**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: Business should take a call to expand business during Summer and the Fall season.

Bike ride is likely when it's cloudy or clear weather

Bike renting is likely in a moderate temperature.

1. Why is it important to use **drop\_first=True** during dummy variable creation?

A: It drops the extra column.

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

A: temp

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

A: 1. Linear relationship between target (cnt) and predictors (9 variables as in the last model )

2. Error terms are normally distributed. With mean cantered around mean.

3. Error terms doesn’t have any pattern and are independent.

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

A: season, weather, temperature

**General Subjective Questions**

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning algorithm . It performs a regression by creating models at target prediction variables based on independent variables. It is mostly used for finding out the relationship between variables and prediction.

1. Explain the Anscombe’s quartet in detail.

Anscombe's Quartet is a group of four data sets which are nearly identical in simple descriptive statistics, but there are some qualities in the dataset that creates an impression that the regression model very different distributions and appear differently when plotted on scatter plots.

1. What is Pearson’s R?

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

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

Scaling: It is a step of data pre-processing technique applied to independent variables to normalize the data within a particular range. It also optimizes the calculations in an algorithm.

Why scaling:

Often data set variables have comparative high magnitude. So to make them uniform scaling is done.

Normalised vs Standardized:

Normalized scaling is done by MinMax which limit the data set within range of 0 to 1. Thus it also addresses outliers.

Standard scaling renders the data set to their normal distribution I.e. mu=0 and SD=1.

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

If there is perfect correlation, then VIF = infinity.

This shows a perfect correlation between two independent variables.

Reason: when R2 =1,

=> 1/(1-R2) =infinity.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.

For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The aim of Q Q plots is to ascertain whether two sets of data belong to the same distribution.