ANSWERS for the 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:

Three of the categorical values have a huge effect on the dependent variable which is Year, Weathersit and season

- Year - according to the data we could clearly see that in 2019 more number of bikes were get rental than 2018
- Weathersit - Similar to the year more number of bikes get rental when there is a clear sky and no rain
- Season - Comparitively low number of bikes been rented in the season of spring

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

ANS:

- Using drop_first= true helps to get rid of the redundancy and avoids the multicollinearity between the features.

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

ANS:

- **Between the temp and atemp has the highest correlation**

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

ANS:

To validate the assumptions of the Linear regression

- Need to plot the Residual plot, need to confirm there is no pattern and it should ideally band around zero
- Check with No Multicollinearity between the feature 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:

- **weathersit**
- **yr**
- **holiday**

1. Explain the linear regression algorithm in detail.

ANS:

There is two type of regression

1. Simple linear regression
2. Multiple Linear Regression

- Simple Linear regression is use single feature to predict the target variable and in multiple linear regression we will use more than one feature to predict the target variable.
- Linear Regression model is the fundamental algorithm used to predicting a continuous target variable based on the independent variable. The main goal of this model is to find the best fit line to get a coefficients for the features that predict the dependent variable.
- Then we will evaluate the model which just trained using R^2 , Adj R^2 method, MSE (Mean squared error), Root mean squared error (RMSE)

2. Explain the Anscombe's quartet in detail.

ANS:

- Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. These data sets have very different distributions so they look completely different from one another when once we visualize the data on scatter plots.

- So we need to visualize the dataset before applying any ML algorithm to the dataset from this we can find the outliers, linear separability of data and some other anomalies.

3. What is Pearson's R?

ANS:

Pearson's correlation coefficient is a measure of the linear relationship between two feature of continuous variables and the value scales between -1 to +1.

- Positive values were represent how one value positively having effects, if A increases B also increases .
- Negative values were represent how one value negatively impact , if A increases B decreases.

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

ANS:

- In Machine Learning scaling is the technique that we have use to change different range values between all the feature comes under similar range.
- We are scaling values due to various reasons those are
    - Algorithm will converge (like gradient descent) faster when numerical features are in a similar scale.
    - And this will avoid the Bias because it will ensure that larger value range doesn't affect the model
    - It will make to understand the relative importance of different feature.
- Normalization (Min-Max Scaling) rescales the features values to fit within a specific range [0,1] and Standardization rescale the feature values so that they have mean of 0 and a standard deviation of 1 b

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

ANS:

- VIF (Variance Inflation Factor) is used to detect the multicollinearity between the features and if this value is high it represent that correlation between the predictor columns are high. However, VIF value could be infinite in case of perfect collinearity between two or more columns and it denotes that is exact linear combinations of other features in the dataset and this will make coefficient unstable and unreliable.

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

ANS:

- Q-Q plots plays the vital role in graphically analysing and comparing two probability distribution by plotting their quantiles against each other. If the two distribution that we are comparing are exactly equal, then the points on the Q-Q plot will perfectly lie on a straight line.
- In linear regression we use this plot to check visually that the Residuals are normally distributed (We assume that LR residuals were normally distributed) the points on the Q-Q plot should roughly follow a straight line.