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)

Effect of Categorical Variables on `cnt`:

- **Season**: Higher rentals in Summer/Fall, lower in Winter.
- **Year**: Increasing trend if bike usage grows over time.
- **Month**: Warmer months (e.g., June) likely see higher rentals.
- **Holiday**: Higher rentals for leisure on holidays, possibly lower if focused on commuting.
- **Weekday**: Higher rentals on weekdays for commuting; weekends may vary.
- **Weather**: Clear weather boosts rentals; adverse weather reduces them.

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)

Importance of `drop_first=True`:

- **Avoids Multicollinearity**: Prevents the dummy variable trap by reducing redundancy.
- **Establishes Baseline**: Allows interpretation of dummy coefficients 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)

temp (temperature) often has the highest positive correlation with cnt because warmer temperatures usually encourage more bike rentals

---

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 on the training set, we can perform the following checks:

1. **Linearity**:
    - **Residual Plot**: Plot residuals vs. predicted values. Residuals should be randomly scattered around zero, indicating a linear relationship.
2. **Homoscedasticity (Constant Variance of Errors)**:
    - **Residual Plot**: Again, check the residual plot to see if the variance remains constant across predicted values. If residuals fan out or form patterns, heteroscedasticity may be an issue.
3. **Normality of Errors**:
    - **Residual Distribution Plot**: Use a histogram or density plot of residuals to check for a normal distribution.
4. **Multicollinearity**:
    - **Variance Inflation Factor (VIF)**: Calculate VIF for each predictor. A high VIF (usually above 5 or 10) indicates multicollinearity, suggesting that some variables may need to be removed.

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)

To identify the top 3 features contributing significantly to the demand for shared bikes in the final model, you can look at:

1. **Coefficient Magnitude**: In the final model summary from `statsmodels`, the features with the largest absolute values of coefficients typically have the most impact on the target variable, assuming they're statistically significant.
2. **p-Values**: Check that the p-values of these features are below a chosen significance level (commonly 0.05), which indicates statistical significance.

From typical bike-sharing datasets, the top features often include:

1. **`temp` (Temperature)**: Higher temperatures generally correlate with increased bike usage.
2. **`yr` (Year)**: This variable often shows an increase in demand over time if bike-sharing adoption grows.
3. **`season` or `weathersit` (Weather Situation)**: Clear weather or favorable seasons tend to have a positive impact on 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)

<Your answer for Question 6 goes here>

---

1. Concept and Equation

In a simple linear regression (one independent variable), the relationship is modeled by a line defined by the equation:

y=β0+β1x+ϵy = \beta_0 + \beta_1 x + \epsilony=β0+β1x+ϵ

- yyy: Dependent variable (target variable we want to predict)
- xxx: Independent variable (feature variable)
- β0\beta_0β0: Intercept (value of yyy when x=0x = 0x=0)
- β1\beta_1β1: Slope (how much yyy changes for a unit increase in xxx)
- ϵ\epsilonϵ: Error term (captures the difference between the actual and predicted values)

In **multiple linear regression** (more than one independent variable), the equation expands to:

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 each βi\beta_iβi represents the coefficient for the corresponding feature xix_ixi.

2. Goal of Linear Regression

The goal of linear regression is to find the best values for β0,β1,…,βn\beta_0, \beta_1, \dots, \beta_nβ0,β1,…,βn (coefficients) that minimize the error between the predicted values (y^\hat{y}y^) and the actual values (yyy).

3. Ordinary Least Squares (OLS)

The most common method for finding the best-fit line in linear regression is the **Ordinary Least Squares (OLS)** method. OLS aims to minimize the **sum of squared errors** (SSE):

SSE=∑i=1n(yi−y^i)2SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2SSE=i=1∑n(yi−y^i)2

Where:

- yiy_iyi: Actual value for observation iii
- y^i\hat{y}_iy^i: Predicted value for observation iii

OLS minimizes the squared differences between the observed values and the predicted values, resulting in a line that best fits the data points.

4. Steps in Linear Regression

1. **Fit the Model**:
   - Using OLS, find the best-fit line by calculating the coefficients β\betaβ that minimize the SSE.
2. **Make Predictions**:
   - Once the model is trained, predictions for new data can be made by plugging the values of xxx into the model.
3. **Evaluate the Model**:
   - **R^2 (Coefficient of Determination)**: Measures how much of the variance in the dependent variable is explained by the model. An R^2 of 1 means the model explains all the variance; an R^2 of 0 means it explains none.
   - **Residual Analysis**: Analyze residuals (errors) to validate assumptions.

5. Assumptions of Linear Regression

Linear regression relies on certain assumptions to produce accurate results:

1. **Linearity**: The relationship between the dependent and independent variables should be linear.
2. **Independence of Errors**: Residuals (errors) should be independent.
3. **Homoscedasticity**: The variance of residuals should be constant across all levels of the independent variables.
4. **Normality of Errors**: Residuals should be normally distributed.
5. **No Multicollinearity**: In multiple regression, independent variables should not be highly correlated.

Violating these assumptions may lead to biased or inefficient estimates.

6. Interpretation of Coefficients

- The coefficient βi\beta_iβi for each feature xix_ixi indicates the average change in yyy for a one-unit increase in xix_ixi, holding all other variables constant.
- The intercept β0\beta_0β0 represents the predicted value of yyy when all xxx values are zero.

7. Advantages and Limitations

- **Advantages**: Linear regression is simple, interpretable, and works well for linearly separable data.
- **Limitations**: It struggles with non-linear relationships, is sensitive to outliers, and may underperform if assumptions are violated.

8. Extensions

- **Polynomial Regression**: Introduces polynomial terms to capture non-linear relationships.

- **Regularized Linear Regression**: Techniques like Lasso and Ridge Regression add penalties to the coefficients, helping address overfitting and multicollinearity.

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>

demonstrate the importance of data visualization when analyzing data. Each dataset in the quartet has nearly identical statistical properties:

- **Mean** of xxx and yyy values
- **Variance** of xxx and yyy
- **Correlation** between xxx and yyy
- **Linear regression line**

However, despite these similarities, the datasets have very different distributions and relationships between xxx and yyy when visualized:

1. **Dataset 1**: A standard linear relationship, as expected.
2. **Dataset 2**: A clear non-linear relationship.
3. **Dataset 3**: A linear relationship with an outlier affecting the regression line.
4. **Dataset 4**: A constant xxx value except for one outlier, distorting the correlation.

Key Takeaway

Anscombe's Quartet shows that relying solely on summary statistics can be misleading. Visualizing data is essential for accurately understanding patterns, relationships, and potential outliers in a dataset.

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** or simply **correlation coefficient (r)**, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.

Formula

The formula for Pearson's R is:

***Pearson's R**, also known as the **Pearson correlation coefficient** or simply **correlation coefficient (r)**, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.*

Formula

The formula for Pearson's R is:

r=∑(xi−x¯)(yi−y¯)∑(xi−x¯)2∑(yi−y¯)2r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}r=∑(xi−x¯)2∑(yi−y¯)2∑(xi−x¯)(yi−y¯)

Where:

- *xix_ixi and yiy_iyi are the individual data points of variables XXX and YYY.*
- *x¯\bar{x}x¯ and y¯\bar{y}y¯ are the means of XXX and YYY.*

Interpretation of Pearson's R

- ***Range**: rrr ranges from -1 to 1.*
  - ***+1**: Perfect positive linear relationship (as XXX increases, YYY increases).*
  - ***-1**: Perfect negative linear relationship (as XXX increases, YYY decreases).*
  - ***0**: No linear relationship.*
- ***Strength of Relationship**:*
  - ***|r| close to 1**: Strong linear relationship.*
  - ***|r| close to 0**: Weak linear relationship.*

Key Points

- ***Linear Relationship**: Pearson's R only captures linear relationships. Non-linear relationships may yield an rrr value close to zero, even if there is a strong non-linear pattern.*
- ***Sensitivity to Outliers**: Pearson's R can be significantly affected by outliers, which may distort the correlation.*

Use Cases

Pearson's R is commonly used in statistics, research, and data science to evaluate the relationship between variables, such as understanding correlations in financial markets, health studies, or social science research.

)

Key Points

- **Linear Relationship**: Pearson's R only captures linear relationships. Non-linear relationships may yield an rrr value close to zero, even if there is a strong non-linear pattern.
- **Sensitivity to Outliers**: Pearson's R can be significantly affected by outliers, which may distort the correlation.

Use Cases

Pearson's R is commonly used in statistics, research, and data science to evaluate the relationship between variables, such as understanding correlations in financial markets, health studies, or social science research.

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>

Why Scaling is Performed:

- **Consistency**: Ensures all features contribute equally to the model.
- **Improves Model Performance**: Certain algorithms (e.g., gradient-based methods) perform better with scaled data.
- **Faster Convergence**: Models converge faster with scaled features.

Types of Scaling:

1. **Normalized Scaling (Min-Max Scaling)**:
    - **Definition**: Scales values to a range, typically [0, 1].
    - **Formula**: (x−min)/(max−min)(x - \text{min}) / (\text{max} - \text{min})(x−min)/(max−min)
    - **Use Case**: Suitable for algorithms sensitive to feature scales (e.g., neural networks, KNN).
    - **Outcome**: Keeps the shape of the distribution but changes the scale.
2. **Standardized Scaling (Z-score Scaling)**:
    - **Definition**: Scales features to have a mean of 0 and standard deviation of 1.
    - **Formula**: (x−μ)/σ(x - \mu) / \sigma(x−μ)/σ, where μ\muμ is mean and σ\sigmaσ is standard deviation.
    - **Use Case**: Suitable for algorithms that assume a Gaussian distribution (e.g., linear regression, SVM).
    - **Outcome**: Centers data around zero and adjusts for variance.

In summary:

- **Normalization** rescales data to a fixed range.
- **Standardization** centers data with zero mean and unit variance.

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>

A **Variance Inflation Factor (VIF)** becomes infinite when there is **perfect multicollinearity** among the predictor variables, meaning one predictor variable can be

perfectly predicted by one or more of the other predictors. This perfect linear relationship causes the denominator in the VIF calculation to be zero, leading to an infinite VIF.

Reasons for Infinite VIF:

1. **Duplicate Features**: When identical or nearly identical columns are included (e.g., two features with the exact same values).
2. **Linear Combinations**: When one feature is a linear combination of others (e.g., Feature A=2×Feature B\text{Feature A} = 2 \times \text{Feature B}Feature A=2×Feature B).

Solution:

To handle infinite VIF, identify and remove or combine the perfectly collinear variables, reducing multicollinearity in the model.

4o

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 compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution.

How a Q-Q Plot Works:

- **X-axis**: Quantiles of the theoretical (expected) distribution, often normal.
- **Y-axis**: Quantiles of the observed (sample) data.

If the observed data follows the expected distribution, the points will roughly align along a 45-degree line. Deviations from this line indicate departures from the expected distribution.

Use and Importance in Linear Regression:

In linear regression, one assumption is that the **residuals (errors) follow a normal distribution**. A Q-Q plot of the residuals helps assess this assumption:

- **Straight Line**: Residuals are normally distributed, supporting the assumption.
- **Curved or Deviated Line**: Indicates non-normality, suggesting issues with the linear model, such as non-linearity or outliers.

Importance:

- **Validates Model Assumptions**: Ensures that normality assumptions of residuals are met, which is critical for accurate confidence intervals and hypothesis tests.
- **Identifies Outliers**: Helps detect outliers or heavy tails in the distribution, which may affect model performance.
