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: • In 2019, bookings were up compared to the previous year, showing growth in the business.

- Fall saw a nice jump in bookings, and across all seasons, there was a clear increase from 2018 to 2019.
- Bookings were lower on non-holidays, likely because people prefer spending time with family during holidays.
- Good weather ("Good" in the notebook) seemed to drive more bookings.
- There was a pretty even split between bookings on working and non-working days.
- Thursdays to Sundays had more bookings than the start of the week.
- Most bookings happened from May to October, with a peak in the middle of the year and a dip toward the end.

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: Dummy variables are used to turn categories into binary values (0 or 1) for analysis. If you have a variable with multiple categories, you create a dummy variable for each category, except one.
For example, if you have a "Color" variable with "Red," "Blue," and "Green," you'd create two dummy variables like "Is_Blue" and "Is_Green." If neither is 1, it means the color is "Red."

You typically create *k - 1* dummy variables to avoid issues in the model, like redundancy or confusing relationships between categories. The constant term in the model picks up the information for the category left out.

In Python's pandas, you can use `drop_first=True` to automatically drop one of the dummy variables.

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: The variable 'temp' shows the strongest correlation with the target variable, as seen in the graph below. Since 'temp' and 'atemp' are redundant, only one is chosen when determining the best fit line.
The model equation is: `cnt = 4491.30 + 998.75 × yr + 178.28 × workingday + 1174.49 × temp − 429.07 × hum − 349.15 × windspeed + 344.84 × Summer + 526.80 × Winter + 234.70 × September + 159.98 × Sunday`

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: Validating linear regression assumptions ensures model reliability. After training the model, here's the process I followed:

1. **Residual Analysis:**
   - **Process:** Check residuals (observed - predicted).
   - **Check:** Should be normally distributed with no patterns.
2. **Homoscedasticity (Constant Variance):**
   - **Process:** Plot residuals vs. predicted values.
   - **Check:** Residual spread should be constant.
3. **Linearity:**
   - **Process:** Scatterplot of observed vs. predicted values.
   - **Check:** Points should align along a diagonal line.
4. **Independence of Residuals:**
   - **Process:** Check residual autocorrelation.
   - **Check:** No pattern when plotted against time/variables.
5. **Multicollinearity:**
   - **Process:** Calculate VIF for predictors.
   - **Check:** VIF should be below 5-10.
6. **Cross-Validation:**
   - **Process:** Test model on a new set or through cross-validation.
   - **Check:** Ensure generalizability.
7. **Overfitting:**
   - **Process:** Test model performance on unseen data.
   - **Check:** Ensure good generalization to new data.

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: From the best fit line equation:

```
cnt = 4491.30 + 998.75 × yr + 178.28 × workingday +
1174.49 × temp − 429.07 × hum − 349.15 × windspeed +
344.84 × Summer + 526.80 × Winter + 234.70 × September +
159.98 × Sunday
```

The three features that significantly influence the demand for shared bikes are:

- **Temperature (temp)**
- **Winter season (winter)**
- **Calendar year (year)**

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)

Total Marks: 4 marks (Do not edit)
Answer: Linear regression models the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting line (or hyperplane for multiple variables) that minimizes the squared differences between observed and predicted values.
Steps in linear regression:

1. **Model Representation:**
   - **Simple Linear Regression:**
     $y=\beta_0+\beta_1 \cdot x+\epsilon$

   - **Multiple Linear Regression:**

     $y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + \dots$

2. **Objective Function:** Minimize the sum of squared errors (SSE) or mean squared error (MSE)

3. **Minimization:** Use techniques like gradient descent to optimize coefficients.

4. **Training the Model:** Train the model on data to adjust coefficients for better predictions.

5. **Prediction:** Use the trained model to predict outcomes for new data.

6. **Evaluation:** Assess model performance using metrics like R2 or MSE.

7. **Assumptions:** Assumes a linear relationship, normally distributed errors, constant error variance (homoscedasticity), and no perfect multicollinearity. If assumptions are violated, consider more advanced techniques.

Linear regression is powerful but requires checking assumptions for reliable results.

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 consists of four datasets with nearly identical summary statistics but distinct graphical patterns. Created by Francis Anscombe in 1973, it highlights the importance of data visualization and the limitations of relying solely on summary statistics.
Although the summary statistics are the same, the graphical differences are notable:

- The first plot shows a clear linear relationship between x and y.
- The second plot reveals a non-linear relationship, making Pearson's correlation less useful.
- The third plot is linear but skewed by an outlier, reducing the correlation.
- The fourth plot shows a high correlation due to a single high-leverage point, even though the rest of the data shows no relationship.

Anscombe's Quartet emphasizes the need to graph data before analyzing it, as visual inspection can reveal important patterns that summary statistics might miss.

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 correlation coefficient (r) measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1:

- r=1: Perfect positive correlation.
- r=−1: Perfect negative correlation.
- r=0: No linear correlation.

A positive value indicates that as one variable increases, the other does too, while a negative value means as one increases, the other decreases.

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 variable values to a similar range or distribution, making them comparable and preventing one variable from dominating others.

Advantages of Scaling:

1. **Equal Weightage**: Ensures all variables contribute equally, avoiding disproportionate influence from larger-magnitude variables.
2. **Convergence**: Machine learning algorithms (e.g., k-NN, SVM, gradient descent) perform better with similarly scaled features, speeding up optimization.
3. **Interpretability**: Improves coefficient interpretation in linear models, as they reflect changes in the dependent variable for a one-unit change in predictors.

Normalized Scaling (Min-Max Scaling) vs. Standardized Scaling (Z-score Normalization):

1. **Normalized Scaling**:
   - Scales values to a specific range (usually [0, 1]).
   - **Advantages**: Useful for unknown or non-Gaussian distributions.
   - **Disadvantages**: Sensitive to outliers.
2. **Standardized Scaling**:
   - Scales values to have a mean of 0 and a standard deviation of 1.
   - **Advantages**: Less sensitive to outliers and preserves distribution shape.
   - **Disadvantages**: Assumes 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 multicollinearity in multiple regression by quantifying how much the variance of regression coefficients is inflated due to correlated predictors.

An infinite VIF indicates perfect multicollinearity, meaning one variable is a perfect linear combination of others. This causes two issues:

1. **Redundant information**: One variable can be predicted from others.
2. **Matrix inversion problems**: The matrix used to compute VIF becomes singular (non-invertible) due to perfect correlation.

To resolve this, remove or combine highly correlated variables, or use dimensionality reduction techniques. This improves model stability and interpretability.

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 (Quantile-Quantile) plot is used to check if a dataset follows a theoretical distribution, like the normal distribution, by comparing observed data quantiles with expected quantiles. If the points form a straight line, the data is well-modelled by the chosen distribution.

Use and Importance in Linear Regression:

1. **Normality Assessment**: Q-Q plots check if residuals are normally distributed. Non-normal residuals can affect the reliability of regression inferences.
2. **Identifying Outliers**: Points deviating from the straight line can indicate outliers, which may influence model parameters.
3. **Model Fit Assessment**: The plot visually checks if the residuals conform to normality, crucial for accurate predictions.
4. **Validity of Statistical Tests**: Normal residuals are important for valid p-values and confidence intervals in hypothesis testing.

Interpretation:

- Points on a straight line suggest normality in residuals.
- Deviations indicate non-normality.

Q-Q plots are essential for diagnosing residuals, identifying outliers, and validating regression assumptions.