Jennifer Mead

9/5/19

IST 652: Scripting for data analysis

Final project: Refugees and asylum seekers to the USA

# Data sources

**Main source:**

* <https://www.kaggle.com/unitednations/refugee-data>, specifically the time series data and asylum seekers csv files.

**Other sources:**

* <https://www.unhcr.org/en-us/returnees.html>
* <https://www.dhs.gov/immigration-statistics/yearbook/2017>
* <https://en.wikipedia.org/wiki/Asylum_seeker>
* <https://en.wikipedia.org/wiki/Refugee>

# Definitions

From Wikipedia:

* “An **asylum seeker** …is a person who flees their home country, enters another country and applies for [asylum](https://en.wikipedia.org/wiki/Right_of_asylum), i.e. the right to international protection, in this other country. An asylum seeker is a type of [migrant](https://en.wikipedia.org/wiki/Immigration) and may be a [refugee](https://en.wikipedia.org/wiki/Refugee), a [displaced person](https://en.wikipedia.org/wiki/Displaced_person), but not an [economic migrant](https://en.wikipedia.org/wiki/Economic_migrant).” (Wikipedia)
* “an asylum applicant must establish that he or she fears persecution from their Government in their home country. Second, the applicant must prove that he or she would be persecuted on account of one of five protected grounds: race, religion, nationality, political opinion, or particular social group” (Wikipedia)
* Note that I found different definitions on different sites.

For the sake of this analysis, I’m **grouping people classified as refugees and asylum-seekers together**. Most of the data I had did not distinguish between the two.

# Questions to answer

* How many people are refugees/asylum-seekers to the USA? How does this differ over time?
* How many refugee/asylum-seekers to the USA were there between 2000 and 2016?
* Where are people coming from? What does this pattern look like over time?
* Which countries have the most between 2000 and 2016? Between 2013 and 2016?
* When people seek asylum/refugee status, are the rejected or recognized (approved)?
* Are there patterns between the countries of origins and the outcome of the decisions?

# Description of the program

In this program I:

* Import modules
* Import the files
* Check the data types for the columns
* Clean and transform the data to have consistent types and naming.
* Group and aggregate the data to find patterns
* Create visualizations for those patterns

Challenge 1:

The data was not granular enough for me to do predictive analysis with it. The data was already aggregated at the country level, not at the level of individuals. I also looked for other data sources, such as the Department of Homeland Security website (which tracks US immigration data), but the data there was worse. There, each interesting characteristic like gender or age were in a separate table, without any way to join them. It was unclear whether this was done this way to protect privacy or be deliberately uninformative.

Challenge 2:

The time\_series.csv file and asylum\_seekers.csv files both had several columns with mixed types that threw warnings when I imported them. Instead of the usual fillna, I had to do extra steps to replace phrases and characters, coerce errors, remove rows, bad values, etc in order to get clean data. This was good practice.

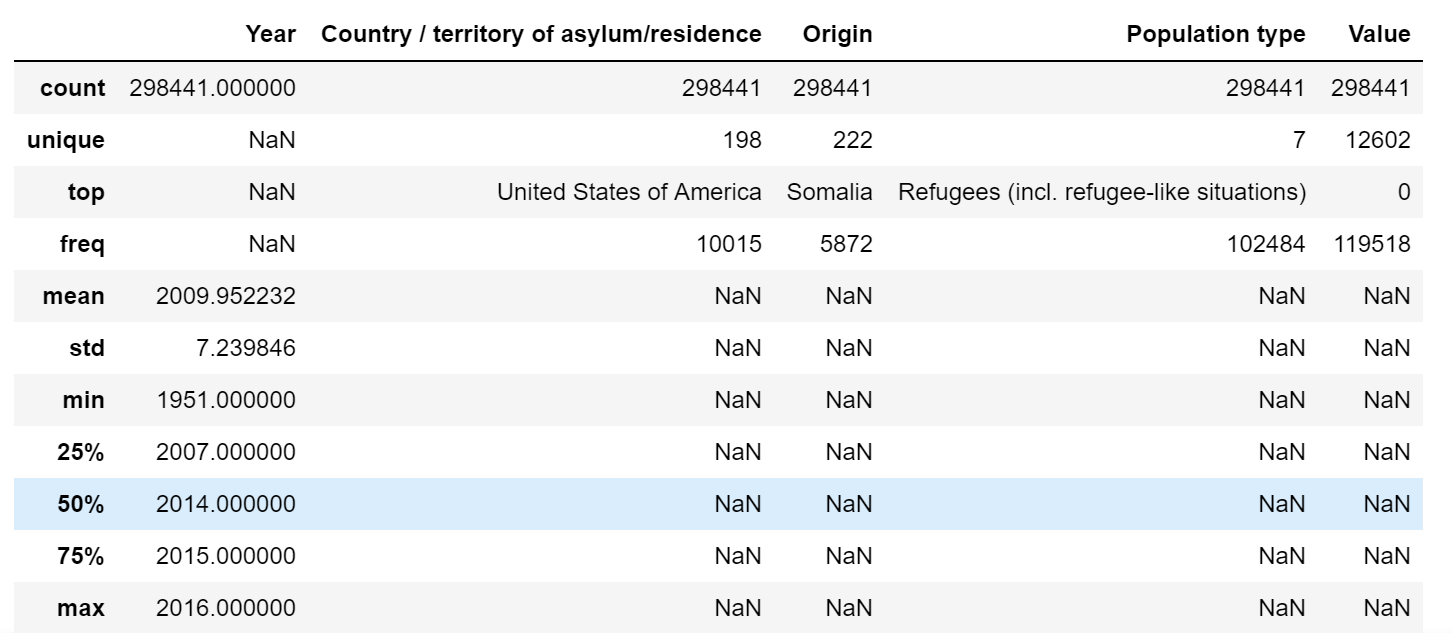
Challenge 3:

I wanted to learn some new visualizations, so I experimented with different graphs using matplotlib and seaborn. I kept the most interesting ones in the code, notably time-series and a heatmap.

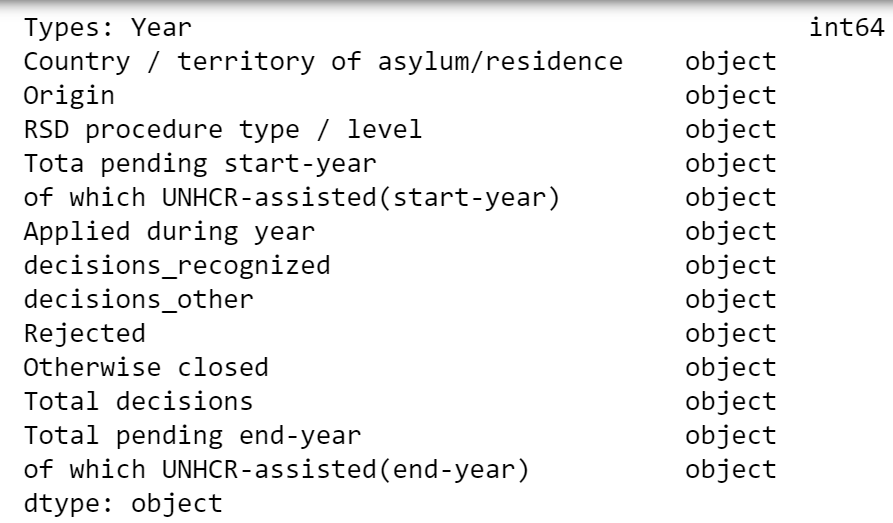
# Data pipeline and pre-processing

Import: I used the pandas function read\_csv for both files.

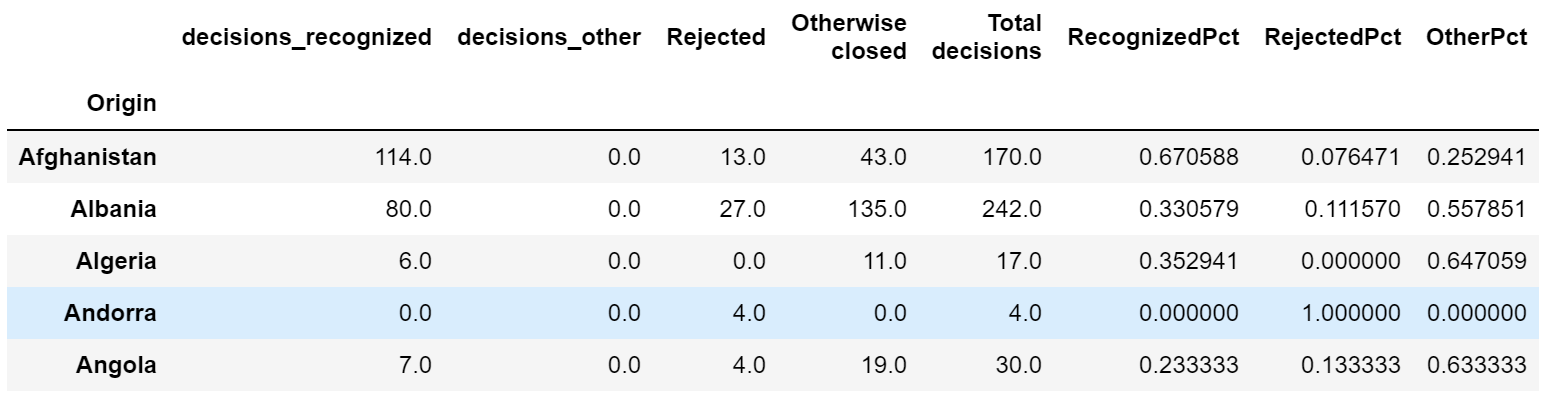
Time\_series.csv fields were:



Asylum-seekers.csv fields were:



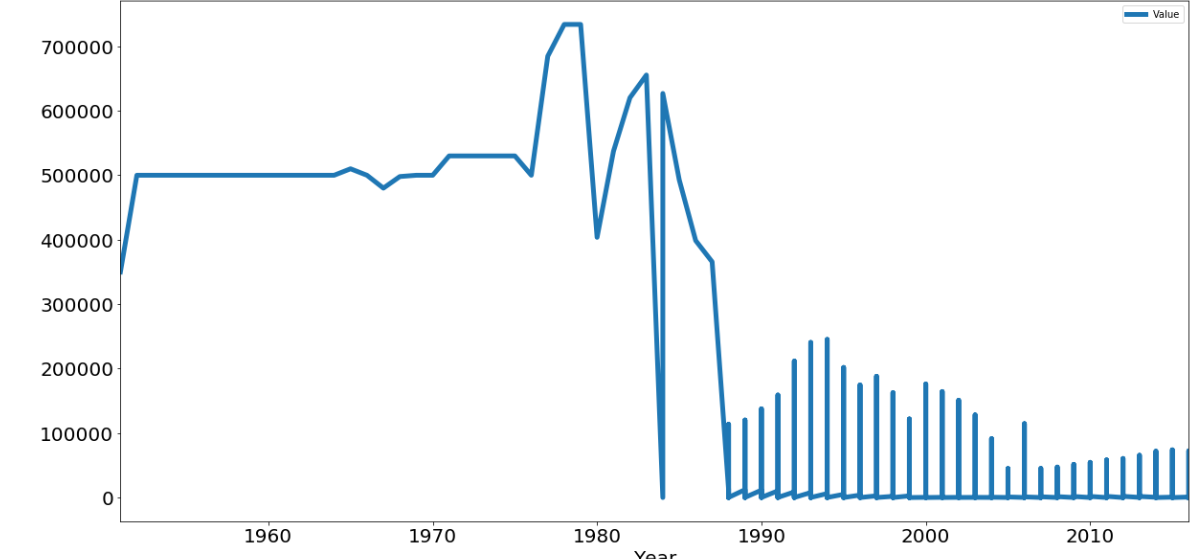
* .describe(include = all) to see which columns had a lot of NaNs and what the range of values were
* .info to see the types and number of values
* .unique to see the possible values for each column. This helped identify mixed types
* Subsetting to choose the USA data and different date ranges to eliminate ranges with inconsistent data
* .astype to change the type
* .dtypes to check the types
* .columns to change the names of the columns
* .replace to shorten values (for example ‘United States of America’ to ‘USA’) and get rid of random characters (‘\*’)
* .fillna to replace nulls with zeros
* For asylum-seekers I also calculated fields with percentages for the different decision outcomes



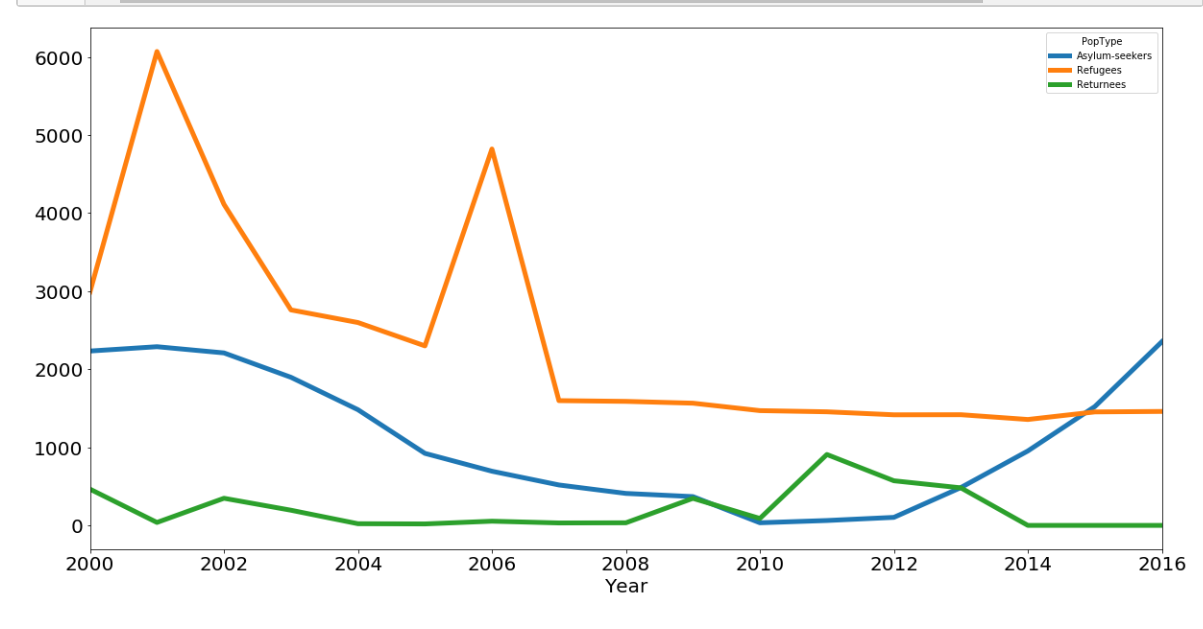
# Conclusions

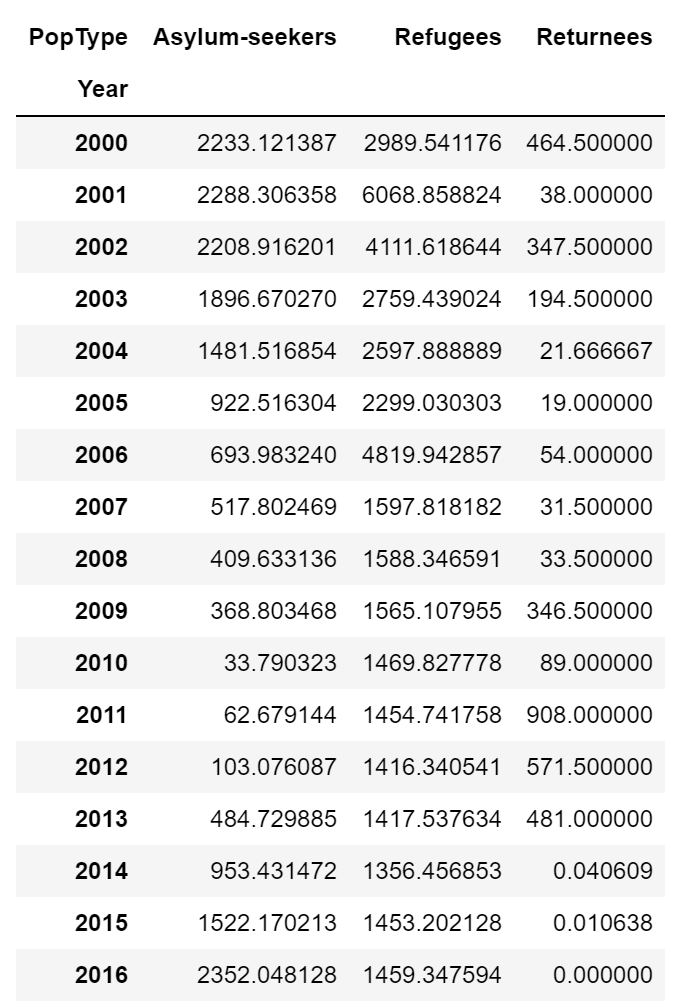
## How many people are refugees/asylum-seekers to the USA? How does this differ over time?

Note that it looks like the reporting methods changed around 1985.



## How many refugee/asylum-seekers to the USA were there between 2000 and 2016?

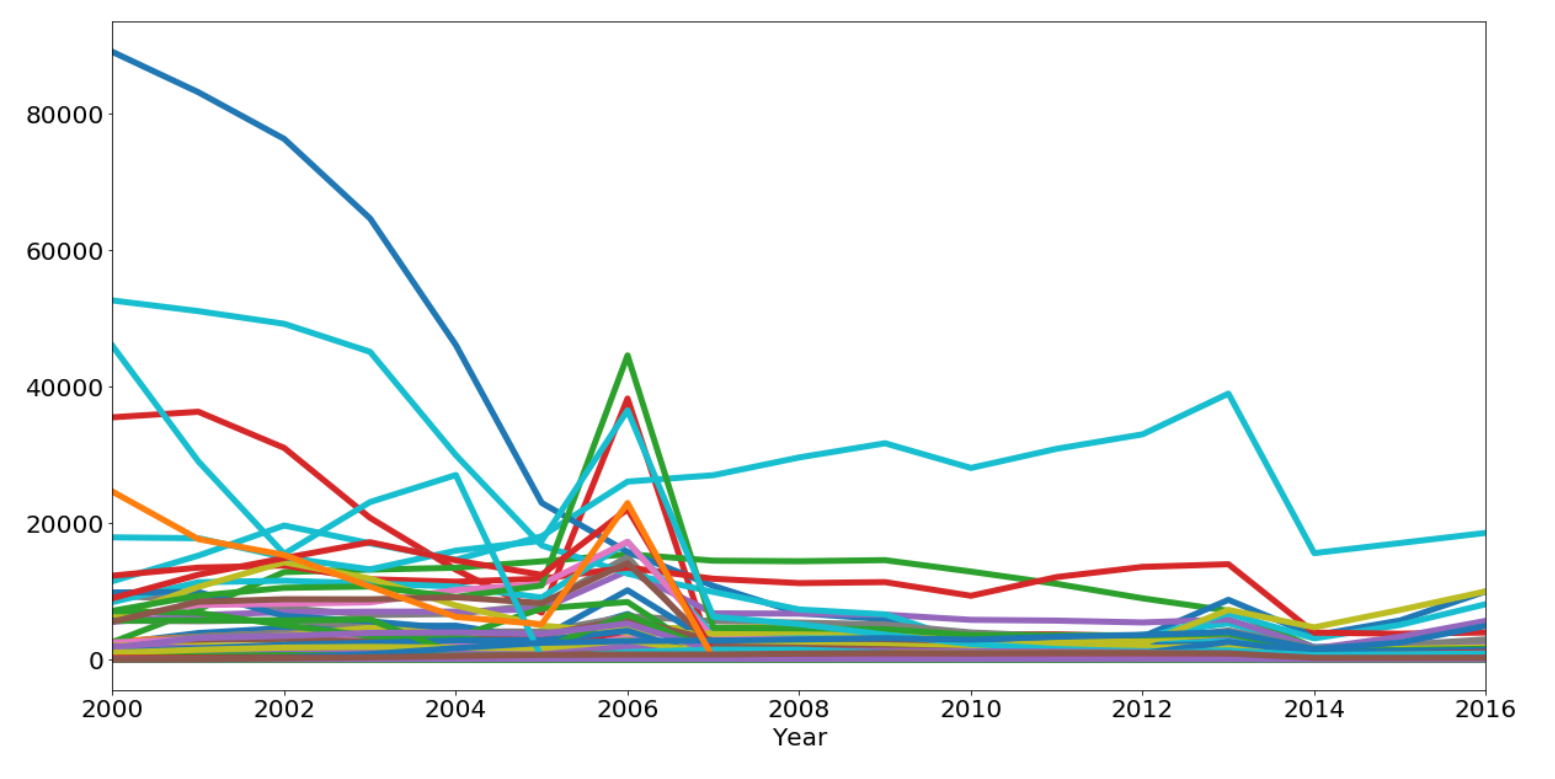




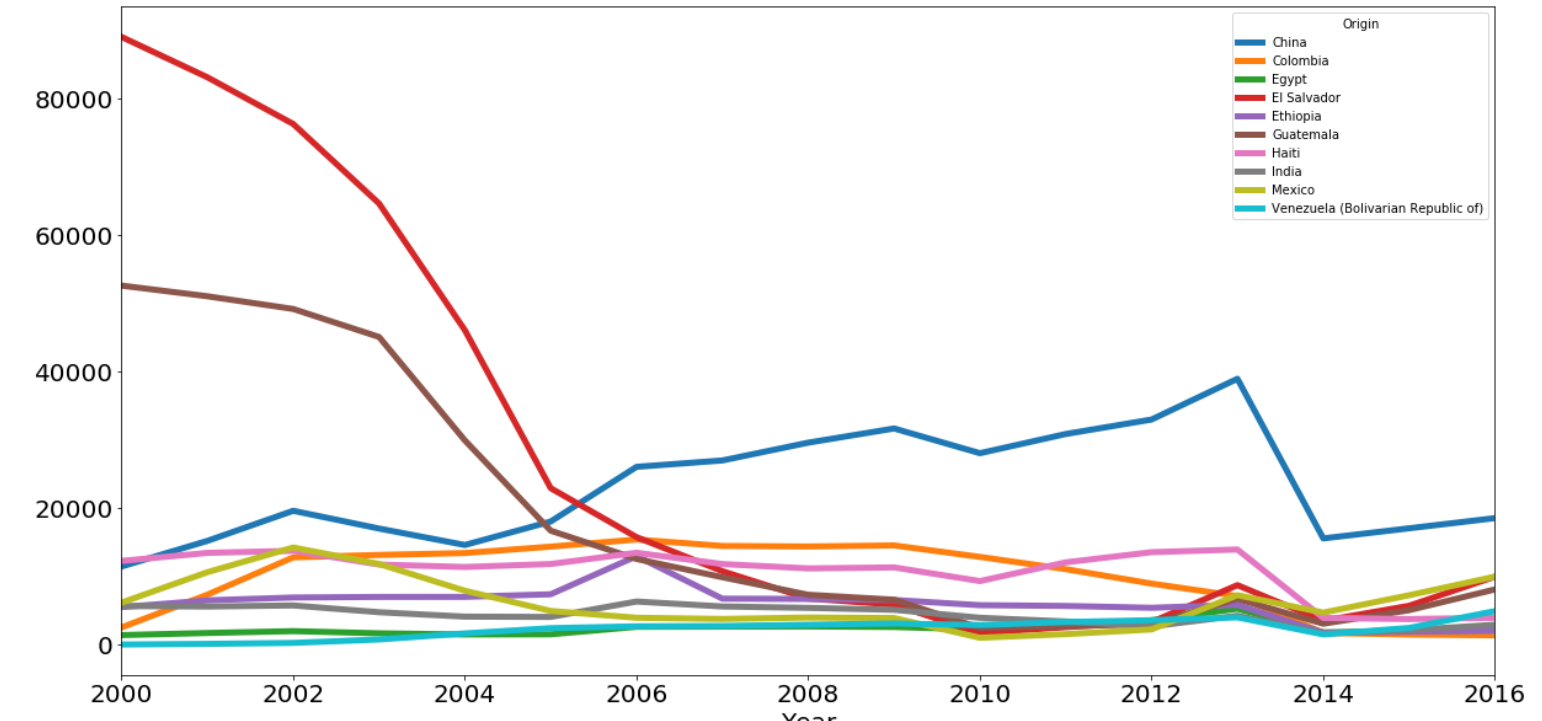
## Where are people coming from? What does this pattern look like over time?

Note that the numbers start off much higher and that there are spikes in 2006 and 2013.

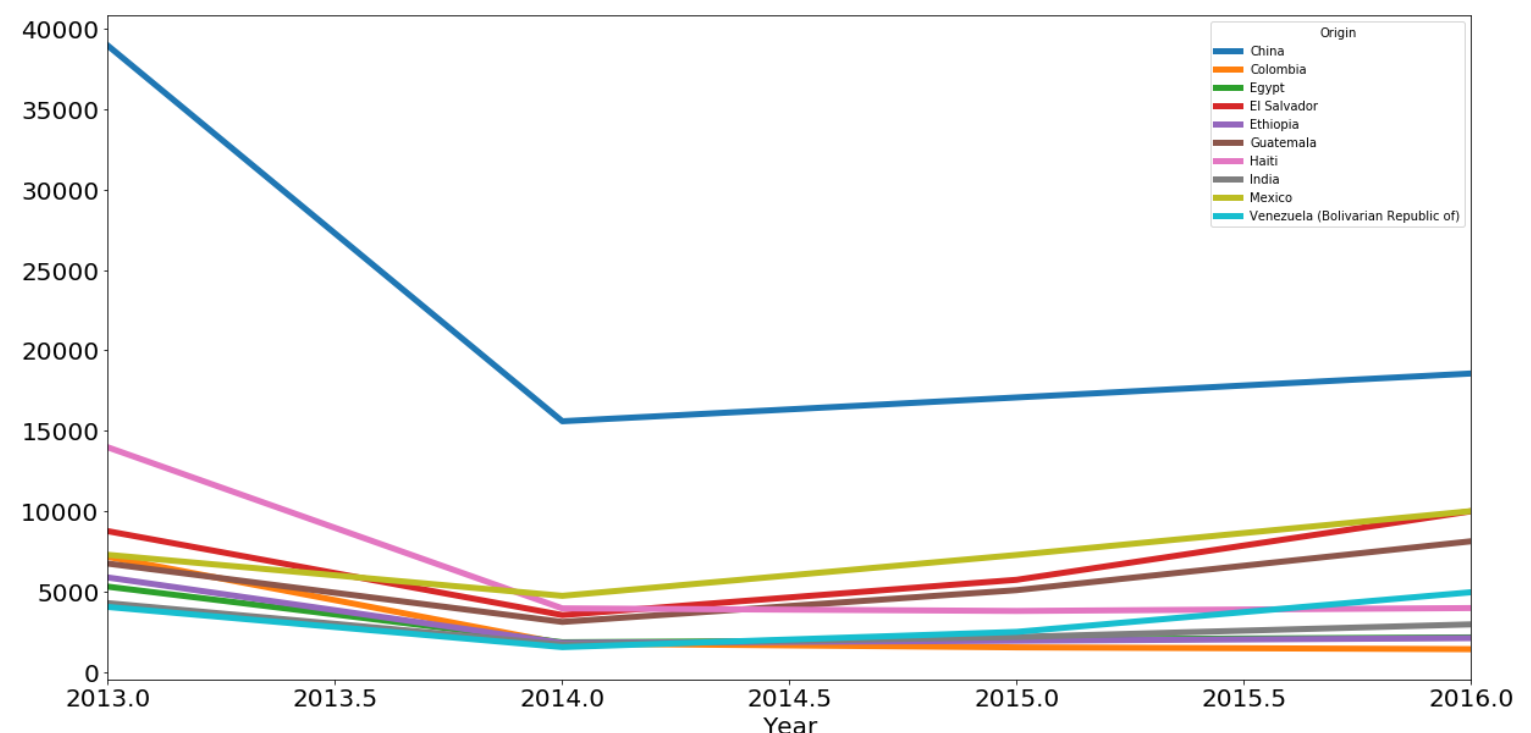
Most countries have nearly no refugees, so I’ll next drill down to the top 10 in recent years.



Which countries have the most between 2000 and 2016? 



## Between 2013 and 2016?



## When people seek asylum/refugee status, are they rejected or recognized (approved)?

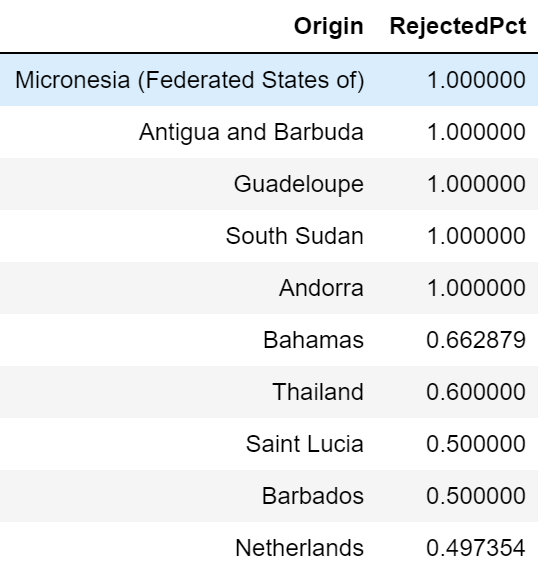
**Top ten countries for recognitions**

This list includes Egypt from the top ten



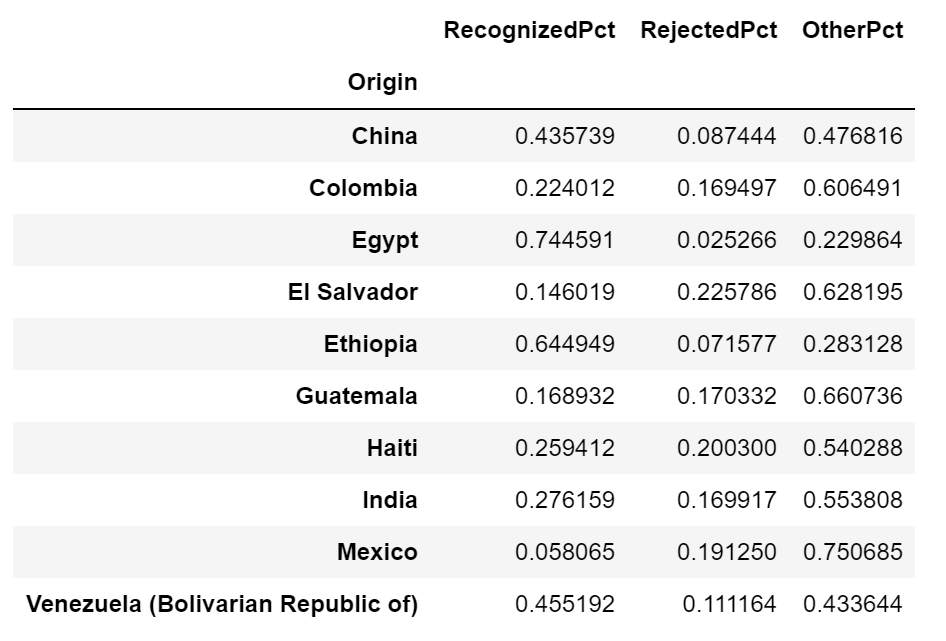
**Top ten countries for rejections**

This does not include any countries from our top ten



## Are there patterns between the countries of origins and the outcome of the decisions?

It’s hard to see the patterns until I add the heat map.



* There seems to be a difference between countries like Ethiopia and Egypt and countries like Mexico, El Salvador, Guatemala, and Colombia.
* Few rejections are given, but a lot of "other" are.
* I couldn't find clear defintions of "other".
* Because the data was old and not very granular, the next step would be trying to find more recent and more detailed data to drill down further in analysis.

