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?

- A Dependent variable is what happens as a result of the independent variable. For example, if we want to explore whether high concentrations of vehicle exhaust impact incidence of asthma in children, vehicle exhaust is the independent variable while asthma is the dependent variable.

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

The **drop_first=True** helps in reduction i=of extra column and thus reduces the correlations created among dummy variables

Thus if we n levels the we need only n-1 columns to create dummy variables

Example: We know we have 3 levels in furnishing status it enough to know about **furnished** and **semi_furnished** as 3rd one (not-furnished) will be inferred from this first 2

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

- Both temperature and a-temperature columns has the highest correlation with the target variable
- Thus its highly correlated to temperature

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

- Validation wa done using R-squared and Adjusted R-Squared and it was done to both training and test data
- Regression Plotting helped in finding the actual line for the plot

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

- Temperature
- Seasons (Summer, Winter)
- Month (August, Sepetember)
- Days like Weekends

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear Regression Algorithm is a machine learning algorithm based on supervised learning
- Linear regression is a part of regression analysis
- THIS HELPS IN PREDICTIVE MODELLING OF THE DATA
- When the data is fed to model it helps in understanding and predicts the possibility of the future based on previous trend
- It shows the linear relationship between a dependent and one or more independent variables
- It supports both numerical and categorical values

2. Explain the Anscombe's quartet in detail.

As the name describes `quartet` we have 4 datasets that have symmetrical simple statistical properties but they will vary when it is visualized/graphed.

Each of the four data set has 11 (x,y) points

3. What is Pearson's R?

Pearson's r is the numerical summary of the strength in the linear association between the multiple variables. If the variables tend to go up and down together, the correlation coefficient will be surely of positive value.

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

- To improve the performance of the machine learning algorithms and also changes the ranges in the data sets we use scaling and there are two commonly used scaling techniques

- Normalization helps to transform features to be on a similar scale and its affected by outliers and used when features of different scales

- Standardized scaling is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score

- The standardized scaling is less affected by outliers and not bounded to a range

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

When VIF is infinite there is perfect correlation between two independent variables and we need to drop the one of the column

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 the scatterplot created by plotting any two sets of columns against each and If both sets of columns form a same distribution, we should see the points forming a straight line

Q-Q plot is the graphical plot that helps in visualizing in linear regression

This helps when the training and test data is received separately and this ensures the data forms a same distribution