## 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? (3 marks)

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

This is because for any feature with multiple levels lets say n levels, all levels can be explained by defining n-1 variables so it is better to drop first dummy variable as it can be easily inferred by seeing values of other n-1 variables.

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

   registered has highest correlation with cnt indicating more number of registered users mean more bike demand.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- error terms should be normally distributed with mean equal to 0.
- error terms should be independent of each other. Again for this, I plot the error terms, this time -with either of X or y to check for any patterns.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   temp
   hum
   windspeed

## General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

# Simple Linear Regression Process

## Step 1: Reading and Understanding the Data

Importing data using the pandas library and running stats summary to see for any nulls and outliers in the data.

## Step 2: Visualizing the Data

Let's now visualise our data using seaborn. We'll first make a pairplot of all the variables present to visualise which variables are most correlated to `cnt`.

## Steps 3 -Steps for Building model:

- Create X and y  (y is kept small case and X is kept largecase)

- Create train and test sets (Logic to choose split is to test sufficients number of records not too less nor too much)

- Train your model on the Training set (i.e learn the coefficients)

- Evaluate the model - both on training set and test set to refine it further

- Use techniques to remove redundant features and optimize model

## Steps 4 -Residual Analysis on Train data and test data:

- Calculate RMSE (mean squared root error) on test data

- Calculate R-Squared on test data and ensure it is in 5% range of Training set data to deduce model has learnt in training and applying similar algo in test data

## Steps 5 -Visualizing the fit on test data

2. Explain the Anscombe’s quartet in detail. (3 marks)

3. What is Pearson’s R? (3 marks)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is done to bring all coefficients under same scale so that they can be inferred easily and their impact on dependent variable. Normalized scaling is plotting

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

(3 marks)

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

(3 marks)