# 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?

In the Bike sharing data set it has one categotical variable is there and we removed that because there is no use on this and we are using the month and year.

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

Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi\_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

**Registered** highly correlated with cnt(0.95)

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

>Check The Error should follow normal distribution

>no Multicolinearity

> **Homoscedasticity**

>Linearity

>Error terms are independent of each other.

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

'weekday','weathersit','casual','registered'

# General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc

1. Explain the Anscombe’s quartet in detail.

**Anscombe’s quartet** comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations

1. What is Pearson’s R?

The **Pearson correlation coefficient (r)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

>Normalization will help to reduce the value min max scalling

>standardization will help to reduce the values b/t mean 0 and sd 1

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

Infinite means which are highly correlated with dependent variable

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

A Q-Q plot, short for quantile-quantile plot, is a type of plot that we can use to determine whether or not the residuals of a model follow a normal distribution. If the points on the plot roughly form a straight diagonal line, then the normality assumption is met.