# 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)

# The analysis of categorical variables provides critical insights into bike rental patterns, especially as Boom Bikes aims to restore and grow demand in a post-pandemic context:

# Seasonal Impact: Fall exhibits the highest rental median, followed by summer, which aligns with optimal weather conditions. Winter sees a decline, likely due to less favorable conditions for biking.

# Yearly Increase: Rentals increased significantly from 2018 to 2019, potentially driven by rising awareness of sustainable transport and greater adoption of bike-sharing services, suggesting a positive trajectory Boom Bikes could leverage as cities reopen.

# Monthly Trends: Demand by month reflects the seasonal influence, with fall months like September experiencing peak rentals. This indicates an opportunity for Boom Bikes to capitalize on predictable high-demand periods.

# Holidays vs. Non-Holidays: Rentals are lower on holidays as people may choose personal vehicles or family time, suggesting demand is stronger during routine commute days.

# Weekday Variations: While the median rentals remain consistent across the week, Saturdays and Wednesdays show more variability. The Saturday trend may imply leisure-based rentals, which can fluctuate based on weekend plans.

# Working vs. Non-Working Days: Both categories have similar medians, but non-working days show wider spread, possibly due to spontaneous leisure plans rather than commuting.

# Weather Condition: Clear weather drives the most demand, as it offers ideal riding conditions with moderate temperatures and low humidity. Boom Bikes can expect demand to dip during light snow or rain, suggesting the need for strategic promotion in adverse weather.

# This analysis underscores how seasonality, weekdays, and weather directly affect demand, providing Boom Bikes with actionable insights to target users based on favorable conditions and high-demand periods effectively.

# 

**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)

# The use of drop\_first=True in dummy variable creation addresses multicollinearity by excluding one category per variable group. This adjustment ensures that each remaining dummy variable conveys distinct information, enabling accurate interpretation and reducing redundancy. For Boom Bikes, this approach is essential for stable model performance and ensures that comparisons among factors, such as seasons or weather conditions, are interpretable without overfitting.

**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)

# In examining the pair-plot, temperature (temp) shows the highest correlation with bike demand (cnt), with a correlation coefficient of 0.63. This correlation reflects that as temperatures rise, outdoor activity and bike rentals increase significantly, underlining temperature as a core predictor for Boom Bikes. Understanding this helps the company forecast demand surges in warmer months and align resource allocation accordingly.

**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 ensure the linear regression model’s assumptions hold:

# Normality of Residuals: The residuals follow a normal distribution with a mean close to zero, as confirmed by a distplot and Q-Q plot. This is crucial for accurate interval estimation and inference.

# Linearity and Homoscedasticity: The residual plot against predicted values showed no pattern or funnel shape, indicating constant variance across predictions. This stability is critical in providing reliable demand forecasts for Boom Bikes.

# Multicollinearity: VIF analysis revealed low VIF values (below 5) across predictors, confirming no multicollinearity. Each variable thus offers unique, additive information without overlapping, essential for understanding distinct effects like season or weather on demand.

# Independence of Errors: The residual plot confirmed that errors are independently distributed, as no significant autocorrelation or pattern was observed, enhancing the model’s reliability in time-based demand predictions.

# Overall Fit: The model achieved an R-squared of 0.84 on training data, explaining 84% of demand variability, and 0.81 on test data, suggesting strong predictive power and alignment with business goals.

# Through rigorous assumption checks, this model proves to be a robust tool for Boom Bikes, capable of predicting demand reliably while helping the company adapt to post-pandemic trends.

**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)

# The top three features contributing to bike demand in the final model are:

# Temperature (temp): With a coefficient of 0.548, temperature is the strongest predictor, as warmer weather encourages outdoor activity and bike rentals. Boom Bikes can anticipate high demand in warmer seasons and adjust availability accordingly.

# Weather Condition - Light Snow/Rain (weathersit): With a coefficient of -0.2838, adverse weather like light snow or rain reduces rentals. Boom Bikes may consider promoting rides or offering discounts during adverse weather to mitigate this drop.

# Year (yr): The positive coefficient of 0.2328 for the year variable reflects a growing trend in bike rentals. This trend, likely due to increased awareness of eco-friendly transport, suggests a sustained rise in demand, especially as people seek cost-effective, green commuting options post-COVID.

# These insights enable Boom Bikes to align their business strategy with data-driven demand patterns, focusing on optimal times and conditions to maximize rentals and customer satisfaction.

# 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)

# Definition and Objective: Linear regression is a supervised learning algorithm that models the relationship between a dependent variable Y and one or more independent variables X. The main objective is to find the best-fitting linear equation that minimizes prediction error.

# Mathematical Equation: The linear relationship is expressed as Y = β0 +β1X + ϵ where β0 is the intercept, β1​ is the slope, and ϵ represents the error term. For multiple linear regression, it generalizes to Y = β0 + β1X1+ β2X2 + ⋯ + βnXn + ϵ.

# Ordinary Least Squares (OLS): OLS is the most common technique to estimate β coefficients by minimizing the Residual Sum of Squares (RSS), defined as RSS = .

# Assumptions: Linear regression relies on certain assumptions:

# -Linearity: The relationship between X and Y is linear.

# -Independence: Observations are independent of each other.

# -Homoscedasticity: Constant variance of residuals.

# -Normality of Residuals: Residuals should be normally distributed.

# Evaluation Metrics: Common metrics for evaluating linear regression include R-squared (measures explained variance) and Mean Squared Error (MSE). High R-squared values indicate a good fit of the model.

# Gradient Descent: In cases where data is large, gradient descent can be used instead of OLS to iteratively adjust parameters, finding the optimal values of β0 and β1​ by moving towards a minimum cost function.

**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)

# Concept and Origin: Anscombe’s quartet, created by statistician Francis Anscombe, consists of four datasets with nearly identical statistical summaries (mean, variance, correlation, and linear regression), despite having distinctly different distributions.

# Datasets and Linear Regression Lines: Each dataset has approximately the same linear regression equation, y = 3 + 0.5x, yet they exhibit different patterns, challenging assumptions made solely on statistical measures.

# Purpose: The quartet demonstrates the importance of visualizing data, as statistical measures alone can be misleading without examining underlying data structures.

# Diversity of Patterns: Each dataset in Anscombe’s quartet shows a unique relationship:

# Dataset 1 follows a simple linear trend.

# Dataset 2 shows a clear non-linear relationship.

# Dataset 3 has a strong linear relationship disrupted by one outlier.

# Dataset 4 forms a nearly vertical line with a single influential point.

# Lesson for Regression Analysis: Anscombe’s quartet underscores the importance of data visualization in regression analysis to detect outliers, identify non-linear relationships, and ensure model assumptions hold.

**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)

# 1.Definition and Purpose: Pearson’s correlation coefficient (denoted as r) measures the linear relationship between two continuous variables, providing insight into the direction and strength of their relationship.

# 2. Range and Interpretation: The value of r ranges from -1 to 1:

# r = 1: Perfect positive correlation.

# r = −1: Perfect negative correlation.

# r = 0 : No linear correlation.

# 3. Formula: Pearson’s r is calculated as:

# where and are the means of x and y, respectively.

# 4. Assumptions: Pearson’s r assumes a linear relationship between X and Y and requires both variables to be normally distributed.

# 5. Importance in Regression: In linear regression, Pearson’s r informs about the strength of the relationship. A high absolute r value suggests a stronger relationship, providing confidence in the linear regression model’s predictions.

**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)

# 1.Definition of Scaling: Scaling is the process of adjusting the range of feature values, bringing all features to a common scale, which is particularly important in algorithms sensitive to the magnitude of values, such as linear regression.

# 2. Importance of Scaling: It improves convergence in optimization algorithms, reduces computation time, and prevents certain features from dominating due to scale differences.

# 3. Normalization (Min-Max Scaling): Rescales feature values to a specific range, usually [0, 1], using the formula:

# ​​

# It preserves the relationships of data points within a feature range and is especially useful when a feature distribution is not Gaussian.

# 4. Standardization (Z-score Scaling): Transforms data to have a mean of 0 and standard deviation of 1, using:

# Standardization is particularly useful for normally distributed data and is a common scaling method in regression.

# 5. Choosing the Method: Normalization is preferred for bounded values, whereas standardization is ideal for features with Gaussian distributions. In linear regression, scaling ensures that gradient descent converges faster and the model is less biased towards variables with large magnitudes.

**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)

# 1.Definition of VIF: Variance Inflation Factor (VIF) quantifies multicollinearity by measuring how much the variance of a regression coefficient is inflated due to linear relationships among predictors.

# 2.Formula: VIF is calculated as:

# where is the R-squared value from regressing predictor j against all other predictors.

# 3. Cause of Infinite VIF: VIF becomes infinite when there is perfect multicollinearity, meaning one predictor is an exact linear combination of others, causing ​ to equal 1. This results in division by zero in the VIF formula.

# 4. Implications: Infinite VIF indicates severe multicollinearity, which can make the model unstable, lead to inaccurate parameter estimates, and reduce predictive power.

# 5.Mitigation Strategies: To address infinite VIF, one can remove or combine highly correlated predictors, use dimensionality reduction techniques (e.g., PCA), or employ regularization techniques (e.g., Ridge regression) to handle multicollinearity.

**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)

# 1.Definition: A Quantile-Quantile (Q-Q) plot is a graphical tool that plots the quantiles of observed data against the quantiles of a theoretical distribution, often a normal distribution, to assess distributional assumptions.

# 2. Interpretation: In a Q-Q plot, if the observed data matches the theoretical distribution, points align along a 45-degree reference line. Deviations from this line indicate discrepancies from the expected distribution.

# 3. Use in Regression: Q-Q plots are essential in linear regression to check if residuals (errors) follow a normal distribution, a key assumption for the reliability of hypothesis tests and confidence intervals.

# 4. Detecting Non-Normality: Patterns in Q-Q plots (e.g., S-shapes, heavy tails) can reveal issues like skewness or kurtosis in residuals, suggesting that transformations or alternative models may be needed.

# 5.Importance: Q-Q plots help validate model assumptions, providing insight into the suitability of the linear model. Non-normality of residuals may indicate that the linear regression model could be improved by adjusting variables or choosing a different model structure.