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 spring season and specific weather conditions (light snow/rain and mist) have statistically significant negative effects on the dependent variable.
The summer and winter seasons do not show strong statistical evidence of affecting the dependent variable.

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 in the model, which can distort the estimation of coefficients. If we include all the categories, one of the dummy variables can be perfectly predicted by the others. This leads to perfect multicollinearity, which makes the model's coefficients non-identifiable

---

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)

According to correlation_matrix, temp and atemp has highest correlation of 0.63 with cnt (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)

1) Checked for multicollinearity: checked if VIF > 5 for any of the predictor
2) build the model again with reduced list of columns
3) Checked residuals if they are independent of each other, Durbin-Watson statistic was 2.003
4) Checked residuals vs predicted values plot
5) Checked Q-Q plot - following the 45 degree line
6) Histogram of residuals normally distributed
7) Checked scale location plot
8) Scaled the test data set and made predictions on that
9) calculated its r2_score
10) compared it with train data set

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:

- **Coefficient**: 2119.12
- **Significance**: This variable has a highly significant positive coefficient (p-value = 0.000), suggesting that as the year progresses, there is an increase in bike demand. Given the magnitude of the coefficient, it is the most significant feature in explaining the demand for shared bikes

Temperature:

- **Coefficient**: 653.03
- **Significance**: Temperature also has a strong positive effect on the bike demand, with a highly significant p-value (p-value = 0.000). Higher temperatures are associated with an increase in the demand for shared bikes, indicating that warmer weather encourages more people to use bikes.

Weather Situation (Light Snow/Rain):

- **Coefficient**: -2391.83
- **Significance**: The weather condition of light snow or rain has a strong negative impact on the bike demand, with a highly significant p-value (p-value = 0.000). This indicates that such weather conditions significantly reduce the demand for bikes.

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 is a statistical method used to model the relationship between a dependent variable yyy and one or more independent variables x1,x2,...,xpx_1, x_2, ..., x_px1,x2,...,xp by fitting a linear equation to the observed data. The model is represented as:

$y=\beta0+\beta1x1+\beta2x2+...+\betapxp+\epsilon$

Key Steps in Linear Regression:

1. **Objective**: The goal is to minimize the **sum of squared residuals** (SSR), which is the difference between observed and predicted values.

Assumptions: Linear regression assumes:

- **Linearity**: The relationship between predictors and target is linear.
- **Independence**: Observations are independent.
- **Homoscedasticity**: Constant variance of residuals.
- **Normality**: Residuals should be normally distributed.

Evaluation: After training, model performance is evaluated using metrics like **R-squared**, which measures how well the model explains the variance in the dependent variable, and **Mean Squared Error (MSE)**, which measures prediction accuracy.

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 (mean, variance, correlation), yet they exhibit very different distributions and relationships between the variables. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before drawing conclusions from statistical analyses.

Key Points about Anscombe's Quartet:

1. **Four Datasets**: The quartet consists of four different datasets, each with 11 data points. All datasets have the same:
    - **Mean of x**: 9
    - **Mean of y**: 7.5
    - **Variance of x**: 11
    - **Variance of y**: 4.12
    - **Correlation between x and y**: 0.82
2. Despite these statistics being identical, the datasets reveal different underlying patterns when visualized.
3. **Differences in Data Patterns**:
    - **Dataset 1**: Shows a strong linear relationship between xxx and yyy.
    - **Dataset 2**: Appears to have a non-linear relationship, with a quadratic shape.
    - **Dataset 3**: Has a near-constant yyy value, except for one outlier that causes a high correlation.
    - **Dataset 4**: Displays a perfect vertical line of data points, with a single outlier.
4. **Lesson from Anscombe's Quartet**:
    - The quartet illustrates that **descriptive statistics like mean, variance, and**

correlation can be misleading and fail to capture the actual structure of the data.
    - **Visualization is crucial** in data analysis, as it can reveal underlying patterns, outliers, and relationships that are not apparent from summary statistics alone.

Anscombe's quartet emphasizes the importance of using both statistical metrics and graphical methods to fully understand and interpret data.

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, or Pearson correlation coefficient, is a measure that quantifies the strength and direction of the **linear relationship** between two continuous variables. It ranges from -1 to 1, where:

1. A value of **1** indicates a **perfect positive linear relationship**, meaning both variables increase together.
2. A value of **-1** indicates a **perfect negative linear relationship**, meaning as one variable increases, the other decreases.
3. A value of **0** suggests **no linear relationship** between the variables.

Pearson's R assumes that the relationship between the variables is linear and that the data is approximately normally distributed. It is sensitive to outliers, which can significantly impact the value of the correlation.

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 adjusting the range or distribution of features in a dataset so that they are on a comparable scale. This is important because many machine learning algorithms perform better or converge faster when the data is scaled.

Why Scaling is Performed:

- **Improves model performance**: Some algorithms, like gradient descent-based methods, require features to be on the same scale to perform optimally.
- **Prevents bias**: Features with larger ranges or units can dominate the model, leading to biased results.
- **Enhances convergence speed**: Algorithms can converge more quickly when all features are

scaled similarly.

Difference Between Normalized and Standardized Scaling:

1. **Normalized Scaling**: This technique rescales the data to a fixed range, usually between 0 and 1. It is useful when we want all features to have the same scale and are particularly important when using algorithms like neural networks that rely on distance metrics.
2. **Standardized Scaling**: This technique transforms the data to have a mean of zero and a standard deviation of one. It is useful when the data follows a normal distribution or when the model relies on distances (like k-nearest neighbors or linear regression), ensuring that the features have equal importance regardless of their original scale.

In summary, normalization rescales data to a specific range, while standardization centers and scales data around a mean of zero with 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>

The value of **Variance Inflation Factor (VIF)** can become **infinite** when there is **perfect multicollinearity** between two or more independent variables in the regression model. This happens because VIF measures how much the variance of a regression coefficient is inflated due to the correlation between the independent variables.

Reasons for Infinite VIF:

1. **Perfect Multicollinearity**: When one independent variable is a perfect linear function of another (e.g., one variable is a direct multiple of another), the model cannot distinguish their individual effects. This causes the variance of the estimated coefficients to be infinitely large.
2. **Linear Dependence**: If the predictors in the model are highly correlated, the inverse of the correlation matrix becomes singular, leading to infinite VIF values.

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 whether a dataset follows a specific theoretical distribution, typically the **normal distribution**. It compares the quantiles of the data with the quantiles of the theoretical distribution.

Use and Importance in Linear Regression:

1. **Assess Normality of Residuals**: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot helps visually check if the residuals follow a normal distribution, which is crucial for valid hypothesis testing and confidence intervals.
2. **Identify Deviations**: The plot helps identify deviations from normality, such as skewness, heavy tails, or outliers. If the points in the Q-Q plot form a straight line, it indicates that the residuals are approximately normally distributed. Deviations from the line suggest non-normality, which may require transformation of the data or the use of non-parametric models.
3. **Validate Assumptions**: Normality of residuals is important for ensuring that the regression model's inferential statistics (like p-values) are reliable. By using a Q-Q plot, you can validate this assumption and decide if any further steps (like data transformation) are needed to improve model accuracy and interpretation.

In summary, a Q-Q plot is a valuable diagnostic tool in linear regression for checking the normality assumption of residuals, which directly impacts the validity of statistical tests and model performance.
