Milestone 2

1. Select a Project:
   1. ~~Feasible given the data~~
   2. ~~Perform initial analyses (exploring and visualizing)~~
   3. ~~From 2015-2020~~
   4. ~~Can use external data~~
2. Google Colab notebook with code to start loading the dataset
3. Steps:

* ~~Acquaint yourself with data~~
* ~~Preprocess the data~~
* ~~Complete necessary descriptive statistic tasks~~

1. Requirements:

* ~~Have a pipeline in place~~
* Fully document pipeline-ESCRIBE MUCHAS (Y BUENAS) DESCRIPCIONES
* ~~Show that we have clear project goals~~

1. Proofs:
   1. ~~Handle whole data size~~
   2. ~~Understand what is in data (formats, distributions, missing values, correlations, ...)~~
   3. ~~Consider ways to enrich, filter, transform data according to needs~~
   4. ~~Plan for analysis and communication is reasonable and sound~~
   5. ~~Explain other choices of analysis and communication that we considered but discarded~~
2. Their evaluation:

* How well the steps are done
* How are the steps documented
* Quality of code
* Documentation of code
* ~~Feasibility of project~~
* Critical awareness of the project
* ~~Explain in clear, reasonable, and thorough way the project idea~~

1. Where:

* ~~Public ADA repository~~

1. What:

* ~~Project proposal-README~~
* ~~Initial analysis notebook~~

1. External sources:

* Additional metadata about 9M speakers in .parquet file (speaker\_attributes.parquet) in Google Drive
* Load it as: df = pd.read\_parquet(<path\_to\_file>).
* Pandas requires pyarrow to read parquet files, which can be installed using conda install pyarrow -c conda-forge.
* Sample code to extract information from speakers form the Wikidata knowledge base (they are very large ~100GB)

1. The deliverable

* Readme:
  + Title
  + 150-word abstract (description of project idea and goals: motivation? What story do we want to tell? Why?)
  + Research questions to address during the project
  + Proposed additional datasets (ideas on how to get, manage, process and enrich them. Show you have read the docs, some examples and you have a clear idea on what to expect. Also, discuss data size and format)
  + Methods
  + Proposed timeline: future lines
  + Organization within the team. List of internal milestones up until milestone 3.
  + Questions for Tas
* Notebook
  + Initial analyses
  + Data handling pipelines
  + Look at correctness, quality of code, and quality of textual descriptions

intro - alvaro

the data - kamil: explain quotebank and external sources proposed by the ADA teachers

load data - kamil

examine, clean up and justifications- alicia

q1, q2- victor

q3 - kamil

q4 - alicia thinks about it

modeling - victor thinks about it

storytelling - alvaro

conclusions - (future)

future lines - alvaro