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 categorical variables, such as season, weathersit, and holiday, likely have a significant impact on the dependent variable (cnt) since they can influence bike-sharing demand. For instance, demand may vary with season due to weather conditions, with higher usage in pleasant seasons, and weather conditions (like clear or rainy weather) could also affect bike rentals, influencing the total demand for bikes. Additionally, holidays or weekends may see increased demand due to more people being free.

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 multicollinearity. When creating dummy variables, one of the categories (usually the first) is dropped to serve as the reference category. This prevents the "dummy variable trap," where the inclusion of all dummy variables would lead to perfect multicollinearity (i.e., the independent variables being perfectly correlated), which can cause issues in model estimation and interpretation. Dropping the first category ensures that the model doesn't redundantly account for that 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)

The variable with the highest correlation with the target variable (cnt) is **temp (temperature).**

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, I performed residual analysis by plotting residuals vs. fitted values and checking for patterns. Additionally, I examined the normality of residuals using a Q-Q plot and assessed homoscedasticity to ensure constant variance of 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)

Based on the final model, the top 3 features contributing significantly towards explaining the demand for shared bikes are:

1. **Temperature (temp)**
2. **Holiday**
3. **Weather situation (weathersit)**

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>

Introduction: Linear Regression is a statistical technique used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship between the dependent and independent variables. The goal is to find the line (or hyperplane in multiple dimensions) that best fits the data, allowing predictions for new data points.

1. **Simple Linear Regression (SLR)**

In Simple Linear Regression, we predict a dependent variable YYY based on a single independent variable XXX. The relationship is represented by the equation:

Y=β0+β1X+ϵY = \beta_0 + \beta_1 X + \epsilonY=β0+β1X+ϵ

Where:

- YYY is the dependent variable.
- XXX is the independent variable.
- β0\beta_0β0 is the intercept.
- β1\beta_1β1 is the slope (coefficient) indicating the change in YYY for each unit change in XXX.
- ϵ\epsilonϵ is the error term.

2. **Multiple Linear Regression (MLR)**

Multiple Linear Regression is an extension of Simple Linear Regression to handle more than one independent variable. The equation is:

Y=β0+β1X1+β2X2+⋯+βnXn+ϵY = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilonY=β0+β1X1+β2X2+⋯+βnXn+ϵ

Where X1,X2,...,XnX_1, X_2, ..., X_nX1,X2,...,Xn are multiple independent variables, and β1,β2,...,βn\beta_1, \beta_2, ..., \beta_nβ1,β2,...,βn are the corresponding coefficients.

3. Model Fitting and Coefficients Estimation

Linear Regression uses methods like Ordinary Least Squares (OLS) to estimate the coefficients β0,β1,...,βn\beta_0, \beta_1, ..., \beta_nβ0,β1,...,βn by minimizing the sum of squared errors (SSE).

The formula for coefficients in MLR is:

β=(XTX)−1XTy\beta = (X^T X)^{-1} X^T yβ=(XTX)−1XTy

Where:

- XXX is the matrix of input features.
- yyy is the target variable.

4**. Assumptions**

Linear Regression makes several assumptions:

- **Linearity**: The relationship between predictors and the target is linear.
- **Independence of errors**: Residuals are independent.
- **Homoscedasticity**: Equal variance of residuals across all levels of independent variables.
- **Normality**: Residuals are normally distributed.
- **No multicollinearity**: Predictors are not highly correlated.

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 created by statistician Francis Anscombe in 1973 to highlight the importance of data visualization and the limitations of relying solely on summary statistics. The quartet consists of four different datasets, each with 11 data points for two variables, XXX and YYY. Despite having identical summary statistics—such as mean, variance, and correlation—each dataset displays drastically different relationships between XXX and YYY when plotted.

The four datasets all share the following summary statistics:

- The mean of XXX and YYY are the same.
- The variance of XXX and YYY is identical across datasets.
- The correlation between XXX and YYY is 0.82 for all datasets.
- The regression line for all datasets is the same: Y=3+0.5XY = 3 + 0.5XY=3+0.5X.

However, when visualized:

1. Dataset I shows a clear linear relationship between XXX and YYY, confirming the regression line's fit.
2. Dataset II has a linear relationship but includes an outlier that greatly affects the regression line, demonstrating the impact of outliers.
3. Dataset III shows a perfect parabolic relationship, where the linear regression model does not fit well, despite the high correlation.
4. Dataset IV contains a strong linear pattern but also features an influential outlier that distorts the regression analysis.

Anscombe's Quartet demonstrates that relying solely on numerical summaries can be misleading. It underscores the importance of data visualization in identifying patterns, outliers, and non-linear relationships, as these can significantly impact the choice of analysis and model selection.

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 the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It provides a value between -1 and 1, where:

- +1 indicates a perfect positive linear relationship (as one variable increases, the other also increases in a perfectly linear manner).
- -1 indicates a perfect negative linear relationship (as one variable increases, the other decreases in a perfectly linear manner).
- 0 indicates no linear relationship between the two variables.

Formula:

The Pearson correlation coefficient is calculated using the formula:

r=n∑xy−∑x∑y[n∑x2−(∑x)2][n∑y2−(∑y)2]r = \frac{n\sum{xy} - \sum{x}\sum{y}}{\sqrt{[n\sum{x^2} - (\sum{x})^2][n\sum{y^2} - (\sum{y})^2]}}r=[n∑x2−(∑x)2][n∑y2−(∑y)2]n∑xy−∑x∑y

Where:

- xxx and yyy are the individual sample points of the two variables.
- nnn is the number of paired scores.

Interpretation:

- **r > 0:** A positive correlation. As one variable increases, the other tends to increase.
- **r < 0**: A negative correlation. As one variable increases, the other tends to decrease.
- **r = 0**: No linear correlation.
- **r ≈ 1** or **r ≈ -1**: A very strong positive or negative linear relationship.

Limitations:

- **Linearity**: Pearson's R only measures linear relationships. It cannot capture non-linear relationships between variables.
- **Outliers**: Pearson's R is sensitive to outliers, which can disproportionately affect the result.

Overall, Pearson's R is a valuable tool for understanding the linear relationship between two continuous variables, but it should be used in conjunction with visual analysis and other methods to assess the overall relationship between variables.

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 refers to the process of adjusting the values of numerical data to fit within a specific range or distribution. This is done to ensure that all features in a dataset have a comparable scale, preventing features with larger ranges from disproportionately affecting machine learning models. Scaling is critical when using algorithms sensitive to feature magnitude, such as k-Nearest Neighbors (k-NN) or Support Vector Machines (SVM).

Scaling is necessary for several reasons:

1. **Equal Influence**: Many machine learning algorithms, especially those based on distances like k-NN or gradient-based methods like neural networks, perform better when features are on similar scales. Without scaling, features with larger numerical ranges could dominate the model.
2. **Optimization Efficiency**: Algorithms like Gradient Descent can converge faster when features are scaled. Large-scale differences between features might slow down the convergence process.
3. **Improved Performance**: Some models, like neural networks, are sensitive to the input data scale and may fail to learn effectively if the data isn't properly scaled.

Normalized vs. Standardized Scaling

1. **Normalized Scaling (Min-Max Scaling):**
   - Definition: This technique scales features to a fixed range, typically [0, 1]. The transformation involves subtracting the minimum value of the feature and dividing by the range (max-min).
   - Formula: Xscaled=X−XminXmax−XminX_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}Xscaled=Xmax−XminX−Xmin
   - Use Case: Normalization is ideal when algorithms need data in a specific range, such as Neural Networks or k-NN.
2. **Standardized Scaling (Z-score Scaling):**
   - Definition: Standardization transforms data so that it has a mean of 0 and a standard deviation of 1. It is less sensitive to outliers than normalization.
   - Formula: Xscaled=X−μσX_{\text{scaled}} = \frac{X - \mu}{\sigma}Xscaled=σX−μ where μ\muμ is the mean and σ\sigmaσ is the standard deviation.
   - Use Case: Standardization is preferred for algorithms assuming data is normally distributed, like Linear Regression and 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>

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. When the VIF becomes infinite, it indicates perfect multicollinearity between two or more independent variables, meaning one variable can be perfectly predicted by another.

The VIF is calculated using the formula:

VIFj=11−Rj2VIF_j = \frac{1}{1 - R_j^2}VIFj=1−Rj21

Where Rj2R_j^2Rj2 is the coefficient of determination when the j-th variable is regressed on all other predictors. If Rj2R_j^2Rj2 is equal to 1, it suggests that one variable is perfectly correlated with the others, leading to an infinite VIF.

Infinite VIF arises when:

1. Perfect Linear Correlation: Two or more variables are perfectly linearly correlated, causing perfect prediction.
2. Redundant Variables: If predictors are nearly identical, multicollinearity becomes extreme.

This results in:

- Unstable Coefficients: The model cannot distinguish between the correlated variables.
- Inaccurate Significance Testing: It becomes difficult to assess the individual impact of variables.

To resolve infinite VIF, remove one of the correlated variables or apply techniques like Principal Component Analysis (PCA) to reduce redundancy.

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 tool used to assess if a dataset follows a specific theoretical distribution, such as a normal distribution. The plot compares the quantiles of the sample data against the quantiles of a reference distribution. If the data follows the reference distribution, the points will lie approximately along a straight line.

How a Q-Q Plot Works:

- The x-axis represents the quantiles of the reference distribution (usually normal).
- The y-axis represents the quantiles of the sample data.
- If the points form a straight line (usually diagonal), this indicates that the data follows the reference distribution.
- If the points deviate significantly from the line, it suggests that the data does not follow the reference distribution.

Use and Importance of a Q-Q Plot in Linear Regression:

In the context of linear regression, a Q-Q plot is commonly used to assess whether the residuals (errors) of the model are normally distributed, which is an important assumption for linear regression.

1. **Normality of Residuals**: One of the key assumptions of linear regression is that the residuals are normally distributed. The Q-Q plot helps visually check this assumption. If the residuals are normally distributed, the Q-Q plot should show points lying on a straight line.
2. **Model Validity**: Violations of the normality assumption (such as heavy skewness or fat tails) might suggest problems with the model, indicating that the linear regression assumptions are not fully met. In such cases, transformation of variables or alternative models may be considered.
3. **Detecting Outliers**: Q-Q plots can also help identify outliers or extreme values in the residuals, which could influence the regression model and lead to misleading conclusions.

In summary, a Q-Q plot is a useful diagnostic tool in linear regression to check the assumption of normally distributed residuals, assess model validity, and identify potential outliers.
