# 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)

# 

**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)

**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)

**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)

**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)

# 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 regression is a *supervised machine learning algorithm* that models relationship between a dependent variable `*y`* and one or more independent variables `x1, x2, ..., xn`. Goal is to find best-fitting straight line (or hyperplane in higher dimensions) that minimizes difference between predicted and actual values of the dependent variable. This is achieved by minimizing the Residual Sum of Squares (RSS), difference between observed values and predicted values. Equation of a Simple Linear Regression is: y=mx+b Where: m is the slope, representing relationship between x and y, b is the intercept, where line crosses y-axis. Linear Regression For multiple variables, model becomes: y=b0+b1x1+b2x2+...+bnxn Linear regression uses techniques like Ordinary Least Squares (OLS) to find best values of coefficients.

# 

**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 set of four datasets that demonstrate importance of visualizing data. Each dataset has nearly identical statistical properties, including the same mean, variance, correlation, and regression line. Lightbox However, their scatter plots reveal distinct patterns: one is linear, another is nonlinear, a third shows an outlier influencing the trend, and last has a vertical cluster with an influential point. This illustrates how relying solely on summary statistics can be misleading, as different datasets with similar metrics can exhibit vastly different distributions and relationships. Visualization ensures a better understanding of the data's true characteristics.

**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)

# In linear regression, Pearson's R (Pearson correlation coefficient), quantifies strength and direction of linear relationship between two continuous variables. It ranges from -1 to +1: \* +1 indicates a perfect positive linear relationship. \* -1 indicates a perfect negative linear relationship. \* 0 means no linear relationship. In context of linear regression, Pearson's R is closely related to the slope of regression line; a higher absolute value of R signifies a steeper slope, indicating a stronger linear association. Below is a formula for calculating the Pearson correlation coefficient (r): \begin{equation*} r = \frac{ n\sum{xy}-(\sum{x})(\sum{y})}{% \sqrt{[n\sum{x^2}-(\sum{x})^2][n\sum{y^2}-(\sum{y})^2]}} \end{equation*} Additionally, Square of Pearson's R, known as coefficient of determination or R², represents proportion of variance in dependent variable that is predictable from independent variable. Relationship underscores importance of Pearson's R in assessing how well regression model explains variability of the data.

**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)

# Scaling adjusts range of independent variables (features) to ensure they contribute equally to the model. This is crucial when features have different units or magnitudes, as unscaled data can lead to biased coefficient estimates and affect model performance. Scaling is performed because => \* Improved Model Performance: Scaling ensures that features with larger ranges don't dominate the regression coefficients, leading to a more balanced model. \* Faster Convergence: Algorithms like gradient descent converge more quickly when features are on similar scales. \* Interpretability: Scaled data allows for more meaningful comparisons between coefficients, aiding in understanding feature importance. Normalized Scaling vs. Standardized Scaling Normalization (Min-Max Scaling): This technique rescales features to a specific range, typically [0, 1]. It's useful when the data doesn't follow a Gaussian distribution and is sensitive to outliers. The formula is: Min–max normalization - Regression Analysis with R [Book] Standardization (Z-score Scaling): This method transforms features to have a mean of 0 and a standard deviation of 1, centering the data around zero. It's effective when the data follows a normal distribution and is less sensitive to outliers. The formula is: Standard Score Calculation & Formulas

**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)

# Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to collinearity with other predictors. *A VIF value of infinity occurs when a predictor variable is perfectly or nearly perfectly correlated with one or more other predictors in the model.* Mathematically, VIF for a variable is calculated as: Image result for he VIF for a variable is calculated as: Where Ri2 ​is coefficient of determination from regressing xi on the other predictors. When Ri2​ approaches 1 (i.e., perfect correlation), denominator becomes nearly zero, causing VIF to approach infinity.

**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 compare the distribution of a dataset to theoretical distribution, typically normal distribution. It plots quantiles of observed data against quantiles of theoretical distribution. If data follows theoretical distribution, points will lie along a straight line. If there are deviations from straight line, it indicates that data does not follow expected distribution. In the context of linear regression, a Q-Q plot is primarily used to assess assumption that residuals (the differences between observed and predicted values) are normally distributed. This is important because one of the key assumptions in linear regression is that the residuals should follow a normal distribution for valid hypothesis testing and reliable coefficient estimates. Importance in Linear Regression: Normality of Residuals: A Q-Q plot helps to visually assess if residuals are normally distributed. If points on plot lie close to a straight line, it indicates that residuals are approximately normal, which is a crucial assumption for valid inference in regression. Model Validity: Deviations from normality (e.g., skewness or heavy tails) can signal model issues, such as incorrect functional form or outliers, which can lead to misleading conclusions or unreliable statistical tests. Outlier Detection: Q-Q plots also help identify outliers, as points far from line suggest data points that deviate significantly from expected distribution.