Rome2rio

One of the sources we need to scrap was the Rome2Rio website. We retrieve the informations on journeys between all European Capitals. In this report, we're going to explain in a first time which data we scrap and the format of these data. In a second time, we're going to develop our scraping method and we're going to finish with some statistics.

I Format of scraped data :

The table below summary all data we scraped and give their type and a Description of what is the column.

Variable_name	Type	Description
Data_Source	string	Website used to scrape
Departure_city	string	Departure city
Arrival_city	string	Arrival city
Nb_bus_taken	int	Number of buses taken during the journey
Nb_train_taken	int	Number of trains taken during the journey
Nb_car_taken	int	Number of cars taken during the journey
Nb_plane_taken	int	Number of planes taken during the journey
Duration	int	Travel time
Price_min	string	Minimum ticket price
Price_max	string	Maximum ticket price

II Scraping method:

We scrape our data into several steps :

- For each trip, recover the departure city, the arrival city and the tag: «route__title» which permits to get :
    - -the travel time
    - - the number of times each means of transport is taken during the trip

We have all this information with the driver « chromedriver »

- Recover also for each trip the tag «route__details» which permits to extract the minimum and the maximum ticket price.

- Apply the algorithm(function) to all European capitals.

This website don't need a robot which scrap data all weeks because there is no comments on this website. If we want the new data, we need to scrap again the entire website. We can scrap the website once a month to have the new journeys informations.

WARNING : One of the things which don’t work on this website is the transition from chromedriver to phantomjs. This code don’t work on OSIRIM platform but work with chromedriver and google chrome.

III Statistics:

The final dataset counts 13674 rows and 72 columns

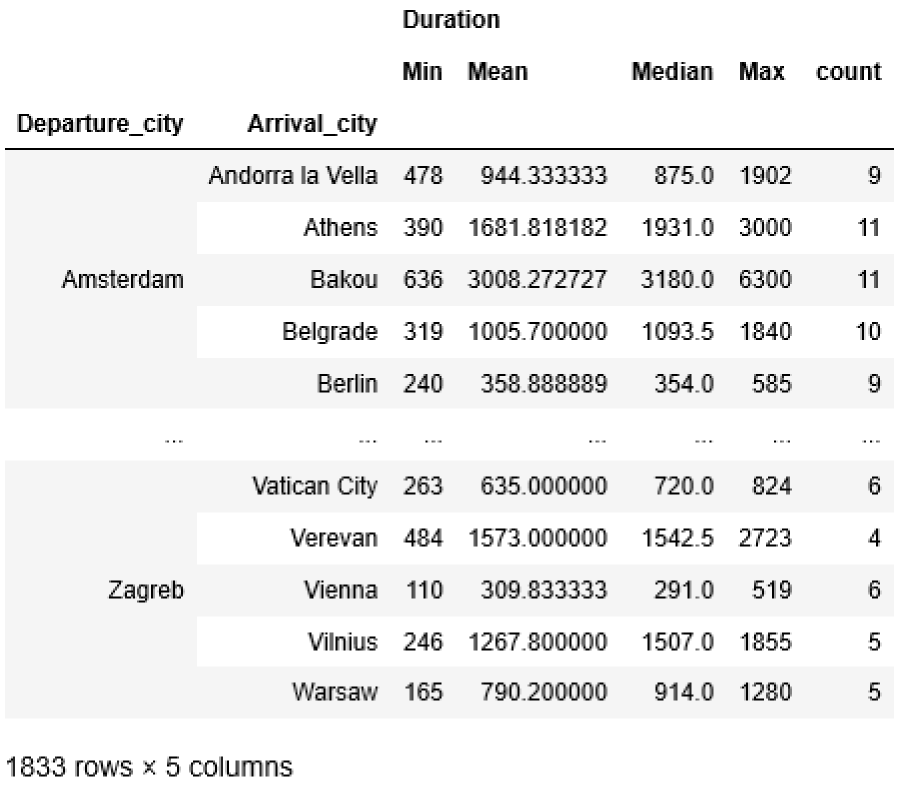
The following table shows for each travel :

-the minimum and the maximum duration

-the average and the median duration,

-the number of trips

For example, They are 9 ways to travel from Amsterdam to Berlin, the minimum duration is 240 minutes and the maximum 585. The average duration is 359 mn and the median 354 mn.



		Duration				
		Min	**Mean**	**Median**	**Max**	**count**
Departure_city	**Arrival_city**					
Amsterdam	Andorra la Vella	478	944.333333	875.0	1902	9
	Athens	390	1681.818182	1931.0	3000	11
	Bakou	636	3008.272727	3180.0	6300	11
	Belgrade	319	1005.700000	1093.5	1840	10
	Berlin	240	358.888889	354.0	585	9
...	...	...	...	...	...	...
Zagreb	Vatican City	263	635.000000	720.0	824	6
	Verevan	484	1573.000000	1542.5	2723	4
	Vienna	110	309.833333	291.0	519	6
	Vilnius	246	1267.800000	1507.0	1855	5
	Warsaw	165	790.200000	914.0	1280	5

1833 rows × 5 columns