Project Data Thieves

Team Falafel

Ulrike Anklam

Bruna Miguel

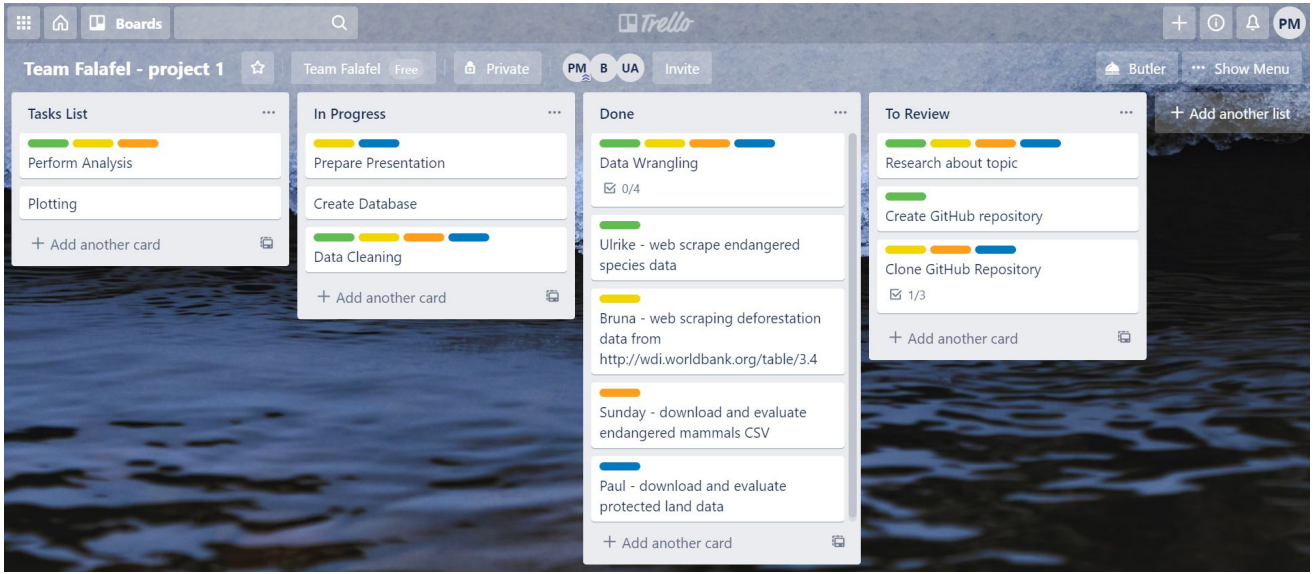
Paul Musco

Akinbuwa Oluwadare Sunday

Project aim

Investigate relationship between endangered animals and land use by country

Organisation



- Each team member sourced and cleaned their own dataset
- Combine individual data into one dataset for group analysis
- Slack channel: communication, Trello: keep track of progress,
- GitHub: shared repository

Techniques

- 2 x CSV download
- 2 x web scraping
- (API)

Data sources

CSV

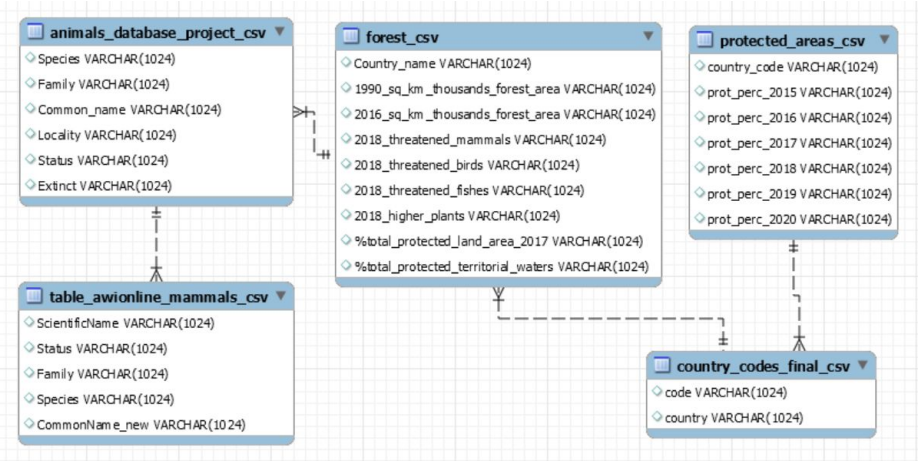
1. Mammals by location classificed by conservation status (Sunday) - https://www.departments.bucknell.edu/biology/resources/msw3/
2. Changes in percentage of protected areas by country (Paul) - https://data.oecd.org/biodiver/protected-areas.htm#indicator-chart

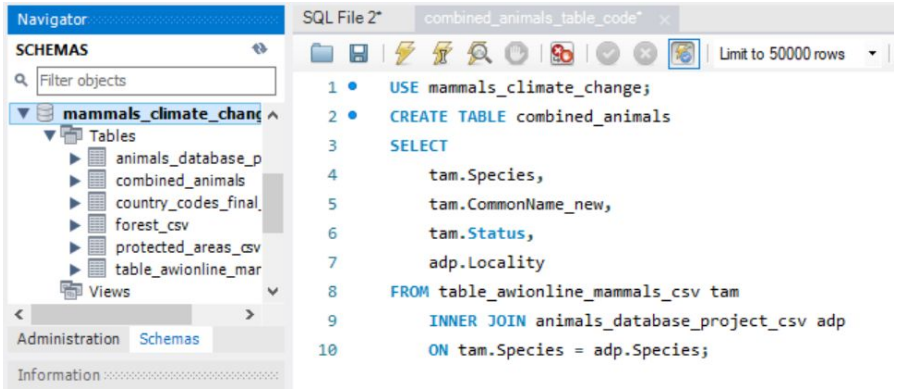
WEB SCRAPING

1. Deforestation & biodiversity (Bruna) - http://wdi.worldbank.org/table/3.4
2. Mammal species & level of endangerment (Ulirke) - https://awionline.org/content/list-endangered-species#mammals

Database schema

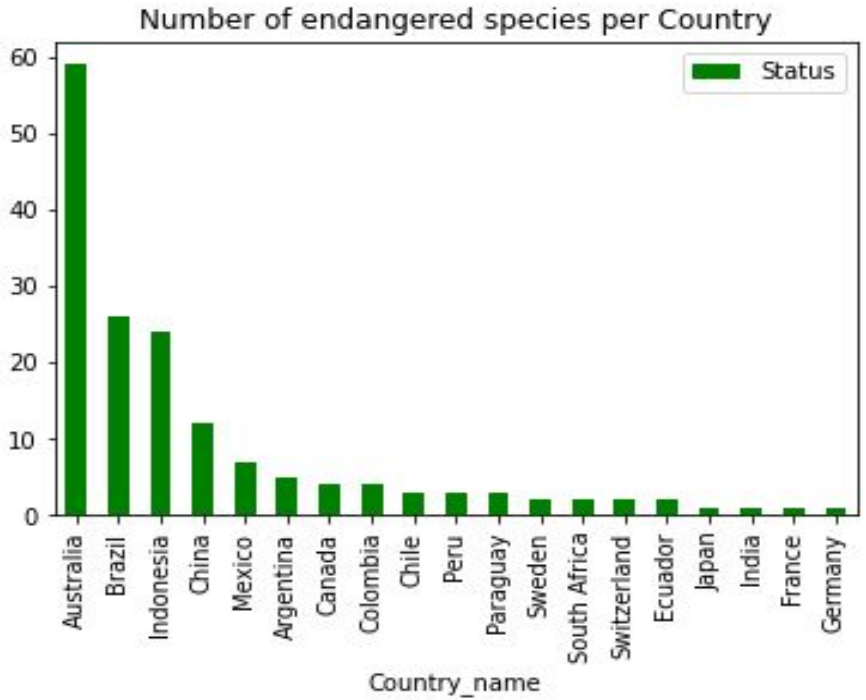
- Import 5 tables from individual data sourcing into MySQL
- The 2 tables relating to animals, and the 3 tables relating to land use combined using INNER JOIN





- 2 combined tables imported into Jupyter Notebook, merged and cleaned to create one large dataframe

Results


Number of endangered species per Country
Status
60
50
40
30
20
10
0
Australia
Brazil
Indonesia
China
Mexico
Argentina
Canada
Colombia
Chile
Peru
Paraguay
Sweden
South Africa
Switzerland
Ecuador
Japan
India
France
Germany
Country_name

Problems

- API's...'<Response[500]>' messages, subscription only APIs, lots of wasted time
- Data alignment...difficult to find data from different sources that matches in terms of dates
- Table merging...'Country_name' & 'country' fields didn't match exactly, so took significant data cleaning to get them to a point where they could be joined

Learnings

- Trying to source data from APIs can be incredibly time consuming (and frustrating)
- Spend additional time on making the data cleaner, allows for easier analysis
- More extensive dataset (fewer than 200 rows in final dataset)
- More practice needed in terms of plotting

Questions?