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)

Categorical variable such as season, weathersit, holiday, weekday and working day has the impact on the cnt variable. Weathersit_cloudy and raniy negatively affects cnt.

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)
Drop_first=True is used to avoid the multicollinearity and provides clear interpretation of coefficients. Ex: one dummy variable is dropped in season.

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)

Atemp and temp has the highest correlation with 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)

Normal distribution of error terms: Plotting the error terms shows the normal distribution
Linear Relationship
Homoscedasticity
Absence of Multicollinearity
Independence of residuals (absence of auto-correlation)
Residuals are normally distributed

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)

yr (year): Indicates a strong positive trend over time.
temp: Higher temperatures are associated with increased counts.
season_summer and season_winter: Both seasons positively affect the dependent variable compared to the baseline season.

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 is machine learning model used to build the relationship between a dependent variable and one or more independent variables. It establishes the linear relationship between the variables.

Equation of line: $Y = B_0 + B_1X_1 + B_2X_2 + ... + B_nX_n + E$

Types of Linear regression

1. Simple linear regression:

Model with one independent variable X. $Y = B_0 + B_1X + E$

2.Multiple Linear Regression

Model with multiple independent variable (X1,X2,....Xn). $Y = B_0 + B_1X_1 + B_2X_2 + ... + B_nX_n + E$

Assumptions:

1.Linear relationship between X and Y
2.Errors terms are normally distributed
3.Error terms are independent of each other
4.Homoscedasticity: Error terms have constant variance.

Evaluate the model:

R-squared: Propotion of variance in Y explained by independent varibales
Adjusted R2 : Accounts the numner of predictors in the model.
Mean squared error: Average of the squared errors.

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 is a set of 4 dataset has nearly identical statistical properties such as mean, variance, correlation coefficient and linear regression equation, but they are vastly different when represented graphically.

The 4 datasets:

Dataset1: shows the Linear relationship.
Dataset2: shows the non-linear relationship.
Dataset3: Has the clear outlier affected the regression line.
Dataset4: shows a vertical line, where most data points have the same x value.

Anscombe's quartet demonstrates that solely relying on the numerical summaries can be

misleading and emphasizes the importance of data visualization for a complete understanding of 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 also known as the pearson correlation coefficient or simply the correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two variables.

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Cov(X,Y): Covariance of the two variables XXX and YYY.
σX sigma_XσX: Standard deviation of XXX.
σY sigma_YσY: Standard deviation of YYY.

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 is the process of transforming the features of a dataset so that they fall within a specific range.
Scaling is performed to improve the model performance, accelerate convergence, prevents feature dominance and ensures interpretability.

Type of scaling:

1. Min Max Scaling:
   Transforms data to fit within a specific range ,usually (0,1)

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2. Standardized Scaling:
   Centers the data around the mean of 0 with standard deviation of 1.

$$x' = \frac{x - \mu}{\sigma}$$

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) quantifies how much the variance of a regression coefficient is inflated due to multicollinearity.

A VIF can become infinite in cases where there is perfect multicollinearity.

1.Perfect Multicollinearity: One predictor variable is a perfect linear combination of the other predictor variables.

2.Deterministics Relationships: If one predictor is entirely determined by a linear combination of the other predictors.

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 plot is a graphical tool used to compare the distribution of a dataset with normal distribution. It helps to assess whether the data follows a specific distribution by plotting the quantiles of the observed data against the quantiles of the reference distribution.

Use of Q-Q plot in linear regression.

1. Residual analysis: After fitting a lr model, compute the residuals. To evaluate whether these residuals are approximately normally distributed.

2.Identifying deviations from normality:

- Linear alignment
- Heavy tails
- Skewness

3.Model Diagnostica: If the residuals are not normally distributed, it may indicate model misspecification, outliers.
