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)

Summer and winter seasons has positive correlation with dependent variable and weather (Light snow and mist) has negative correlation.

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)

We can identify the categorical variables with n-1 values. So it's advisable to drop drop_first=true, So that we can avoid too many feature variables to reduce load during model building.

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 , atemp has high correlation with target variable.

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)

We will do the residual analysis and identify the difference between y_train and y_train_pred in a graph to see the pattern.

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)

Year, atemp, season are the 3 features which are significantly contributed.

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

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fitting line that minimizes the difference between the predicted values and the actual values.

Here's a step-by-step breakdown:

1. **Data Preparation:**
    - **Data Collection:** Gather relevant data for the dependent and independent variables.
    - **Data Cleaning:** Handle missing values, outliers, and inconsistencies.
    - **Feature Engineering:** Create or transform features to improve model performance.
2. **Model Specification:**
    - **Linear Equation:** Define the linear relationship between the dependent variable (y) and independent variables (x1, x2, ...):
    - `y = mx1 + bx2 + ... + c`
        - `m`: Slope coefficients
        - `b`: Intercept
        - `c`: Constant term
3. **Model Training:**
    - **Least Squares Method:** Find the optimal values for the coefficients (m, b, c) that minimize the sum of squared errors between the predicted and actual values.
    - **Gradient Descent:** An iterative optimization algorithm to find the optimal coefficients by adjusting them in the direction of the steepest descent of the error function.
4. **Model Evaluation:**
    - **R-squared:** Measures the proportion of variance in the dependent variable explained by the model.
5. **Model Interpretation:**
    - **Coefficient Interpretation:** Analyze the coefficients to understand the impact of each independent variable on the dependent variable.
    - **Statistical Significance:** Use hypothesis testing to determine if the coefficients are statistically significant.
6. **Model Prediction:**
    - Once the model is trained, use it to make predictions on new, unseen data.
    - Input the values of the independent variables into the equation and calculate the predicted value of the dependent variable.

Key Points:

- **Simple Linear Regression:** Involves one independent variable.
- **Multiple Linear Regression:** Involves multiple independent variables.
- **Assumptions:** Linearity, independence of errors, homoscedasticity, and normality of errors.
- **Overfitting and Underfitting:** Avoid these issues by using techniques like regularization and cross-validation.

By following these steps and considering the assumptions of linear regression, you can build effective models to make accurate predictions.

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 famous set of four datasets that, despite having nearly identical statistical properties, look very different when visualized. This quartet was created by Francis Anscombe in 1973 to highlight the importance of data visualization and the limitations of relying solely on summary statistics.

Key Points about Anscombe's Quartet:

1. **Identical Statistical Properties:**
    - Each dataset has the same mean, variance, correlation coefficient, and linear regression line.
    - This might lead one to believe that the datasets are very similar.
2. **Distinct Visual Representations:**
    - When plotted, the four datasets reveal very different patterns:
        - **Dataset I:** A clear linear relationship between x and y.
        - **Dataset II:** A quadratic relationship with a single outlier.
        - **Dataset III:** A perfect linear relationship except for one outlier.
        - **Dataset IV:** A nearly constant x-value with varying y-values.

The Importance of Data Visualization:

Anscombe's quartet underscores the importance of visualizing data before drawing conclusions. While summary statistics can provide a quantitative overview, they may not capture the underlying patterns and nuances of the data. Data visualization helps to:

- **Identify outliers and anomalies:** Visual inspection can reveal unusual data points that might not be evident from summary statistics.
- **Detect non-linear relationships:** Plots can uncover patterns that are not linear, such as quadratic or exponential relationships.
- **Understand the distribution of data:** Visualizations can help identify skewness, kurtosis, and other distributional characteristics.
- **Explore relationships between variables:** Scatter plots can reveal correlations and trends that might not be apparent from numerical summaries.

In Conclusion:

Anscombe's quartet serves as a powerful reminder that data visualization is an essential tool in data analysis.

By combining statistical analysis with visual exploration, we can gain deeper insights into data and make more informed decisions.

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 a linear relationship between two variables. It is denoted by the letter "r".

Key characteristics of Pearson's R:

- **Range:** It can take values between -1 and +1.
- **Interpretation:**
  - **+1:** Perfect positive correlation (as one variable increases, the other increases proportionally).
  - **-1:** Perfect negative correlation (as one variable increases, the other decreases proportionally).
  - **0:** No linear correlation between the variables.

When to use Pearson's R:

- **Linear Relationship:** It is most suitable for measuring linear relationships between variables.
- **Continuous Variables:** Both variables should be continuous (e.g., numerical).
- **Normal Distribution:** The variables should ideally be normally distributed.
- **Outlier Sensitivity:** Be cautious of outliers as they can significantly impact the correlation coefficient.

Limitations of Pearson's R:

- **Non-linear Relationships:** It may not be suitable for non-linear relationships (e.g., exponential or logarithmic).
- **Outliers:** Outliers can distort the correlation coefficient, leading to misleading results.
- **Causation:** Correlation does not imply causation. A high correlation between two variables does not necessarily mean that one causes the other.

By understanding the strengths and limitations of Pearson's R, you can effectively use it to analyze the relationships between variables in your data.

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 transforming numerical features to a common scale. It's crucial for many machine learning algorithms as they assume features have similar scales. **Normalized scaling** scales features to a specific range (e.g., 0 to 1), while **standardized scaling** scales features to have 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 VIF value of infinity indicates perfect multicollinearity, meaning that one or more independent variables are perfectly correlated with each other. This occurs when one variable is a linear combination of others, leading to unstable and unreliable regression coefficients.

This can happen due to various reasons, such as:

- **Redundant features:** Including multiple features that convey the same information.
- **Data entry errors:** Mistakes in data entry can create artificial correlations.
- **Data transformations:** Certain transformations can introduce 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 plot is a graphical tool to assess the normality of residuals in linear regression. It compares the quantiles of the observed residuals to the quantiles of a theoretical normal distribution. A straight line on the Q-Q plot indicates normality. Deviations from the line suggest non-normality, which can affect the model's validity and predictions.
