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)

Categorical variables provide valuable insights into bike rental patterns:

1. Season:
   Rentals are highest during summer and fall, likely due to favourable weather conditions. The demand drops in winter, likely due to colder temperatures and fewer outdoor activities.
2. Weather Situation (weathersit):
   Rentals are significantly higher on clear days, while they decrease on misty/cloudy days and drop further during rain or snow, indicating weather strongly influences user behaviour.
3. Holiday:
   Rentals are lower on holidays, as people may not commute to work or follow routine schedules. In contrast, non-holidays show higher demand due to regular commuting.
4. Working Day:
   Bike rentals are higher on working days, primarily driven by commuters, and decrease on non-working days, aligning with reduced travel needs.
5. Weekday:
   Rentals are relatively steady across weekdays, with a slight dip on weekends, suggesting fewer commutes but possibly more leisure rides.

Overall, these variables highlight that bike demand is closely tied to seasonality, weather conditions, and user routines such as workdays and holidays.

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 multicollinearity by removing one dummy variable, ensuring the model remains stable and coefficients are interpretable relative to a reference category.

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 among the numerical variables, temp (temperature) has the highest correlation with the target variable (cnt, total bike rentals). This indicates that bike rentals increase as the temperature rises, reflecting a strong positive relationship.

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, the following steps were performed:

1. Linearity:
    - Checked the linear relationship between predictors and the target variable by plotting predicted values vs. actual values.
    - Ensured no clear non-linear patterns exist in the residuals.
2. Homoscedasticity:
    - Plotted residuals vs. predicted values to confirm constant variance of residuals.
    - Looked for any funnel shapes or patterns indicating heteroscedasticity.
3. Normality of Residuals:
    - Plotted a histogram or KDE plot of residuals to check if they follow a normal distribution.
    - Verified using a Q-Q plot for normality.
4. Multicollinearity:
    - Checked Variance Inflation Factor (VIF) values for predictors; ensured VIF values were below 5 to avoid multicollinearity.
5. Independence of Errors:
    - Examined residuals to ensure no autocorrelation (errors are independent).

These above steps confirmed the model satisfies the assumptions of Linear Regression, ensuring its reliability and validity.

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 3 features contributing significantly to shared bike demand are:

1. Temperature (temp): Higher temperatures lead to more rentals due to favourable biking conditions.
2. Year (yr): Bike rentals increased in 2019, reflecting the growing popularity of bike-sharing services.
3. Season (season_Summer, season_Fall): Rentals peak in summer and fall, driven by favourable weather for outdoor activities.

These features capture weather, temporal trends, and seasonality, which strongly influence bike demand.

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 fundamental algorithm in supervised learning used for predicting continuous outcomes based on input features. The goal is to establish a linear relationship between independent variables (features) and a dependent variable (target).

Key Concepts in Linear Regression

1. Linear Relationship:
- The assumption is that there exists a linear relationship between the input variables X and the output y. The relationship can be represented mathematically as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

Where:
- y: Dependent variable (target)
- $x_1, x_2, ..., x_n$: Independent variables (features)
- β_0: Intercept (value of y when all $x_i = 0$)
- $\beta_1, \beta_2, ..., \beta_n$: Coefficients (slopes of the line for each feature)
- ε: Error term or residual (difference between predicted and actual y).

2. Objective:
- The primary goal is to estimate the coefficients ($\beta_0, \beta_1, ..., \beta_n$) such that the error between the predicted and actual values is minimized.

Steps in the Linear Regression Algorithm

1. Define the Hypothesis Function:
- The hypothesis function predicts y based on the input features:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

- $\hat{y}$: Predicted value.

2. Formulate the Cost Function:
- The cost function measures the error of the predictions. For linear regression, we use the Mean Squared Error (MSE):

$$J(\beta_0, \beta_1, ..., \beta_n) = (1/m) \Sigma (\hat{y}_i - y_i)^2$$

Where:
- m: Number of training examples.
- $\hat{y}_i$: Predicted value for the i-th example.
- y_i: Actual value for the i-th example.

3. Optimize the Coefficients:
- The goal is to find values of $\beta_0, \beta_1, ..., \beta_n$ that minimize the cost function. This can be achieved using:
  - Analytical Method: The Normal Equation.
  - Iterative Method: Gradient Descent.

Methods for Coefficient Optimization

1. Normal Equation (Closed-Form Solution):
- For a feature matrix X and target vector y:

$$\beta = (X^T X)^{-1} X^T y$$

- This directly computes the optimal coefficients without iteration but can be computationally expensive for large datasets.

2. Gradient Descent:
- Iteratively updates the coefficients to minimize the cost function:

$$\beta_j := \beta_j - \alpha (\partial J / \partial \beta_j)$$

Where:
- α: Learning rate (step size).

∂J/∂βj: Partial derivative of the cost function with respect to βj.

Assumptions of Linear Regression

1. Linearity: The relationship between independent and dependent variables is linear.
2. Independence: Observations are independent.
3. Homoscedasticity: Constant variance of residuals.
4. Normality of Residuals: Residuals should follow a normal distribution.
5. No Multicollinearity: Independent variables should not be highly correlated.

Types of Linear Regression

1. Simple Linear Regression:
- Only one independent variable (x1).
- Equation: $y = \beta0 + \beta1x1 + \varepsilon$.
2. Multiple Linear Regression:
- Multiple independent variables (x1, x2, ..., xn).
- Equation: $y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta nxn + \varepsilon$.

Advantages

- Easy to interpret and implement.
- Computationally efficient for small datasets.
- Works well for linearly separable data.

Disadvantages

- Assumes a linear relationship, which might not hold for complex datasets.
- Sensitive to outliers.
- Poor performance with multicollinearity or highly correlated features.

Applications

- Predicting house prices based on features like size, location, and amenities.
- Estimating sales based on marketing spend.
- Forecasting trends in financial markets.

By understanding the above principles, you can effectively implement and interpret linear regression in practical scenarios.

>

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 created by statistician Francis Anscombe in 1973. These datasets demonstrate that datasets with identical statistical properties can have very different distributions and relationships when visualized. This emphasizes the importance of data visualization in statistical analysis.

Each dataset in Anscombe's quartet consists of 11 data points, and they all share the following statistical properties:

1. Mean of xxx: 9.0
2. Mean of yyy: 7.5
3. Variance of xxx: 11.0
4. Variance of yyy: 4.1
5. Correlation between xxx and yyy: 0.816
6. Linear Regression Line: y=3+0.5xy = 3 + 0.5xy=3+0.5x

Despite these identical properties, each dataset looks very different when plotted.

Why is Anscombe's Quartet Important?

1. Detecting Outliers and Patterns:
    - Visualization reveals non-linear relationships, clusters, and outliers that summary statistics cannot detect.
2. Model Selection:
    - Helps decide whether a linear model is appropriate or if other models (like polynomial regression) should be used.
3. Educational Value:
    - It's a classic example demonstrating why data scientists and statisticians must go beyond summary statistics.

Anscombe's quartet is a timeless reminder that visualization is as important as numerical analysis in data science! >

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)

< Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It is widely used in data analysis to assess how strongly two variables are correlated and in which direction (positive or negative) the correlation occurs.

Formula for Pearson's R

The formula for Pearson's correlation coefficient is:

$R = \Sigma(Xi - \bar{X})(Yi - \bar{Y}) / \text{sqrt}(\Sigma(Xi - \bar{X})^2 * \Sigma(Yi - \bar{Y})^2)$

Where:

- Xi: Individual values of variable X.
- Yi: Individual values of variable Y.
- $\bar{X}$: Mean of variable X.
- $\bar{Y}$: Mean of variable Y.
- n: Number of data points.

Properties of Pearson's R

1. Range:
- The value of R lies between -1 and +1.
  - R = +1: Perfect positive linear correlation.
  - R = -1: Perfect negative linear correlation.
  - R = 0: No linear correlation.
2. Direction:
- A positive R indicates that as one variable increases, the other variable also increases.
- A negative R indicates that as one variable increases, the other variable decreases.
3. Magnitude:
- The closer R is to +1 or -1, the stronger the linear relationship.
- The closer R is to 0, the weaker the linear relationship.

Assumptions of Pearson's R

1. Linear Relationship: Assumes that the relationship between the two variables is linear.
2. Continuous Data: Both variables should be continuous (e.g., height, weight, temperature).
3. No Outliers: Outliers can significantly affect the value of R.
4. Homoscedasticity: The variance of one variable should be similar across the range of the other variable.

Interpreting Pearson's R

- R = 0.70 to 1.00: Strong positive correlation.
- R = 0.30 to 0.69: Moderate positive correlation.
- R = 0.00 to 0.29: Weak positive correlation.
- R = -0.29 to 0.00: Weak negative correlation.
- R = -0.69 to -0.30: Moderate negative correlation.
- R = -1.00 to -0.70: Strong negative correlation.

Example Calculation

Suppose we have two datasets:
Variable X: [2, 3, 5, 7, 9]
Variable Y: [4, 5, 7, 10, 15]
Steps to calculate Pearson's R:
1. Compute the means of X and Y ($\bar{X}$ and $\bar{Y}$).
2. Compute the deviations ($Xi - \bar{X}$) and ($Yi - \bar{Y}$).
3. Compute the numerator: $\Sigma(Xi - \bar{X})(Yi - \bar{Y})$.
4. Compute the denominator: sqrt($\Sigma(Xi - \bar{X})^2 * \Sigma(Yi - \bar{Y})^2$).
5. Divide the numerator by the denominator to obtain R.

Applications of Pearson's R

1. Research and Academia: Understanding relationships between variables (e.g., income and education level).
2. Business and Marketing: Identifying relationships between sales and advertising spend.
3. Healthcare: Studying the relationship between physical activity and health outcomes.
4. Social Sciences: Examining correlations between socioeconomic factors and life expectancy.

Limitations of Pearson's R

1. Linear Relationship: Only measures linear relationships; cannot detect nonlinear correlations.
2. Sensitive to Outliers: Outliers can skew the results and provide misleading interpretations.
3. Causation vs. Correlation: Pearson's R does not imply causation; it only indicates correlation.

By understanding Pearson's R and its properties, you can effectively use it to analyze linear relationships in datasets and draw meaningful insights**.**>

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 is a data preprocessing technique used in machine learning to adjust the values of features so they are on the same scale. Scaling ensures that no single feature dominates others due to its large magnitude, which can adversely affect the performance of certain algorithms.

Why is Scaling Performed?

1. Improves Model Performance:
- Some machine learning algorithms, such as gradient descent-based models (e.g., logistic regression, support vector machines) and distance-based models (e.g., k-nearest neighbors, k-means clustering), are sensitive to the scale of features. Scaling helps these algorithms converge faster and perform better.
2. Ensures Fairness:
- Without scaling, features with larger magnitudes can disproportionately influence the model's predictions, leading to biased results.
3. Reduces Numerical Instability:
- Features with vastly different ranges can lead to numerical instability during computations.

4. Speeds Up Training:
- Scaling helps algorithms converge more quickly during optimization.

Types of Scaling

1. Normalized Scaling:
- Normalization transforms the data to a specific range, usually [0, 1], by adjusting the minimum and maximum values.
- Formula: X_normalized = (X - X_min) / (X_max - X_min)
- Use Cases:
  - Useful when the distribution of data does not follow a Gaussian (normal) distribution.
  - Commonly used in image processing and when features need to be bounded.

2. Standardized Scaling:
- Standardization adjusts the data to have a mean of 0 and a standard deviation of 1, resulting in a z-score distribution.
- Formula: X_standardized = (X - μ) / σ
  - Where:
    - μ: Mean of the feature.
    - σ: Standard deviation of the feature.
- Use Cases:
  - Useful when data follows a Gaussian distribution or for algorithms that assume data is normally distributed (e.g., linear regression, principal component analysis).

Difference Between Normalized Scaling and Standardized Scaling

Aspects	Normalization	Standardization
Purpose	Rescales data to a fixed range (e.g., [0, 1])	Centers data around 0 with unit variance
Formula	(X - X_min) / (X_max - X_min)	(X - mean) / std_dev
Range	Fixed range (e.g., [0, 1] or [-1, 1])	No fixed range, mean = 0, std = 1
Impact on Outliers	Sensitive to outliers	Less sensitive to outliers
Data Assumptions	No assumptions about data distribution	Assumes data is normally distributed
Use Cases	When comparing proportions or percentages	When using models that require normally distributed data
Example Algorithms	k-NN, k-Means, Neural Networks	SVM, Logistic Regression, PCA

Summary

Scaling is a critical preprocessing step to improve the performance and stability of machine learning models. Whether to use normalization or standardization depends on the specific algorithm and dataset characteristics. Normalization is suitable for bounded values, while standardization is best for normally distributed 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)

< The value of Variance Inflation Factor (VIF) can become infinite in cases where there is perfect multicollinearity among the independent variables. This occurs when one or more predictors in a regression model are an exact linear combination of others.

Why Does VIF Become Infinite?

1. VIF Formula:
The VIF for a predictor X_i is calculated as:
VIF = 1 / (1 - R^2_i),
where R^2_i is the R^2 value obtained by regressing X_i on all other predictors.

2. Perfect Multicollinearity:
If $R^2\_i = 1$, it means the predictor X_i is perfectly explained by the other predictors. Substituting $R^2\_i = 1$ into the VIF formula results in an infinite VIF value, indicating perfect redundancy in the predictors.

When Does This Happen?

1. Duplicate Features:
If one feature is an exact duplicate or a linear combination of another feature.
Example: X2 = 2 * X1.

2. Dummy Variable Trap:
When all dummy variables for a categorical feature are included, leading to perfect collinearity.
Example: Including all dummy variables for a feature with 3 categories instead of dropping one.

3. Highly Correlated Variables:
If two or more predictors have extremely high correlations (near 1), it can result in a near-infinite VIF.

How to Handle Infinite VIF?

1. Remove Redundant Variables:
Identify and drop duplicate or highly correlated features.

2. Drop One Dummy Variable:
Use drop_first=True during one-hot encoding to avoid the dummy variable trap.

3. Regularization:
Techniques like Ridge Regression can handle multicollinearity by adding penalties to large coefficients.

4. Correlation Matrix:
Check for high correlations between features and remove one of the highly correlated variables.
>

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 a theoretical distribution, such as the normal distribution. It helps to visually assess whether the data follows a specific distribution.

- The Q-Q plot plots the quantiles of the dataset against the quantiles of the reference (theoretical) distribution.
- If the data follows the reference distribution, the points in the Q-Q plot will align closely along a 45-degree line.

How to Interpret a Q-Q Plot

1. Points on a Straight Line:

- The data matches the reference distribution (e.g., normally distributed data if compared to a normal distribution).

2. Points Deviate Systematically:
   - Indicates deviations from the reference distribution (e.g., skewness, heavy tails).
3. Points Curve Upward or Downward:
   - Suggests non-normality (e.g., data may be skewed or have outliers).

Use of Q-Q Plot in Linear Regression

In linear regression, the Q-Q plot is primarily used to check the assumption that residuals are normally distributed. Normality of residuals is important for the validity of:

1. Hypothesis Tests:
   - Many statistical tests in regression, like ttt-tests for coefficients and FFF-tests for overall model significance, assume that residuals are normally distributed.
2. Confidence Intervals:
   - Normal residuals ensure accurate confidence and prediction intervals.

Steps to Use a Q-Q Plot in Linear Regression

1. Obtain Residuals:
   - Residuals are the differences between the observed values and the predicted values from the regression model.
2. Generate the Q-Q Plot:
   - Plot the quantiles of the residuals against the theoretical quantiles of a normal distribution.
3. Interpret the Plot:
   - If residuals align along the 45-degree line, the assumption of normality holds.
   - Systematic deviations indicate potential issues with the model or data, such as outliers, skewness, or heavy tails.

Importance of Q-Q Plot in Linear Regression

1. Validates Model Assumptions:
   - A Q-Q plot ensures the residuals meet the normality assumption, which is crucial for reliable hypothesis testing and confidence intervals.
2. Detects Deviations:
   - Identifies issues like skewness, kurtosis, or outliers in the residuals that might impact model accuracy.
3. Guides Model Improvement:
   - If residuals are non-normal, transformations (e.g., log or square root) or alternative models (e.g., non-linear regression) may be necessary.

>