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)

1. **Year (yr):**
   Bike rentals were significantly higher in 2019 compared to 2018, indicating strong business growth. Across all seasons, there was a noticeable rise in booking counts from 2018 to 2019.

2. **Season (season):**
   The Fall (Autumn) season exhibited the highest demand, with the highest median rental counts. In contrast, Spring had the lowest rental numbers. This suggests that people prefer biking during the cooler, pleasant fall weather.

3. **Month (mnth):**
   Rentals peaked between May and October, particularly in September. The number of bookings increased steadily from the start of the year, peaked mid-year, and then declined towards December, possibly due to colder weather and snowfall.

4. **Holiday (holiday):**
   Interestingly, bookings were lower on holidays, likely because people prefer to spend time at home with family. Conversely, on non-holidays, there was an increase in bookings, indicating that bikes are frequently used for commuting.

5. **Weekday (weekday):**
   Rental demand remained relatively consistent throughout the week, but Thursday, Friday, Saturday, and Sunday saw higher bookings compared to the early weekdays. This could reflect a mix of commuting and leisure activities.

6. **Working Day (workingday):**
   The analysis showed a minimal difference in bookings between working and non-working days. The median number of rentals typically ranged between 4000 and 6000, suggesting consistent usage patterns regardless of the day type.

7. **Weather Situation (weathersit):**
   Clear and partly cloudy weather conditions resulted in the highest rental counts, while bookings dropped sharply during heavy rain or snow. This highlights how favorable weather plays a significant role in influencing rental activity.

✓ **Conclusion:**

From the analysis of categorical variables, we can infer that bike rentals are highest during the summer and fall seasons, with September and October showing peak activity. Bookings were more frequent in 2019, particularly on Thursdays, Saturdays, and Sundays. Rentals also increased during non-holidays, and clear weather conditions were key drivers of demand. Overall, the data indicates a positive trend in bike-sharing usage, influenced by seasonal, temporal, and weather-related factors.

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)

1. **Prevents Multicollinearity**: It removes one dummy variable to avoid perfect multicollinearity (dummy variable trap), ensuring that no dummy variable can be predicted from the others.

2. **Simplifies Model Interpretation**: The dropped category acts as a reference point, making it easier to interpret how other categories affect the dependent variable in comparison.

3. **Improves Model Stability**: Eliminating redundant variables helps produce stable coefficient estimates and reduces inflated standard errors in regression models.

4. **Ensures Proper Model Functioning**: Regression models may fail or give incorrect results if multicollinearity exists; drop_first=True prevents this issue.

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 **'temp'** shows the strongest correlation with the target variable. Since **'temp'** and **'atemp'** are highly correlated and provide redundant information, only one of them is chosen when determining the best-fit line.

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. Normality of Error Terms (Residuals):
    - The assumption of normality for the residuals is checked using histograms and Kernel Density Estimation (KDE) plots. A histogram and KDE plot of the error terms (Y_test - Y_pred) were created, allowing us to visually inspect whether the error terms follow a normal distribution.
2. Homoscedasticity (Constant Variance of Errors):
    - The assumption of homoscedasticity (constant variance of residuals) was validated by plotting the residuals against the predicted values. If the residuals exhibit a random scatter around zero without any clear patterns, the assumption of homoscedasticity is satisfied.
3. No Multicollinearity (Correlation between Independent Variables):
    - Variance Inflation Factor (VIF) was calculated to check for multicollinearity between the features. A VIF value greater than 10 indicates high multicollinearity, but all features in the dataset had VIF values below 10, which means multicollinearity was not an issue.
4. Linearity of Relationships:
    - The relationship between independent variables and the target variable (cnt) was examined using various plots such as box plots and pair plots. The goal was to

ensure that the predictors had a linear relationship with the target variable, which is a fundamental assumption of linear regression.

5. Independence of Errors:
    - Independence of errors is a key assumption. This assumption was indirectly validated by examining the residuals. The residuals should not exhibit any patterns over time, which would indicate autocorrelation. While this is not directly tested in the code, the absence of autocorrelation was inferred from the plots.

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)

- ✓ The three most significant features are:
    1. **temp** – Temperature
    2. **yr** – Year
    3. **weathersit** – Light Snow and Rain

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)

- Linear regression is a **supervised learning algorithm** used to predict a **continuous target variable (Y)** based on one or more **independent variables (X)**. It assumes a **linear relationship** between the dependent and independent variables.
- The general equation of linear regression is:

    $$Y=\beta0+\beta1X+\epsilon$$

    Where:

    - Y = Dependent variable (what we are predicting)
    - X = Independent variable (feature/input)
    - β0 = Intercept (the value of Y when X = 0)
    - β1 = Slope of the line (how much Y changes for a unit change in X)
    - ϵ = Error term (difference between actual and predicted values)
- **Assumptions of Linear Regression**
    - **Linearity**: The relationship between X and Y is linear.
    - **Independence**: Observations are independent of each other.
    - **Homoscedasticity**: Constant variance of errors across all levels of X.
    - **Normality of Errors**: Errors (residuals) are normally distributed.
    - **No Multicollinearity** (for multiple linear regression): Independent variables are not highly correlated.
- **Formula for Coefficients:**
    - Slope (β1\beta_1):
        - β1=∑(X−X⁻)(Y−Y⁻)∑(X−X⁻)2\beta_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}

    - Intercept (β0\beta_0):
        - β0=Y¯−β1X¯\beta_0 = \bar{Y} - \beta_1\bar{X}
- **Performance Metrics**:
    - **R^2 (Coefficient of Determination)**: Measures how well the regression line fits the data.

    - **MSE (Mean Squared Error)**:
      MSE=1n∑(Y−Y^)2MSE = \frac{1}{n}\sum (Y - \hat{Y})^2

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)

- **Definition:**
    - Anscombe's Quartet is a collection of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and linear regression line) but very different distributions and appearances when graphed.
- **Purpose**:
    - Demonstrates the importance of data visualization.
    - Shows that relying solely on statistical summaries can be misleading.
- **Datasets**:
    - Each dataset has:
        - The same mean of x and y.
        - The same variance of x and y.
        - The same correlation between x and y.
        - The same linear regression line.
- **Visualization**:
    - When plotted, the datasets reveal:
        - Dataset 1: Linear relationship, fitting well with the regression line.
        - Dataset 2: Non-linear relationship, requiring a different model.
        - Dataset 3: Linear with an outlier influencing the regression.
        - Dataset 4: Vertical clustering with one influential outlier.
- **Conclusion**:
    - Anscombe's Quartet emphasizes that statistical metrics alone cannot capture the full story of data. Visualization is crucial to identify patterns, outliers, and the appropriateness of models.

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)

- **Pearson's R**, also known as the **Pearson correlation coefficient**, is a statistical measure that calculates the strength and direction of the **linear relationship** between two continuous variables. It ranges from **-1 to +1**:

- **+1** indicates a perfect positive linear correlation (as one variable increases, the other increases).
- **0** indicates no linear correlation.
- **-1** indicates a perfect negative linear correlation (as one variable increases, the other decreases).
- It is commonly used to understand how closely related two variables are in datasets.

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 the process of transforming features in a dataset so that they have a similar range or distribution. This ensures that no single feature dominates the model due to its scale, improving model performance and convergence, especially in algorithms like k-NN, SVM, and gradient descent-based models.
- **Why is scaling performed?**
  - **Improves model performance** by ensuring all features contribute equally.
  - **Speeds up convergence** in optimization algorithms.
  - **Prevents bias** toward features with larger magnitudes.
- **Difference between Normalized Scaling and Standardized Scaling:**
  - **Normalized Scaling (Min-Max Scaling):** Rescales features to a fixed range, usually **[0, 1]**. Formula:
  - Xnorm = X−Xmin / Xmax−Xmin
  - Useful when data doesn't follow a normal distribution.
- **Standardized Scaling (Z-score Standardization):**
  - Centers the data around **mean = 0** and **standard deviation = 1**.
  - **Formula**:
    - Xstd = X − μ / σ
- Useful when data follows a normal distribution or when algorithms assume Gaussian distribution.

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)

- The **Variance Inflation Factor (VIF)** measures how much the variance of a regression coefficient is inflated due to **multicollinearity** among independent variables.

- A **VIF value becomes infinite** when there is **perfect multicollinearity**, meaning one independent variable is an exact linear combination of one or more other variables. In this case, the model cannot distinguish the unique contribution of each variable, causing the denominator in the VIF formula to approach zero, leading to an infinite value.
- This typically happens when:
    - **Duplicate variables** or perfectly correlated features are included in the model.
    - There's a **linear dependency** between variables, making the matrix inversion in regression impossible.
- To resolve this, you can remove or combine highly 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)

- A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the **quantiles** of a dataset against the **quantiles** of a theoretical distribution, usually the **normal distribution**. It helps assess whether the data follows a specific distribution.
- **Use and Importance in Linear Regression:**
    - **Checking Normality of Residuals:** In linear regression, one key assumption is that the **residuals** (errors) are **normally distributed**. A Q-Q plot helps visualize this by plotting the residuals against a normal distribution. If the points fall along the 45-degree reference line, the residuals are approximately normal.
    - **Detecting Outliers:** Points that deviate significantly from the line indicate **outliers** or **non-normality**, which can affect model performance.
    - **Model Validation:** Ensuring normality of residuals validates the reliability of **hypothesis tests** (like t-tests and F-tests) used in regression, as these tests assume normally distributed errors.
- In summary, Q-Q plots are essential for diagnosing whether linear regression assumptions hold, thereby ensuring accurate and reliable model results.
