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)

We can infer the following

- During the months of October and September there is demand in bike. Likewise, when the temperature is warm
- Demand decreases specially for the conditions when it rains or snows or if its misty or cloudy conditions
- Month march is redundant as march comes under Spring season
- Demand improved significantly in the year 2019.

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 removes one dummy variable to prevent confusion for the model. This helps the model avoid unnecessary duplication of information, making it easier to understand the impact of other categories compared to the one that's removed.

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)

Columns temp and atemp have highest 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)

Once the model is trained and shows strong performance metrics—such as high R-squared and adjusted R-squared values, low VIF, and statistically significant p-values for the features—it's important to perform residual analysis on the trained dataset.

Residuals should ideally be randomly scattered with a mean of zero. This indicates that the model captures the underlying patterns in the data effectively. Additionally, checking whether the adjusted R-squared value on the test set is close to that of the training set helps confirm that the model generalizes well and avoids overfitting.

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)

1. Yr: Year has significant impact on bike demand
2. Temp: Temperature has positive impact on bike demand
3. Oct and Sep: October and September months; there is 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>
Linear Regression is a supervised learning algorithm used to predict a continuous target variable YYY based on one or more predictor variables (X1,X2,…,XnX_1, X_2, \dots, X_nX1,X2,…,Xn). The goal is to fit a linear relationship between the input variables (independent variables) and the output (dependent variable).

- **Simple Linear Regression**:
    - Involves only **one independent variable** (XXX).
    - The formula is: Y=b0+b1X+eY = b_0 + b_1 X + eY=b0+b1X+e
    - Where:
        - b0b_0b0: Intercept (value of YYY when X=0X = 0X=0).
        - b1b_1b1: Coefficient (slope of the line representing the relationship between XXX and YYY).
        - eee: Error term, representing noise in the data.

- **Multiple Linear Regression**:
    - Involves **more than one independent variable** (X1,X2,…,XnX_1, X_2, \dots, X_nX1,X2,…,Xn).
    - The formula is: Y=b0+b1X1+b2X2+⋯+bnXn+eY = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_n X_n + eY=b0+b1X1+b2X2+⋯+bnXn+e
    - Each coefficient (bib_ibi) represents the slope or effect of the respective independent variable on YYY, holding all other variables constant.

Advantages of Linear Regression:

1. **Simplicity**: Easy to implement and interpret.
2. **Efficiency**: Computationally inexpensive, especially for small datasets.
3. **Linear Interpretability**: Coefficients clearly show the relationship between predictors and the target variable.
4. **Works Well for Linearly Separable Data**: Ideal for data where a linear relationship exists between independent and dependent variables.

Disadvantages of Linear Regression:

Overfitting with High Dimensionality: Adding too many predictors can lead to overfitting, especially with limited data.

Sensitive to Outliers: Outliers can significantly affect the coefficients, leading to a poor fit.

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)

Anscombe's Quartet is a set of **four datasets** that share nearly identical descriptive statistical properties, such as **means, variances, R-squared values, correlation coefficients, and linear regression equations**. However, these datasets display dramatically different patterns when plotted as scatter plots.

Created by statistician **Francis Anscombe in 1973**, the quartet highlights the importance of **visualizing data** to avoid being misled by summary statistics alone.

Key Characteristics

- Each dataset in Anscombe's Quartet consists of **11 x-y data pairs**.
- The datasets share the same:
  - **Mean** of xxx and yyy.
  - **Variance** of xxx and yyy.
  - **Correlation coefficient** between xxx and yyy.
  - **Linear regression equation**.
- Despite identical summary statistics, scatter plots reveal distinct relationships between xxx and yyy in each dataset, showcasing unique variability patterns and outlier effects.

Purpose of Anscombe's Quartet

Anscombe's Quartet serves as a powerful illustration of:

1. **The Importance of Visualization**:
   - Data visualization helps uncover trends, outliers, and patterns that might not be evident from statistical summaries.
2. **Limitations of Summary Statistics**:
   - Relying solely on statistics such as mean, variance, or correlation can lead to incorrect conclusions about the data.
3. **Exploratory Data Analysis (EDA)**:

- Visualization is a critical step in EDA to ensure accurate interpretation and decision-making.

Anscombe's Quartet reminds analysts and data scientists to complement statistical analysis with visualization techniques for a deeper and more accurate understanding of their 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)

1. **Pearson's r** is a statistical measure that quantifies the strength and direction of the **linear relationship** between two continuous variables, ranging from −1 (perfect negative correlation) to +1 (perfect positive correlation).
2. The formula for r involves the covariance of two variables divided by the product of their standard deviations, making it unitless and comparable across datasets.
3. A value of r=0 indicates no linear relationship, while values closer to −1 or +1 represent stronger linear relationships.
4. While Pearson's r is simple and effective, it only measures linear relationships, is sensitive to outliers, and does not imply causation 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)

Scaling is a technique used to adjust the range of feature values in a dataset. From a machine learning perspective, scaling ensures that all features contribute equally to the model's performance, particularly when features have widely differing scales or units.

Scaling enhances model performance by providing an unbiased contribution of features to the model. It is also crucial for handling features with different units and ensuring faster convergence during optimization.

Key Differences Between Normalization and Standardization

1. **Range of Values**:
   - **Normalization** scales data to a fixed range, typically from **0 to 1** (or sometimes −1-1−1 to 111).
   - **Standardization**, on the other hand, does not have a fixed range. It adjusts data based on its distribution, scaling it to have a **mean of 0** and a **standard deviation of 1**.
2. **Sensitivity to Outliers**:
   - **Normalization** is sensitive to outliers, which can distort the scaled range and lead to skewed data.

- **Standardization** is less affected by outliers because it centers the data by subtracting the mean and scales it using the standard deviation, which accounts for the overall spread.

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)

- **Perfect Multicollinearity**: VIF becomes infinite when one independent variable is an exact linear combination of one or more other variables in the dataset.

- **Mathematical Explanation**: The formula for VIF is VIF=11−R2VIF = \frac{1}{1 - R^2}VIF=1−R21. If R2=1R^2 = 1R2=1 (perfect correlation), the denominator becomes zero, leading to VIF=∞VIF = \inftyVIF=∞.

- **Causes of Perfect Multicollinearity**:

  - Duplicate variables or near-identical features.
  - Strong linear relationships, such as Temperature in Celsius vs. Temperature in Fahrenheit.
  - Including all dummy variables for a categorical feature without dropping one category (dummy variable trap).

- **Effects**: Infinite VIF prevents the model from estimating coefficients properly, making it unstable and unreliable.

- **Solutions**:

  - Remove one of the redundant or highly correlated variables.
  - Drop one dummy variable to avoid the dummy variable trap.
  - Use dimensionality reduction techniques like PCA to combine correlated variables.

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)

- **Definition**:

A Q-Q plot (Quantile-Quantile plot) compares the quantiles of the observed data (e.g., residuals) to a theoretical distribution, such as a normal distribution, to assess whether the data follows the expected distribution.

- **Purpose in Linear Regression**:

It is used to check if the residuals of the regression model are normally distributed, which is a key assumption in linear regression.

- **Interpretation**:

  - If the residuals follow a normal distribution, the points on the Q-Q plot will align along a 45-degree diagonal line.
  - Deviations from the line indicate non-normality, which can affect model reliability.

- **Importance**:

  - Normality of residuals is essential for valid hypothesis testing and accurate p-values in regression.
  - It helps identify patterns like skewness (asymmetry) or heavy tails (outliers) in the residuals.

- **When to Use**:

After fitting a regression model, use a Q-Q plot to validate the normality assumption and ensure that the residuals meet the necessary conditions for valid model inference

---