## **Lab 10**

### Pandas and Plotting

**Tasks**

1. Data Preparation
   1. If you look at the data frame, you notice that column “link”, “logo” takes too much space. Remove those two from the data frame. Also, remove “city” column since there are too many unique values.
   2. You can rename variables to your preferred name since some variables have long text.
   3. Create a new data frame which contains Top 400 universities for each year (2017 to 2022). Hence there will be total 2400 observations. You might notice that the variable “rank” is in the format of ‘factor’. This might cause further problem throughout the analysis. Replace the “rank” variable to include integers from 1 to 400.
2. Top 100 universities from the Top 6 countries.
   1. Show the count of where top 400 universities are located across the region in year 2022.
   2. Check how many universities from the United States ranked top 400 in year 2022. Calculate a percentage of it out of top 400 universities.
   3. Identify top 6 countries which have the most universities ranked in TOP 100 in 2022.
   4. Create a data frame that shows the average score of universities for each country from year 2017 to 2022. Arrange the average score in descending order. (to take the mean remove the NA from the data frame).
   5. Create an appropriate plot that shows the changes in average score of universities over time (2017 to 2022) for countries you have identified in 2.3.
3. University in Australia
   1. Create a table that shows changes in rank and score for Monash University from 2017 to 2022.
   2. Provide a multiple bar chart that visualizes the changes in ‘score’ of Australian universities that are ranked in top 100 from 2017 to 2022.
   3. Provide a multiple bar chart that visualizes the changes in ‘rank’ of Australian universities from 2017 to 2022. Note that your y-axis should be reversely plotted since the lower rank is better.
4. Visualize the data to explore relationship between university score and the number of international students for top 100 universities from 2017 to 2022.
5. Produce an appropriate plot to exhibit a comparison of the distribution of ranking across the size of the universities. Use data for top 400 universities in year 2022. Mind the order of the x-axis.
6. Create a plot that shows a distribution of ‘average score’ from each country covering top 400 universities in year 2022. To enhance the aesthetic of the plot, use ‘region’ to identify how universities from each region is evaluated.
7. Lastly, generate a plot that shows a comparison of distribution top 400 universities’ score between university type using year 2022 data.