# 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)

# Effect of Categorical Variables on Bike Demand:

# Season: Bike demand varies significantly by season. Winter shows lower demand while summer and fall drive higher rentals.

# Weathersit: Clear weather has the highest demand while light snow or rain negatively affects demand.

# Weekday: Weekends generally show higher bike rentals, indicating more leisure usage.

# Holiday: Bike demand is slightly lower on holidays.

# 

**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)

# It prevents the dummy variable trap by dropping one category, avoiding perfect multicollinearity. This ensures the model remains interpretable and avoids redundancy.

**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)

**temp** has the highest correlation with **cnt.** This indicates that warmer temperatures drive higher bike usage.

**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)

# Validation of Linear Regression Assumptions:

# Linearity – Checked through residual vs fitted plots.

# Normality of Errors – Validated using Q-Q plots.

# Homoscedasticity – Assessed through random scatter of residuals.

# Multicollinearity – Addressed by calculating VIF and dropping atemp.

**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)

# Top 3 Features in Final Model:

# Weathersit\_Light Snow / Rain: -2119 (Adverse weather reduces demand)

# Yr: +1989 (2019 shows higher demand than 2018)

# Season\_Spring: -975 (Spring exhibits lower demand compared to other seasons)

# 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 models the relationship between a dependent variable (Y) and one or more independent variables (X)

# The goal is to minimize the error by finding the line of best fit

# Y = β0 ​+ β1​X1 ​+ β2​X2 ​+ ... + βn​Xn ​+ ϵ

# The model assumes linearity, independence, homoscedasticity, and normality of residuals.

**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 consists of four datasets with nearly identical statistical properties but different distributions. It highlights the importance of visualizing data rather than relying solely on summary statistics.

**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 correlation coefficient measures the linear relationship between two variables.

# It values ranges from -1 to 1:

# 1 : Perfect positive correlation

# -1 : Perfect negative correlation

# 0 : No 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>

1. **Scaling** adjusts the range of features, ensuring no variable dominates the model.
2. **Normalization** scales data between 0 and 1.
3. **Standardization** scaled data to have mean 0 and standard deviation 1.
4. Scaling is critical for distance based models **(eg; KNN, SVM)**

**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>

# Infinite VIF occurs when one feature is a perfect linear combination of others causing multicollinearity. This can be resolved by dropping correlated variables or and applying principal component analysis (PCA)

**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 the quantities of residuals against a theoretical normal distribution.

# Purpose –

# It validates the normality assumption in regression.

# Importance-

# Non-linearity in the plot may indicate that residuals deviate from normality.

# Addressing this improves model reliability.