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)

I found “cnt” as the dependent variable based on the following listed effects – Year, Season, and Temperature seem to be driving factors in increasing the bike ride counts.

Season:

Demand is highest in the “Season 3”, likely due to favorable weather.
Season 1 (Winter) has the lowest bike rental demand

Year:

Demand increased in 2019 compared to 2018, likely due to service expansion or increased popularity

Temperature:

Warmer temperatures drive higher demand, as biking is more comfortable in favorable weather.

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)

Setting drop_first=True ensures:

No redundant columns (avoids multicollinearity).
Clearer interpretation of coefficients.
A more stable and efficient 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)

From the pair-plot and correlation analysis:

The variable **"atemp"** (**feeling temperature)** has the highest positive correlation with the target variable **"cnt" (demand for shared bikes),** with a correlation coefficient of **"0.63".**

This suggests that higher perceived temperatures are associated with increased bike demand. which also shows a strong positive relationship **(0.63).**

Conversely, windspeed has a negative correlation **(-0.23),** indicating that higher wind speeds tend to decrease 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)

Assumptions of Linear Regression:

Linear: The relationship between the independent variables and the dependent variable.

Validation: I am validating through the following Plots and Factors

- **Predicted Vs Actual Count plot**
  - > Shows that residuals are mostly scattered randomly around zero.
- **Correlation Heatmap**
  -> Ensures strong linear correlations between independent variables and the dependent variable.
- **Multicollinearity (Variance Inflation Factor (VIF))**
  -> temp: 1223.51
  -> atemp: 1445.29
  -> hum: 30.93
  -> registered: 37.24

  -> High VIF values indicate severe multicollinearity. Dropping one of the highly correlated variables (e.g., "temp" or "atemp").
- **Durbin-Watson Test**
  -> The Durbin-Watson Test statistic (2.07) is close to 2, indicating no significant autocorrelation in the residuals.

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)

3 Features:

Casual: Coefficient = 1.00
->Indicates a strong, direct relationship between casual users and bike demand.

Registered: Coefficient = 1.00
->Another strong, direct relationship, showing that registered users are a key driver of total bike demand.

Year (yr): Coefficient = 0.00000000000133

->The impact is negligible compared to the other two variables.

The features casual and registered dominate the model's explanation of bike demand due to their direct measurement of user count.

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>

Step 1: Initialization

Define the hypothesis function and

Initialize the coefficients

Step 2: Loss Function

The most common loss function is the Mean Squared Error (MSE)

Step 3: Optimization

The algorithm computes the optimal coefficients using methods like:

Ordinary Least Squares (OLS)

Gradient Descent (iterative approach)

Step 4: Model Evaluation

Evaluate the performance of the model using metrics like:

Coefficient of determination: Measures the proportion of variance in RMSE (Root Mean Squared Error): Square root of MSE for 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 group of four datasets devised by statistician Francis Anscombe in 1973. It demonstrates how datasets with identical statistical properties (e.g., mean, variance, correlation, and regression line) can exhibit vastly different distributions when visualized. This highlights the importance of data visualization in statistical analysis.

Anscombe's quartet was designed to illustrate:

1. The importance of visualizing data: Statistical summaries can be misleading if data distribution or relationships are not visualized.
2. The limitations of relying solely on descriptive statistics: Summary statistics like mean, variance, or correlation do not capture the entire structure or pattern of data.

Plotted Four Datasets:

Dataset 1: A Typical Linear Relationship

- yyy increases linearly with xxx.
- This dataset adheres to assumptions of linear regression.
- The relationship is straightforward and well-modeled by the regression line.

Dataset 2: A Curved Relationship

- yyy exhibits a non-linear, quadratic relationship with xxx.
- A straight line is not a good fit for this dataset, and the residuals would show systematic patterns.

Dataset 3: Outlier-Driven Relationship

- Most data points align horizontally, except for one extreme outlier.
- The outlier heavily influences the regression line and the correlation coefficient, distorting the analysis.

Dataset 4: Vertical Relationship with a Single Influential Point

- Most data points form a vertical line (no variation in xxx).
- A single influential point determines the slope and correlation, misleading the interpretation.

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 (or Pearson Correlation Coefficient) is a statistical measure of the linear relationship between two continuous variables. It quantifies both the strength and the direction of the relationship.

Quantifying Relationships:

- Measures the correlation between variables like temperature and season.

Feature Selection:

- Identifies relevant predictors for regression models by examining their correlation with the target variable.

Hypothesis Testing:

- Tests the significance of the correlation coefficient.

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>

What is Scaling?
Scaling is a preprocessing technique used in machine learning to adjust the range of independent variables (features) to ensure that they contribute equally to the model. It modifies the feature values without distorting their relationships, making them suitable for algorithms sensitive to magnitude differences.

Why is Scaling Performed?

1. To Ensure Algorithm Efficiency:
    - Many machine learning algorithms (e.g., Gradient Descent, Support Vector Machines, K-Nearest Neighbors) are sensitive to the scale of features.
    - Large magnitude differences can lead to longer convergence times or suboptimal results.
2. To Prevent Dominance by Larger Features:
    - Features with larger ranges (e.g., age in years vs. salary in thousands) can disproportionately influence model training.
3. To Improve Interpretability:
    - Scaling transforms feature values into comparable ranges, improving the clarity of feature importance.
4. For Distance-Based Algorithms:
    - Algorithms like K-Means, K-Nearest Neighbors, and PCA rely on Euclidean distances, which are impacted by differing feature scales.

Key Differences:

Aspect	Normalization (Min-Max)	Standardization (Z-Score)
Range	Transforms data to [0,1][0, 1][0,1] or [−1,1][-1, 1][−1,1].	No fixed range; mean = 000, std = 111.
Use Case	For bounded data or non-Gaussian distributions.	For Gaussian distributions or algorithms needing z-scores.
Effect of Outliers	Sensitive to outliers as it uses min/max.	Less sensitive due to use of mean/std.
When to Use	Neural Networks, clustering.	Regression, PCA, algorithms using gradients.

Which Scaling Method to Use?

- Normalization: Use when you know the data does not follow a normal distribution and the algorithm benefits from bounded input.
- Standardization: Use when the data is approximately Gaussian, or when outliers are present.

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 Variance Inflation Factor (VIF) becomes infinite due to perfect multicollinearity among the predictors in a regression model.

Based on the given Dataset

"atemp" (VIF = 64.30) and "temp" (VIF = 63.27) exhibit extremely high multicollinearity

If I remove - "**atemp**" the updated VIF values show significant improvement

"temp" now has a **VIF of 1.22**, indicating minimal collinearity

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>

What is a Q-Q Plot?

A Q-Q Plot (Quantile-Quantile Plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution (e.g., a normal distribution). It compares the quantiles of the observed data with the quantiles of the expected theoretical distribution.

How Does a Q-Q Plot Work?

1. X-Axis: Quantiles of the theoretical distribution.
2. Y-Axis: Quantiles of the observed data.

If the data follows the theoretical distribution:

- The points in the Q-Q plot will lie approximately on a straight diagonal line (the line y=xy = xy=x).

If the data deviates from the theoretical distribution:

- The points will deviate systematically from the diagonal line, revealing skewness, kurtosis, or other patterns.

Use of Q-Q Plot in Linear Regression

In linear regression, several assumptions need to be validated to ensure the model's reliability. The Q-Q plot is a crucial diagnostic tool for checking one of these assumptions: normality of residuals.

1. Checking Normality of Residuals:
    - One key assumption of linear regression is that the residuals (errors) are normally distributed.
    - A Q-Q plot of the residuals helps identify whether they deviate significantly from a normal distribution.
    - If the residuals follow a normal distribution, the points in the Q-Q plot will align closely with the diagonal line.
2. Identifying Deviations from Normality:
    - Heavy Tails: If the points curve away from the line at the ends, the residuals may have heavier tails than a normal distribution.
    - Skewness: If the points systematically deviate from the line on one side, the residuals may be skewed.
    - Outliers: Individual points far from the line may indicate outliers.
3. Improving Model Assumptions:
    - If the Q-Q plot reveals non-normality, data transformations (e.g., log or square root) or robust regression techniques might be needed to address this issue.

Importance of Q-Q Plot in Linear Regression

1. Model Validity:
    - Verifying the normality of residuals ensures valid hypothesis tests (e.g., ttt-tests for coefficients).
2. Interpretability:
    - Many metrics and statistical tests in regression rely on normally distributed residuals for accurate interpretation.
3. Detecting Problems Early:
    - Non-normality in residuals can indicate model misspecification, the need for variable transformation, or the presence of outliers.
