# 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)

# From my analysis of the categorical variables, these are my inferences

# Season: Autumn has the highest bike rental count.

# Year: More rides in 2019 compared to 2018.

# Holiday: Non-holidays see higher average riders.

# Weekdays: Tuesday to Friday have higher average riders.

# Working Day: Slightly higher average riders on working days.

# Weather: Clear weather attracts the most riders, followed by mist and light snow; heavy rain sees very low ridership.

# These factors collectively influence bike rental usage patterns.

# 

**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 during dummy variable creation is important to avoid multicollinearity in regression models. By dropping the first category of each categorical feature, we ensure that the dummy variables are independent, preventing redundancy and ensuring the model's coefficients are interpretable. This helps in maintaining the integrity of the regression analysis.

**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)

# Looking at the pair-plot among the numerical variables, the variable temp has the highest correlation with the target variable count.

**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)

# After building the model on the training set, the assumptions of Linear Regression were validated as follows:

# Linearity: The relationship between the independent variable temp and the dependent variable count was checked to ensure it was linear.

# Multicollinearity: Variance Inflation Factor (VIF) values were calculated for all variables, ensuring that each VIF was less than 5, indicating no multicollinearity.

# Normality of Errors: The residuals were analyzed using a histogram and a Q-Q plot to confirm that they were normally distributed, with errors concentrated around zero.

# These steps confirmed that the model met the key assumptions of Linear Regression.

**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)

# Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

# Temperature (temp): This variable has the highest correlation with bike demand, indicating that higher temperatures generally lead to increased bike rentals.

# Working Day: Whether it is a working day or not significantly affects bike demand, with more rentals typically occurring on working days.

# Season: The season plays a crucial role, with autumn showing the highest bike rental counts, followed by winter.

# These features collectively help in understanding and predicting the demand for 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)

# <Your answer for Question 6 goes here>

# Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting linear equation that describes this relationship. Here’s a detailed explanation:

# 1. Model Representation:

# Simple Linear Regression: Involves one independent variable. The model is represented as: y=β0+β1x+ϵy=β0​+β1​x+ϵ where ( y ) is the dependent variable, ( x ) is the independent variable, ( \beta\_0 ) is the intercept, ( \beta\_1 ) is the slope, and ( \epsilon ) is the error term.

# Multiple Linear Regression: Involves multiple independent variables. The model is represented as: y=β0+β1x1+β2x2+…+βnxn+ϵy=β0​+β1​x1​+β2​x2​+…+βn​xn​+ϵ

# 2. Assumptions:

# Linearity: The relationship between the dependent and independent variables is linear.

# Independence: Observations are independent of each other.

# Homoscedasticity: Constant variance of the error terms.

# Normality: The error terms are normally distributed.

# 3. Estimation of Coefficients:

# The coefficients (( \beta\_0, \beta\_1, \ldots, \beta\_n )) are estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of the squared differences between the observed and predicted values: Minimize∑i=1n(yi−y^i)2Minimize∑i=1n​(yi​−y^​i​)2 where ( y\_i ) are the observed values and ( \hat{y}\_i ) are the predicted values.

# 4. Model Evaluation:

# R-squared: Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

# Adjusted R-squared: Adjusts the R-squared value based on the number of predictors in the model.

# p-values: Test the null hypothesis that the coefficient is equal to zero (no effect).

# F-statistic: Tests the overall significance of the model.

# 5. Validation:

# Residual Analysis: Check the residuals to ensure they meet the assumptions of linear regression.

# Cross-Validation: Split the data into training and testing sets to evaluate the model’s performance on unseen data.

# 6. Interpretation:

# The coefficients (( \beta )) represent the change in the dependent variable for a one-unit change in the independent variable, holding all other variables constant.

# The intercept (( \beta\_0 )) represents the expected value of the dependent variable when all independent variables are zero.

# Linear regression is a foundational algorithm in machine learning and statistics, widely used for its simplicity and interpretability.

**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 is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how statistical properties can be misleading if not visualized.

# Here are the key points about Anscombe's quartet:

# 1. Identical Statistical Properties:

# Each dataset in the quartet has the same mean, variance, correlation coefficient, and linear regression line.

# 2. Different Graphical Representations:

# Despite having similar statistical properties, the four datasets look very different when plotted:

# Dataset I: Appears as a simple linear relationship with some random noise.

# Dataset II: Forms a perfect curve, indicating a non-linear relationship.

# Dataset III: Contains an outlier that significantly affects the regression line.

# Dataset IV: Shows a vertical line with one outlier, indicating a different type of relationship.

# 3. Importance of Visualization:

# Anscombe's quartet highlights the importance of visualizing data before performing statistical analysis. It shows that relying solely on summary statistics can be misleading and that graphical representations can reveal underlying patterns, outliers, and relationships that are not apparent from the statistics alone.

**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, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the strength and direction of the association between the variables. The value of Pearson's R ranges from -1 to 1:

# +1 indicates a perfect positive linear relationship.

# -1 indicates a perfect negative linear relationship.

# 0 indicates no linear relationship.

# Pearson's R is calculated using the formula: r=∑(xi−xˉ)(yi−yˉ)∑(xi−xˉ)2∑(yi−yˉ)2r=∑(xi​−xˉ)2∑(yi​−yˉ​)2​∑(xi​−xˉ)(yi​−yˉ​)​

# where ( x\_i ) and ( y\_i ) are the individual data points, and ( \bar{x} ) and ( \bar{y} ) are the means of the ( x ) and ( y ) variables, respectively.

# Pearson's R is widely used in statistics to determine the strength and direction of the linear relationship between two continuous variables.

**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 is the process of transforming the features of a dataset so that they are on a similar scale. This is important in machine learning because many algorithms are sensitive to the scale of the input data.

# Why Scaling is Performed:

# Improves Model Performance: Algorithms like gradient descent converge faster with scaled data.

# Prevents Dominance: Ensures that no single feature dominates others due to its scale.

# Enhances Interpretability: Makes the model coefficients more interpretable.

# Difference Between Normalized Scaling and Standardized Scaling:

# Normalized Scaling:

# Definition: Rescales the data to a fixed range, usually [0, 1].

# Formula: Xnorm=X−XminXmax−XminXnorm​=Xmax​−Xmin​X−Xmin​​

# Use Case: Useful when the data does not follow a Gaussian distribution and when the scale of the data is important.

# Standardized Scaling:

# Definition: Centers the data around the mean with a unit standard deviation.

# Formula: Xstd=X−μσXstd​=σX−μ​ where ( \mu ) is the mean and ( \sigma ) is the standard deviation.

# Use Case: Useful when the data follows a Gaussian distribution and when the relative distance between data points is important.

# Both methods are essential for preparing data for machine learning models, depending on the specific requirements of the algorithm and the nature of the data.

**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>

# The value of the Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the independent variables in a regression model. This means that one of the independent variables is an exact linear combination of one or more of the other independent variables. In such cases, the denominator in the VIF formula, which involves the R-squared value of the regression of one independent variable on the others, becomes 1, leading to division by zero and resulting in an infinite VIF. This indicates that the model cannot distinguish the individual effect of the perfectly collinear variables.

**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 (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, typically the normal distribution. It compares the quantiles of the data against the quantiles of the theoretical distribution.

# Use and Importance in Linear Regression:

# Assessing Normality of Residuals: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps to visually check this assumption.

# Interpreting the Plot:

# If the residuals are normally distributed, the points on the Q-Q plot will lie approximately along the 45-degree reference line.

# Deviations from this line indicate departures from normality, such as skewness or kurtosis.

# Model Validation: Ensuring that the residuals are normally distributed validates the linear regression model, making the statistical inferences (like confidence intervals and hypothesis tests) more reliable.

# In summary, a Q-Q plot is a crucial diagnostic tool in linear regression for validating the normality assumption of residuals, thereby ensuring the robustness of the model.