# NM2207

Session 08

Challenges

# Overview of what we will do today:

* Data subsetting
* Charts

Before attempting the challenges, you are expected to have watched and coded along with the Lecture videos. A tutorial is meant to practice the skills presented in the video lecture, and show you more applications of it. Tutors will explain the challenges and answer the questions you may have**.**

**The challenges are due to be completed at the end of class each week for full credit which is also attendance. Submitting by midnight of the same day accounts for half the credit.**

**Summary of learnings (Chapter 2 of Hands-on Data Viz)**

* Reading a dataset
* Manually preparing it for exploration
* Manually visualizing trends
* Putting it into a chart!

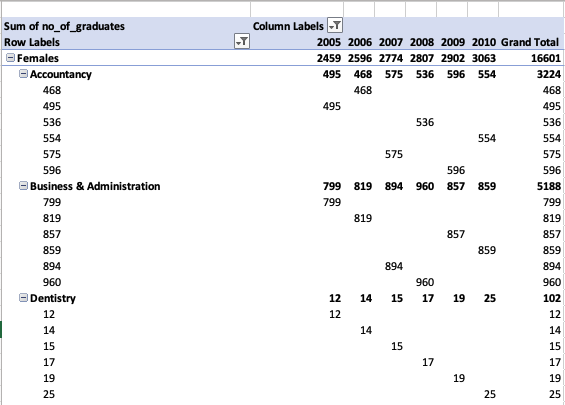
**Aim of this practice:**

We are going to prepare a dataset for visualization. We are going to look at a dataset detailing the number of graduates from different courses from 1993 to 2014. From this, we want to be able to tell, at one glance, which course produced the most number of female graduates between the years 2005 to 2010, as well as which course had the least female graduates.

* Reading a dataset
  + Go to https://data.gov.sg/dataset/graduates-from-university-first-degree-courses-by-type-of-course.
  + Click on “Download” on the top right corner. We will be using the CSV file in the folder which you have downloaded. Open the file using excel.
* Cleaning the dataset
  + There are a few variables present in this data set. What are they? Fill in your answers below:
    - Variables in this data: Year, Sex, Type of course and Number of graduates.
  + Filter the data. We only want to keep the following 8 courses:

1. Accountancy
2. Business & Administration
3. Dentistry
4. Health Sciences
5. Information Technology
6. Law
7. Medicine
8. Natural, Physical & Mathematical Sciences

* Subset the data:
  + We only want to see female students who graduated between the years 2005-2010.
* Using colors to see trends
  + Now that the number of graduates have been coloured on a scale from low to high, we can easily compare the number of graduates from each course and have a sense of which course produced the most and least female graduates
* Finally, we want to compare which course produced the most female graduates over the span of 5 years, between 2005 to 2010.
  + First, we need to group the course together so that we can visualise our data better. Go to Data 🡪 Sort 🡪 Sort by values (Order A to Z). Select type\_of\_course for column. Now, we can easily see which course produced the most and least number of female graduates between 2005 to 2010.
  + Fill in your answers:
    - Course with most female graduates: Natural, Physical & Mathematical Sciences
    - Course with least female graduates: Dentistry
* Did you find that it was difficult to conclude who had the most female graduates? We will have to do even better data visualization to tell.
  + First, let us create a pivot table (Insert 🡪 Pivot Table). Then, we select our range of data, and create the table in a new worksheet.
  + Now, we have to sort our data into the correct field. Explore around to see where each variable should go, until you are able to render a table that provides you with the answer to our opening problem. Remember to make use of the filter tool within the table itself to filter out only the data you want. [Hint: If your table is not showing the numbers correctly, check that the settings for calculating value is correct. We want to use the ‘sum’ function to add the numbers together. The ‘count’ function is used to calculation text. You can change these calculations by going to the "PivotTable Options" tab and clicking the "Fields, Items & Sets" button.
  + Your final result should look something like this **(and focus only on the rows in bold)**:



* **Now that we have this table, we can accurately answer the problem at hand**
  + Fill in your answers:
    - Course with most female graduates: Natural, Physical & Mathematical Sciences
    - Course with least female graduates: Dentistry
* Visualize this in Excel. Copy the rows into a fresh tab. Select the rows. Insert->Bar chart. Is this what you want to show in your web page? Pick the color scheme you like using the tools on excel.
* Now, use the graduate\_barChart.html to replace the data in the bar chart with the rows in bold! Follow the comment prompts in the js file.

**Reflection**

* Based on the chart you drew in Part 2, what have you learned females in different academic disciplines? Does this surprise you? Why/why not?
* What are the errors you encountered today, and how did you address them? Add a table showing this in the html body of your page.