**Web Scraping and Social Media Scraping Project**

Participant: Olga Marut

Analysed domain: <https://www.worlddata.info/capital-cities.php>

Abovementioned <https://www.worlddata.info> domain contains various data about countries from all over the world. There are many tables describing time zones or coronavirus statistics available on it and more descriptive data divided by continents and other geographical units as well. In the project attention was given to a */capital-cities.php* subdomain. Data collected by prepared scrapers describe capital cities from all over the world and contain some basic information about countries they belong to. Text data like the region of the world country lies in or name of a country in the local language were scraped but some numerical data like export and import volumes or area of a country were gathered as well.

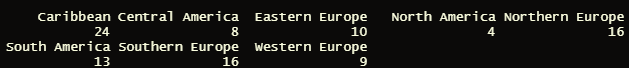
As it was expected by project rulings scrapers employing three different web scraping packages were prepared: Beautiful Soup, Scrapy and Selenium. Generally, all of them search for list of all countries available on the website firstly and then gather data from links corresponding to each of countries listed in the previous step.

Data scraped by each of the scrapers contain the following information for each of capitals:

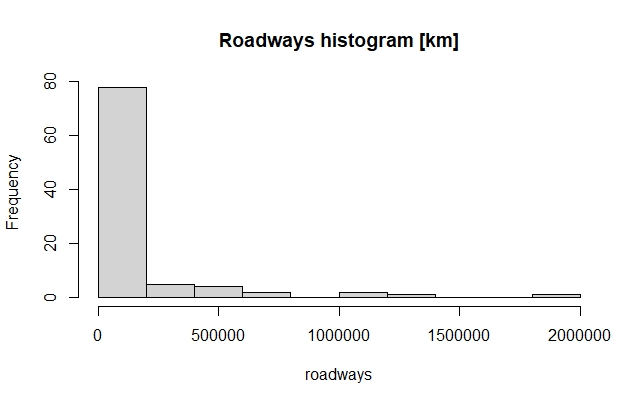
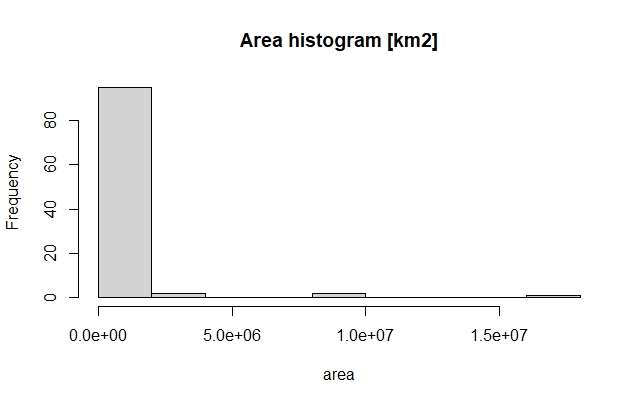
* capital: name of the capital city
* name: name of a country capital belongs to
* region: region of a world country lies in
* area: surface of a country expressed in squared kilometres
* local\_name: name of a country written in a local language
* export: volume of country’s exports in millions of dollars
* import: volume of country’s imports in millions of dollars
* airports: number of country’s airports
* birth\_rate: country’s percentage birth rate
* death\_penalty: information if death penalty is allowed/ not practiced/ banned (if it is banned, year since it is not allowed is given)
* roadways: number of kilometres of country’s roads

Short statistics of scraped data:

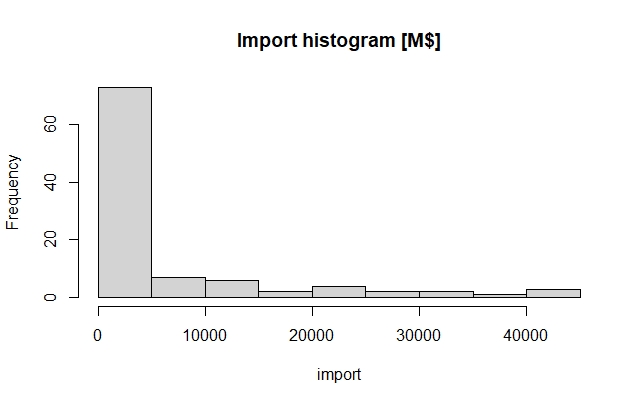
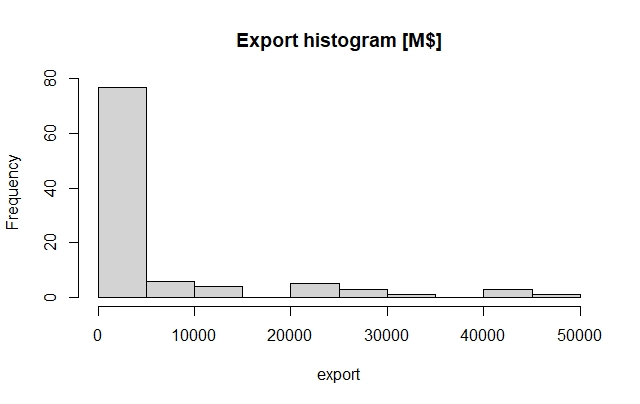
* **region**: number of instances for each region was counted. As we can see, countries analysed in the dataset most commonly belong to Caribbean and most seldomly to North America. It is needed to be mentioned that this disproportion results from the fact that only first 100 pages were scraped and countries were listed in order of continents they belong to, so f. ex. Asian countries are missing.



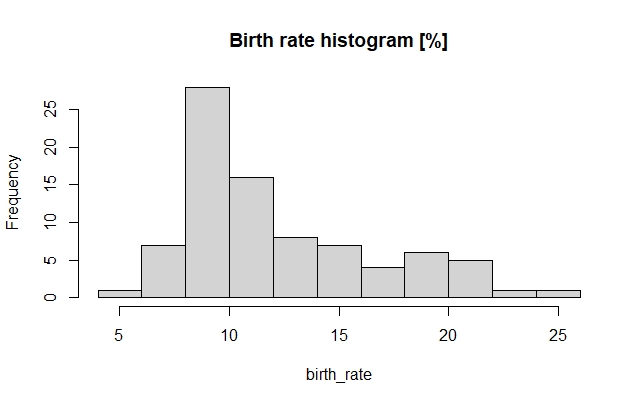
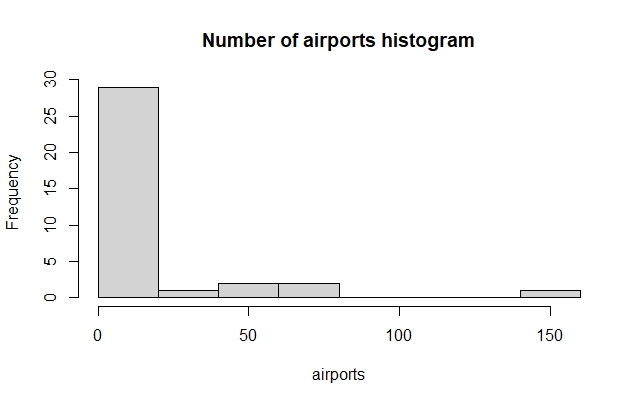
* **area and roadways**: histograms



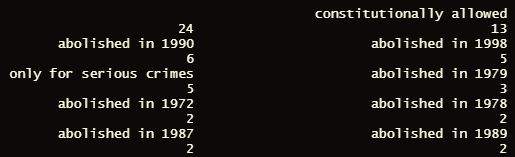
* **export and import**: histograms



* **airports and birth rates**: histograms



* **death penalty**: below number of instances for most common 10 statuses of death penalty in analysed countries is listed. As we can see, apart from missing data (as they were not provided for all countries), constitutionally allowed status is most common. Obviously, this happens only when for countries where death penalty was abolished, the year of abolishment is distinguished. All in all, the vast majority of analysed countries doesn’t allow death penalty.



* **capital, name and local\_name** were excluded from the statistical analyses as they are text data which is not comparable to each other. F. ex. region which is text type of data, too was analysed but it has repeatable values. But three above mentioned variables can be useful too. If we need to compare if country’s name is the same in English as in local language, we can use name and local\_name and compare them.

Scrapers comparison:

Even though all three scrapers are based on the same method – firstly list of all countries was prepared and then data for each country was scraped – they are characterised by different time of performance.

Selenium took over 24 minutes to gather data for 100 countries. It was definitely the slowest one. This happens as selenium needs to open browser and each of scraped pages separately. What is more, sometimes Google Adds were appearing when selenium scraper wanted to click on country’s subdomain. This consumes even more time for scraper to gather data, as it need to reserve some time for a pop up to load and then to be dismissed by the program. To be more precise, selenium ‘list of links’ for all countries was built in a slightly different way than in case of other scrapers, as links were found by searching for buttons that program can click on, not by creating list of links in csv file.

Scrapy occurred to be much faster as it took around 10 seconds to scrap the same information from 100 pages. Two spiders were prepared to achieve the data analysed in this project. First of them scraps links for all countries from main website to a csv file and second one scraps data from each country’s separate website and creates a table in pandas exported to the csv file at the end, too.

Beautiful soup also did the job much faster than Selenium but longer than Scrapy, as it took it around 50 seconds to gather information. It was prepared in a similar way to Scrapy program as it firstly creates list of links in a csv file and then gather information from listed websites. The output is also stored in a csv file. In contrast to Scrapy scraper, the BS one does everything inside one python file.

The project was fully prepared by myself.