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:

- Pleasant weather in summer and fall likely contributes to higher rental counts. Spring might have more unpredictable weather, while winter could have very cold days that discourage bike rental
- The median rental count is higher on working days (1) compared to non-working days (0). This suggests that more bikes are rented during the workweek.
- The bike rental in 2019 it increased compared to 2018
- In terms of weather, clear weather conditions are most favorable for bike rentals, The box is also slightly wider, indicating greater variability in rental counts during cloudy conditions and significantly lower demand for bike rentals during light snow/rain
- The boxplot shows a clear seasonal pattern in bike rentals. The median rentals increase from winter to summer and then decrease again. August and September have the highest median rentals.
- Bike rentals appear to be consistently high throughout the workweek, with a possible slight dip on weekends.

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:

- Setting drop_first=True removes one dummy column (usually the first category alphabetically or numerically).
- By doing this, the model avoids the dummy variable trap and eliminates redundancy.
- The removed category becomes the reference category, meaning its effect is implicitly captured in the model's intercept and comparisons with other categories.

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:

- Temp and atemp is having the highest correlation with the target variable which is 0.99

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:

- First checked outliers, then checked the coorelation between the variables and removed some of the variables based on the high correlation
- Then based on dummy variables were created using pandas and drop_first = True

- Then ran the model wherein, I got the p-value and VIF, wherein the p-value threshold was kept to 0.05 and removed all VIF values which were greater than 5.
-

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:

- The top 3 features contributing were : Year , seasonality and month (especially September)

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)

- Its about finding the relation between a dependent variable (which is continuous in nature) and independent variable.
- For eg: Finding out offers and discounts effects on sales of a product. Here, the Sales will be continuous in nature (basically numerical) and offers and discounts will be independent variable.
- In Linear regression we will find how data points effect the relation and how linear they are that is which data points we can fit in a linear fashion and which are the variables that are affecting in maintaining this linearity (When X increases, Y also increases and vice versa when there is a decreasing trend)
- Now, how we can use ? We can do it by Simple linear regression which is Y = B0 + B1X1, wherein B1 is slope , X is variable, B0 is intercept and Y is output, it follows Y = mx+c theory
- If there are many variables then we use Multiple Linear Regression which is Y = B0 + B1X1 + B2X2 …… BnXn , where in B is Regression Coefficient , X is independent variable and Y is dependent Variable

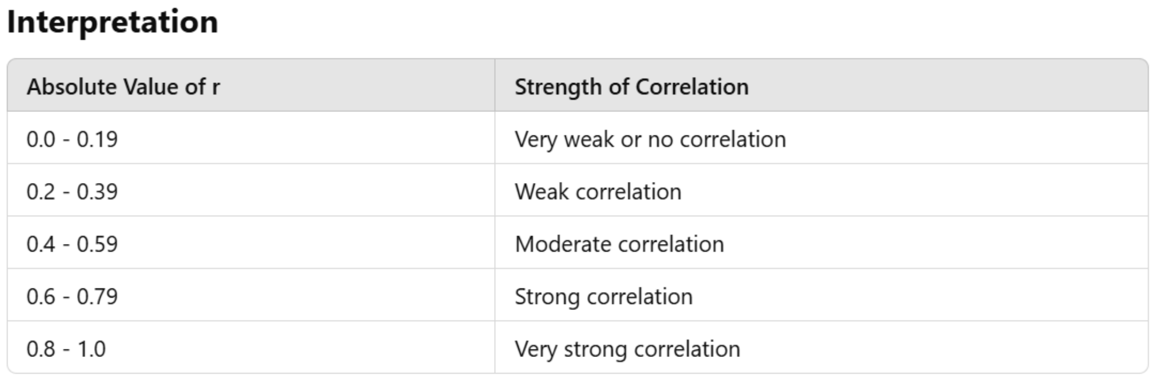
---

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 demonstrates that relying solely on statistical measures can be misleading. By emphasizing the importance of visual data analysis, it encourages analysts to:
- Investigate patterns in the data.
- Look for outliers, trends, and non-linearities.
- Ensure the statistical methods applied are appropriate for the data structure.
- This dataset remains a cornerstone example in data visualization and exploratory data analysis (EDA).

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 quantifies the strength and direction of a linear relationship between two continuous variables. It is one of the most widely used correlation metrics.



Interpretation

Absolute Value of r	Strength of Correlation
0.0 - 0.19	Very weak or no correlation
0.2 - 0.39	Weak correlation
0.4 - 0.59	Moderate correlation
0.6 - 0.79	Strong correlation
0.8 - 1.0	Very strong correlation

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 transforms numerical features to a similar range. It's like adjusting the units of measurement so all features contribute more equally to analysis.
- Features with larger values can dominate machine learning models, biasing them. Scaling prevents this.
- Scaling can help optimization algorithms converge faster
- **<u>Normalized vs. Standardized Scaling:</u>**
    - Normalized Scaling (Min-Max Scaling): Scales features to a range between 0 and 1. Useful when you know the data's exact bounds. More sensitive to outliers.
    - Standardized Scaling (Z-score Normalization): Scales features to have a mean of 0 and a standard deviation of 1. Less sensitive to outliers. Commonly used when the data distribution is approximately normal

---

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 VIF value can become infinite (or extremely large) when:
  1. Perfect Multicollinearity (R2=1R^2 = 1R2=1):
    - If an independent variable is a perfect linear combination of other independent

variables, then R2 = 1

2. Exact Duplicate Columns in Data:
    - If the dataset has identical or highly similar columns, their correlation becomes 1, leading to infinite VIF.
3. High Dependence Between Features:
    - Even if exact collinearity doesn't exist, very high correlation (R2≈1) leads to very high VIF values (e.g., VIF>10).

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 visually compares data quantiles to a theoretical distribution (like normal).
- In linear regression, it checks if residual errors are normally distributed—a key assumption. Deviations from a straight line on the Q-Q plot indicate non-normality, potentially invalidating the regression model's results.
-
