Data Manipulation Assignment

### **Problem Statement:**

Sam is done with exploring the data and has understood the structure of the data properly. Now, he has to go ahead and manipulate the data to extract useful information from the data.

Sam will start off by using the select() function from the dplyr package to extract some specific columns from the dataset

## Questions on select() function:

1. Extract these individual columns:
   1. Extract the 5th column & store it in ‘customer\_5’
   2. Extract the 15th column & store it in ‘customer\_15’
2. Extract the column numbers 3,6,9,12,15 & 18 and store the result in ‘customer\_3\_multiple’
3. Extract all the columns from column number-10 to column number-20 and store the result in ‘c\_10\_20’
4. Extract all the columns which start with letter ‘P’ & store it in ‘customer\_P’
5. Extract all the columns which end with letter ‘s’ & store it in ‘customer\_s’

Now, Sam has to use the filter() function from the dplyr package to filter out data on the basis of some conditions

## Questions on filter() function:

1. Extract all the customers whose Internet Service is “DSL” & store the result in ‘customer\_dsl’
2. Extract all the customers whose Contract type is ‘Month-to-month’ & store the result in ‘customer\_month’
3. Extract all the male senior citizens whose Payment Method is Electronic check & store the result in ‘senior\_male\_electronic’
4. Extract all those customers whose tenure is greater than 70 months or their Total charges is more than 8000$ & store the result in ‘customer\_total\_tenure’
5. Extract all the customers whose Contract is of two years, payment method is Mailed check & the value of Churn is ‘Yes’ & store the result in ‘two\_mail\_yes’

Now, Sam has to use the sample\_n(), sample\_frac() & count() functions from the dplyr package to do random sampling of the dataset and get a count of levels of different categorical columns

## Questions on sample\_n(), sample\_frac() & count():

1. Extract 333 random records from the customer\_churn dataframe & store the result in ‘customer\_333’
2. Extract 1000 random records from the customer\_churn dataframe & store the result in ‘customer\_1000’
3. Randomly extract 23% of the records from the customer\_churn dataframe & store the result in ‘customer\_23\_percent’
4. Get the count of different levels from the ‘PaymentMethod’ column
5. Get the count of different levels from the ‘Churn’ column

Now, Sam has to use the summarise() & group\_by() functions from the dplyr package to get summarized results

## Questions on summarise() & group\_by():

1. Get the median, variance & standard deviation for the ‘tenure’ column
2. Get the median, variance & standard deviation for the ‘MonthlyCharges’ column
3. Get the standard deviation of ‘tenure’ & group it w.r.t ‘PaymentMethod’ column
4. Get the median of ‘MonthlyCharges’ & group it w.r.t ‘Contract’ column
5. Get the variance of ‘TotalCharges’ & group it w.r.t ‘InternetService’ column