Email Laura:

lauraruis@live.nl

Schedule

	Monday	Tuesday	Wednesday	Thursday	Friday
Week 1	HC:11:00-13:00	TA: 12:30	HC: 9:00-11:00 H: 13:00-16:00	TA: 10:00 I: 13:00-16:00	
Week 2	HC:11:00-13:00 J: 13:00-16:00	TA: 12:30 M: 13:00-16:00	HC: 9:00-11:00 J: 11:00-14:00	TA: 10:00	
Week 3	HC:11:00-13:00	TA: 12:30	HC: 9:00-11:00	TA: 12:00	
Week 4	HC:11:00-13:00	TA: 12:30	AIExpo	TA: 10:00	Pres: 10:00-12:00

Group Log

WEEK 1
4/06/18

- Orientation. Chosen dataset is "University World Ranking". This looked like an interesting dataset, with a lot of room for further questions; How can you map quality; noise in data; different countries with different values. - All
- Created Shared Drive
- Created ShareLatex Report - AH (https://www.sharelatex.com/2654235839zdtppgdncjqp)
- Created Google.Colab - AH
- Created GitHub repository - SO
- Created Schedule - TS
- started scraping - All

05/06/18

- Preparation meeting -All
- Meeting TA - All
    - add Laura to communications

        - goal next meeting: Finish scraping and clean data

    Problems:

        - dataset not visible via normal html source, probably a javascript component
            - solved by storing all contents in drive, will still scrape those documents

- Assigned attribute types to data - JC
- Start scraping the data from THE site. starting with 2018
- Put all university rankings in pandas dataframe and replaced ‘n/a’ values with mean - TS

07/06/18

- Intercepted JSON file from THE website, which means the data doesn't have to be scraped. intercepted years 2011 to 2018 (https://www.timeshighereducation.com/sites/default/files/the_data_rankings/world_university_rankings_2018_limit0_369a9045a203e176392b9fb8f8c1cb2a.json)
- Descriptions of all colums in db
- cleaned data:
    - double data: 'rank_order' and 'scores_overall_rank' the same; deleted scores_overall_rank
    - dropped 'apply_link', since only present in 2 out of 1103
    - dropped 'member_level' & 'record_type'; information for the original site, so not useful for us
    - dropped aliases; same as name
- Problems:
    - Replace missing values with mean or leave them empty?
    - (for mean percentage of male students, per country, take ratio of each uni, calculate how many male students there are, add these up and divide by total amount of students in that country)
- goals for monday:
    - finish deliverables week 1

09/06/18

- Take a look at 2016 data and clean it the way 2018 data was cleaned - JC

10/06/18

- Finished filling in missing data with correct mean in 2018 data - TS

WEEK 2

11/06/18

- een heleboel shizzle
- get data from local file in git, not an external url. This data isn't dynamic, so it doesn't pose a problem. It makes the data faster accessible and faster interpretable by the algorithm.
- dataanalysis; form general ideas about what the abnormalities are in the datasets, what is interesting, what we want to investigate further, etc - All

12/06/18

- Combine data to get the rank per university over the course of the years in one dataframe - TS

13/06/18

- analysed data; sought connections between different variables of the data -JC

14/06/18

- Put data about the change in rank - TS
- Andere dataset gescraped, met het totale aantal universiteiten in elk land -JC

16-17/06/18

- looked at data per country, to see whether that would bring interesting ideas to the table
- searched for other interesting occurrences in data

WEEK 3
18/06/18

- looked at the concept of clustering and regression, and whether it is useful for our dataset

19/06/18

- Summarized preliminary findings, and added them to the report

21/06/18

- made a mockup version of the site. Understand what we want to understand what we want to show on the site. Understand how we want things to look on the site. This also gave a proper outlook on what has been done, and what still has to be further looked at and worked out.
    - final answers original 3 questions
    - apply regression, and clustering
    - combine everything in website
    - get website working

22-23-24/06/18

- data on site
- Site op server
    - probleem; bokeh server can't be embedded in html. This means that any plot using the bokeh server, css/html can't be added to the design.
- looked at and plotted the consistent universities and countries, looking at the variance of the ranks through the years.
- looked at (lack of) anomalies, and considered the reason for this
- went through data to find (interesting) patterns, and visualize them
- applied univariate regression, following the example on canvas. To get an idea what multivariate regression could be interesting

WEEK 4

25/06/18

- tie up loose ends
- made sure we have a concrete plan to get a solid website and analysis ready thursday evening

26/06/18

- Finished regression
    - problem; what does the MSE *really* mean?
- Fixed that site can be embedded with bokeh server

27/06/18

- AIExpo
- pre-final touches website and data analysis
    - Googled interpretation for MSE, different stackexchange pages explained it clearly; the unit of the MSE is the unit of the variable you are trying to predict, squared.
    - Typed technical report
    - assemble and prettify site

28/06/18

- finalfinal touches