# 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)

# Due to Categorical variables able to distinguish the target variables dependent variable and drop the unnecessary variables if the categories found are more than two levels and drop the dummy values if not necessary

# 

**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: < When creating dummy variables in machine learning, using the drop\_first=True parameter is important to avoid the dummy variable trap. This issue arises when you create dummy variables for a categorical variable that has more than two levels (categories). Without dropping the first category, you end up with multicollinearity in the dataset.> (Do not edit)

**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)

# Among the numerical values yr has the highest correlation with the target variable

**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: < After building a linear regression model on the training set, it's important to validate its assumptions to ensure the model's reliability and the correctness of its predictions. There are several key assumptions in linear regression that should be checked:

# 1. Linearity:

# Assumption: The relationship between the independent and dependent variables should be linear.

# > (Do not edit)

**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)

# After completion of the final model found yr , temp, atemp are the top 3 features contributing significantly towards explaining the demand of the shared bikes

# 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)

# <Linear Equation

# ****Simple Linear Regression****:

# y=m1+m​2x+c

# y: dependent variable

# x: Independent variable

# m1: Intercept(the value of y when x=0)

# m2: Coefficient (Slope of the line) c: Error term

# Multiple Linear Regression:

# y=β0+β1x1+β2x2+…+βnxn+ϵy = \beta\_0 + \beta\_1x\_1 + \beta\_2x\_2 + \ldots + \beta\_nx\_n + \epsilony=β0​+β1​x1​+β2​x2​+…+βn​xn​+ϵ

**Objective**: The goal is to minimize the **Residual Sum of Squares (RSS)**:

RSS=∑i=1n(yi−y^i)2RSS = \sum\_{i=1}^{n}(y\_i - \hat{y}\_i)^2RSS=i=1∑n​(yi​−y^​i​)2

Here, yiy\_iyi​ is the actual value, and y^i\hat{y}\_iy^​i​ is the predicted value.

# >

**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)

# < Anscombe’s quartet is a collection of four datasets that have nearly identical simple statistical properties, such as the mean, variance, correlation coefficient, and linear regression line. Despite their statistical similarity, the datasets differ significantly when visualized graphically, illustrating the importance of visualizing data alongside numerical analysis.>

**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)

# < Pearson’s R, also known as the Pearson correlation coefficient, is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables. It is denoted by rrr and ranges from -1 to 1: >

**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)

< Normalized scaling and standardized scaling are two distinct methods used for feature scaling, differing in their approach and applications. Normalized scaling rescales data to a fixed range, typically between [0, 1] or [-1, 1], by focusing on the minimum and maximum values of the dataset. This method is highly sensitive to outliers, as extreme values can significantly impact the scaling process. It is best suited for scenarios where maintaining the range of data is important, such as in image processing.

On the other hand, standardized scaling centers the data around a mean of 0 and adjusts the scale so the standard deviation becomes 1. It focuses on the mean and standard deviation of the data, making it less sensitive to outliers compared to normalization. Standardization is particularly useful when the data follows a Gaussian distribution or when algorithms like Principal Component Analysis (PCA) are employed, which rely on variance analysis >

**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)

# < The Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity in the dataset. This means that one predictor variable in the regression model can be expressed as an exact linear combination of one or more other predictors.>

**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)

# < A Q-Q plot (quantile-quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. It compares the quantiles of the sample data to the quantiles of the chosen theoretical distribution.>