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: Season, weather and working days have positive impact on the bike rental count. The usage of bike are high during summer and winter, where as spring showed low usage. Usage of rental bikes are low during light precipitation and no usage at all during heavy precipitation. Though Holidays showed more spread of bike renters but the mean bike renters are slightly higher during working days. (Do not edit)

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: Importance of using drop_first = True:

1. Avoid Multicollinearity: In dummy encoding, if all categories are included as dummy variables, their values add up to 1, making one category perfectly predictable from the others. By dropping one category, the remaining categories are independent, preventing multicollinearity.
2. Simplified Interpretation of Model Coefficient : the dropped category becomes the baseline or reference category, and the coefficients of the remaining categories represent their impact relative to this baseline
3. Dropping one dummy variable reduces feature count and helps the model generalize better and avoid over fitting (Do not edit)

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: temperature and feel like temperate have positive correlation and windspeed has negative correlation (Do not edit)

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: 1. Through VIF where all the feature should be well below 5 and
2. through Residual Analysis by plotting a histogram to check if we are getting a normal distribution curve (Do not edit)

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: weather (precipitation) [+ve coefficient], working day[+ve coefficient] and windspeed[-ve coefficient] (Do not edit)

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 model the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship, represented as: y= Mo +M1x1+M2x2+….+Mnxn + e. Here, Mo is the intercept, M1, M2… are the coefficients and e being the error term. The algorithm minimizes the sum of squared errors between predicted and actual values. It fits the best fit line by finding the optimal coefficients.

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 consists of four data sets with nearly identical statistical properties but vastly different distributions and appearances when visualized. It demonstrates the importance of visualizing the data in addition to statistical summaries as it can be misleading at times.

Outliers can distort statistical measures like mean, variance, correlation, and regression lines. A good fit requires understanding the data's distribution. Not all data can be described by a linear model. Relying solely on statistical metrics like R square without visual inspection can lead to incorrect conclusions. Hence, visualization before statistical interpretation is very important.

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, is also known as Pearson's correlation coefficient. It is a statistical measure used to quantify the linear relationship between two continuous variables. It provides both the strength and direction of the relationship.

1.Range :

The value of R lies between -1 and 1.

- R=1: Perfect positive linear relationship.
- R=−1: Perfect negative linear relationship.
- R=0: No linear relationship.

2.Direction:

- Positive R: As one variable increases, the other tends to increase.

- Negative R: As one variable increases, the other tends to decrease

3.Strength: The closer R is to ±1, the stronger the linear relationship

Limitations:

- Pearson's R measures only linear relationships and cannot capture nonlinear dependencies.
- Sensitive to outliers, which can distort the value.
- Does not imply causation; a high R indicates correlation, not causality

---

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: Scaling is a preprocessing technique used in data analysis and machine learning to adjust the range of features so that they are comparable and lie within a specific range. This ensures that no single feature dominates the model due to its scale

Scaling is important because of below points:

- Many machine learning algorithms are sensitive to the scale of input features. Large values may dominate smaller ones, biasing the model.
- Algorithms like gradient descent converge faster when features are scaled to a similar range.
- In algorithms like k-means clustering or k-nearest neighbours, scaling ensures all features contribute equally to distance computations.
- Scaling avoids numerical instability in models, especially when using matrix operations.

Difference Between Normalized scaling and Standardized scaling:

Normalised Scaling	Standardised scaling
Rescales within [0,1]	Rescales to a mean of 0 and std. dev. of 1
Formula : (X-Xmin)/(Xmax -Xmin)	Formula : (X-mean)/std. dev.
Sensitive to outliers	Less sensitive to outliers but impacted
Easier to interpret	Centered around zero so useful in statistical analysis

---

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 with other predictors in the model.

A VIF of infinity occurs when the Ri square value for a feature is 1. This means that the feature Xi is perfectly linearly correlated with one or more other features in the model. In other words, there is perfect multicollinearity between the predictors.

Possible cases of Perfect multicollinearity:

- It could be redundant feature
- Perfect multicollinearity can also occur if a feature is a direct mathematical function of another feature
- Sometime there can be perfect multicollinearity while including dummy variables as well where one dummy variable can be perfectly predicted by other dummy 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 Quantile-Quantile (Q-Q) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, often the normal distribution. It plots the quantiles of the observed data against the quantiles of a specified reference distribution. If the data points on the plot fall approximately along a straight line, it suggests that the data follows the chosen distribution.

Use and Importance of Q-Q plot:

- Normality of residuals is a key assumption in linear regression. A Q-Q plot is commonly used to visually inspect whether the residuals of the model follow a normal distribution.
- If the residuals are not normally distributed, it can indicate that the linear regression model might not be the best fit, potentially affecting hypothesis tests and confidence intervals.
- A non-linear pattern in the Q-Q plot may suggest that the residuals are not normally distributed.
- Outliers will appear as points that are far from the line, which might indicate influential data points that could affect the regression model's performance.
- A Q-Q plot helps in model diagnostics by providing insights into how well the model assumptions hold. If the residuals are non-normal, it may signal the need for transformations or different modeling approaches
- For improving Model accuracy: A Q-Q plot can help in diagnosing and improving the model by suggesting potential adjustments or transformations to improve the accuracy and reliability of predictions.
