**Data Science & Modelling Foundation Capstone Project – R**

**Data**

Please download the Capstone Dataset from your email and store it in your preferred drive location. Data are about colleges and universities in the US. It contains both numeric and categorical data. Be sure to check it.

**Data Preprocessing Tasks**

1. There are some categorical data like region, college type, campus setting, and Carnegie classification. Convert these categorical names to dummy codes using the command that we showed in the class.
2. There are some columns which are totally unnecessary, try to drop them. Which of you find most unnecessary for our analysis and which one you want to drop.
3. Are there any missing values? Are the missing values in one column or multiple columns? If there are in multiple columns use the appropriate function method to fill the cell with na values?
4. We have rank data for the universities. But rank data are mostly uninformative. So, we want to create a bracket of ranking like high, medium, and low ranking? How would you do that?
5. Similarly, we have student to faculty ratio. We also want to create three buckets of student to faculty ratio: High, Medium, and Low? Do that.
6. If you think you want covert any categorical variable to dummy codes, carry on and do so in your project.
7. We want to create a different data set and store in our local machine . The dataset will contain only those values associated with high student to faculty ratio. Perform this. And save the data set in your machine in CSV format.

**Visualization Tasks**

1. Create a bar plot that will show student population by region and also create the box plot for this. In the box plot also show the mean value.
2. Create a box plot that will show median salary of universities by Carnegie classification.
3. Create a scatter plot of student teacher ratio and median salary of universities and fit a line that will depict the relation between these two variables.
4. Create a scatter plot between campus setting and percent of student financial aid and highlight the dot of the scatter by campus setting: urban or rural?
5. Can you create multiple scatterplots in R that will show the following:
6. Relation between student teacher ratio and median salary
7. Relation between student faculty ratio and percent students in Financial aid