Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- Weathersit and season plays a very vital role as a categorical variable -

The weather is clear most of the seasons except winter where there is more Rain. Misty weather is almost the same across seasons. High chances are less bike riders during winter season. Chances of higher number of bike rides must be for a spring stroll as well.

1.There is a dip in the bike riders during the winter months starting from Sept to Jan.
2. The number of bikers increase from the spring season and a very less dip during Fall

2. Why is it important to use drop_first=True during dummy variable creation?

The main use of dummy variables is to convert categorical variables into 0 and 1s. Which means if there is a categorical variable Gender - Male, Female, Other
We first split Gender column into three columns and then use one -hot encoding meaning,

1. If gender is Male - value will be 1 0 0
2. If gender is Female - value will be 0 1 0
3. If gender is Other - value will be 0 0 1

Now we drop first, because if male is 1 it is a male, if female is one it 1, if both are a zero we can conclude the gender is other and hence drop the extra column.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Temperature had the highest correlation with cnt. It almost acted like the target variable, hence had to be dropped from the model.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

1. atemp indicates that comfort in perceived temperature is a primary factor driving bike rentals.
2. yr shows that the bike-sharing program is seeing significant growth year over year.
3. weathersit_Light Rain highlights the negative effect of bad weather on bike rentals, making it a key factor in predicting low demand days.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features). The aim is to fit a linear equation that best predicts the target variable, meaning a line that will explain all the features in the best possible way.
- **Equation of a Simple Linear Regression:**

  $y=\beta0+\beta1x+\epsilon y$

  y: Dependent variable (target)

  β0: Intercept (value of y when x=0x = 0x=0)

  β1: Coefficient of the independent variable (slope)

  x: Independent variable (feature)

  ϵ: Error term (residual)

- The goal is to estimate the values of β0\beta_0β0 and β1\beta_1β1 that minimize the difference between the actual values yyy and the predicted values y^\hat{y}y^ using the least squares method, which minimizes the sum of squared residuals

2. Explain the Anscombe's quartet in detail.

- Anscombe's Quartet is a group of four datasets that have nearly identical statistical properties (mean, variance, correlation, etc.) but exhibit significantly different patterns when plotted.

Importance of Anscombe's Quartet:

- **Identical Statistics**: All four datasets have the same mean, variance, correlation coefficient, and regression line.
- **Different Graphs**: Despite similar statistics, each dataset behaves differently when plotted, illustrating different relationships (linear, non-linear, outliers, etc.).
- Anscombe's Quartet emphasizes that relying solely on summary statistics (like mean, correlation, or variance) can be misleading. Visual inspection of data is crucial to understanding the underlying patterns and relationships.

3. What is Pearson's R?

- Pearson's R is a measure of the linear correlation between two variables. It quantifies the strength and direction of a linear relationship.
- It ranges from -1 to 1 and 0 meaning no linear relationship
- Pearson's R only measures linear relationships, so it may not capture nonlinear associations.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- Scaling refers to transforming data so that it falls within a specified range or follows a particular distribution.
- Algorithms that compute distances between points (e.g., KNN, SVM) or rely on gradients (e.g., neural networks) are sensitive to the scale of features. Features with larger ranges can dominate others during model training.
- Normalization (Min-Max Scaling): Scales features to a range between 0 and 1 (or other specified range). It helps normalize outliers as well, hence preferred more than the standardized scaling.
- Standardization (Z-score Scaling): Transforms features to have a mean of 0 and a standard deviation of 1

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- Variance Inflation Factor (VIF) measures multicollinearity in a regression model. It quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other variables.
- Infinite VIF occurs when there is perfect multicollinearity, meaning one variable is a perfect linear combination of one or more other variables. This happens when two or more features are highly correlated, causing the matrix inversion required in regression to fail.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot is a graphical tool to assess if a dataset follows a specified distribution, typically a normal distribution.

Use and Importance in Linear Regression:

1. Normality of Residuals: In linear regression, one assumption is that residuals (the differences between observed and predicted values) should be normally distributed. A

Q-Q plot compares the quantiles of the residuals against the quantiles of a normal distribution.

2. How to Interpret:
    - If residuals follow a normal distribution, the points on the Q-Q plot should fall roughly along the 45-degree line.
    - Deviations from the line suggest departures from normality, which could indicate model misspecification or the presence of outliers.
3. Importance: Normality of residuals is crucial for accurate hypothesis testing and confidence intervals in linear regression. If residuals deviate from normality, transformations or robust methods may be necessary.