This is the form your group must fill out for the project proposal (due May 5 2019 by 11:59pm CET).

El nombre y la foto asociados a tu cuenta de Google se registrarán cuando subas archivos y envíes este formulario. ¿No es tuya la dirección analucy.bejarano.m@gmail.com? Cambiar de cuenta

\*Obligatorio

Title of your project proposal \*

What makes a top University in USA in 2017?

Group member 1's name \*

Anna Alfieri

Group member 2's name \* \*

Ana Lucy Bejarano Montalvo

Background and motivation \*

**Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.**

As students ourselves, we needed to research which institution to attend. Rankings are always a practical source for exploring options and many undergraduates rely on them. Nevertheless, different agencies use different scoring methods and the procedures are not transparent. This limit the decision to a simple selection that does not allow to identify which are the key characteristics that a top university should have.

Since American colleges always occupy the top 10 positions of all rankings, we decided to focus only on the American system, but we believe this analysis could be extended also to other countries

On one hand, with this analysis we hope to help our siblings and friends in the decision process of where to attend college. It could be beneficial for them to know which are the key factors that they should look for instead of simply relying on rankings.

On the other hand, the course “Data Driven Business” help us realize that the current non-transparent methods can lead to biased decision. In this sense we would like to use Data Science tools to contribute to critical thinking among students.

**What are the scientific and inferential goals for this project? What would you like to learn and accomplish? List the benefits. What are some optional features (features or calculations which you consider would be nice to have, but not critical)?**

The goal of this project is to identify which are the most important characteristics that lead to a top USA Universities in 2017 and predict their rankings.

To identify the central features, we want to answer the following questions:

The top 50 colleges in USA

**Institution characteristics**

1. Have an international environment?
2. Generally, have large campus?
3. Are in a particular region?
4. Are mostly private universities?
5. Have a similar gender or ethnicity composition?

**Educational offerings**

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

**Cost and investment**

1. Are the most expensive considering tuition and fees?
2. Are the ones that expend the most in their libraries?

**Application and admission**

1. Receive more applications compared with the other universities?
2. Have the lowest rate of admission related to the number of applicants?
3. Have mostly a non-open admission policy?
4. Usually Requires GPA, rank and records to admit students or others test, for example foreign language?

**Effective teaching**

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

Other questions may come up when analyzing the data.

**Benefits**

* Students and/or families can decide themselves which is their top university considering the information available
* The data presented will reduce the unintended bias generated by the ranking companies
* People can verify the information directly from the available files

**From where and how are you collecting your data?**

The data has been collected mainly downloading files available in the web page of the Integrated Postsecondary Education Data System (IPEDS) . The rankings will be extracted by scraping the Times higher Education and The QS World University Ranking web page. Links presented below.

* Integrated Postsecondary Education Data System (IPEDS) <https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx>
* Wall Street Journal/Times Higher Education College Rankings 2017: https://www.timeshighereducation.com/rankings/united-states/2017#!/page/0/length/25/sort\_by/rank/sort\_order/asc/cols/stats
* The QS World University Ranking

<https://www.topuniversities.com/university-rankings/world-university-rankings/2016>

**List the statistical and computational methods you plan to use.**

Data cleaning and tidying, data transformation (dyplr package), joining tables, data exploration and visualization (ggplot package), creation of a predictive model, comparison of model’s results to different rankings’ sources.

The methods that we plan to use for prediction are: Classification tree, random forest and unit weighted model. We plan to compare the performance of these three models considering the accuracy but also their performance outside a training set.

N.B. This list is not exhaustive. Changes will be made depending on the results obtained from the exploratory analysis of the data.

Schedule/timeline \*

**Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.**

**Deadlines**

May 14: Data loading and scraping , cleaning, tidying and transformation.

May 19: Data visualization and exploratory analysis (Part I) and complete the “project update”.

May 22: Finalize data visualization and decide which variables to include for the prediction model.

May 26: Built the prediction model and compare the results to different rankings’ sources.

May 30: Review of the report and preparation of the oral presentation

**Responsibilities**

**Data loading**

Download data files with the variables selected -> Team member 1

Extract ranking from webpages -> Team member 2

**Wrangling data and analysis**

Institution characteristics, Educational offerings, Effective teaching-> Team member 1

Cost and investment, Application and admission -> Team member 2

**Modeling**

Classification three and random forest -> Team member 1

Weighted model -> Team member 2

Comparation of model -> Team member 1 and 2

**Final Analysis**

Team member 1 and 2

N.B. the schedule and responsibilities may be updated/modified according to the progresses made.