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)

Weathersit:
'Clear, Few clouds, Partly cloudy' weather conditions are likely to show higher rental counts. Adverse conditions like 'Mist + Cloudy' and 'Light Snow, Light Rain + Thunderstorm' may result in lower rental counts due to discomfort and safety concerns.

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** during dummy variable creation is important to avoid the issue of multicollinearity in regression models.

By using **drop_first=True**, you drop the first category and create only n-1 dummy variables. This allows the model to avoid redundancy and multicollinearity, while still capturing the information from the categorical variable.

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)

Tempeature (**temp**) has the highest correlation with the target variable **cnt**.

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)

Here are the key assumptions of Linear Regression and how you can validate them:
Linearity:
Assumption: The relationship between the independent variables and the dependent variable should be linear.
Validation: Plot the residuals (errors) against the predicted values. If the relationship is linear, there should be no clear pattern in the residuals.

Homoscedasticity:
Assumption: The variance of the residuals should be constant across all levels of the independent variables.
Validation: Check the spread of residuals against the predicted values. If the residuals have

constant variance, the spread should be uniform across the plot.

Normality of Residuals:
Assumption: The residuals should be normally distributed.
Validation: Plot a histogram or Q-Q plot of the residuals. The residuals should roughly follow a normal distribution.

Independence of Residuals:
Assumption: The residuals should be independent of each other.
Validation: Use the Durbin-Watson test to check for autocorrelation in the residuals.

No Multicollinearity:
Assumption: The independent variables should not be highly correlated with each other.
Validation: Variance Inflation Factor (VIF) for each independent variable. VIF values above 5-10 indicate multicollinearity.

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 towards explaining the demand of the shared bikes are **Temp, yr & season_winter.**

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>

Linear Regression Algorithm: An In-depth Explanation

Linear Regression is most commonly used predictive modeling techniques. Its primary goal is to model the relationship between a dependent variable (target) and one or more independent variables (predictors).

1. Conceptual Foundation

- **Objective**: To find the best-fitting line (or hyperplane in multidimensional space) that predicts the dependent variable based on the independent variables.
- **Equation**: The relationship is modeled using a linear equation:

$$Y=\beta0+\beta1X1+\beta2X2+...+\betanXn+\epsilon$$

2. Types of Linear Regression

- **Simple Linear Regression**: Involves a single independent variable.

$$Y=\beta0+\beta1X+\epsilonY = \beta_0 + \beta_1X + \epsilon$$

- **Multiple Linear Regression**: Involves two or more independent variables.

$$Y=\beta0+\beta1X1+\beta2X2+...+\betanXn+\epsilon$$

3. Assumptions of Linear Regression

- **Linearity**: The relationship between the dependent and independent variables should be linear.
- **Homoscedasticity**: The variance of residuals (errors) should be constant across all levels of the independent variables.
- **Normality**: The residuals should be normally distributed.
- **Independence**: The residuals should be independent of each other.
- **No Multicollinearity**: The independent variables should not be highly correlated with each other.

4. Estimating Coefficients

- **Ordinary Least Squares (OLS)**: The most common method for estimating the coefficients.
    - The OLS method minimizes the sum of the squared differences between the observed values and the predicted values.
    - The objective is to find β values that minimize the cost function:

$$Cost Function=\sumi=1n(Yi−Y^i)^2$$

- Yi: Actual values
- Y^i: Predicted values

5. Model Evaluation

- **R-squared (**R^2**)**: Represents the proportion of the variance in the dependent variable that is predictable from the independent variables.
    - Ranges from 0 to 1, where 1 indicates perfect prediction.
- **Adjusted R-squared**: Adjusted for the number of predictors in the model.
- **Mean Squared Error (MSE)**: Average of the squared differences between the observed and predicted values.
- **Root Mean Squared Error (RMSE)**: Square root of the MSE, providing error in the same units as the dependent variable.
- **Mean Absolute Error (MAE)**: Average of the absolute differences between observed and predicted values.

6. Interpreting Coefficients

- **Intercept (β0)**: The expected value of YY when all predictors Xi are zero.
- **Slope (βi)**: The change in YY for a one-unit change in Xi, holding other predictors constant.
- **Significance (p-values)**: Determine if the predictors are significantly contributing to the model.

7. Steps to Build a Linear Regression Model

- **Step 1: Data Preparation**: Clean the data, handle missing values, encode categorical variables, and standardize features if necessary.
- **Step 2: Train-Test Split**: Split the data into training and testing sets.
- **Step 3: Model Building**: Fit the linear regression model to the training data.
- **Step 4: Model Validation**: Validate the model assumptions using residual plots and statistical tests.
- **Step 5: Model Evaluation**: Evaluate the model's performance using metrics like R-squared, RMSE, MAE.
- **Step 6: Interpretation**: Interpret the model coefficients to understand the relationship between predictors and the target variable.

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 set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed.

Key Points of Anscombe's Quartet

Each dataset in Anscombe's quartet consists of eleven (x, y) points. The four datasets share some common statistical properties:

- **Mean of x**: Approximately the same for all datasets.
- **Variance of x**: Approximately the same for all datasets.
- **Mean of y**: Approximately the same for all datasets.
- **Variance of y**: Approximately the same for all datasets.
- **Correlation between x and y**: Approximately the same for all datasets.
- **Linear regression line**: $y = 3 + 0.5x$ (approximately the same for all datasets).

Despite these similarities in summary statistics, plotting the datasets reveals that they have very different distributions and relationships between the variables.

The Four Datasets

1. **Dataset 1**:
   - Appears as a fairly typical dataset with a linear relationship and a moderate amount of scatter around the regression line.
2. **Dataset 2**:
   - Features a perfect linear relationship between x and y, but with one outlier that greatly influences the regression line.
3. **Dataset 3**:
   - Displays a strong non-linear relationship (a quadratic shape) despite having similar statistics to the other datasets.
4. **Dataset 4**:
   - Consists of a vertical line (a single unique value of x) with one outlier, showing how a single data point can greatly affect correlation and regression analysis.

Importance

Anscombe's quartet emphasizes the importance of:

- **Visualizing data**: Graphs can reveal patterns, trends, and outliers that are not apparent from summary statistics alone.
- **Examining outliers**: Outliers can have a disproportionate impact on statistical measures and should be carefully considered.
- **Understanding data distribution**: Similar statistics can describe very different datasets, underlining the need for thorough data exploration.

Anscombe's quartet remains a powerful example in statistics and data analysis, reminding us to always visualize and critically assess our data before drawing 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 known as Pearson's correlation coefficient, is a measure of the strength and direction of association between two continuous variables. It quantifies how well the values of one variable predict the values of another

Key Points of Pearson's R

1. **Range**:
   - Pearson's R ranges from -1 to 1.
   - R=1R = 1: Perfect positive correlation. As one variable increases, the other increases proportionally.

- R=−1R = -1: Perfect negative correlation. As one variable increases, the other decreases proportionally.
- R=0R = 0: No correlation. There is no linear relationship between the variables.

2. **Interpretation**:
   - **Positive Correlation**: Values closer to 1 indicate a strong positive relationship, where both variables increase together.
   - **Negative Correlation**: Values closer to -1 indicate a strong negative relationship, where one variable increases as the other decreases.
   - **No Correlation**: Values around 0 indicate little to no linear relationship.
3. **Calculation**:
   - Pearson's R is calculated using the following formula:

$$r=(\sum(xi-\bar{x})(yi-\bar{y}))/ (\sum(xi-\bar{x})^2 * \sum(yi-\bar{y})^2)$$

- xi and yi are the individual sample points.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, respectively.

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 numerical features in a dataset to a common scale without distorting differences in ranges or distributions. It ensures that no feature dominates the learning process due to its magnitude.

Scaling Performed due to below,

1. **Improves Model Performance** – Many machine learning models (e.g., gradient descent-based algorithms like linear regression, logistic regression, SVM, and neural networks) perform better when numerical features are on a similar scale.
2. **Speeds Up Convergence** – Algorithms like gradient descent converge faster when features are scaled.
3. **Prevents Dominance of Large-Scale Features** – Features with large magnitudes can overshadow smaller ones, affecting model accuracy.
4. **Essential for Distance-Based Models** – Algorithms like K-Means, K-Nearest Neighbors (KNN), and Principal Component Analysis (PCA) rely on distances, which are affected by scale differences.

Difference Between Normalization and Standardization

Normalization (Min-Max Scaling)

- Scales data between 0 and 1 (or -1 to 1)
- Sensitive to outliers
- When features have different ranges but no extreme outliers
- Deep learning models, KNN, K-Means

Standardization (Z-score Scaling)

- Centers data around mean 0, standard deviation 1
- Less sensitive to outliers
- When data follows a normal distribution or has outliers
- Logistic regression, linear regression, SVM, PCA

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>

This typically happens due to perfect multicollinearity in our data. When two or more predictor variables in a regression model are perfectly correlated (or very nearly so), it causes the VIF to skyrocket to infinity.

VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity. When there's perfect correlation, the model can't distinguish between the variables and essentially tries to divide by zero, which mathematically results in an infinite value.

If you're encountering infinite VIFs, you'll need to address the multicollinearity issue. This can be done by:

- **Removing one of the correlated variables**: If two variables are highly correlated, consider keeping only one.
- **Combining variables**: Create a new variable that is a combination of the correlated ones.
- **Principal Component Analysis (PCA)**: Use PCA to reduce dimensionality and eliminate 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)

<Your answer for Question 11 goes here>

A Q-Q (Quantile-Quantile) plot is a graphical tool to help you understand if your data follows a certain distribution—usually a normal distribution

Q-Q Plot in Linear Regression:

Usage:

- **Checking Normality Assumption**: Linear regression assumes that the residuals (the differences between the observed and predicted values) are normally distributed. A Q-Q plot is used to check this assumption.
- **Visualizing Deviations**: It helps in identifying deviations from normality, such as skewness or kurtosis.

How to Interpret a Q-Q Plot:

1. **Perfect Normality**: If your data is perfectly normal, the points will lie on a straight line at a 45-degree angle.
2. **Deviations from Normality**: Points that deviate from the straight line suggest departures from normality.

Importance in Linear Regression:

- **Model Validation**: Ensures that the residuals are normally distributed, which is crucial for making valid statistical inferences.
- **Detecting Outliers**: Helps in identifying outliers that can affect the regression model.
- **Improving Model**: Provides insights to improve your model, such as transforming variables to achieve normality.