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?

Observations for the categorical variables,

- **'Mnth'**: Majority of the booking occurred in the month 7,8,9.
- **'season'**: Season 3 has the highest percentage of bookings.
- **'Weatehrsit':** weathersit 1 has the maximum number of bookings.
- **'workingday':** workingday is good indicator which affects the number of bookings.

Therefore, the above predictor variables can be considered for our Lr predictive analysis.

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

drop_first= True in dummy variable creation reduces the extra column created during the dummy variable creation and encoding.

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

The 'temp' and 'atemp' variables have the correlation with the target variable.

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

Validated the linear regression assumption, error terms should be normally distributed with mean =0, through a distribution plot of the train error.

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

The top 3 features contribution significantly towards explaining the demand of the shared bikes,

1. Temp
2. yr
3. weathersit_3

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features.

The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables.

The equation provides a straight line that represents the relationship between the dependent and independent variables.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four data set having identical descriptive statistics properties like mean, variance, correlations but have different representations when we plot on a scatter plot.

It is used to illustrate the importance of EDA and drawbacks of only depending on the summary statistics.

3.What is Pearson's R?

Pearson's R correlation coefficient is most common way of measuring a linear relation between two variables. It ranges from -1 to 1.

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

- Scaling is a data prep-processing step which is applied to independent variables to normalize a data to a particular range.
- If the collected data set contains features highly varying in magnitudes, units and range. Scaling is performed.
- Normalized scaling/minmax scaling bring all the data in the range of 0 to 1.
- Standardized scaling brings the data to standard normal distribution with mean 0 and standard deviation 1.

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

If there is a perfect correlation between the independent varialbles , the value of VIF is returned as inf.

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

The quantile-quantile plot is a graphical method for determining whether two samples of data came from population with a common distribution like normal or exponential.

Q-Q plots can be used for residual analysis in linear regression to validate the residuals follow a normal distribution which is an important assumption in linear regression.