

Equation of a Line: The mathematical model for linear regression is:

y=β0+β1x1+β2x2+…+βnxn+ϵy = \beta\_0 + \beta\_1 x\_1 + \beta\_2 x\_2 + \ldots + \beta\_n x\_n + \epsilony=β0​+β1​x1​+β2​x2​+…+βn​xn​+ϵ

yyy: Dependent variable (target)

xix\_ixi​: Independent variables (features)

β0\beta\_0β0​: Intercept (value of yyy when all xi=0x\_i = 0xi​=0)

βi\beta\_iβi​: Coefficients (weights) of the independent variables

ϵ\epsilonϵ: Error term (difference between the actual and predicted values)

Objective:

To find the best-fitting line (or hyperplane for multiple variables) by minimizing the error between predicted and actual values.

**Question 7.** Explain the Anscombe’s quartet in detail. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 7 goes here>

Anscombe's quartet demonstrates four datasets with identical statistical properties (mean, variance, correlation) but different visual patterns. It emphasizes the importance of visualizing data to understand underlying structures and outliers.

**Question 8.** What is Pearson’s R? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 8 goes here>

Pearson's R measures the linear correlation between two variables, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation). A value near 0 indicates no linear correlation.

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 9 goes here>

Scaling adjusts feature magnitudes to improve algorithm performance.

- Normalization scales values between 0 and 1.

- Standardization centers data to mean 0 and variance 1.

Scaling ensures all features contribute equally during modeling.

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 10 goes here>

# VIF becomes infinite when perfect multicollinearity exists, meaning one variable is a perfect linear combination of others. This often happens with redundant features.

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 11 goes here>

A Q-Q plot compares residuals to a theoretical normal distribution. If points lie on a straight line, residuals are normally distributed, validating the assumption of normality in linear regression.