Project 3 : Data Cleaning

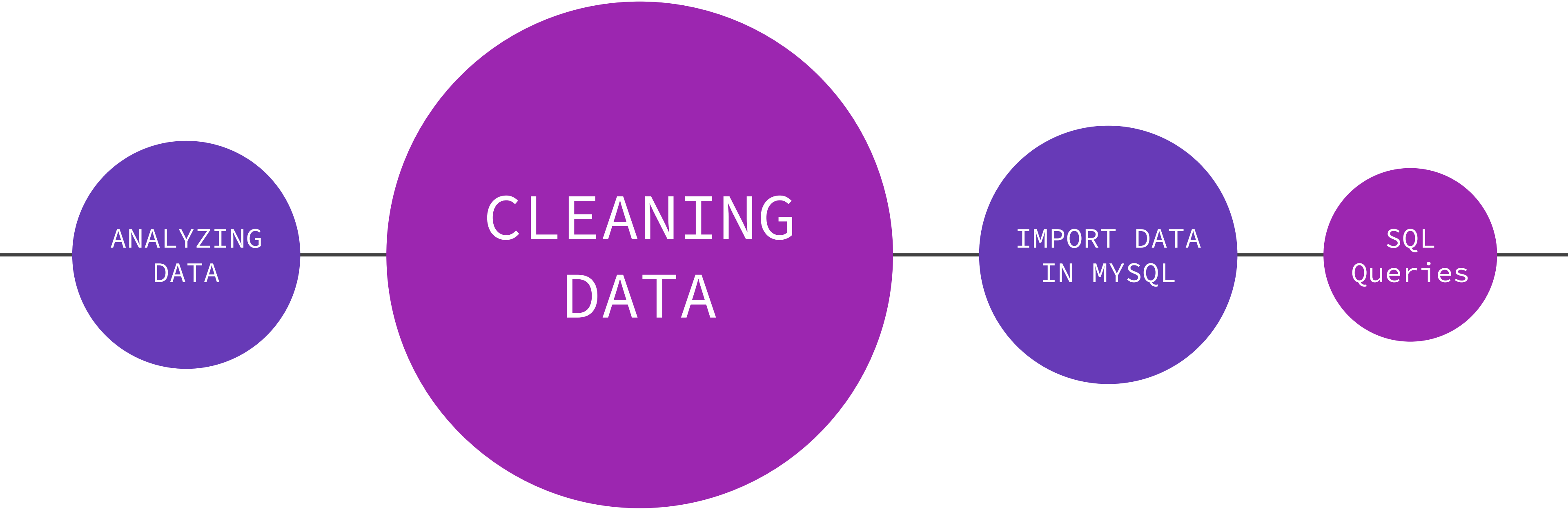
Kaci BOURGUA & Vincent WAKIM

Description of the dataset :

- TABLE ON PRIMARY EDUCATION AROUND THE WORLD
- With 12 COLUMNS (Countries, Regions, Sub-Regions, Income Group, Education Index …)
- 88 rows

PROCESS :

Process



ANALYZING DATA

CLEANING DATA

IMPORT DATA IN MYSQL

SQL Queries

Process

- Cleaning
- Time column to a single format
- Some typos
- Coherence and data consistency
- Some years were superior to 2023
- Columns
- 4 columns with more than 15% of missing values, we decided to drop them, instead of dropping more rows

- Replacing missing values
- The idea was to preserve as much rows as possible, since the dataset is already short. Depending on the case, we decided to fill the values depending on each case
- Region
- Sub-region
- Income groups

FIRST DATA SET :

- 12 columns
- 88 rows
- 29 rows with missing values (NaN)
-

2ND DATA SET :

- 8 columns
- 83 rows
- 0 rows with missing values

CHALLENGES :

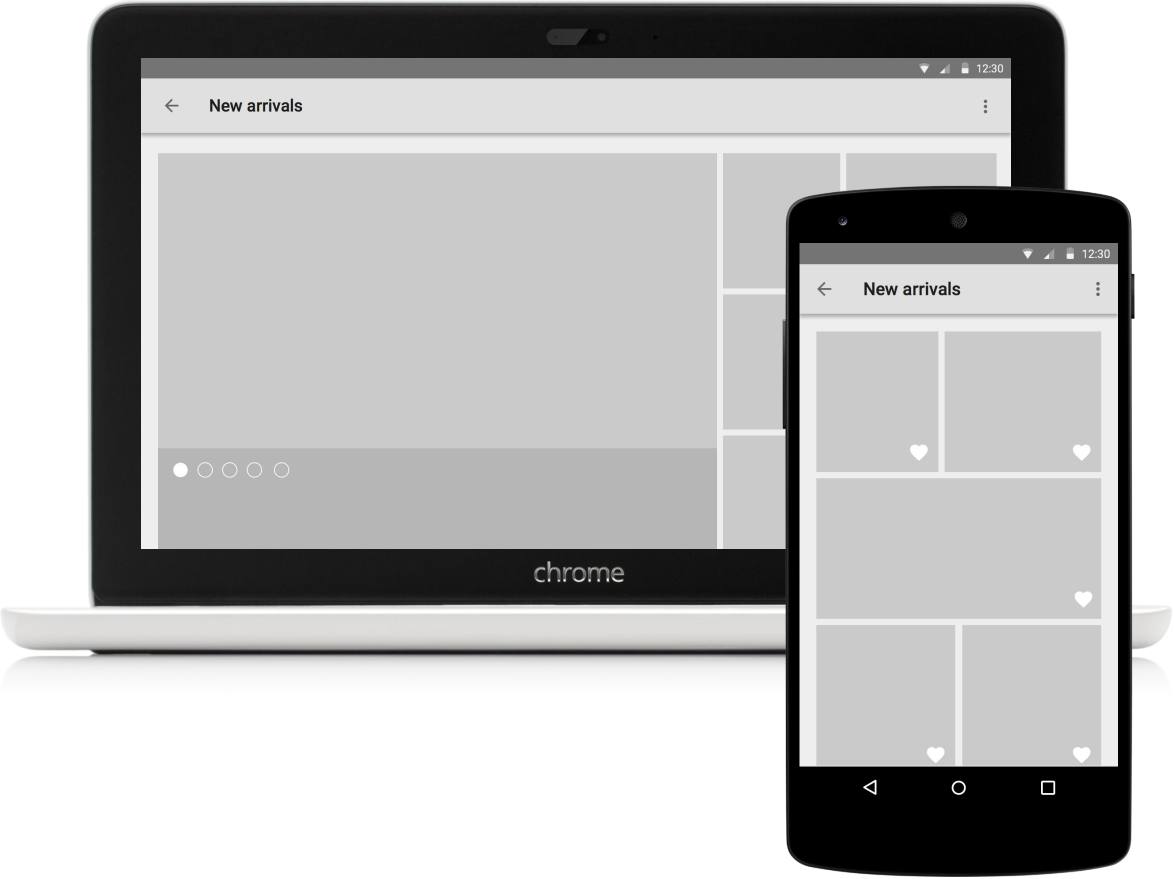
- THE DATASET WASN'T TOO LONG SO WE HAD TO LOOK AT EVERY DATA AND PICK THE MISSING ONES TO EITHER DELETE THEM OR CLEAN THEM

Learnings :

Cleaning data isn't always an easy task.

Lots of attention and knowledge of the data needed.

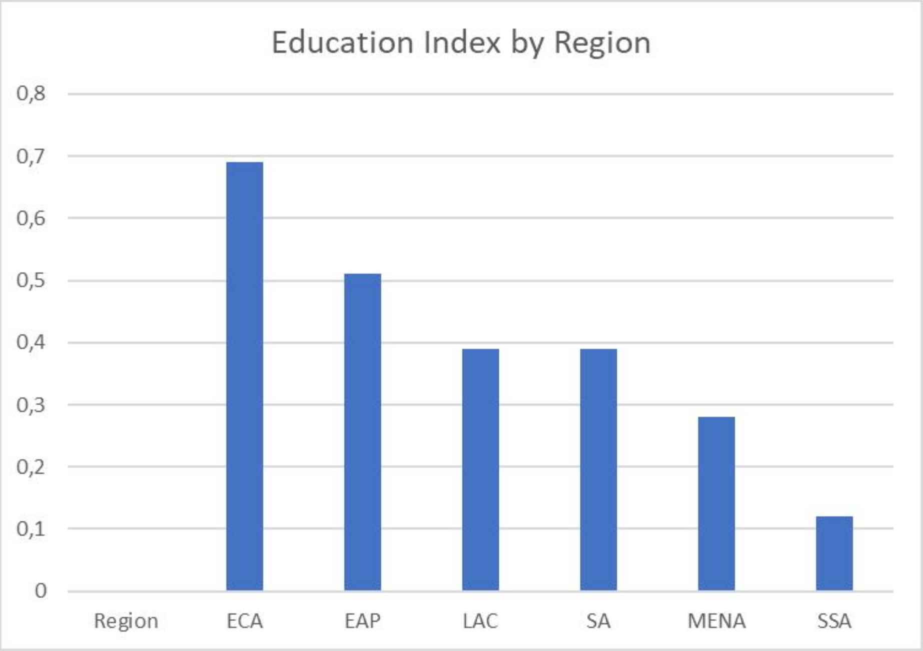
Making decisions based on the nature and relevance of the data

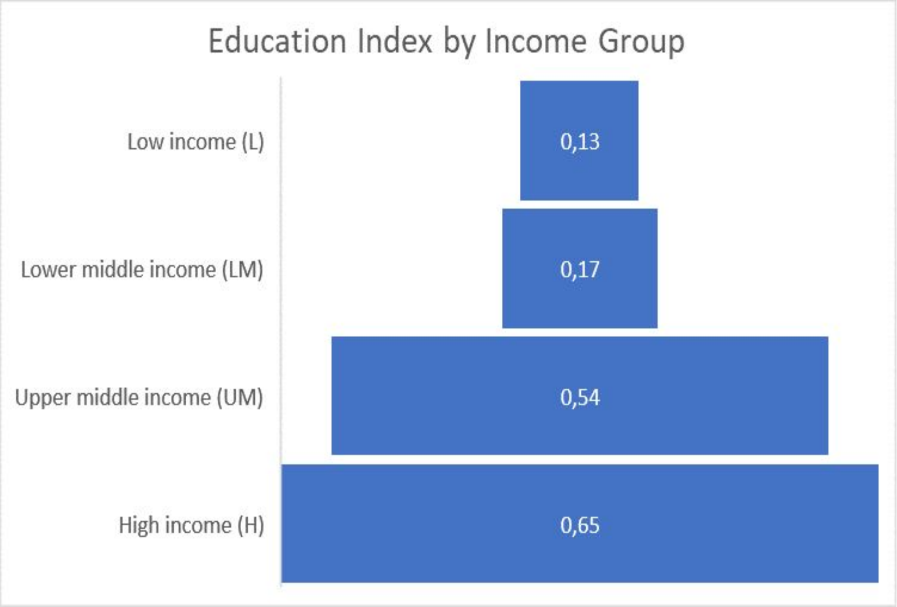


IMPROVEMENTS

CONNECT MYSQL WITH PANDAS TO MAKE OUR SQL QUERIES IN PYTHON.

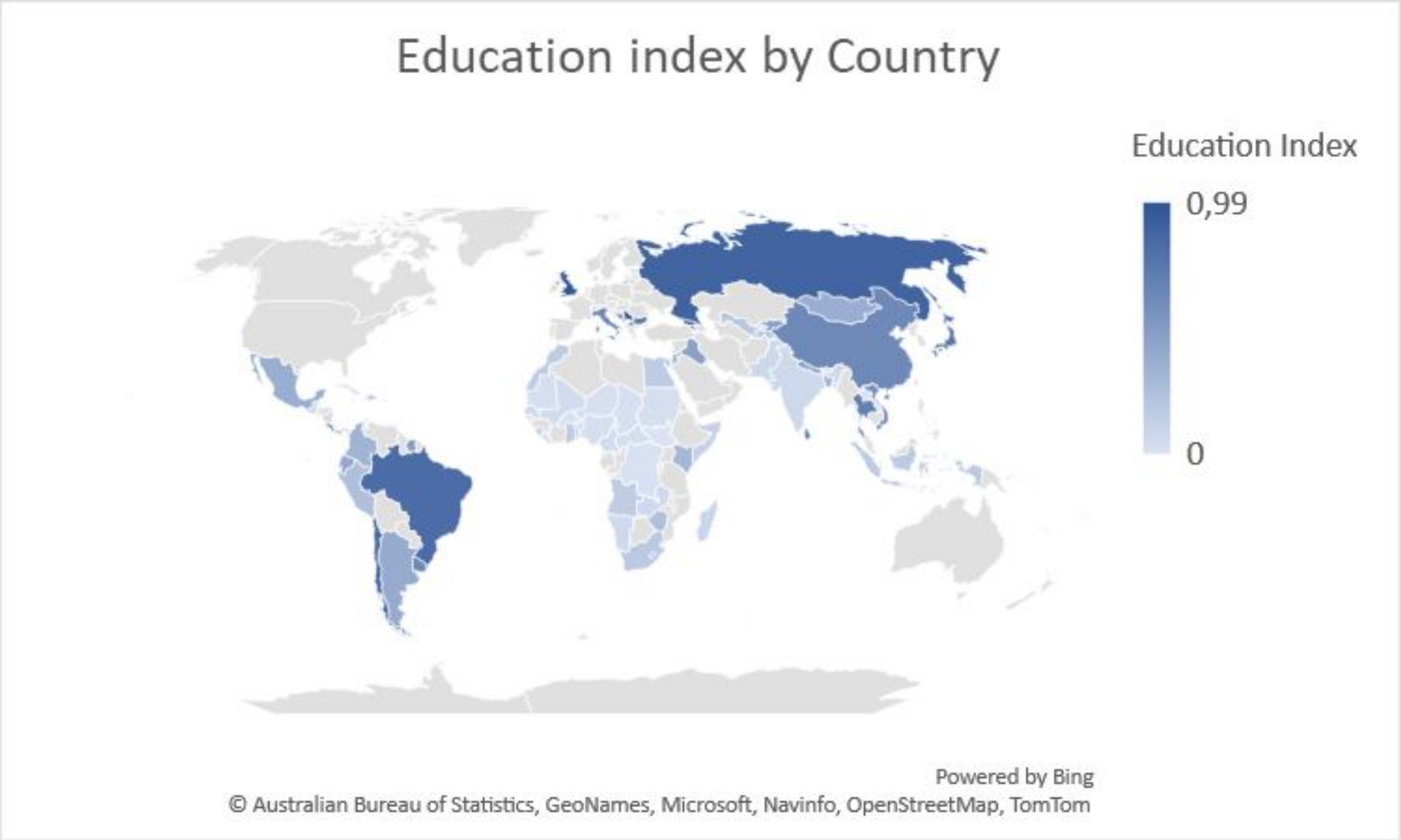
Results of the SQL queries





Europe & Central Asia, East Asia Pacific, Latin America & Caribean, South America,
Middle-East North Africa, Subsaharian Africa

Results of the SQL queries



Education index by Country
Education Index
0,99
0
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, OpenStreetMap, TomTom

Thank you for your patience

any questions ?