Data Visualization with Python

Avery Jan

10-30-2022

Outline

Introduction
Methodology
Datasets
Parts of the Project

- Part 1_Dataset and Line Plots
- Part 2_Area_Histogram_Bar
- Part 3_Pie_Box_Subplot_Scatter_Bubble
- Part 4_WaffleChart_RegressionPlot
- Part 5_Complex Subplots
- Part 6_Maps

Introduction

In this project, a variety of visualizations were created with python and the visualizations were organized into six parts. Most, but not all, visualizations are included in this PDF file. The data used for this project came from three sources: (1) Real-world datasets (2) Data generated within the code (3) A dataset from a scientific study. Among these datasets, the two real-world datasets were read into Pandas DataFrames and underwent data preparation steps wherever necessary to ready them for creating visualizations. The data from the other two sources do not require any data preparation and were used directly without being saved into a dataframe. Several Python libraries were used throughout this project. Matplotlib, Seaborn, and Folium were used to create visualizations, Pandas was used to store data, and NumPy was used to generate and analyze data to be used for creating visualizations. The majority of the project was completed on IBM Skills Network and the remaining work was done on Codio hosted by Cornell University.

Methodology

Data Sources: Noted alongside the charts

Platforms: Codio, IBM Skills Network

Python Libraries:

Matplotlib, Seaborn (charts)
Folium (maps),
Numpy, Pandas (data)

Data Preparation:

1. Remove unwanted columns.
2. Rename columns to more meaningful names.
3. Convert column labels to string to avoid confusion introduced by having an integer as the column name.
4. Add a new column “Total” to hold the sum of data of all years.
5. Set the country name as the index to facilitate the selection of data by country.
6. Create a list of the names of year columns in type 'string' to be used for creating visualizations.

Datasets

Real World Datasets

(1) Canadian Immigration Dataset (used in Parts 1 - 4 and Part 6)

- This is a subset of “International migration flows to and from selected countries”, which contains annual data on the flows of international immigrants as recorded by the countries of destination. The data presents both inflows and outflows according to the place of birth, citizenship or place of previous / next residence both for foreigners and nationals pertaining to 45 countries. This Canadian Immigration Dataset is the subset with the destination being Canada.

(2) San Francisco Crimes Dataset (used in Part 5)

- A dataset regarding the records of crimes in San Francisco in 2016.

Dataset Generated within Code (generated using Python NumPy Library)

(1) The route of a fictitious trail, the JB trail, Part 5_A
(2) The cross section of a plate with various levels of stress exerted on it, Part 5_B

Scientific Study Dataset (stress_strain.csv, used in Part 5_B)

- A dataset that includes the data from a study of the stress and strain of plates

Part 1_Dataset and Line Plots

Examine the Canada Immigration dataset.
Check the dataset for missing value.
Prepare the data for plotting visualizations
Get statistical summary of the cleaned dataset.
Create Line Plots.

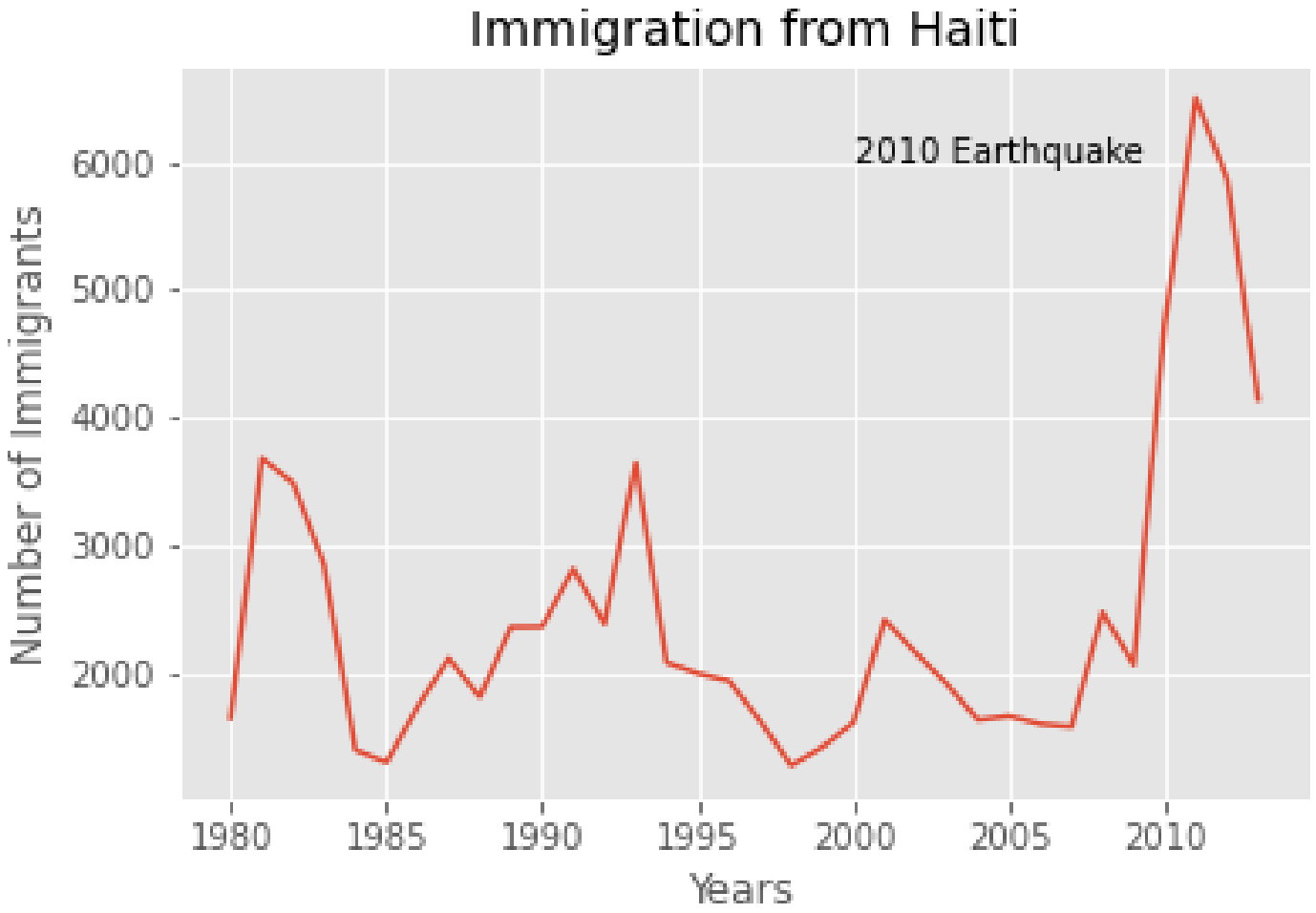
Canada Dataset (Original)

	Type	Coverage	OdName	AREA	AreaName	REG	RegName	DEV	DevName	1980	...	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
0	Immigrants	Foreigners	Afghanistan	935	Asia	5501	Southern Asia	902	Developing regions	16	...	2978	3436	3009	2652	2111	1746	1758	2203	2635	2004
1	Immigrants	Foreigners	Albania	908	Europe	925	Southern Europe	901	Developed regions	1	...	1450	1223	856	702	560	716	561	539	620	603
2	Immigrants	Foreigners	Algeria	903	Africa	912	Northern Africa	902	Developing regions	80	...	3616	3626	4807	3623	4005	5393	4752	4325	3774	4331
3	Immigrants	Foreigners	American Samoa	909	Oceania	957	Polynesia	902	Developing regions	0	...	0	0	1	0	0	0	0	0	0	0
4	Immigrants	Foreigners	Andorra	908	Europe	925	Southern Europe	901	Developed regions	0	...	0	0	1	1	0	0	0	0	1	1

Canada Dataset (Cleaned)

Country	Continent	Region	DevName	1980	1981	1982	1983	1984	1985	1986	...	2005	2006	2007	2008	2009	2010	2011	2012	2013	Total
Afghanistan	Asia	Southern Asia	Developing regions	16	39	39	47	71	340	496	...	3436	3009	2652	2111	1746	1758	2203	2635	2004	58639
Albania	Europe	Southern Europe	Developed regions	1	0	0	0	0	0	1	...	1223	856	702	560	716	561	539	620	603	15699
Algeria	Africa	Northern Africa	Developing regions	80	67	71	69	63	44	69	...	3626	4807	3623	4005	5393	4752	4325	3774	4331	69439

Line Chart



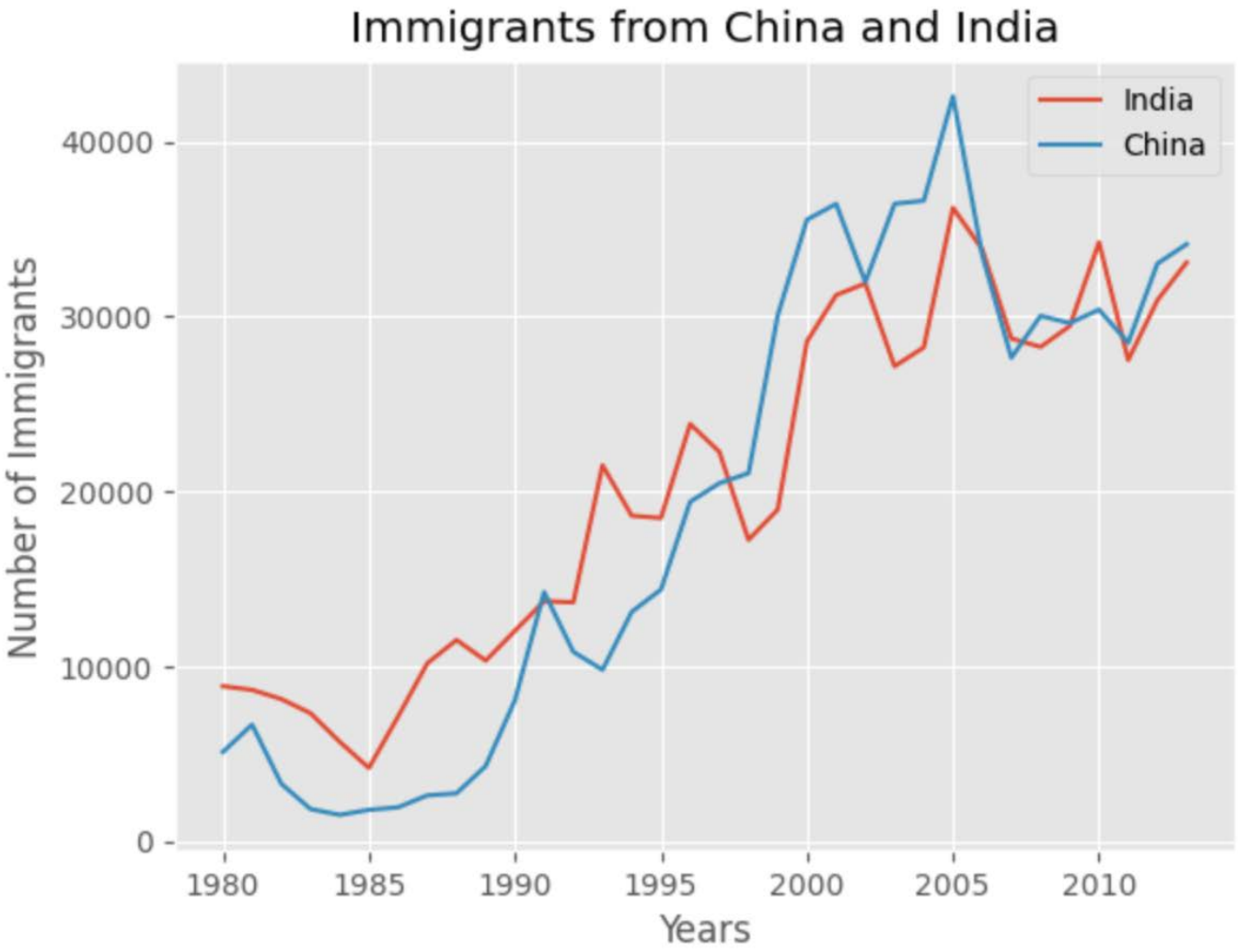
File: Part 1

Library: Matplotlib

Original Data Source:
All Countries
For reference only
International migration flows to and from selected countries: The 2015 revision

Canada subset
(used in this project)
International migration flows to and Canada

Line Chart



File: Part 1

Library: Matplotlib

Original Data Source:
All Countries
For reference only
International migration flows to and from selected countries: The 2015 revision

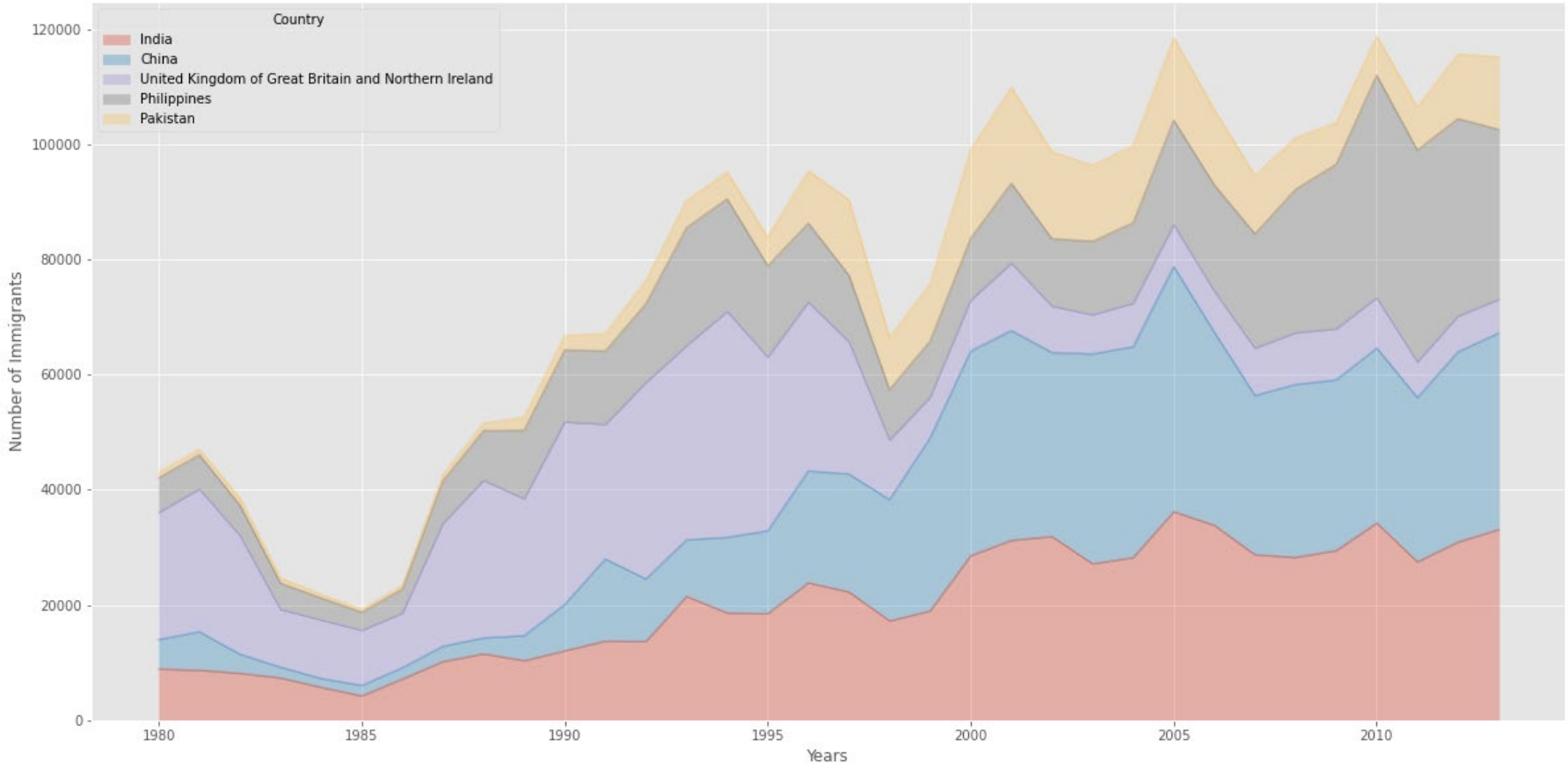
Canada subset
(used in this project)
International migration flows to and Canada

Part 2_Area_Histogram_Bar

Prepare the data for plotting visualizations .
Create stacked and unstacked Area Plots.
Create stacked and unstacked Histograms.
Create a Horizontal Bar Chart and a Vertical Bar Chart.

Area Chart (stacked)

Immigration Trend of Top 5 Countries



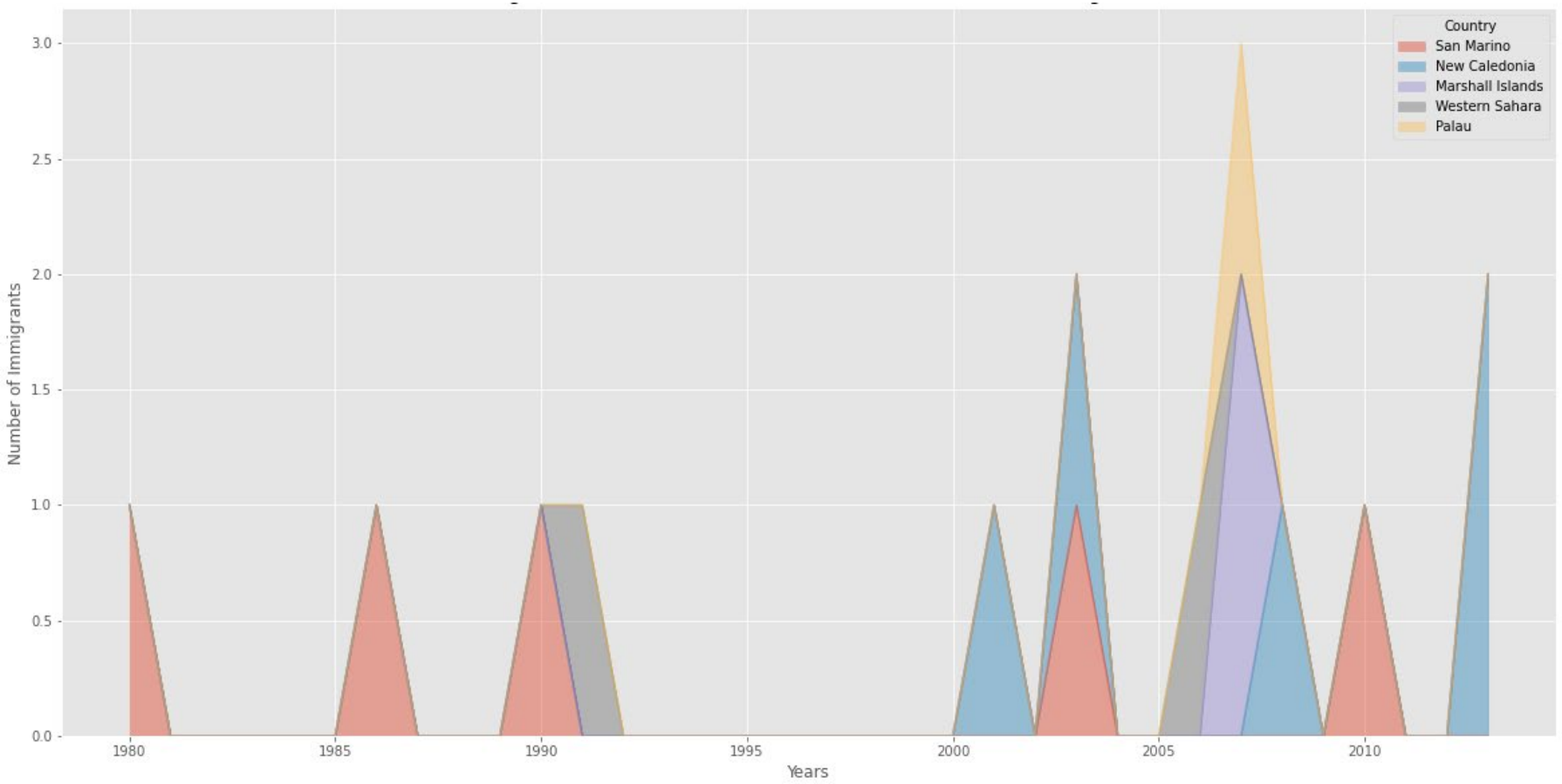
Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

File: Part 2

Area Chart (stacked)

Immigration Trend of 5 Countries with Least Contribution to Immigration



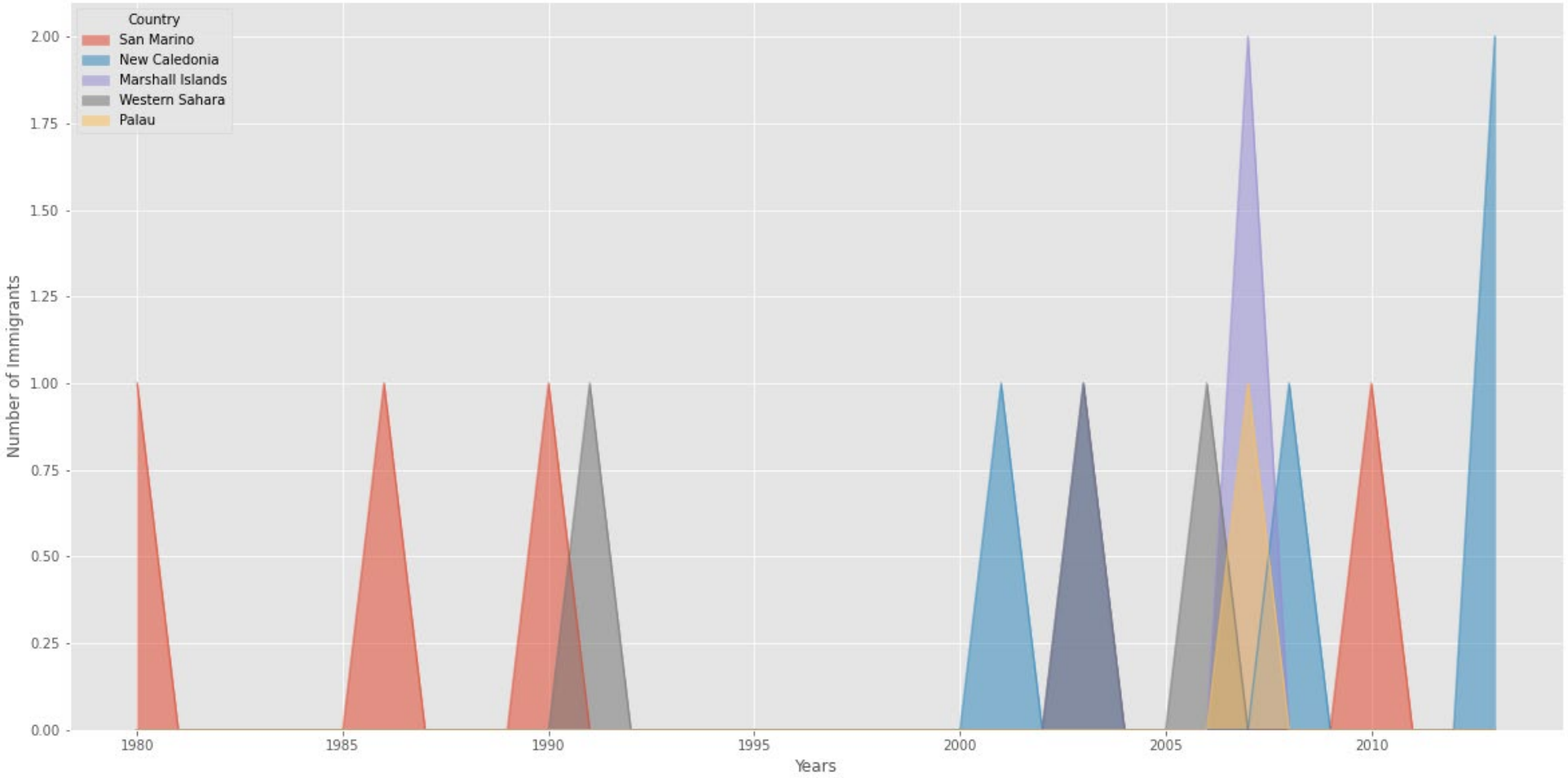
Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

File: Part 2

Area Chart (unstacked)

Immigration Trend of 5 Countries with Least Contribution to Immigration

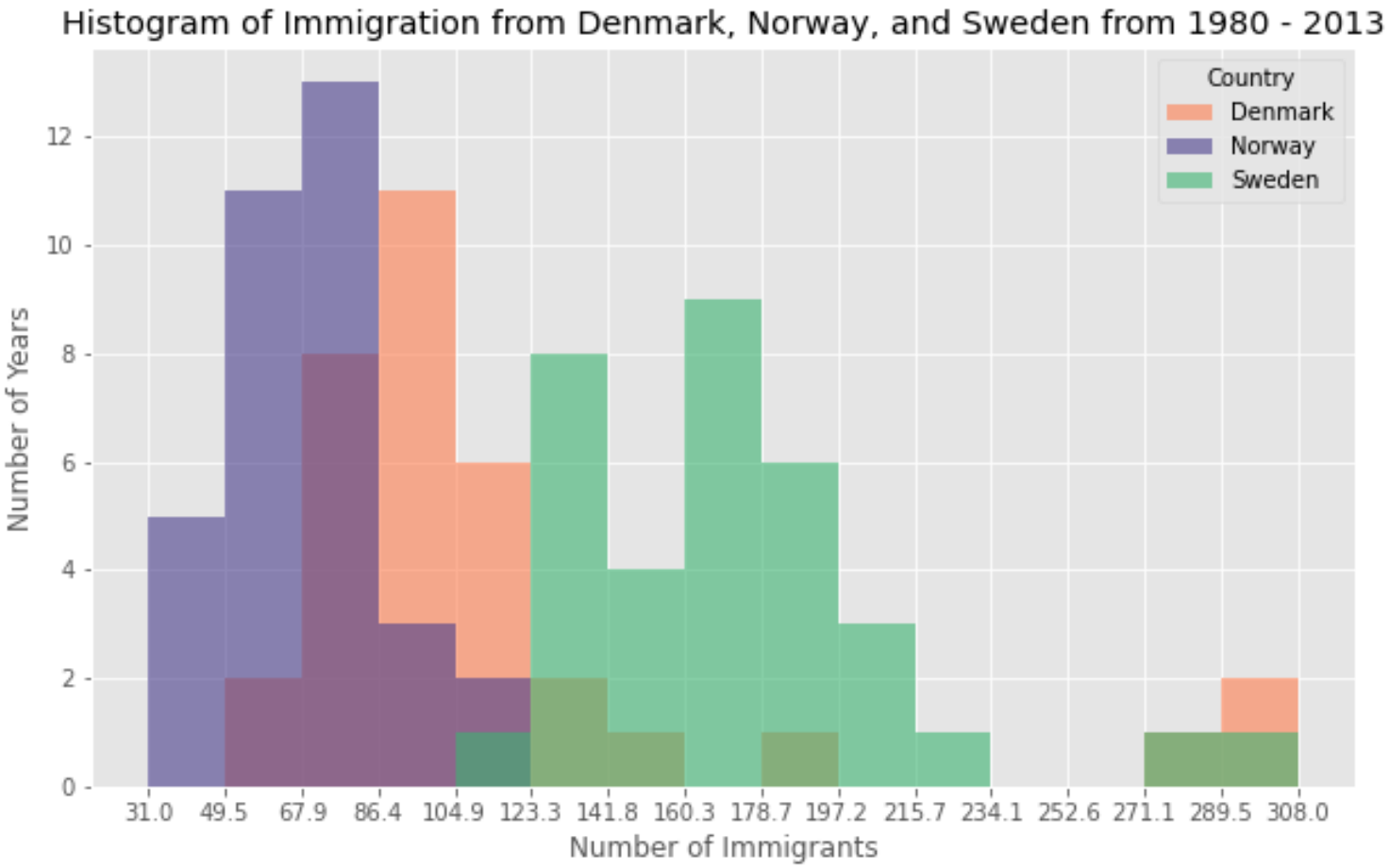


Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

File: Part 2

Histogram (unstacked)

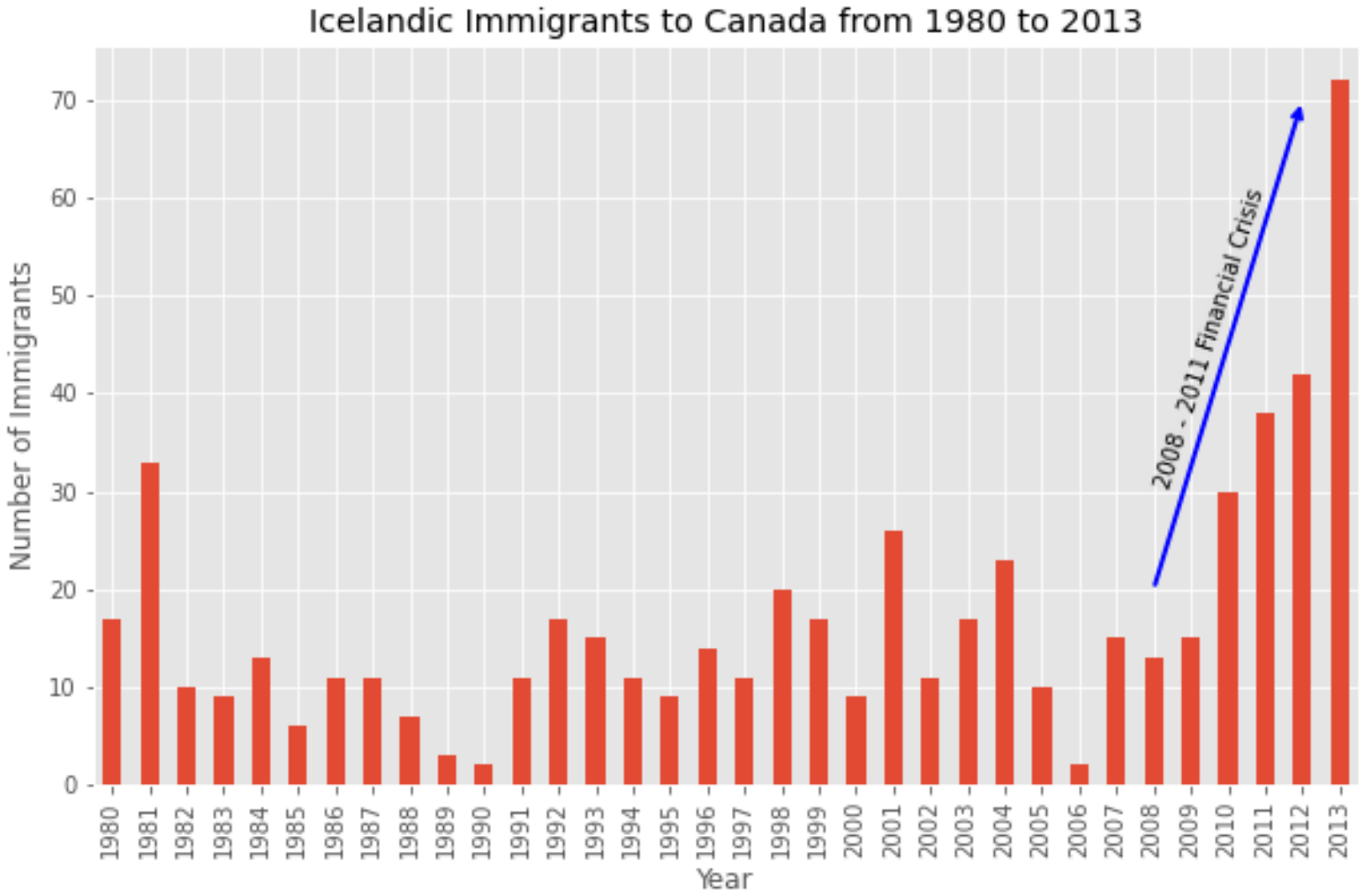


File: Part 2

Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

Column Chart (Vertical Bar Chart)



File: Part 2

Library: Matplotlib

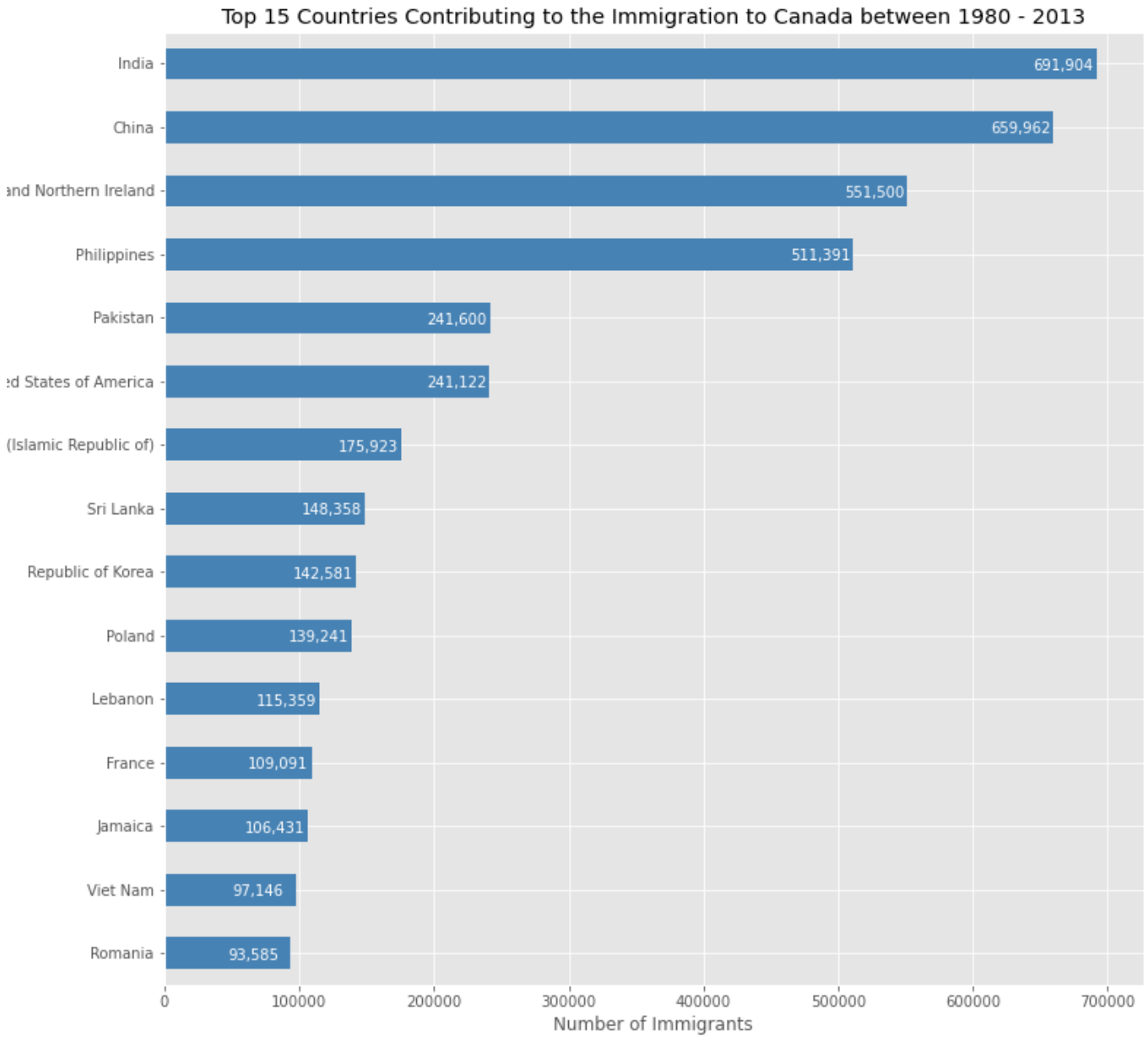
Data Source: Canada dataset
International migration flows to Canada

Bar Chart

File: Part 2

Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada



Discussion Part 1 and Part 2

Data

- The purposes of storing a large real-world dataset in a *Pandas DataFrame* is to take advantage of the functions in Pandas DataFrame to facilitate the data preparation, data analysis, and data visualization.

Charts

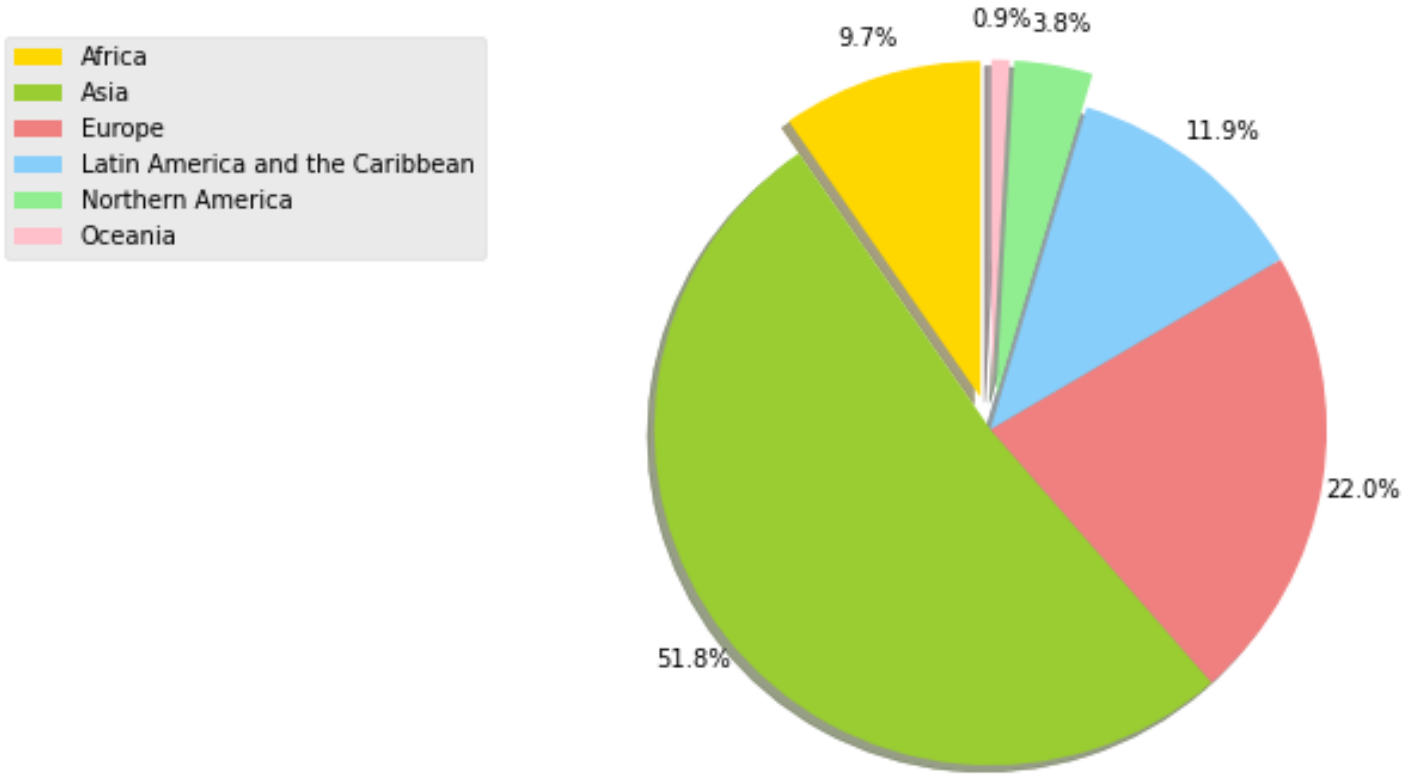
- ***Line chart*** and ***Area Chart*** provide a quick way to see the trend of data and to correlate events with the data. For example, after the 2010 earthquake in Haiti, there is a spike of Haitians immigrated to Canada.
- ***Histogram*** is commonly used to show the shape and spread of the data like how the number of immigrants from each of the three Scandinavia countries to Canada from 1980 to 2013 are spread out.
- The advantage of ***Bar Chart*** and ***Column Chart*** over other chart types is that the human eye has evolved a refined ability to compare the length of objects as opposed to angle or area. These two types of charts are particularly useful in analyzing time series data.

Part 3_Pie_Box_Subplot_Scatter_Bubble

Prepare the data for plotting visualizations.
Create a Pie Chart.
Create Box Plots.
Create simple Subplots. (Each subplot having a pair of the same type of plots.)
Create a Scatter Plot with the best fit line.
Create a Bubble Plot.

Pie Chart

Immigration to Canada Grouped by Continent [1980 - 2013]

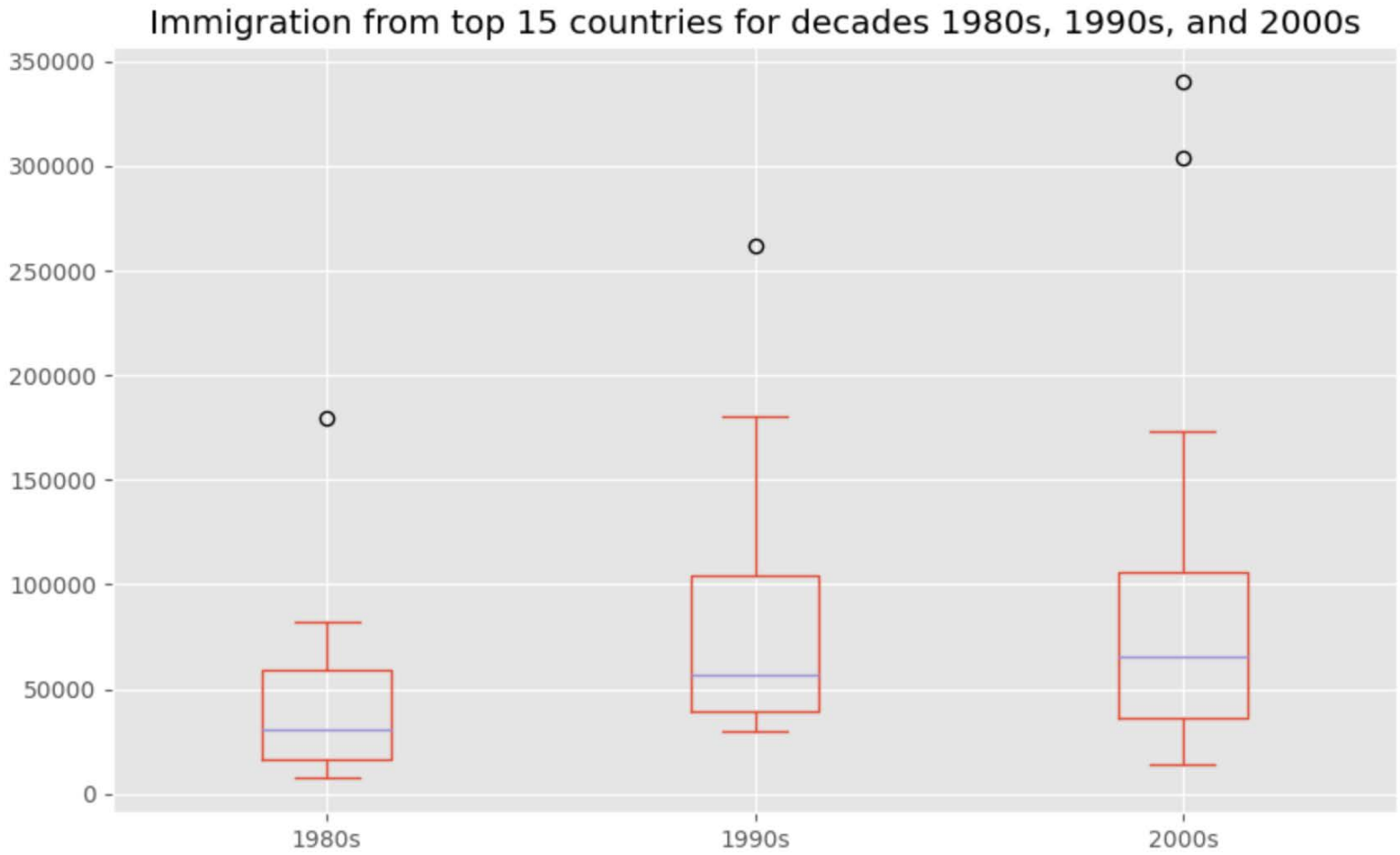


File: Part 3

Library: Matplotlib

Data Source: Canada dataset
International migration flows to and Canada

Box Plot

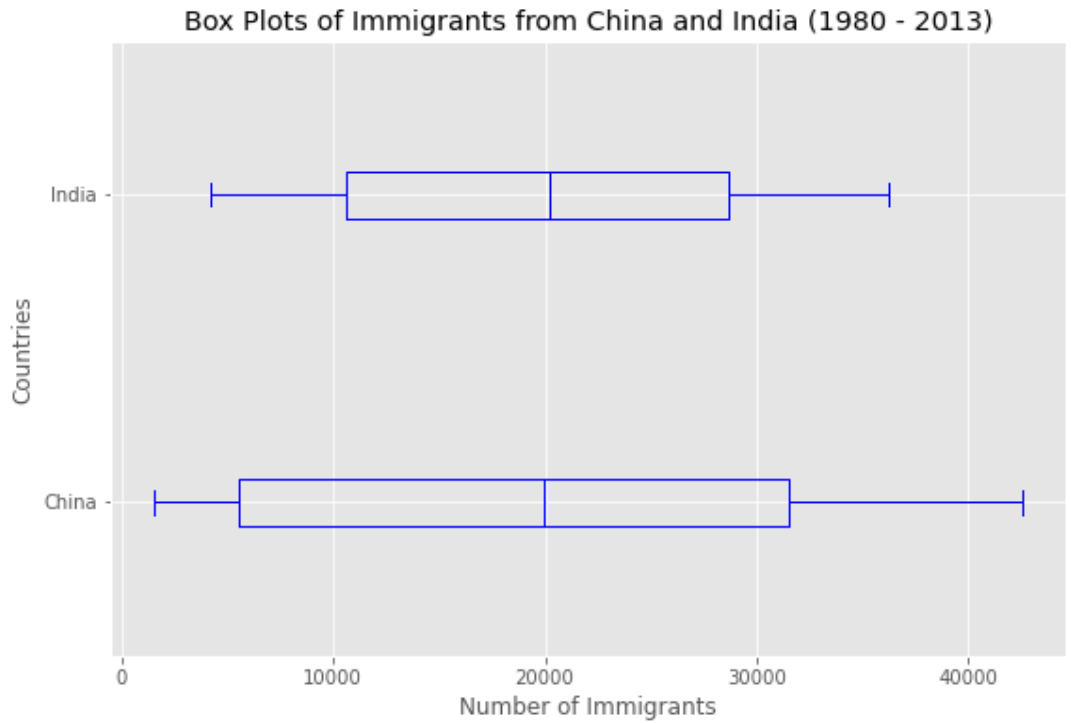


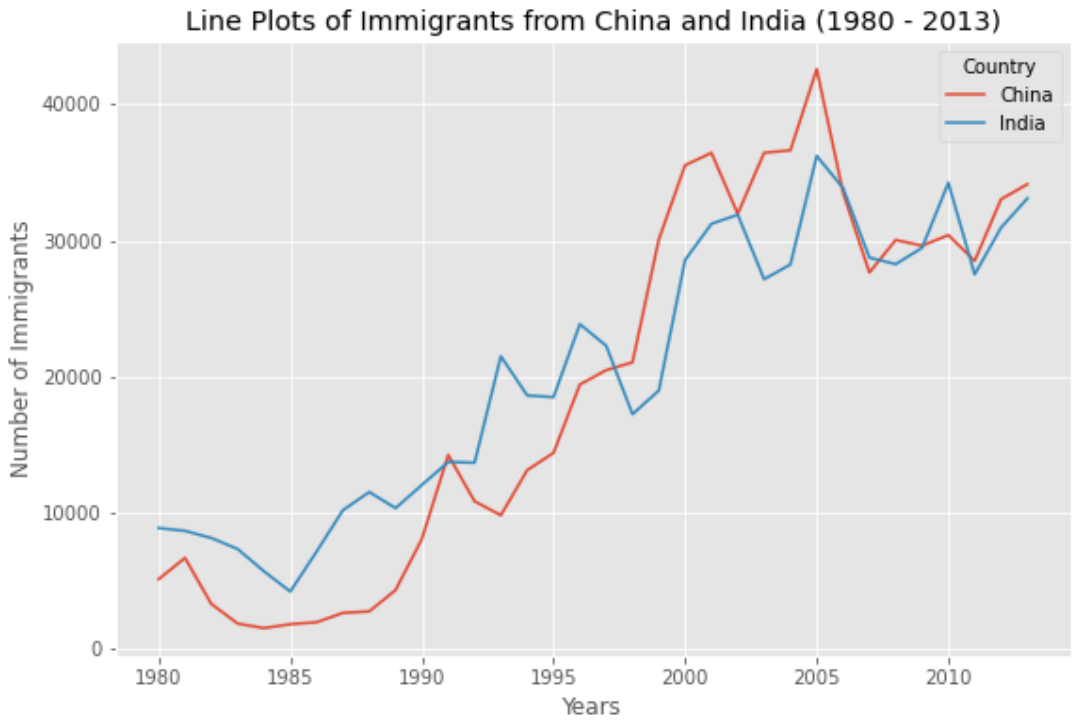
File: Part 3

Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

Subplots



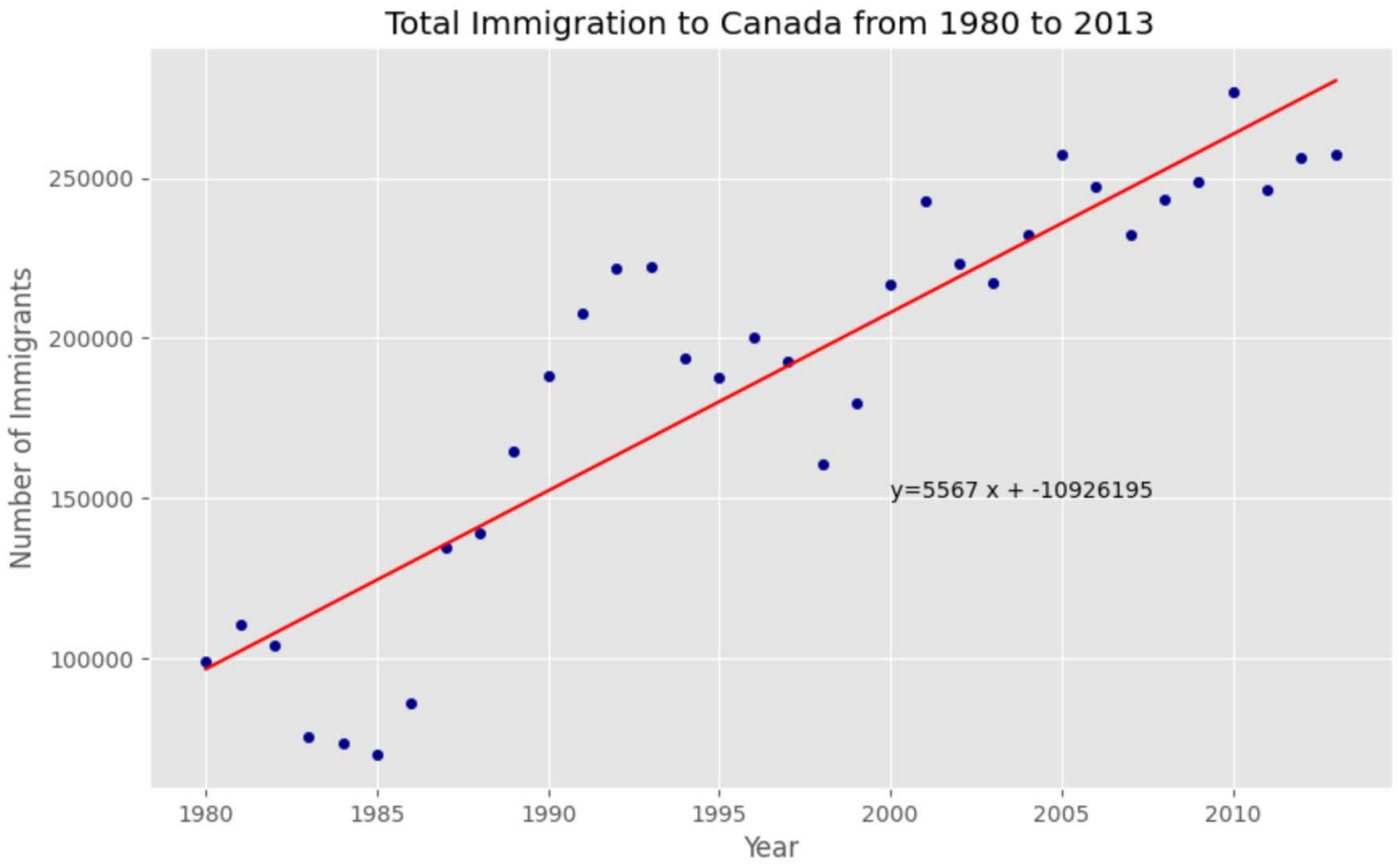


Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

File: Part 3

Scatter Plot



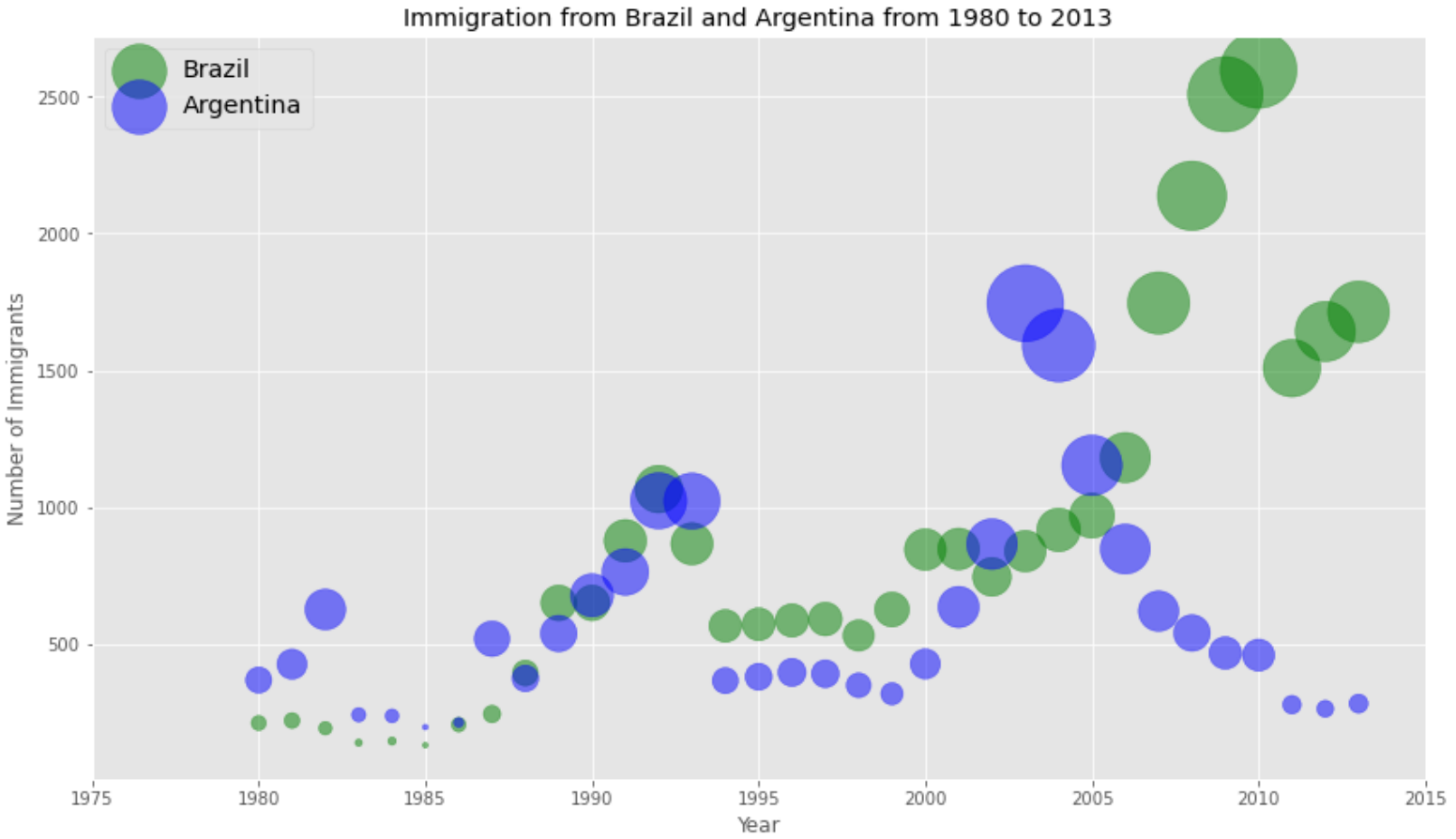

```
'No. Immigrants = 5567 * Year + -10926195'
```

File: Part 3

Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

Bubble Plot



File: Part 3

Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

** The size of a bubble represents the normalized number of immigrants for that data point.*

Discussion Part 3

- ***Pie Chart*** is visually striking and easy to understand the relative proportion for each category of data at a single glance. Example: We can eyeball that more than 50% of total number of immigrants came from Asia, without knowing the exact percentage.
- ***Box Plot*** can summarize data from multiple sources and display the results in a single graph, making the decision-making easier and more effective. Example: The number of immigrants from India has a narrower range than that from China.
- ***Subplots*** are used when data on different types of charts need to be considered together. Example: A box plot and a line chart are used together in the same figure to compare immigrants from India and China. This provides more insights into the trend and pattern of immigration from these two countries.

Discussion **Part 3** cont'd

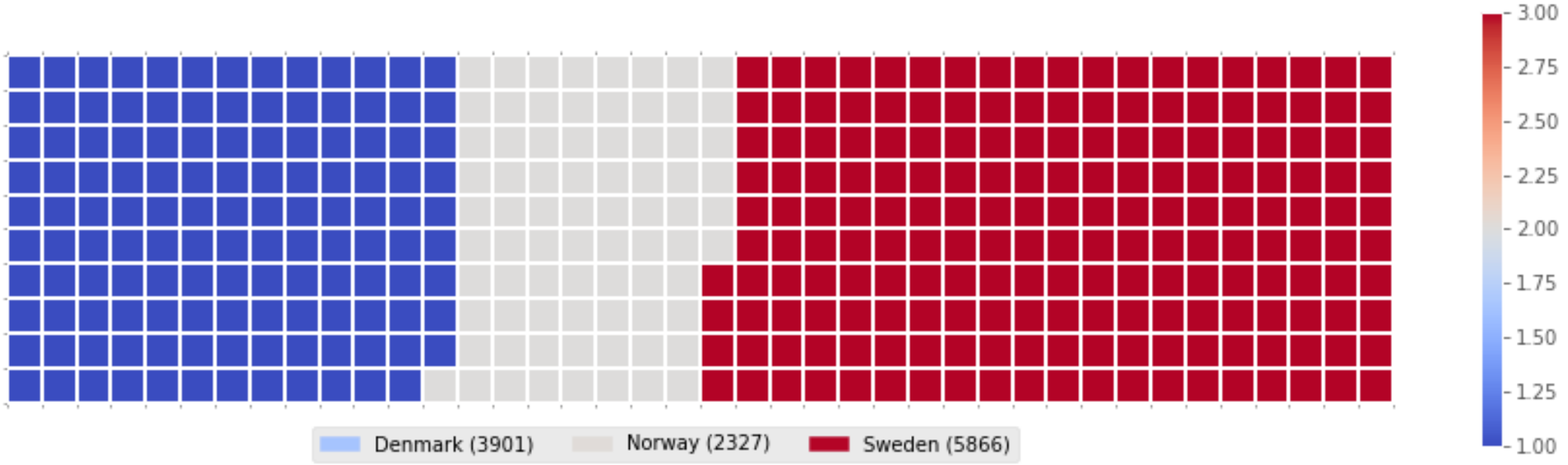
- ***Scatter chart*** is best for revealing: (1) Whether the relationship between two variables are linear or not. (2) How the data fluctuate. Example: The number of immigrants to Canada generally continued to rise as time went by. However, a temporary peak appeared between 1992 and 1993. Then, the typical trend resumed around 1995.
- ***Regression plot*** shows the relationship between dependent and independent variables and can be used for prediction. Example: The number of immigrants to Canada roughly grew with time in a linear fashion. If this pattern holds, the derived formula of the regression line would be appropriate for forecasting the number of immigrants for several more decades to come.
- ***Bubble plot*** presents the relationship between three variables and shows the changes in the trends. Comparison between two series of data can be made by using different colors to represent different series of data. Example: On the bubble plot, immigrants from Argentina peaked around 2002. This is due to a great depression in Argentina, 1998-2002. Once the country revered from that, the emigration from that country began to die down. On the other hand, its neighbor, Brazil, suffered an increased rate of violent crimes at that time. Also, Canada had kept the door wide open for immigrants then. These caused a large inflow of immigrants from Brazil.

Part 4_WaffleChart_RegressionPlot

Prepare the data for plotting visualizations.
Define a function to create waffle charts.
Create a Waffle Chart.
Create Regression Plots using Seaborn Library.

Waffle Chart

Immigration from Denmark, Norway, and Sweden to Canada from 1980 - 2013



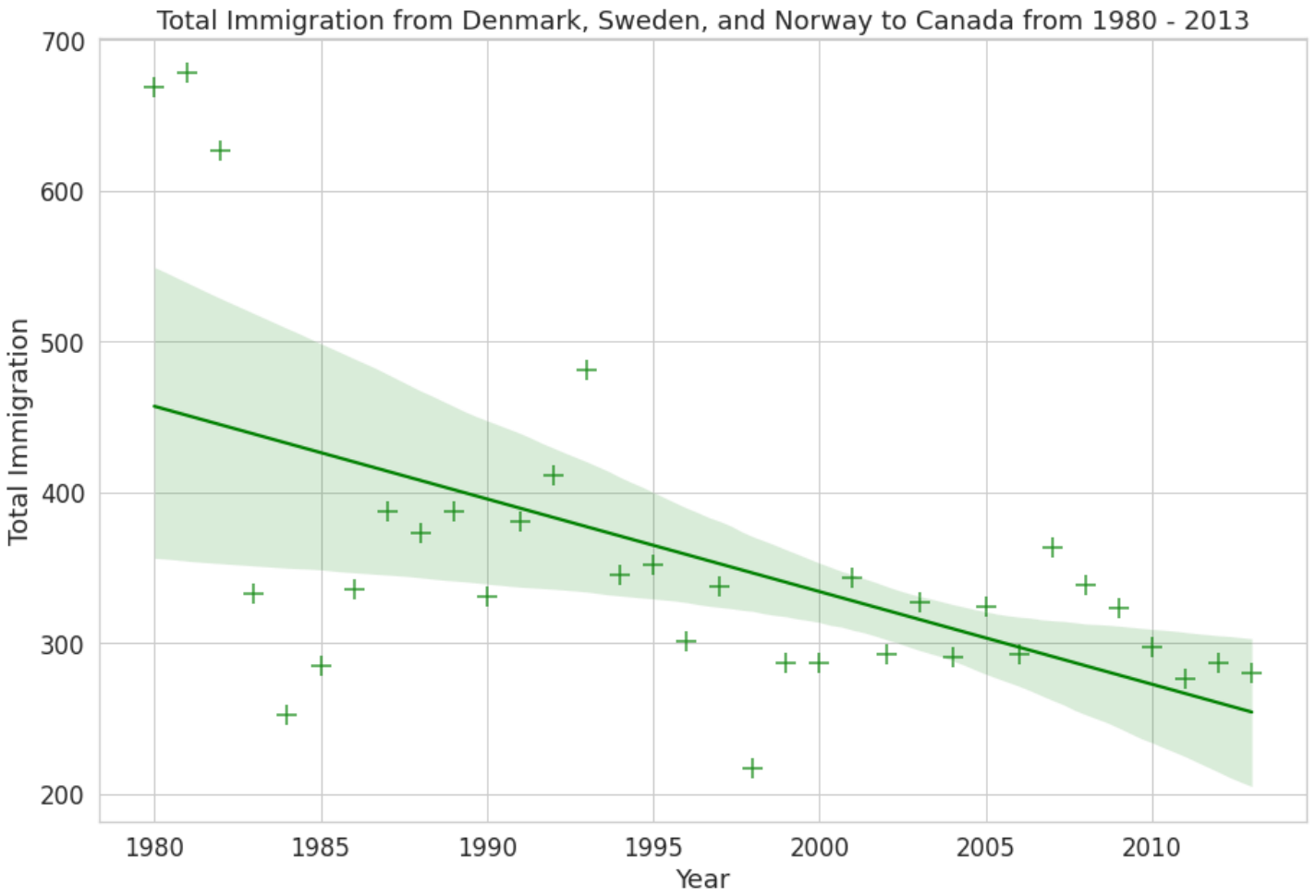
Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

File: Part 4

Regression Plot (Seaborn)

Immigration from Denmark, Norway, and Sweden to Canada from 1980 - 2013



File: Part 4

Library: Seaborn

Data Source: Canada dataset
International migration flows to Canada

Discussion Part 4

- ***Waffle Chart*** shows progress towards a target or a completion percentage. It is great for presenting data when describing the proportions or parts of a whole is important. It is particularly beneficial to distinguish one of the categories compared has very few squares in it. Example: On the waffle chart of the three Scandinavian countries, the squares within the waffle of each country really highlight the quantitative differences among these countries.
- **Using *Seaborn* to create *Regression Plot*** is much simpler than using Matplotlib. Seaborn creates a scatter plot with a linear fit on top of it in one step while Matplotlib accomplishes these separately. In addition, Seaborn plots the confidence level (the shaded area) along the length of the regression line, but Matplotlib doesn't do this.

Part 5_Complex Subplots

Section A: Mapping a Trail

Generate the coordinates, the grid, and the elevations along a trail for mapping the trail.
Create Subplot 1: The bird's eye view of the trail (Contour Plots and Line Plot).
Create Subplot 2: The trail profile at Pup's Peak on the trail (Fill_between Plot).

Section B: Plate Fracture Study

Gather data for subplots.

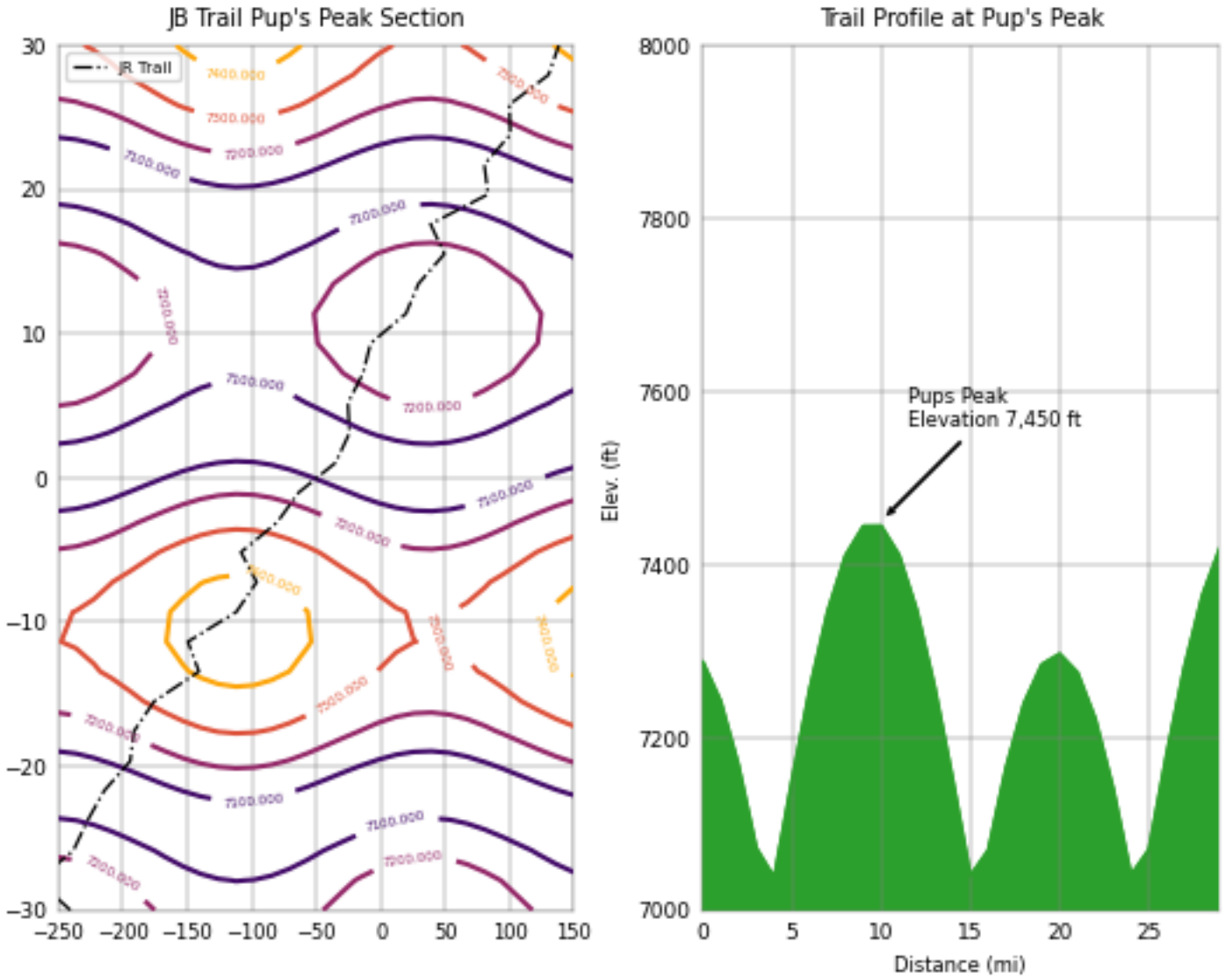
- Subplot 1: Displacement versus Time
  - Define a function to calculate the moving average of displacement with time.
  - Generate data for Subplot 1.
- Subplot 2: Stress vs. Strain Relationship
  - Load csv file of stress and strain data.
  - Get data from each column of the loaded dataset and save it in a variable.
  - Calculate the values of stress above and below the estimated values to show the 95% confidence level.
- Subplot 3: Stress on Plate Cross Section

Create blank subplots, place them in the figure, and adjust the spacing between subplots.

Create the visualizations of each subplot.

- Subplot 1: Create a Scatter Plot and a Line Plot, both plotted against the variable ‘time.’
- Subplot 2: Create two Line Plots (actual and estimated) to show the relationship of stress and strain.
- Subplot 3: Create an Imshow Colorplot.

Complex Subplots



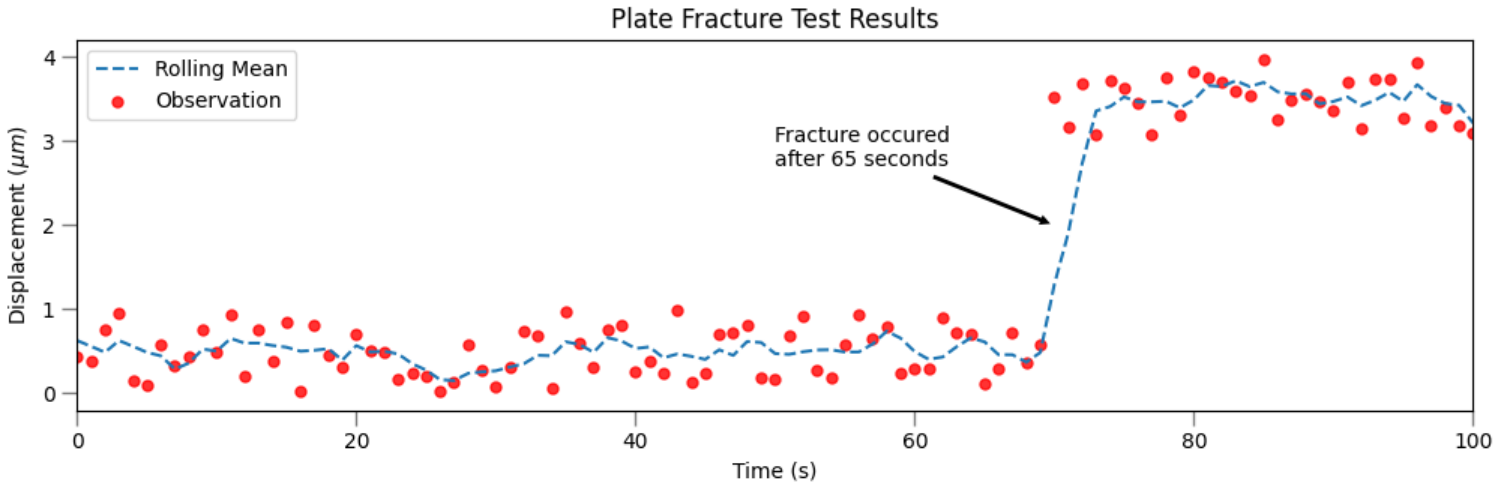
File: Part 5

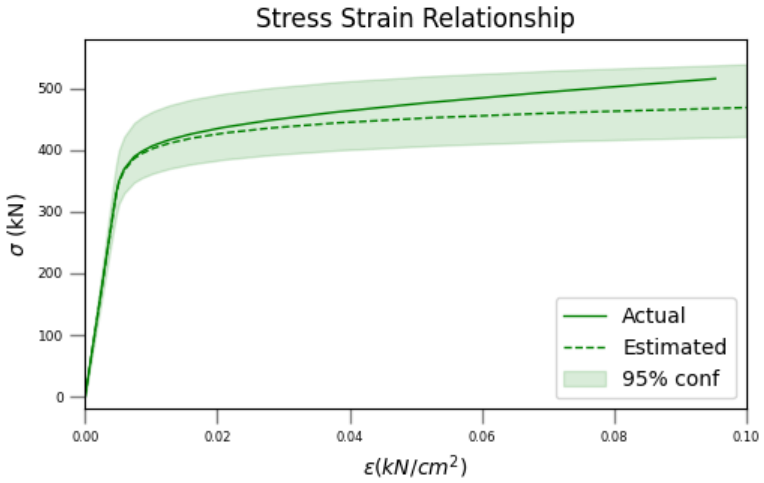
Library: Matplotlib

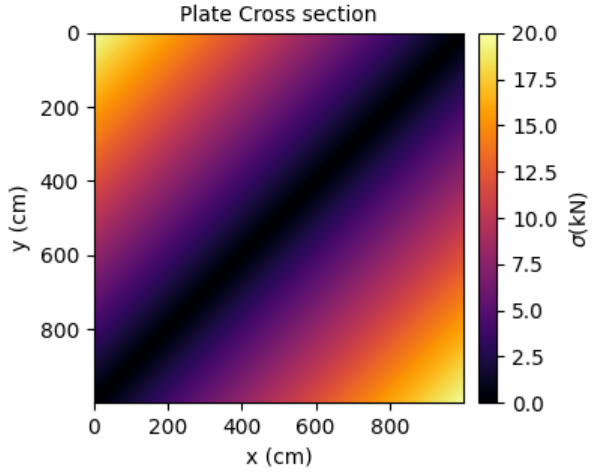
Data Source:

Generated in the code

Complex Subplots







File: Part 5

Library: Matplotlib

Data Source:

stress_strain.csv (in folder)

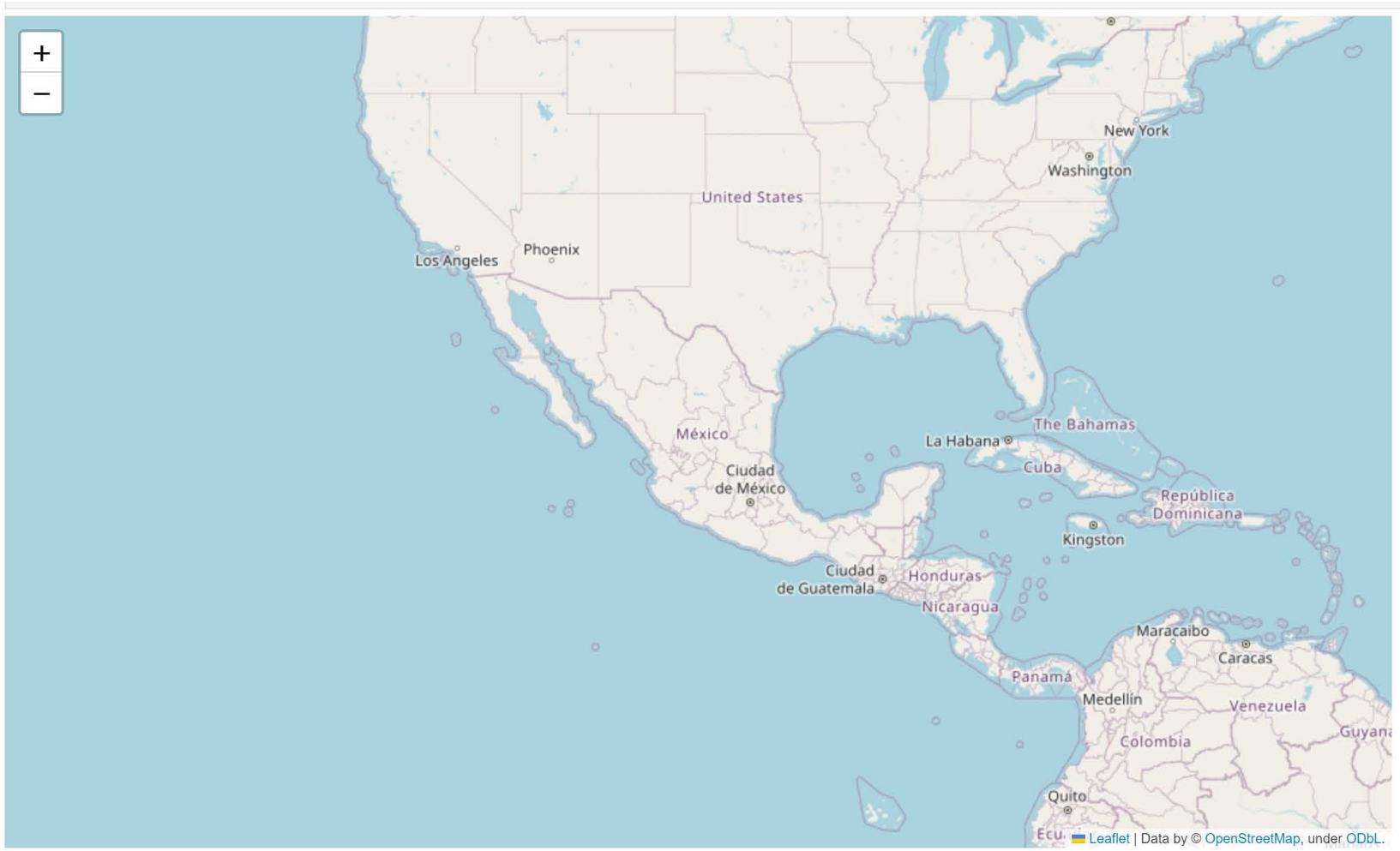
Discussion Part 5

- ***Complex Subplots*** are helpful when multiple aspects of the data that need to be considered altogether are presented in different chart styles. Example: In the case of the stress vs. strain study, the scientists want to consider stress and strain relationship (regression plots) and plate cross section (inshow color plot) while they are looking at the plate fracture results (scatter plot and line chart).

Part 6_Maps

Create a Basic Map.
Create a Stamen Toner Map.
Create a Stamen Terrain Map.
Create a Map with Markers.
Create a Map with Marker Clusters.
Create a Choropleth Map.

Basic Map



File: Part 6

Library: Folium

Stamen Toner Map

File: Part 6

Library: Folium

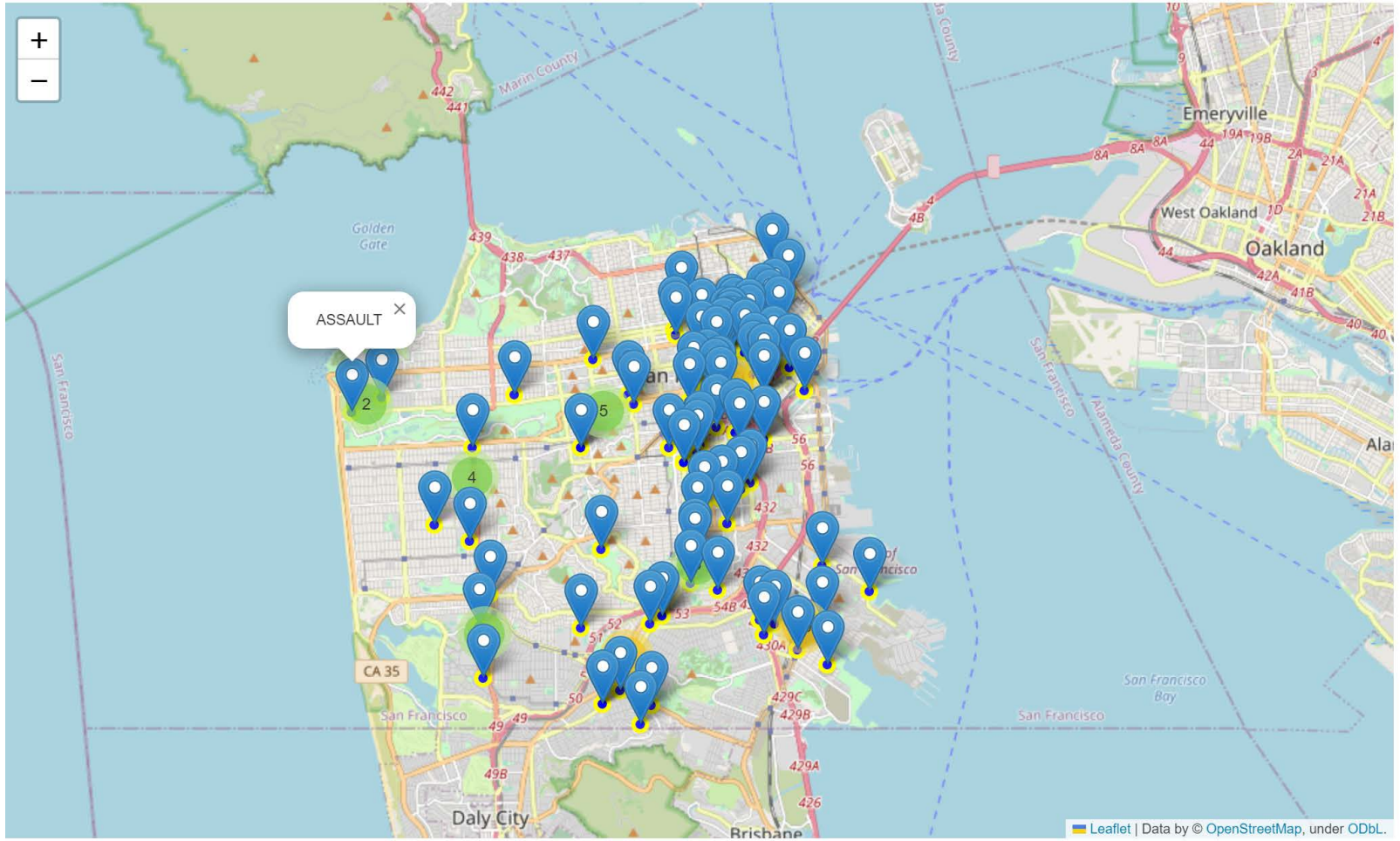
Stamen Terrain Map

File: Part 6

Library: Folium

Map with Markers

Crimes around San Francisco, 2016



*Information like crime categories appears when clicking on a pop-up marker.

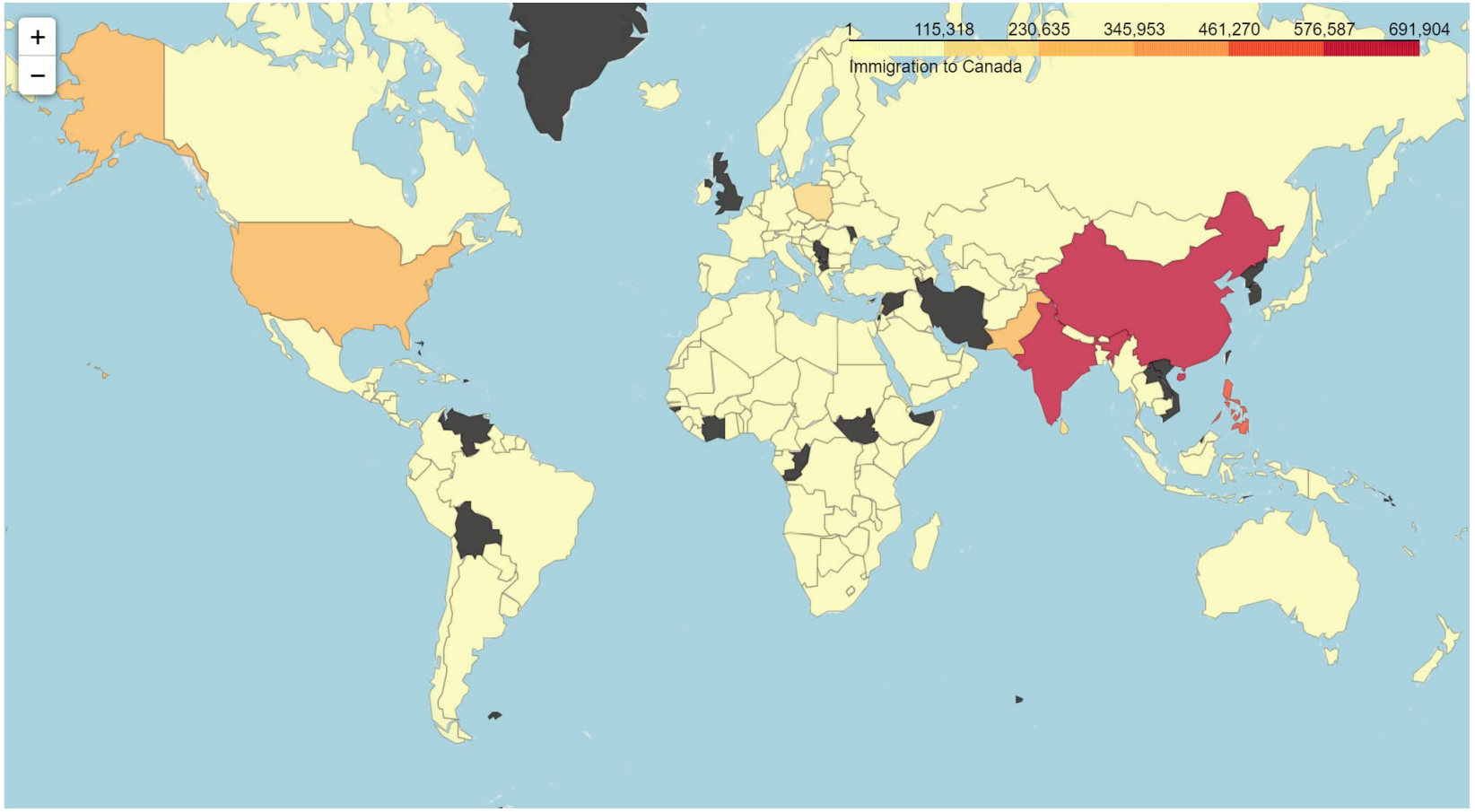
File: Part 6

Library: Folium

Data Source: Police Department Incidents-Previous Year 2016

Choropleth Map

International Migration Flows To Canada



File: Part 6

Library: Folium

Data Source:
Canada dataset
International migration flows to Canada

Discussion Part 6

- All maps shown were created using Folium, a powerful Python library.
- ***Stamen Toner Map*** is used for data mashups and for exploring and visualizing river meanders and coastal zones.
- ***Stamen Terrain Map*** features hill shading and natural vegetation colors. It showcases advanced labeling and linework generalization of dual-carriageway roads.
- ***Map with Markers*** offers an immediate view of the data attached to the markers. Example: Crime categories were added next to the circle markers on the San Francisco map.
- ***Choropleth* Map** uses levels of shading/color to represent a range of values. It is visually effective because a large amount of information and general patterns can be seen on the same map. Example: On the choropleth map of immigrants to Canada, China and India (in red color) can be easily spotted as two countries contributing to the immigrant inflow to Canada the most.