Dataset Report:

Canada Immigration Stats 1980 - 2013

03/21/2022
Data Analytics + Visualizations

Shay Hegarty
Kevin Meehan

TABLE OF CONTENTS:

INTRODUCTION

The dataset I have chosen to analyze is a related to Canadas's immigration statistics from 1980 up until 2013, this is the current most up to date version of this dataset. The dataset goes into detail on how many immigrants are entering Canada each year, what country they're from and other demographic information such as region, type, coverage, area and whether or not they are coming for a developed region or developing region. This version of the dataset I am working with was released in 2015 from the United Nations Populated Divisions, Department of Economic and Social Affairs and it presents data pertaining to 45 countries that relate to Canada immigration specifically. My reason for choosing this dataset is that I am planning a move over to Canada later this year on a working holiday visa so diving into this dataset to learn more about their immigration figures is very appealing to me.

Canada seems to on the constant rise for immigrants each year and they have so many different routes you can go down to gain entry to the country such as a student visa, working professional, express entry etc. I wanted to dive into this dataset to try and establish some trends within the data, what are the most common countries people are coming from? Are there certain years where the counts are very low or very high, can we determine the reason for this from the data alone?

As mentioned above the dataset I am using was retrieved from the United Nations Department of Economic and Social Affairs website, the dataset is free to use and is open to the public to download and review, I have referenced the license agreement[1]. The dataset was downloaded as a CSV (comma separated values) file. This makes it a lot easier to work with as I can open the file at the start to get a feel for how its presented. I also have the option to start to clean the data in excel or I can import it directly into my Python notebook using Pandas library to read the csv and then start the data cleaning process.

If the file, I was using was an excel file I would install the openpyxl (xlrd) which pandas requires in order to read excel files.

WORKING WITH THE DATA

The libraries I will be using to help with the data cleaning and data analysis will be mainly pandas and NumPy. Pandas is used for data manipulation, data analysis and to calculate statistics and most importantly it will help you answer questions you have about your data and then once the date is cleaned and visualized it has the capabilities to transform the worked dataframe back into a csv or another desired format[2]. NumPy stands for numerical Python and consists of multidimensional array objects and routines for processing these. NumPy is extremely useful when it comes to mathematical and logical operations[3].

The original dataset contains two tabs, one is broken down by each specific country and lots of other information and another tab which lists immigration stats for the area and continent. The only tab we will be working with is the 'Canada by Citizenship' tab as this lists out all the specific country information and also contains the area/continent. Since I went ahead and opened the CSV file before I imported it into Python, I learned some valuable information. The first 20 rows of data will need to be skipped as it pertains to information about the dataset itself, the first header that I need starts at row 21. There is also a total row at the bottom which I will not include when importing as I will be calculating any necessary totals manually within Python. To do thing when importing I will be using 'skiprows' and then include my range and 'skipfooter' to not include the bottom rows.

When it comes to cleaning the data there are a number of columns that I will be dropping as I feel its duplicated information. The dataset lists every country the immigrant is from but also includes an area number, region name, region number and a developer region number. All these columns can be dropped as I will mainly be focusing on the Countries and the region the immigrants are from. There are also two columns relating to the immigrant's status, these are type and coverage. These will also be dropped as all the figures we are working with are all already related to immigrants, so we don't need a column confirming that status.

Once I get all the columns that I no longer require dropped I can run a head command to get a look at our dataframe now. I will need to update a few of the column's names so it's easier to understand, for example the column that includes all the countries is currently named 'OdName' so we will need to rename this column along with a few others. To do this I will use the datframe.rename function. Once I have my columns renamed, I can run another head query to verify it's all correct. The next task I want to complete is updating the index to 'Country' rather than having it a number. This will make it a lot easier to search for a specific country's stats using the .loc method. To do this I will use the function dataframe.set_index and specify country and with inplace is equal to true to overwrite the current dataframe.

Another task I want to do is to add in a Total column. The original dataset had a total row at the bottom, but I feel it will be easier to work with having a total column. In order to create a new column labeled 'Total' I will use the following function, dataframe['Total'] = dataframe.sum(axis=1). This will create our new column and also add up all the figures in each row to give us the actual total which will be very useful when visualizing the data.

Once I get to this stage the data is in a good shape to start using some basic visualizations to get a quick view of the data and test my hypothesis on any questions I had.

VISUALIZING THE DATA

The main library I will be using in Python to help visualize the data will be Matplotlib. Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python, it makes visualization extremely easy and helps make the audience understand the data they are looking at much better[4].

The opportunities for visualization when it comes to my dataset are endless. I have a time frame along with consistent yearly figures so I can create simple line graphs, or bar charts showing the top countries immigrants are leaving, a map showing the highest number of countries immigrants are coming from etc.

Another library I will be incorporating is Folium. Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of leaflet.js library[5]. Folium is extremely useful when it comes to geographical data which my dataset ties into perfectly. One of my goals for this dataset is to incorporate a Choropleth map using Folium to represent the countries with the highest number of immigrants into Canada. This map will be extremely useful when presenting my findings to an audience as it almost incorporates every piece of info from the dataset itself and will give a great representation of the dataset.

I also plan on creating a few visualizations through a data visualization tool such as Tableau or PowerBI. I will be trying to create most of my visualizations through Python itself but if some of the visualizations I have in mind aren't possible in Python I will try and demonstrate them through a data visualization tool. An interactive visualization will be very useful when exploring this type of data, I can create the graphs themselves and add filters or slicers to then break the data down extremely easily.

RESULTS

Success will be measured in various stages throughout the project. The first goal was to find an appropriate dataset that was relevant to me and had a genuine interest into diving deeper into and reviewing. By choosing this Canada immigration stats dataset I feel as I have already succeeded in that area. The next steps will be measured by importing the data into Python successfully and then subsequently cleaning the data which in my case mainly involves removing unnecessary columns, checking for duplicate data or corrupted rows/columns. Once all that has been done and the dataframe in Python is looking clean and ready for the next stage I will start to create my visualizations. Starting with basic ones such as a line graph over a time period into a bar chart, then into some map graphs using Folium and finally if appropriate create a few interactive visualizations with Tableau or PowerBI if necessary.

Some of the questions I would like to walk away from this project having answered:

- *What are the top countries immigrants are coming from?*
- *Are there certain years where immigration numbers are very high?*
- *Similarly, are there certain years where immigration numbers are very low?*
- *Have immigration numbers decreased or increased over the years?*
- *Are there any noticeable spikes with the numbers, possibly caused by a natural disaster in a country and immigration numbers go up the following year?*
- *Establish any trends within the data relating to where immigrants are coming from*
- *Another question I would like to answer is how helpful a tool such as tableau will be compared to using visualizations solely in Python*

My expected results for this dataset is that we are going to see very large consistent numbers from countries with some of the highest populations, these could include China, India, and possibly the United States and United Kingdom could creep into that list. I would expect that immigration numbers are consistently rising every year for every country. For some countries who are more prone to natural disasters I would assume the immigration rates for these countries would spike around those times and then start to settle down again, but this will be confirmed once we start exploring the data.

REFERENCES

[1] 'United Nations Population Division | Department of Economic and Social Affairs'. Accessed 21 March 2022. https://www.un.org/en/development/desa/population/migration/data/empirical2/migrationflows.asp.

[2] ' Python Pandas Tutorial: A Complete Introduction for Beginners'. Accessed 21 March 2022. https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/.

[3] Team, Great Learning. 'What Is Numpy in Python | Python Numpy Tutorial'. GreatLearning Blog: Free Resources what Matters to shape your Career!, 11 January 2022. https://www.mygreatlearning.com/blog/python-numpy-tutorial/.

[4] 'Matplotlib — Visualization with Python'. Accessed 21 March 2022. https://matplotlib.org/.

[5] 'Folium — Folium 0.12.1 Documentation'. Accessed 21 March 2022. https://python-visualization.github.io/folium/.