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)

I have analyzed the categorical columns using boxplots and bar plots. Here are some key insights from the visualizations:

- The fall season witnessed the highest number of bookings. Additionally, bookings increased significantly from 2018 to 2019 across all seasons.
- The majority of bookings occurred between May and October, with a rising trend from the beginning of the year until mid-year, followed by a decline toward the year-end.
- Clear weather conditions led to higher booking numbers, which is expected.
- More bookings were observed on Thursdays, Fridays, Saturdays, and Sundays compared to the beginning of the week.
- On non-holidays, bookings were relatively lower, which is reasonable as people tend to spend holidays at home with family.
- Booking volumes remained almost the same on both working and non-working days.
- The year 2019 saw a higher number of bookings than 2018, indicating positive business growth.

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)

Setting drop_first = True is crucial as it eliminates the extra column generated during dummy variable creation, thereby minimizing correlations among dummy variables.
Syntax:

drop_first: bool, default False

This parameter controls whether to generate k-1 dummy variables from k categorical levels by removing the first category.
For example, if a categorical column has three unique values (A, B, and C), creating dummy variables for all three is unnecessary. If a value is neither A nor B, it is implicitly C. Therefore, the third variable can be omitted without losing any information.

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' variable has the highest correlation with the 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)

I have validated the assumptions of the Linear Regression Model based on the following five criteria:

- Normality of Error Terms: The error terms should follow a normal distribution.
- Multicollinearity Check: There should be minimal or no significant multicollinearity among the independent variables.
- Linear Relationship Validation: A linear relationship should be evident between the independent and dependent variables.
- Homoscedasticity: The residual values should not exhibit any discernible pattern, indicating constant variance across all levels of the independent variable.

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 top three features that significantly contribute to explaining the demand for shared bikes are:

- Temperature (temp)
- Holiday
- Windspeed

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 supervised learning algorithm used to model the relationship between a dependent variable (YY) and one or more independent variables (XX). It aims to find the best-fitting line that minimizes prediction errors.

Key Steps:

---

1. Collect & Prepare Data – Handle missing values and preprocess features.
2. Define Hypothesis – Assume a linear relationship.
3. Estimate Parameters – Use Ordinary Least Squares (OLS) or Gradient Descent to minimize the error.
4. Evaluate Model – Use MSE (Mean Squared Error) and R^2 (R-squared) for performance assessment.

5. Make Predictions – Apply the trained model to predict new data points.

Assumptions:

- Linearity (linear relationship exists).
- No Multicollinearity (independent variables are not highly correlated).
- Normality of Error Terms (residuals are normally distributed).
- Homoscedasticity (constant variance of residuals).

Advantages:

✅ Simple, interpretable, and efficient.

✅ Works well for linear relationships.

Disadvantages:

❌ Assumes linearity, sensitive to outliers, and prone to multicollinearity.

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

Anscombe's Quartet is a set of four datasets that have nearly identical statistical properties (mean, variance, correlation, and regression line) but have very different distributions when visualized. It was created by Francis Anscombe to emphasize the importance of data visualization in statistical analysis.

Key Insights:

- All four datasets have the same mean, variance, correlation, and regression line.
- However, their scatter plots reveal very different patterns, including linear, nonlinear, and outlier-influenced distributions.
- This demonstrates that relying solely on summary statistics can be misleading.

Conclusion:

Anscombe's Quartet highlights the necessity of data visualization along with statistical measures to truly understand the underlying patterns in 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 the Pearson Correlation Coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1.

Formula:

r=∑(Xi−X¯)(Yi−Y¯)∑(Xi−X¯)2∑(Yi−Y¯)2r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i -

\bar{X})^2 \sum (Y_i - \bar{Y})^2}}r=∑(Xi−X¯)2∑(Yi−Y¯)2∑(Xi−X¯)(Yi−Y¯)

Interpretation of Values:

- r=+1r = +1r=+1 → Perfect positive correlation (as one variable increases, the other also increases).
- r=−1r = -1r=−1 → Perfect negative correlation (as one variable increases, the other decreases).
- **r=0r = 0r=0 ** → No correlation (no linear relationship between variables).

Key Points:

✅ Measures only linear relationships between variables.

✅ Does not imply causation (correlation ≠ causation).

✅ Affected by outliers, which can distort the correlation value.

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 in Machine Learning

Scaling is the process of transforming numerical features to a common range to improve model performance and convergence, especially for algorithms that rely on distance-based calculations (e.g., KNN, SVM, Gradient Descent).

Why is Scaling Performed?

- Prevents features with larger values from dominating the model.
- Improves convergence speed in optimization algorithms.
- Ensures equal weightage for all features.

  Difference Between Normalization & Standardization
- Normalization (Min-Max Scaling):
  - Formula: X′=X−Xmin⁡Xmax⁡−Xmin⁡X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}X′=Xmax−XminX−Xmin
  - Scales values between 0 and 1 (or -1 and 1).
  - Useful for non-Gaussian distributions and deep learning models.
- Standardization (Z-score Scaling):
  - Formula: X′=X−μσX' = \frac{X - \mu}{\sigma}X′=σX−μ
  - Transforms data to have mean = 0 and standard deviation = 1.
  - Best suited for normally distributed data and linear models.

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>

Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity, meaning one independent variable is a perfect linear combination of other variables. This causes the denominator in the VIF formula to be zero, leading to an infinite value.

Reasons for Infinite VIF:

- Duplicate or highly correlated features in the dataset.
- Dummy variable trap (e.g., not using drop_first=True in one-hot encoding).
- Linear dependency between variables.

Solution:

- Remove or combine highly correlated variables.
- Use Principal Component Analysis (PCA) or feature selection techniques.

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** is a graphical tool used to compare the **distribution of a dataset** with a **theoretical distribution** (usually normal). It plots the quantiles of the actual data against the quantiles of a normal distribution.

Use & Importance in Linear Regression:

- **Checks Normality of Residuals:** Helps verify if residuals follow a normal distribution (a key assumption in linear regression).
- **Identifies Deviations:** If points deviate from the diagonal line, the data may be **skewed** or **heavy-tailed**.
- **Detects Outliers:** Extreme deviations indicate potential outliers that may affect model performance.

<Your answer for Question 11 goes here>
