Our ETL processes

We decided to study the refugee population in North America with the focus on the USA and Canada and see the trends based on the country of origin. In our exercise, we focused on the countries based on the refugee population registered in destination countries.

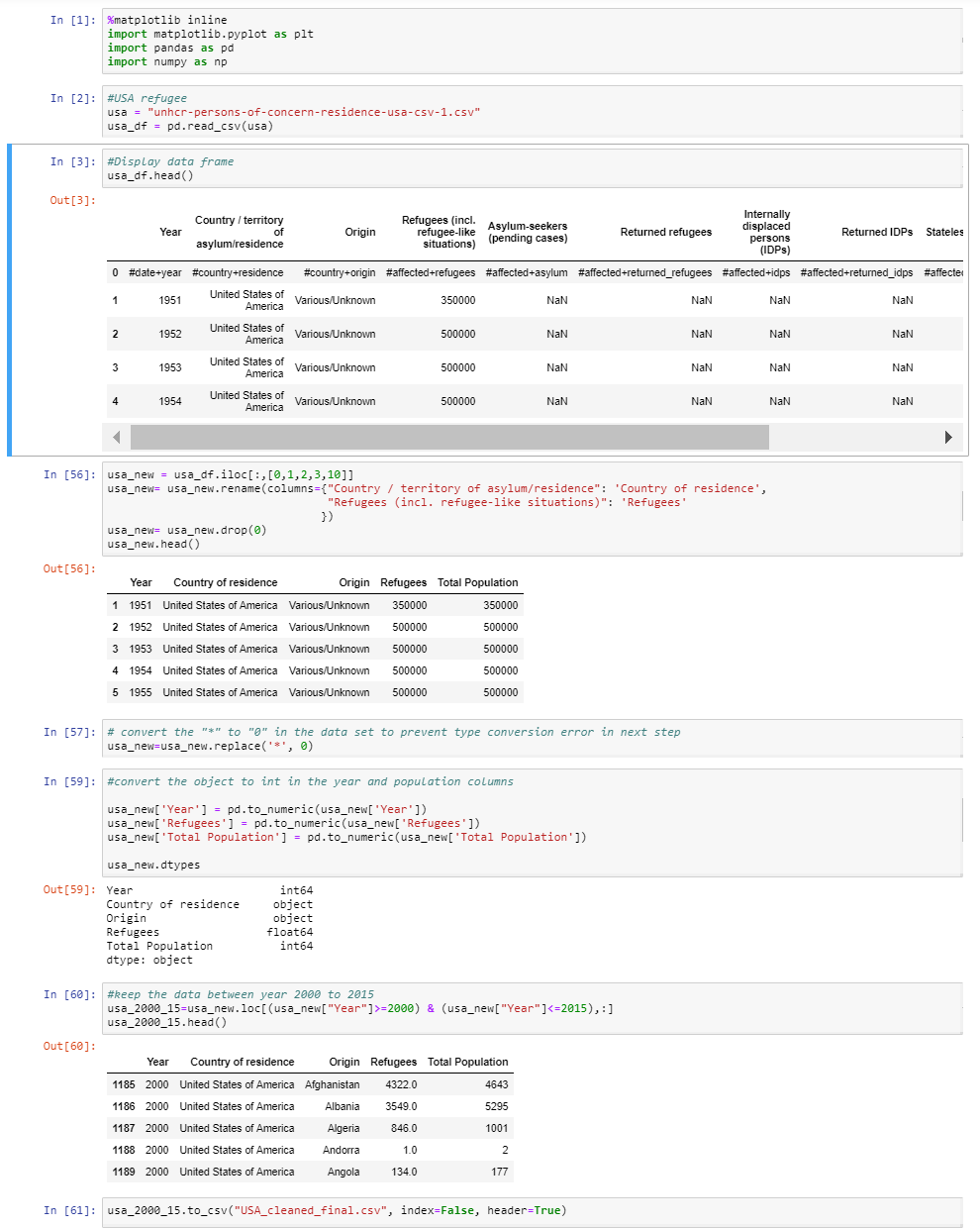
For USA and Canada, data extracted from the UN database available on data.world website. Web scraping technique was used to gather the population table for the countries of origin.

Our goal was to create a data set containing the clean and related data in a manner that we could perform investigation about the population growth or decline, forecasting etc.

1. Gather the refugee data to Canada and USA ([https://data.world](https://data.world/)):

* UNHCR's populations of concern residing in Canada
* UNHCR's populations of concern residing in the USA)

1. Scrap the population of the country of origin from the web (<https://photius.com>)
2. Data cleanup from the UN data
   1. “Python” was used to perform the data cleaning
   2. Download the UN raw data from data.world website
   3. Import the CSV file to pandas and audit the data
   4. Data cleanup
      1. Select the required columns (year, Country/territory of asylum/residence, Origin, Population) from the raw data and create a new data frame.
      2. Data were in “Object format.” To sort and calculate the population conversion was required.
      3. Some of the data field in the year and population columns were a special character “\*” and prevented the conversion.
      4. Extract the data from the year 2000 to 2015.



1. Scrap the world population for the year 2000 and 2015
   1. Pandas and BeautifulSoup were used to scrap the info.
   2. Study the webpage to find the location of the required data.
   3. Imported the entire population info.
   4. Extract the years 2000 and 2015 from the data frame.
   5. Create a secondary data frame



1. Combining the codes
   1. Data frame index was changed from digit to country of origin
   2. “Groupby” technique was used based on country of origin and year to get the aggregate data for 15 years in one column
   3. Three data frame were merged to create a single data source ready for analysis
   4. Data exported to CSV to make it accessible to other stakeholders

The final code and related CSV files are accessible at: