# 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)

# Based on the analysis I did, here are the key insights about how categorical variables (season and weather) affect bike rentals (dependent variable):

# Seasonal Impact:

# Summer and Fall seasons show higher median rental counts

# Winter has lower rental numbers, likely due to cold weather

# There is significant variation in rentals across seasons, indicating strong seasonal influence

# Weather Impact:

# Clear weather shows highest rental numbers

# Rentals decrease significantly during adverse weather (Light Snow/Rain, Mist)

# Heavy Rain sees lowest rental counts, suggesting weather is a strong predictor

# Higher variability in rentals during Clear weather compared to other conditions

# Overall Effect:

# Both categorical variables show clear patterns of influence on bike rentals

# Weather appears to have more immediate impact on daily rentals

# Seasonal effects likely capture longer-term patterns in usage

# The relationships appear logical and align with expected behavior (better weather/warmer seasons = more rentals)

# 

**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 because of below reasons:

# 1.Avoid Perfect Multicollinearity

# Prevents the "dummy variable trap"

# When all dummy variables sum to 1, one variable becomes redundant

# Creates perfect multicollinearity between independent variables

# 2.Statistical Necessity

# With k categories, only k-1 dummy variables are needed

# The dropped category becomes the reference/base category

# All other categories are interpreted relative to this base

# 3.Model Interpretation

# Makes interpretation more straightforward

# Effects are measured as differences from the base category

# Reduces noise in the model

# 4.Computational Efficiency

# Reduces number of features

# Improves model training speed

# Reduces memory usage

# 5.Model Stability

# Prevents matrix singularity issues

# Ensures model parameters are uniquely determined

# Improves numerical stability of calculations

# Without drop\_first=True, models may fail to converge or produce unreliable results due to perfect multicollinearity.

**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 analysis of numerical variables, here are the key findings about correlations with the target variable (cnt):

# Highest Positive Correlation:

# temp (temperature) shows the strongest positive correlation with cnt

# This suggests bike rentals increase significantly as temperature rises

# The relationship appears to be somewhat linear

# The temperature variable (temp) having the highest correlation coefficient suggests it will be one of the most important predictors in the model for bike rental demand.

**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)

# To validate the assumptions of Linear Regression after building the model on the training set, the following steps are typically taken:

# Linearity:

# Check scatter plots of residuals vs. predicted values to ensure no clear pattern.

# Ensure the relationship between independent and dependent variables is linear.

# Homoscedasticity:

# Plot residuals vs. predicted values to check for constant variance.

# Look for a random scatter without funnel shapes or patterns.

# Normality of Residuals:

# Use Q-Q plots to check if residuals follow a normal distribution.

# Perform statistical tests like the Shapiro-Wilk test.

# Independence of Errors:

# Check for autocorrelation in residuals using the Durbin-Watson test.

# Ensure residuals are independent over time.

# Multicollinearity:

# Calculate Variance Inflation Factor (VIF) for each predictor.

# Ensure VIF values are below a threshold (commonly 10).

# Outliers and Influential Points:

# Identify outliers using standardized residuals.

# Use leverage and Cook's distance to detect influential points.

# These steps help ensure the linear regression model's assumptions are met, leading to more reliable and valid results.

**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):

# Strong positive correlation with bike rentals.

# Indicates higher demand as temperature increases.

# Hour (hr):

# Significant impact on bike rentals, capturing daily usage patterns.

# Reflects peak hours of bike usage.

# Season (season):

# Captures seasonal variations in bike demand.

# Different seasons show varying levels of bike rentals, with summer and fall typically having higher demand.

# These features are crucial in predicting bike rental demand due to their strong and consistent influence on the target variable.

# 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. Here's a detailed explanation:

# Key Concepts

# Dependent Variable (Y):

# The outcome or target variable we aim to predict.

# Independent Variables (X):

# The predictors or features used to explain the dependent variable.

# Linear Relationship:

# Assumes a linear relationship between the dependent and independent variables.

# Model Equation

# The linear regression model can be expressed as: [ Y = \beta\_0 + \beta\_1X\_1 + \beta\_2X\_2 + \ldots + \beta\_nX\_n + \epsilon ]

# ( \beta\_0 ): Intercept term.

# ( \beta\_1, \beta\_2, \ldots, \beta\_n ): Coefficients for each independent variable.

# ( \epsilon ): Error term (residuals).

# 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.

# No Multicollinearity:

# Independent variables are not highly correlated with each other.

# Steps in Linear Regression

# Data Collection:

# Gather data with the dependent and independent variables.

# Data Preprocessing:

# Handle missing values, encode categorical variables, and normalize/standardize data if necessary.

# Model Fitting:

# Use methods like Ordinary Least Squares (OLS) to estimate the coefficients ( \beta ).

# Model Evaluation:

# Assess the model using metrics like R-squared, Adjusted R-squared, Mean Squared Error (MSE), etc.

# Assumption Validation:

# Check the assumptions of linear regression using residual plots, VIF for multicollinearity, and statistical tests.

# Ordinary Least Squares (OLS)

# OLS is the most common method to estimate the coefficients:

# Minimizes the sum of the squared differences between observed and predicted values.

# Finds the best-fitting line by reducing the residual sum of squares (RSS).

# Interpretation

# Coefficients:

# Represent the change in the dependent variable for a one-unit change in the independent variable, holding other variables constant.

# Intercept:

# The expected value of the dependent variable when all independent variables are zero.

# Applications

# Used in various fields like economics, finance, biology, and social sciences.

# Helps in predicting outcomes, understanding relationships, and making informed decisions.

# 

# Linear regression is a foundational algorithm in machine learning and statistics, providing a simple yet powerful tool for predictive modeling and analysis.

**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>

**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>

**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>

# Anscombe's quartet consists of four datasets with nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression line) but very different distributions when graphed. Created by Francis Anscombe in 1973, it demonstrates the importance of visualizing data before analysis.

# Key Characteristics

# Each dataset has:

# Same mean for x and y.

# Same variance for x and y.

# Same correlation between x and y.

# Same linear regression line.

# The Four Datasets

# Dataset I: Linear relationship with random noise.

# Dataset II: Quadratic relationship, not suitable for linear regression.

# Dataset III: Linear relationship with an influential outlier.

# Dataset IV: Vertical points with one influential point affecting the regression line.

# Importance

# Graphical Analysis: Visualizing data reveals patterns and anomalies.

# Outliers: Outliers can significantly affect statistical measures.

# Context Matters: Summary statistics alone can be misleading.

# Model Appropriateness:

# Different datasets may require different models despite similar statistics.

**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 VIF becomes infinite due to perfect multicollinearity among independent variables. This happens when one variable is an exact linear combination of others, causing the denominator in the VIF calculation to be zero.

# Reasons

# Perfect Multicollinearity:

# Exact linear relationship between variables.

# Example: Including both x and 2x.

# Dummy Variable Trap:

# Including all dummy variables for a categorical variable without dropping one category.

**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.

# Use in Linear Regression

# Assess Normality of Residuals:

# Plots the quantiles of the residuals against the quantiles of a normal distribution.

# Helps check if residuals are normally distributed, an assumption of linear regression.

# Identify Deviations:

# Points should lie approximately along a straight line if residuals are normally distributed.

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

# Importance

# Model Validity:

# Ensures the residuals meet the normality assumption.

# Normal residuals imply valid hypothesis tests and confidence intervals.

# Detect Outliers:

# Points far from the line may indicate outliers or influential data points.

# Helps in diagnosing and addressing anomalies.

# Improves Interpretation:

# Validates the use of linear regression.

# Ensures reliable and interpretable results.

# A Q-Q plot is crucial for validating the normality assumption of residuals in linear regression, ensuring the model's reliability and accuracy.