**Schedule/timeline update**

**Following the timeline you proposed in your project proposal, please give us an update regarding finished activities, possible modifications and/or additional activities.**

Our team succeeded at downloading all the IPEDS datasets and scraping the lists of rankings from the web page of “Times Higher Education College Rankings 2017” and “QS World University Ranking”.

Our database is composed of 12 .csv files (our core datasets) , 12 .xlsx files with the variables’ descriptions and 2 .xlsx files containing universities’ rankings.

To upload all the data at once and avoid repetition, the lapply function has been used.

Using a loop, we assigned to each variable of the core datasets a label indicating its description. These labels allowed us to identify quicker and in a more practical way which columns to use for our analysis. Columns that did not contain valuable information have been eliminated.

In addition to the expected work, the data required extensive cleaning. This procedure was necessary in order to have a key (universities names) that could be used to join our databases to the lists of rankings. For this new task, we applied tools learned doing text mining for this semester’s company project.

As for the tidying, some files have been tidied before plotting them and for others the tiding will be done in the plot’s code. The choice will depend on the characteristics of the graphs, as well as the dataset’s usage frequency.

**Task accomplished:**

Loading and scrapping data

Cleaning variable corresponding to University names to link files

Initial phase of data visualization and analysis

\*The tasks were completed according to the schedule previously presented.

**Preliminary results**

**Briefly summarise your current findings. Is the project going as expected? Will it be possible to achieve the goals you set in the proposal?**

**Initial findings:**

All top universities have admission tests and take into account school records. Moreover, it was interesting to discover that apart from the usual requirements, having a recommendation letter is also often mandatory. To assess these information, a bar chart was drawn.

Furthermore, the acceptance rate seems to be really low for top universities. Lastly, on average, the top 15 universities have more teachers compared to the number of students than non-top 15 universities. Nevertheless, there are some outliers. As an example, the university of California based in Barkley is part of the top 15 universities but have a really high student/faculty ratio, or Rice University, which is ranked as 87th but has a really low ratio.

**Follow-up of the project proposal:**

As suggested by the previous schedule, the project will now focus on answering and analyzing research questions to have a better understanding of which variables play a bigger role in shaping top universities. Then, we will use these characteristics to build our own rankings and compare these to the original ones. Cleaning the data took more time than anticipated, nevertheless the time invested in this first part pays off. We expect to achieve the goal we set in our proposal, however currently the number of research questions are more focused considering that they will be used for the prediction task.

**Institution characteristics**

1. Have an international environment?
2. Generally, have large campus and are mostly private universities?

**Educational offerings**

1. Offer high degrees such as Post-master's certificate and Doctor's degree?

**Cost and investment**

1. Are the most expensive considering tuition and fees?

**Application and admission**

1. Have mostly a non-open admission policy?
2. Have the lowest rate of admission related to the number of applicants?
3. Usually have more requirements that non-top universities?

**Effective teaching**

1. Have a low number of teachers per student?
2. Have the highest graduation rate among the universities?
3. Provide high salaries to their instructional staff?