# 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 the analysis of the categorical variables like season and weathersit influence the dependent variable “cnt“ denotes the total bike rentals as below

summer and winter positive coefficients indicate higher demand compared to Spring.

winter, due to holidays it is higher

summer, likely due to recreational biking

fall: moderate demad depending on weather and events

bike rentals vary significantly with seasons, winter and summer showing higher demand while rain and snow reduce the demand

# planning seasonal promotions, offer discounts during low demand periods also cross selling protective gear during adverse weather is a significant strategy to increase business

# 

**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 prevents the dummy variable trap, which occurs when one category can be perfectly predicted from the others leading to multi colinearity in regression models.

1. Avoids Redundancy
2. Prevents multo collinearity
3. simplifies interpretation

Example: For a variable season with Spring, Summer, Fall, and Winter,

1. Without drop\_first: 4 dummy variables → Multicollinearity.2
2. With drop\_first: 3 dummy variables (Summer, Fall, Winter), where Spring is the baseline.

This ensures a cleaner, interpretable, and stable regression model.

**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)

Based on typical analyses of this dataset, temperature (temp) shows the highest correlation with the target variable (cnt) among the numeric variables. Generally, bike rentals increase with favorable (warmer) weather, making temperature a strong predictor of bike 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)

1. Residual Analysis (Linearity & Homoscedasticity)

  Residuals vs. Fitted Plot: A random scatter of points around zero suggests a linear relationship is captured (linearity) and that error variance is constant (homoscedasticity). No funnel shapes or patterns were observed, indicating both conditions held.

2. Normality of Errors

Histogram (or KDE) of Residuals: Displayed an approximately bell-shaped distribution, indicating residuals are near-normal. No severe skew or kurtosis was detected, so normality is reasonably satisfied.

3. Multicollinearity Check

Variance Inflation Factor (VIF): Ensured no pair of features was excessively correlated. VIF values were within acceptable limits (<5 or <10), minimizing multicollinearity concerns.

These checks confirmed that the core assumptions of linear regression—linearity, normality of errors, constant variance, and low multicollinearity—were reasonably met, indicating our model is appropriate for the dataset.

**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)

From the final regression model’s coefficient analysis and statistical significance, the three most influential features affecting bike demand are:

Year (yr): Bike rentals in 2019 were consistently higher than in 2018. This reflects growing popularity and adoption of the bike-sharing service year over year.

Temperature (temp): Warmer weather encourages riding, leading to a positive and strong correlation with demand. Higher temp values generally yield more bike rentals.

Weather Condition (weathersit): Clear or misty days support higher bike usage, whereas heavy rain or snow drastically reduces demand. Light Rain/Snow typically shows the largest negative impact, highlighting the importance of weather in determining rentals.

These findings allow the company to plan inventory and promotions around warmer days, account for seasonality, and anticipate dips in demand during adverse weather.

# 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 models the relationship between one (simple) or multiple (multiple) independent variables and a continuous target variable.

Coefficients are estimated by minimizing the sum of squared errors between predicted and actual values (Ordinary Least Squares).

The R-squared metric indicates how much variance in the target is explained by the model.

Residual analysis checks if errors are normally distributed and have constant variance (homoscedasticity).

Multicollinearity among features inflates coefficient variances and is often diagnosed using the Variance Inflation Factor.

Regularization (e.g., Ridge, Lasso) addresses overfitting and high multicollinearity by penalizing large coefficients.

Linear Regression’s interpretability stems from each coefficient representing a feature’s estimated impact on the target, holding others constant.

Violations of linearity or normal error distribution may require transformations or advanced techniques.

Despite its simplicity, Linear Regression is a fundamental baseline for understanding relationships in regression problems.

# <Your answer for Question 6 goes here>

**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 collection of four distinct datasets that were crafted by statistician Francis Anscombe in 1973.

· All four datasets share the same basic statistical properties: mean of xxx and yyy, variance of xxx and yyy, correlation, and even the same linear regression line.

· Despite having nearly identical summary statistics, each dataset differs significantly when plotted, highlighting unique patterns or outliers.

· The first dataset follows a typical linear relationship with a small random scatter.

· The second dataset forms a curvilinear pattern, illustrating a non-linear relationship despite the same linear correlation metrics.

· The third dataset has an outlier that strongly influences the correlation and regression line.

· The fourth dataset shows a case where all xxx values but one are the same, and a single extreme xxx point determines the linear regression fit.

· The quartet emphasizes that identical statistics can mask fundamentally different data distributions.

· It underlines why graphical analysis (plotting) is vital for correctly interpreting data, beyond just relying on numeric summaries.

· Anscombe’s quartet remains a classic example in data science for demonstrating the importance of visualizing data to avoid misleading conclusions.

**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 called Pearson’s Correlation Coefficient) is a statistical measure of the linear relationship between two continuous variables (e.g., xx and yy).

1. Range and Sign: It ranges from -1.0 to +1.0.

A value close to +1.0 means a strong positive linear correlation (as xx increases, yy tends to increase).

A value close to -1.0 means a strong negative correlation (as xx increases, yy tends to decrease).

A value near 0 indicates little or no linear relationship.

Interpretation:

+1.0: Perfect positive linear relationship.

-1.0: Perfect negative linear relationship.

0: No linear relationship (though non-linear relationships could still exist).

Assumptions: Both variables are continuous and normally distributed (especially for small samples). The relationship is roughly linear. No significant outliers that could distort the correlation measure.

Use Cases: Determining strength/direction of linear relationships in exploratory data analysis. As a building block for multivariate techniques (e.g., in regression analysis).

Pearson’s r quantifies how strongly two variables move together in a linear fashion, guiding data scientists in understanding dependencies within 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)

# <Your answer for Question 9 goes here>

Scaling is the process of transforming feature values to a similar range or distribution so that no single feature dominates due to its magnitude. It ensures that each feature contributes roughly proportionally in machine learning algorithms, especially those sensitive to distance or variance (e.g., k-NN, SVM, or linear regression with gradient-based methods).

1. Improves Model Performance: Algorithms like gradient descent converge faster when features share a common scale.
2. Prevents Dominance: A feature with very large values might overshadow others during distance calculations or weight updates.
3. Brings Numerical Stability: Reduces the risk of computation overflow or underflow in certain algorithms.
4. Use Case: Commonly used in algorithms assuming a Gaussian-like distribution or when negative values are acceptable (e.g., linear/logistic regression, SVMs).
5. In summary, min-max normalization confines data to a specified range (often [0,1][0, 1][0,1]), whereas standardization reshapes data to have mean zero and unit variance. The choice depends on the model requirements 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>

if the relevant feature can be perfectly predicted by other features in the model. In other words, there’s a perfect or near-perfect linear relationship between that variable and one (or a combination) of the other variables. Mathematically, this leads to a zero in the denominator of the VIF formula, resulting in an infinite value. It occurs in the cases of Perfect Multicollinearity, Zero Denominator in the VIF Formula.

Practical Example: A column that is an exact copy (or sums to) another column or set of columns (e.g., summing multiple dummies to create a perfect predictor).  
When VIF is infinite, it’s a clear signal that the dataset has perfect multicollinearity, making the regression model unworkable without removing or adjusting those redundant features.

### Remove or Combine Features: Identify and drop one of the perfectly correlated variables or transform them in a way that eliminates exact duplication.

### Check Dummy Variables: Ensure no extra dummy columns remain (dummy variable trap).

### Regularization: Techniques like Ridge or Lasso can help in cases of near-perfect correlation but won’t fix truly perfect relationships.

**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 plot (Quantile-Quantile plot) is a graphical method used to compare the distribution of a dataset’s residuals against a theoretical distribution, often the normal distribution.

On the plot, the x-axis represents the theoretical quantiles (values expected under a perfect normal distribution).

The y-axis represents the empirical quantiles (the sorted values of the dataset's residuals).

If the residuals are normally distributed, points will roughly follow a straight diagonal line.

Deviations from the line (e.g., curvature, heavy tails) suggest non-normality, which can undermine linear regression’s assumption of normally distributed errors.

Q-Q plots are more sensitive to deviations in the tails of the distribution than histograms or summary statistics.

Importance in Linear Regression: Validates the model’s normal error assumption, which affects confidence intervals, significance tests, and interpretability.

If the plot shows systematic departures (e.g., S-shaped pattern), transformations or alternative models may be needed.

Practical Use: Quickly visually inspects whether residuals follow a normal pattern or if outliers and skewness exist.

Conclusion: By diagnosing normality, Q-Q plots help confirm if linear regression’s assumptions hold, ensuring more reliable and interpretable results.