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?

Ans:

Some of the categorical variables have influence on the dependent variable if not at par with the numerical variable but can't be ruled out that they are insignificant. The Year and Weather Situation in this case along with specific seasons. And looking at this dataset from the intuition side. Logically if someone has to ride a bike would be for commute or fun ride situations. And where weather , holiday, workingday, Seasonality would affect the decision to ride a bike or use other means of commute.

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

Ans:

*A: Prevents Multicollinearity**:** By removing one column, you eliminate the perfect correlation between the dummy variables.*

B: Improves Model Interpretability: It becomes easier to interpret the coefficients as they directly compare other categories to the baseline.

C: Reduces Computational Cost: Fewer features can lead to faster model training and evaluation.

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

***Ans:** Temp or atemp(Temperature felt)(It's a variable that have TEMP and Humidity in it)*

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

The major contributors have a high linear relationship with the target variable.

Residuals in the train and test dataset were normally distributed.

Multicollinearity was under threshold in the final model by using VIF.

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

***Ans:** 1: Temperature, 2: Year, 3: Light Snow Weather(negatively)*

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is a statistical method used to model the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). The goal is to find the best-fitting line (or hyperplane in case of multiple variables) that represents the relationship between the variables.

Types of Linear Regression

- *Simple Linear Regression: Involves one independent variable and one dependent variable.*
- *Multiple Linear Regression: Involves multiple independent variables and one dependent variable.*

Assumptions of Linear Regression

- *Linear relationship between the variables.*
- *Independence of errors.*
- *Homoscedasticity (constant variance of errors).*
- *Normality of residuals.*
- *No multicollinearity (high correlation between independent variables).*

Limitations

- *Linear regression assumes a linear relationship, which might not always be the case in real-world data.*
- *Sensitive to outliers.*
- *Cannot handle non-numeric data directly.*

2. Explain the Anscombe’s quartet in detail.

Ans:

Anscombe's quartet is a set of four small datasets that have nearly identical statistical properties but look very different when plotted.

It was created by statistician Francis Anscombe in 1973 to illustrate the importance of graphical data analysis before performing statistical analysis.

The Four Datasets

1. *Dataset I: This dataset exhibits a clear linear relationship between x and y with some scatter around the line. It's a typical example of a dataset suitable for linear regression.*

2. *Dataset II: This dataset shows a clear non-linear relationship between x and y. A linear regression model would be a poor fit for this data.*
3. *Dataset III: This dataset appears linear, but there is one outlier point that significantly influences the regression line.*
4. *Dataset IV: This dataset has constant x values for all but one point, resulting in a horizontal line when plotted.*

3. What is Pearson's R?

Ans: *Pearson's r is a statistical measure that quantifies the linear relationship between two continuous variables.*

It provides information about both the strength and direction of the relationship.

Key Characteristics:

- *Range: Values range from -1 to 1.*
- *Interpretation:*
  - *-1: Perfect negative correlation (as one variable increases, the other decreases)*
  - *0: No correlation between the variables*
  - *1: Perfect positive correlation (as one variable increases, the other increases)*
- *Strength of Relationship:*
  - *Closer to -1 or 1 indicates a stronger linear relationship.*
  - *Values closer to 0 indicate a weaker linear relationship*

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: *Scaling is a data preprocessing technique applied to independent variables to normalize their values within a specific range. This process is crucial for many machine learning algorithms to function effectively.*

Scaling is performed to improve algorithm performance, it helps in convergence when features are on a similar scale. It also prevents dominance of features. Features with larger values can dominate the learning process, leading to biased models. Scaling helps to balance the influence of different features. In models like linear regression, scaled features make it easier to compare the magnitude of coefficients.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: *Have not observed it yet. ;) However, since it identifies multicollinearity in a feature variable to an extreme extent that the entire data points of a feature can be explained by one or more of other independent variables in a dataset. Idea is to remove variables that have high VIF with infinity that need to be removed at the very first iteration.*

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression ?
Ans: *Q-Q plot is a graph which compares a normal distribution percentiles with residuals quartiles on a 2 D graph. It has a reference line at 45 degrees which depicts a perfect normal distribution.*
Q-Q plot helps understand residuals are normally distributed. Which is one of the main assumptions of linear regression. This is a graph which is plotted for the residuals from training or test data or both. If the residuals are not normally distributed, it can affect the validity of hypothesis tests and confidence intervals. A Q-Q plot helps visualize whether this assumption is met or not.