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)

From the analysis of the categorical variables in the dataset, we can infer that these variables can have significant effects on the dependent variable by influencing its distribution or outcomes.

Inference from categorical variables:

1. ***Season: 32% of bookings occurred in season3, followed by season2 (27%) and season4 (25%), making season a strong predictor.***
2. ***Month (mnth): Months 5-9 showed higher bookings (10% each) with medians over 4000, indicating a trend and predictive value.***
3. ***Weather (weathersit): 67% of bookings happened during 'weathersit1', followed by 30% in 'weathersit2', making it a reliable predictor.***
4. ***Holiday: 97.6% of bookings occurred on non-holidays, showing bias, making it unsuitable as a predictor.***
5. ***Weekday: Bookings ranged between 13.5%-14.8% across all weekdays with similar medians, suggesting minimal or no influence; the model can decide its relevance.***
6. ***Working Day: 69% of bookings occurred on working days, with medians close to 5000, indicating strong predictive potential.***

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 prevents the dummy variable trap by removing redundancy and multicollinearity, ensuring the model has independent 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: <Your answer for Question 3 goes below this line> (Do not edit)

The pair plot shows that temp, atemp have highest correlation with target variable cnt .

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)

After building the model, we validated the Linear Regression assumptions:

1. ***Linearity: Checked residual vs. fitted plots for no patterns.***
2. ***Normality: Verified residuals followed a normal distribution using Q-Q plots.***
3. ***Homoscedasticity: Ensured constant variance in residuals.***
4. ***Multicollinearity: Checked VIF values (should be <5).***
5. ***Error Independence: Used the Durbin-Watson statistic (value ~2 confirms independence)***

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)

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- ***Temperature (temp) .***
- ***Weather Situation 3 (weathersit_3)***
- ***Year (yr)***

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 models the relationship between a dependent variable and one or more independent variables using a linear equation:

Y=ϐ0+ϐ1X1+⋯+ϐnXn+ϵY = \beta_0 + \beta_1X_1 + \dots + \beta_nX_n + \epsilonY=ϐ0 +ϐ1X1+⋯+ϐnXn+ϵ

- ***YYY is the dependent variable, X1,X2,…X_1, X_2, \dotsX1,X2,… are the independent variables, and ϐ0,ϐ1,…\beta_0, \beta_1, \dotsϐ0,ϐ1,… are the model coefficients.***
- ***The goal is to minimize the sum of squared errors (SSE) between the observed and predicted values.***

The coefficients are typically found using Ordinary Least Squares (OLS). After training, the model can predict YYY for new inputs.

Key Assumptions:

---

1. ***Linearity: The relationship is linear.***
2. ***Independence: Residuals are independent.***
3. ***Homoscedasticity: Constant variance of residuals.***
4. ***Normality: Residuals are normally distributed.***

Linear regression can be simple (one predictor) or multiple (multiple predictors).

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 consists of four datasets with nearly identical summary statistics (mean, variance, correlation, and regression line) but vastly different distributions.

Purpose:

- ***Highlights the importance of data visualization to avoid misleading conclusions from statistical measures alone.***

Key Insights:

1. ***Datasets have the same mean and variance for xxx and yyy.***
2. ***Correlation and linear regression are identical.***
3. ***Visualizing the data reveals distinct patterns (e.g., outliers, nonlinear trends).***

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:

Pearson's R, also known as the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1:

- ***1 indicates a perfect positive correlation,***
- ***-1 indicates a perfect negative correlation,***
- ***0 indicates no linear correlation.***

It is used to understand how closely related two variables are.

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>

What is Scaling?

Scaling transforms features to a similar scale to avoid one feature dominating the model.

Why scaling is performed:

1. **Improves model performance**
2. **Faster convergence for gradient-based algorithms**
3. **Prevents bias from features with larger ranges.**

Key Differences:

Feature	**Normalization**	**Standardization**
Formula	**Min-Max scaling**	**Z-score formula**
Range	**[0, 1]**	**Mean = 0, Std = 1**

Feature	Normalization	Standardization
Outlier Sensitivity	**Sensitive to outliers**	**Less sensitive to outliers**
When to Use	**Fixed range scaling**	**Normal distribution or 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>

The Variance Inflation Factor (VIF) can be infinite if there is perfect multicollinearity between two or more independent variables. This means that one feature can be exactly predicted by a linear combination of other features in the model. When this happens:

1. ***Determinant of the correlation matrix becomes zero, leading to a division by zero in the VIF formula.***
2. ***Perfect correlation between variables makes the model unstable, as the variance of the regression coefficients increases infinitely.***

In practice, this signals that one or more predictors are redundant and should be removed from the model to avoid collinearity issues.

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 (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a specific theoretical distribution, typically the normal distribution. In the plot, the quantiles of the dataset are plotted against the quantiles of the normal distribution. If

the data follows a normal distribution, the points on the plot will lie approximately on a straight line.

Use and Importance in Linear Regression:

- ***Normality Check: In linear regression, residuals (errors) should ideally follow a normal distribution. A Q-Q plot helps assess this assumption by visually checking if the residuals align with the normal distribution.***
- ***Model Validation: If the residuals deviate significantly from the straight line, it may indicate non-normality, which could lead to invalid conclusions in regression analysis. This helps in diagnosing potential problems in the model.***