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)

1. Fall season had a greater number of counts of bike rentals.
2. 2019 had more bike rentals.
3. Both working day and weekend/holiday has the same amount of bike rentals.
4. September has the highest bike rentals and almost the same on each day of the week.
5. Most number of bike rentals/sharing happens when the weather is clear or partly cloudy.

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)

1. Avoids Multicollinearity: By dropping one dummy variable column, the redundancy is eliminated. The remaining columns are sufficient to represent all categories uniquely.
2. Ensures Model Stability: Multicollinearity can cause unstable regression coefficients and inflate the variance, making the model harder to interpret.
3. Does not Lose Information: Dropping one column does not lose information since the dropped category becomes the reference category and its effect is captured in the intercept term of the regression model.

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 is highly correlated to the target variable count.

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)

1. I made the residual analysis which gave the graph which was like normal distribution.
2. Checked the VIF values which were less than 5 as specified.
3. R-Squared and Adjusted R-Squared values were check which measures the proportion of variance explained by the model.
4. The P-value was checked to be less than 0.05.

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)

For my model the top 3 features are windspeed, summer and year.

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 one of the simplest and widely used techniques in data analysis. It creates the relationship model between the target variable and one or more independent variables.
Type of Linear Regression:
1. Simple Linear Regression: It has only one independent variable.
   Equation: y = B0 + B1x
2. Multiple Linear Regression: It has multiple independent variables.
   Equation: y = B0 + B1x1 + B2x2 + …. + Bpxp

Certain Assumptions made while using the Linear Regression are:
1. Linear Relationship between variables.
2.Constant variance of errors (homoscedasticity).
3. Residuals are normally distributed.
4. No 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)

It is a set of four datasets devised by Francis Anscombe in 1973. It helps in visualizing the data and not depending solely on the summary of the statistics provided by the model. It is widely used during the exploratory Data analysis and the need for geographical representations like scatter plots before making decisions based on statistical measures.

Key Features:

1. Identical Statistical summaries
   a. Mean of x = 9.0
   b. Mean of y = 7.5
   c. Variance of x = 11.0
   d. Variance of y = 4.12
   e. Correlation between x and y = 0.816
   f. Regression Equation y = 3 + 0.5x
2. Different Distributions: The datasets differ significantly when plotted, revealing distinct relationships between x and y.

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)

Also known as the Pearsons correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. Calculated using the formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

- x_i and y_i: Individual data points for the two variables x and y
- $\bar{x}$ and $\bar{y}$: Means of x and y
- r: Pearson correlation coefficient

Assumptions of Pearson's:
1. Both variables are continuous and normally distributed.
2. The relationship between the variables in linear.
3. The variance of variable is constant across the values of the other.
4. Absence of significant outliers that could distort the 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 is the data preprocessing technique used in the ML and ML Modelling to adjust the range of features or variables. It ensures that the numerical features contribute equally to the model and prevents one feature from dominating others due to its larger range of values.

Why Scaling:

1. Improves the Model Performance: Many ML Algorithms are sensitive to the range of feature values.
2. Equal Contribution of features: Ensures that all features are treated equally, especially in distance-based models like KNN and clustering.
3. Stability: Improves the stability during calculations by avoiding large numerical discrepancies.
4. Compatibility of Algorithms

Difference between normalized and standardized scaling:

Normalized Scaling	**Standardized Scaling**
Rescales values to specific range (0,1)(-1,1)	Centers data around mean and scales by SD.
Changes both scale and range of data	Centers data at 0 with SD as 1.
Used case is when data has a bounded range or requires proportion scaling.	Used case is when data follows a Gaussian Distribution or requires zero-centered features.
Application: Min Max Scaling for image data or feature scaling for neural networks.	Standardization for SVM, logistic regression or PCA.

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 quantifies the multicollinearity in the dataset by measuring how much the variance of a regression coefficient is inflated due to linear dependence among predictors.
The value becomes infinite when perfect multicollinearity exists.

Reasons:

1. Perfect Multicollinearity: If one predictor is perfectly correlated with another, its VIF will be infinite.
2. Duplicated Predictors: Including identical or highly similar predictors.
3. Redundant Categorical Variables: When using one hot encoding for categorical variables, not dropping one dummy variable can lead to perfect multicollinearity.
4. Improper Model Design: Using derived variables that are functions of existing 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)

The Quantile – Quantile plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution.

Importance of Q-Q plots in linear regression:

1. Validating Assumptions: Linear regression assumes residuals are normally distributed for accurate hypothesis testing and confidence levels. The plot helps validate this assumption.
2. Identifying Deviations: Detects heavy tails, skewness or other departures from normality that might affect model performance.
3. Model Diagnostics: Helps identify whether transformations or robust statistical methods are needed to handle non-normal residuals.
4. Understanding Outliers: Points far from the 45-Degree line indicate potential outliers or influential data points.
