Data Visualization with Python

Avery Jan

5-22-2022

Outline

Introduction

Methodology

Datasets

Basic Charts (**Files:** Part 1 and Part 2)

Line chart, Area chart, Histogram, Column Chart, Bar Chart

Specialized Charts (**File:** Part 3)

Pie Chart, Box Plot, Subplot, Scatter Plot, Regression Plot (Matplotlib), Bubble Plot

Advanced Charts (**Files:** Part 4 and Part 5)

Waffle Chart, Regression Plot (Seaborn), Complex Subplots

Geospatial Charts (**File:** Part 6)

Maps: Stamen Toner, Stamen Terrain, Map with Markers, Choropleth

Discussion

Introduction

In this project, basic, specialized, advanced, and geospatial visualizations were created with python. Data from three sources were used in this project: (1) Real-world datasets (2) Data generated within the code (3) A dataset from a scientific study. The two real-world datasets were read into Pandas DataFrames. Then, if necessary, these data underwent data wrangling steps specified in the code (see the code files) that built the visualization. On the other hand, the data from the other two sources were used directly. This project consists of six parts (see code files for each part). In this file, the part in which a visualization was constructed is noted next to the visualization. Only selected visualizations created in each code file are included here. These visualizations are grouped into four sections: (a) Basic Charts (b) Specialized Charts (c) Advanced Charts (d) Geospatial Charts. In each section, the applications of these visualizations are discussed at the end of the respective section.

Methodology

Data Sources: Noted alongside the charts

Platforms: Codio, Jupyter Notebook

Python Libraries:

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

Data Wrangling:

Remove unwanted columns.
Rename columns to more meaningful names.
Add a new column “Total” to hold the sum of data of all years.
Transpose Dataframe.
Change data types.

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 6)

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

Dataset Generated within Code (generated and used in Part 5_A)

-Using Python NumPy Library to generate the route of a fictitious trail, the JB trail.

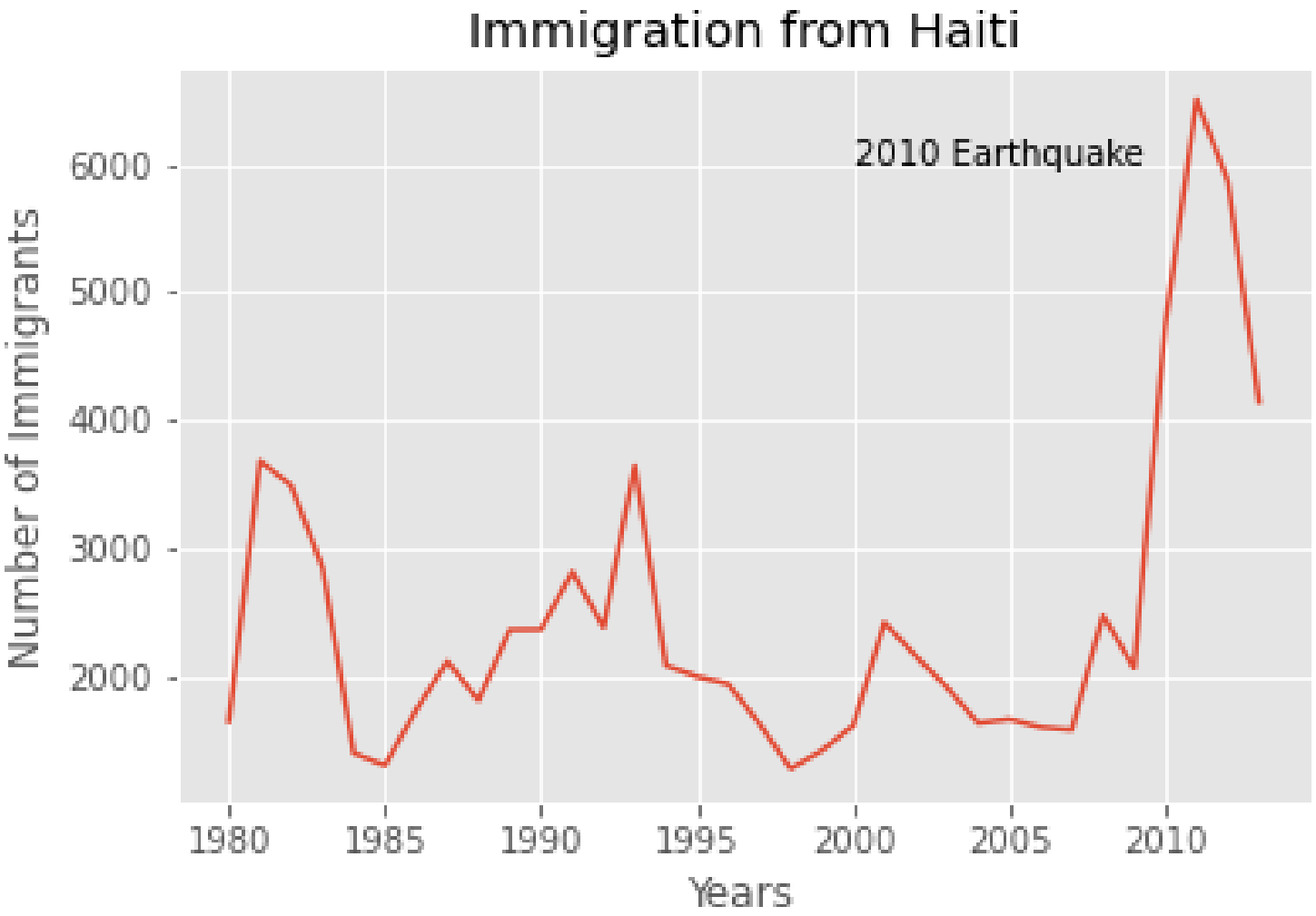
Scientific Study Dataset (stress_strain.csv, used in Part 5_B)

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

Basic Charts

(Files: Part 1 and Part 2**)**

Line Chart



File: Part 1

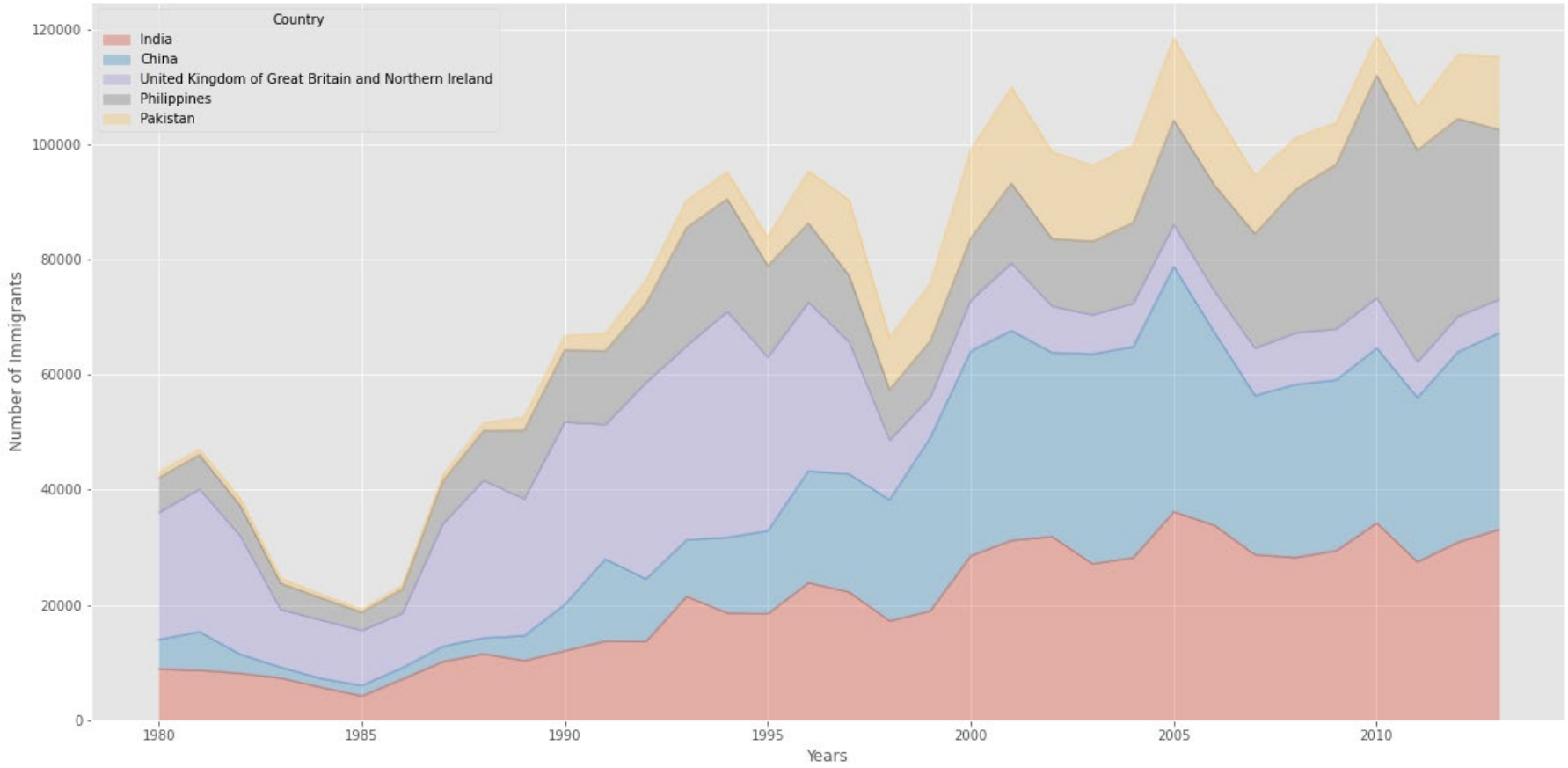
Library: Matplotlib

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

Area Chart (The artist layer, stacked)

Immigration Trend of Top 5 Countries



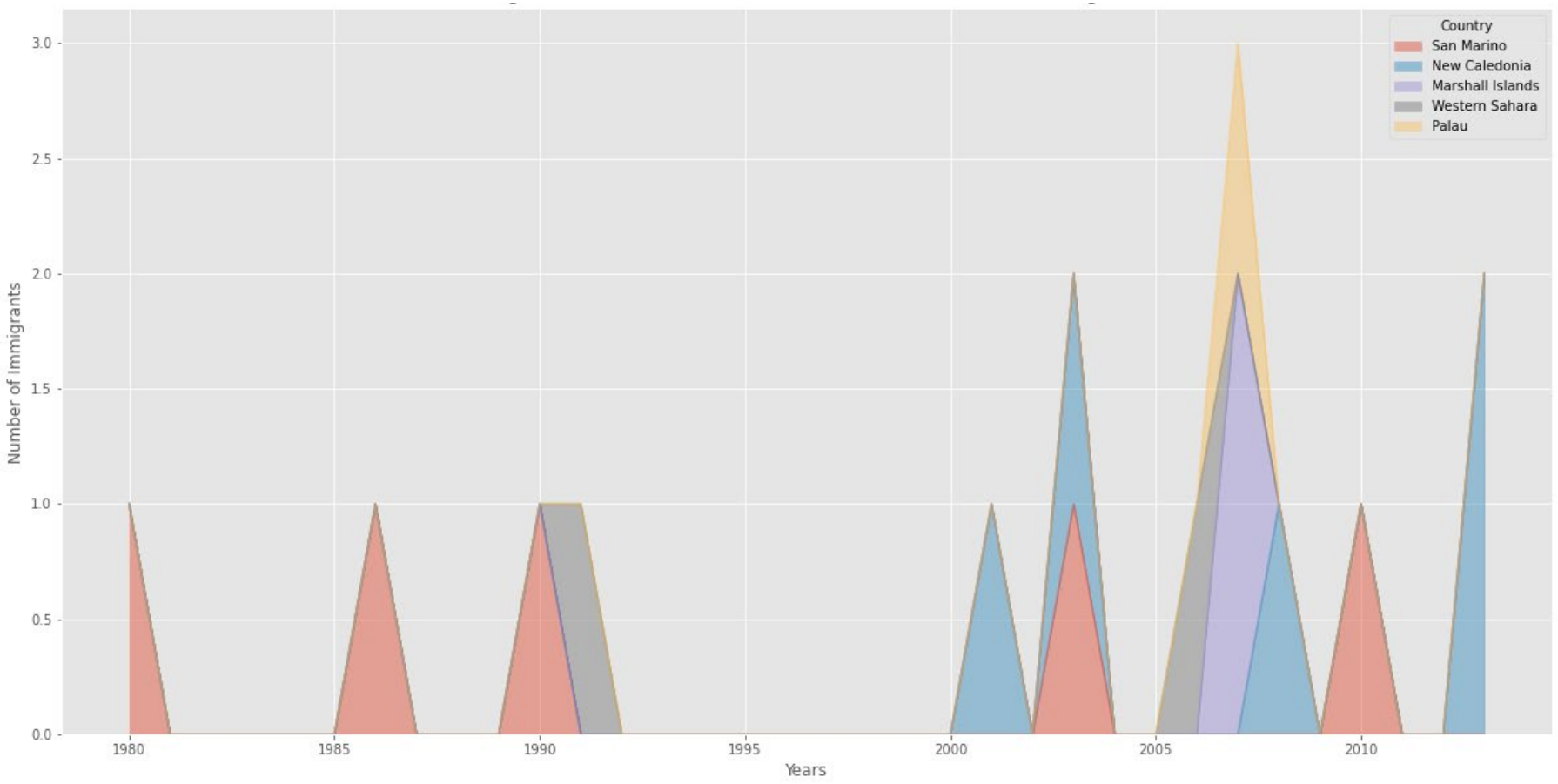
Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

File: Part 2

Area Chart (The script layer, stacked)

Immigration Trend of 5 Countries with Least Contribution to Immigration



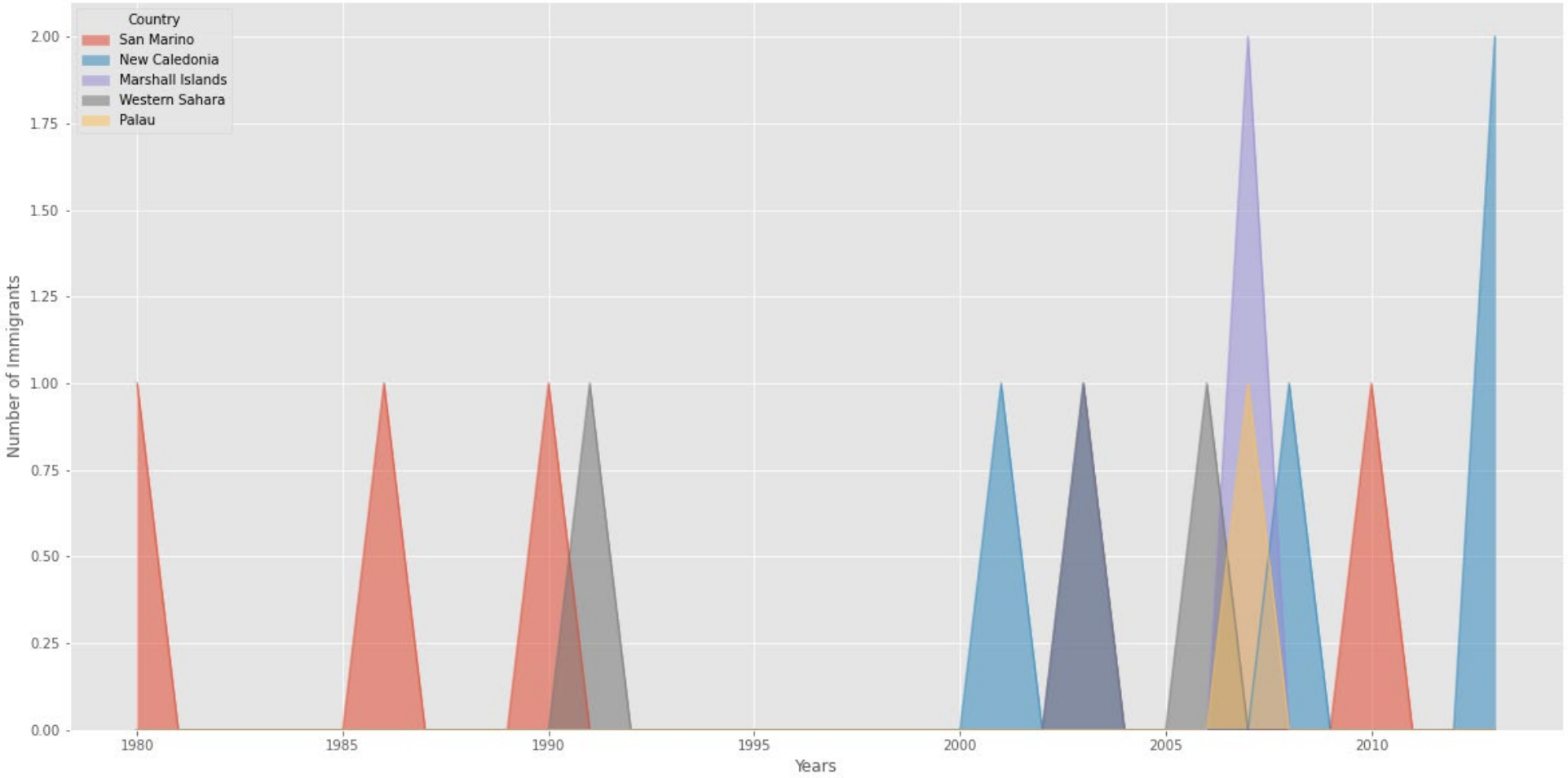
Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

File: Part 2

Area Chart (The artist layer, unstacked)

Immigration Trend of 5 Countries with Least Contribution to Immigration

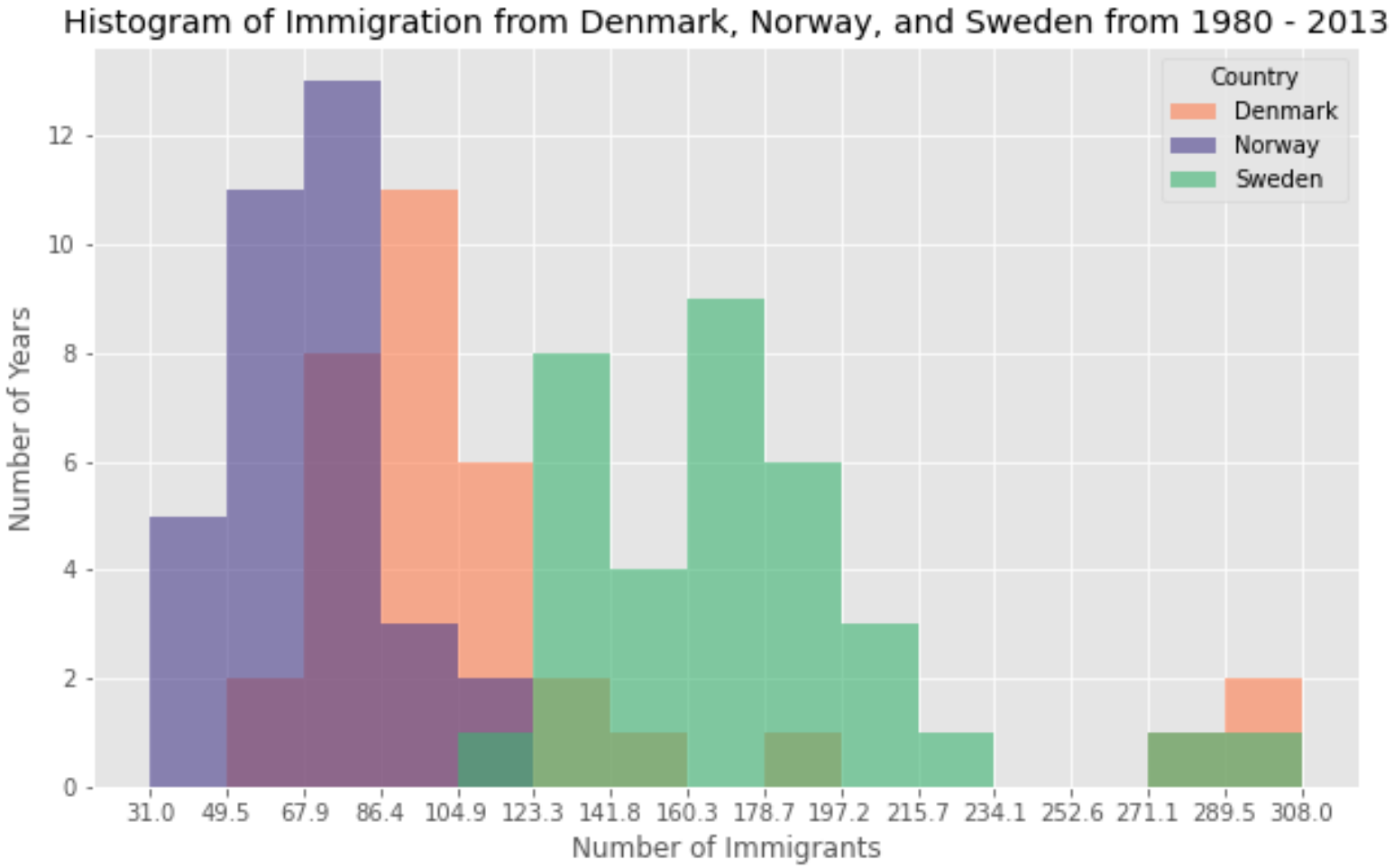


Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

File: Part 2

Histogram

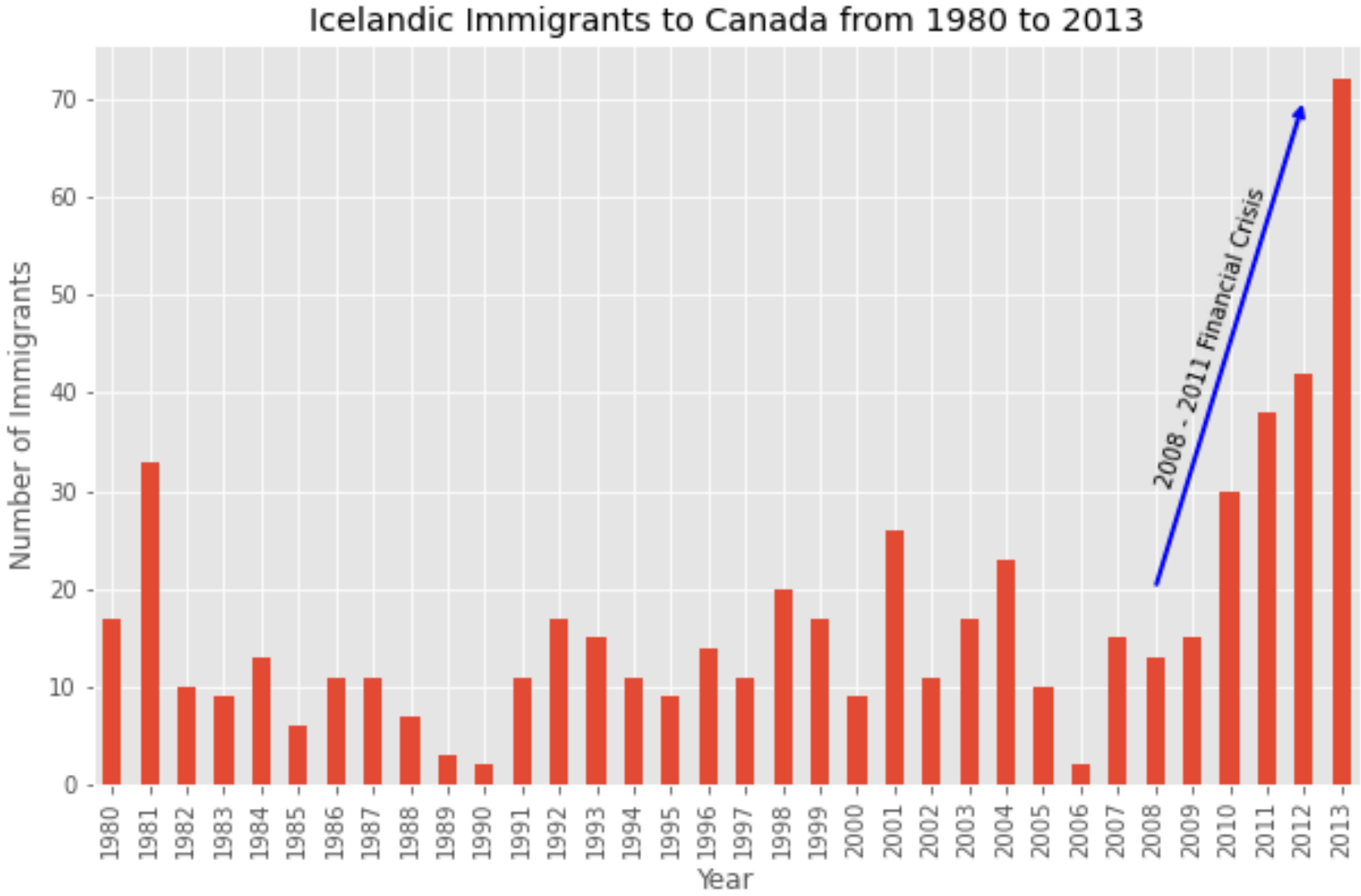


File: Part 2

Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

Column Chart (Vertical Bar Chart)



File: Part 2

Library: Matplotlib

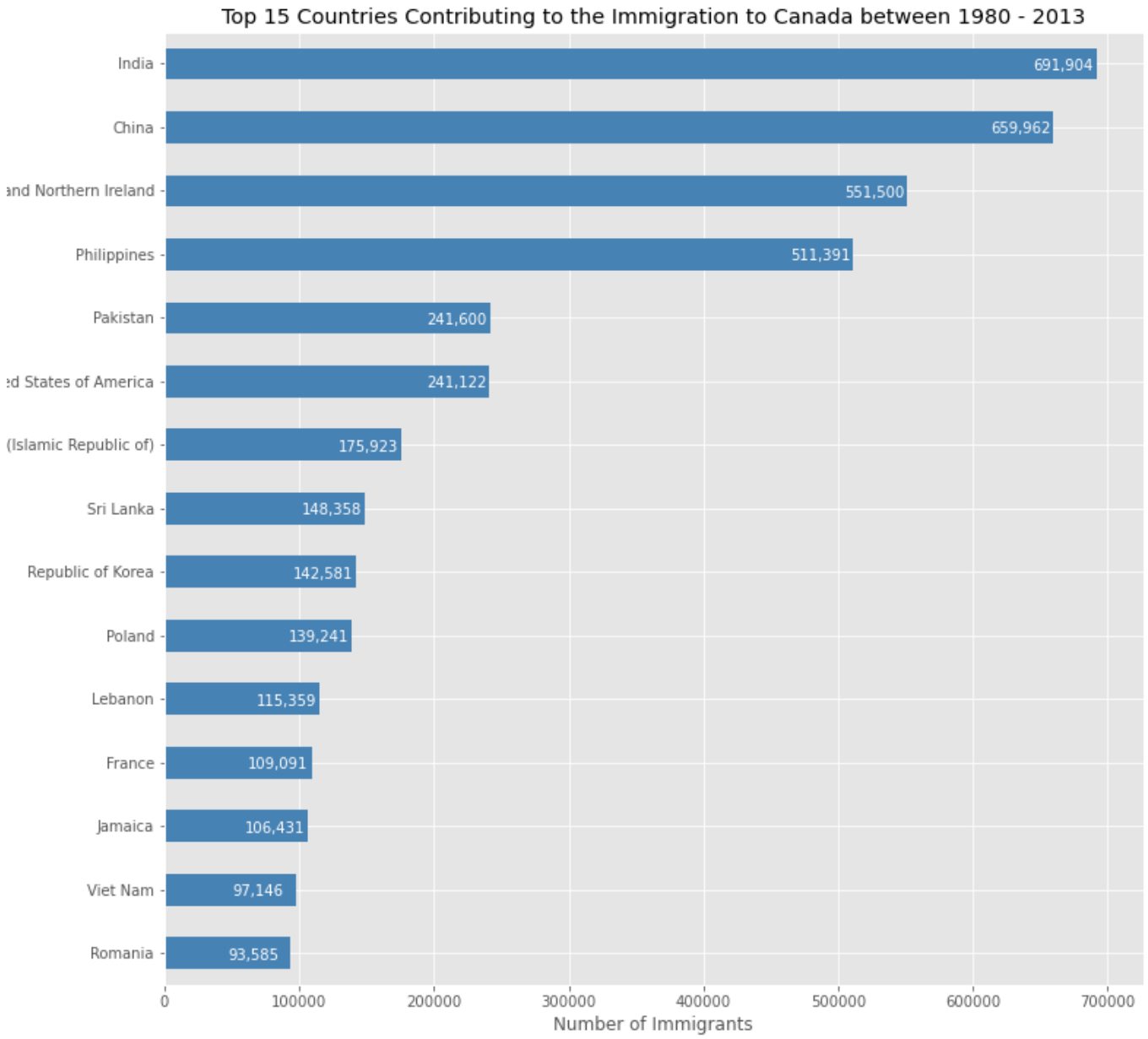
Data Source: Canada dataset
International migration flows to Canada

Bar Chart

File: Part 2

Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada



Discussion Data Prep and Basic Charts

Data Preparation

- The purposes of storing a large real-world dataset in a *Pandas DataFrame* are: (1) to easily obtain subsets for creating specific visualizations and analysis (2) to facilitate data wrangling tasks.

Basic Charts

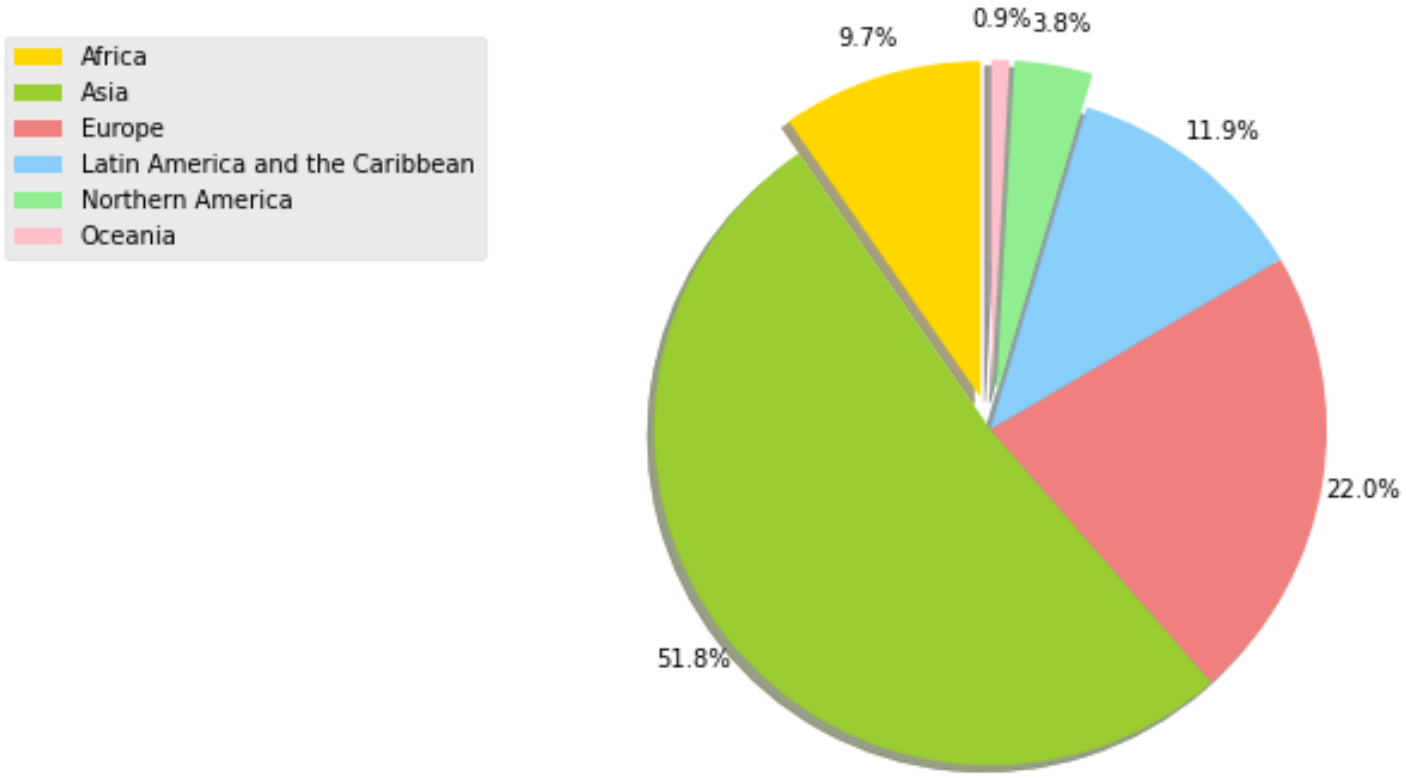
- ***Line chart*** and ***Area Chart*** provide a quick way to see the trend of data and correlate events with the data. For example, after the 2010 earthquake in Haiti, there is a spike of Haitians immigrated to Canada.
- ***Histogram*** is commonly used to show the shape and spread of the data like how the number of immigrants from 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.

Specialized Charts

(File: Part 3**)**

Pie Chart

Immigration to Canada Grouped by Continent [1980 - 2013]

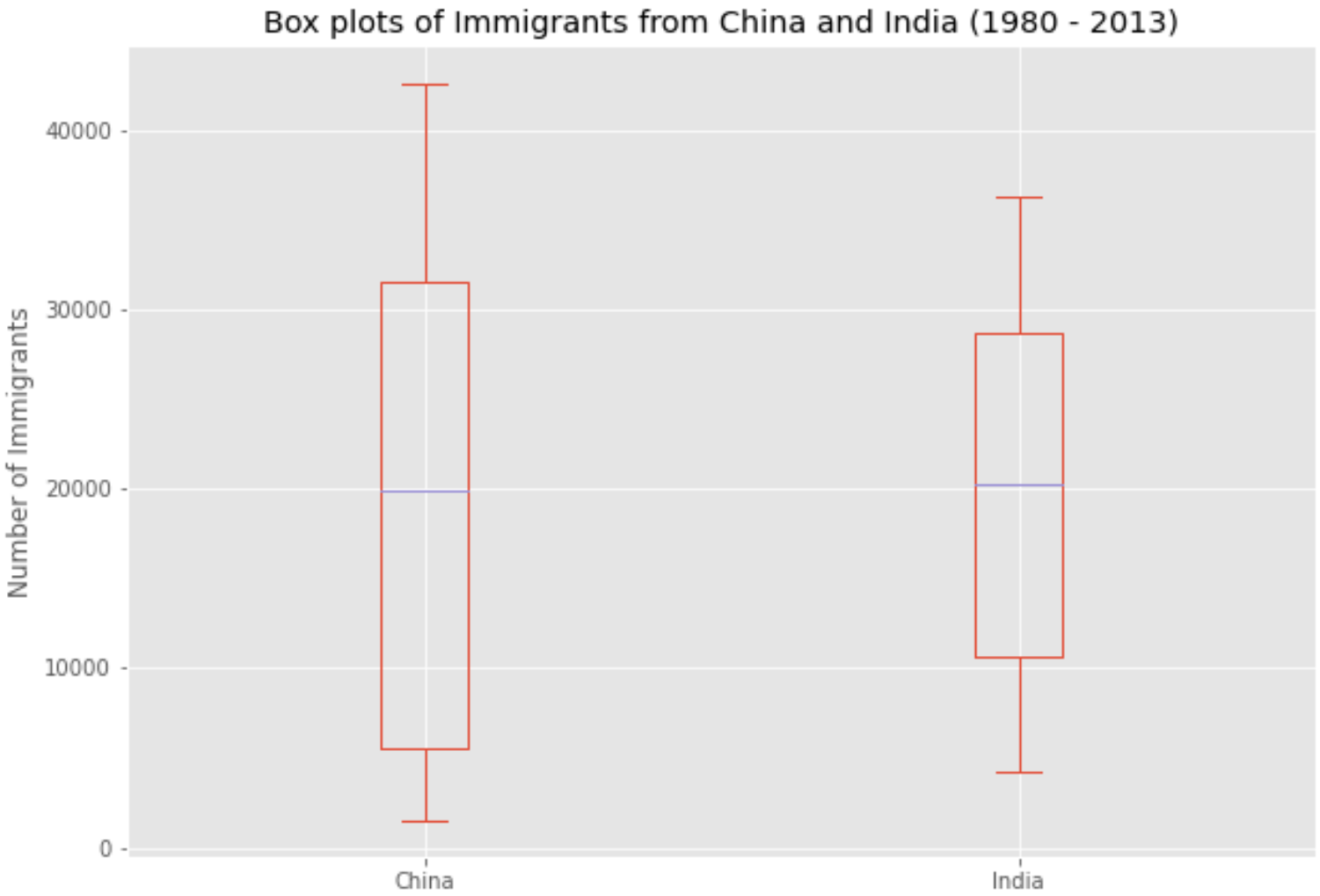


File: Part 3

Library: Matplotlib

Data Source: Canada dataset
International migration flows to and Canada

Box Plot

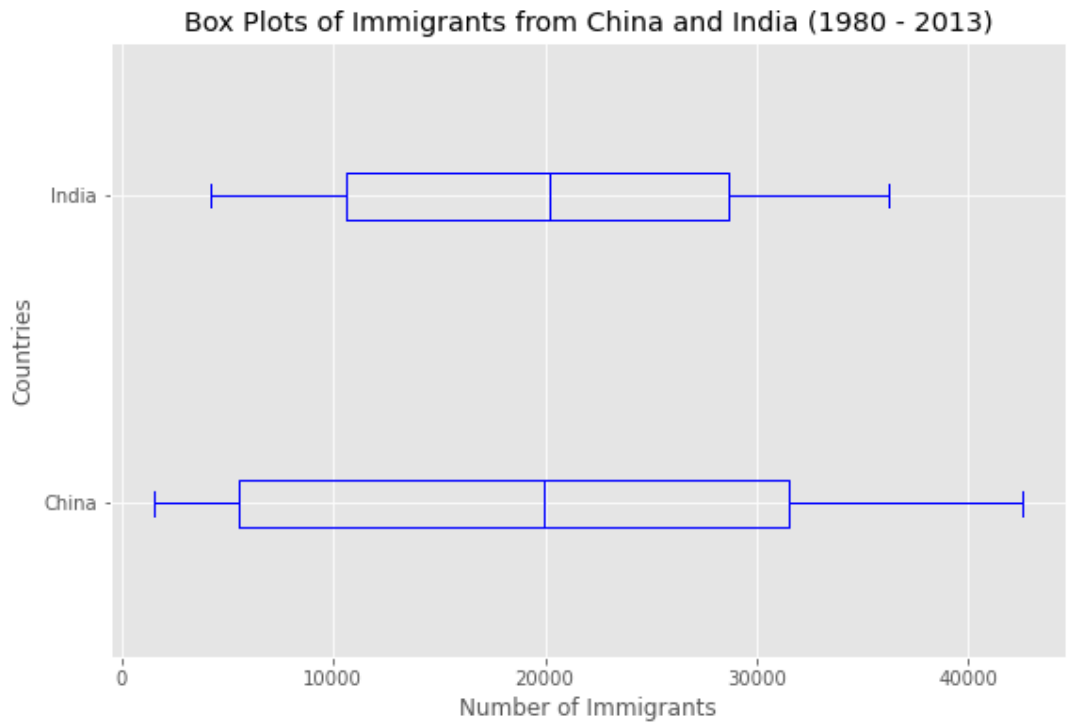


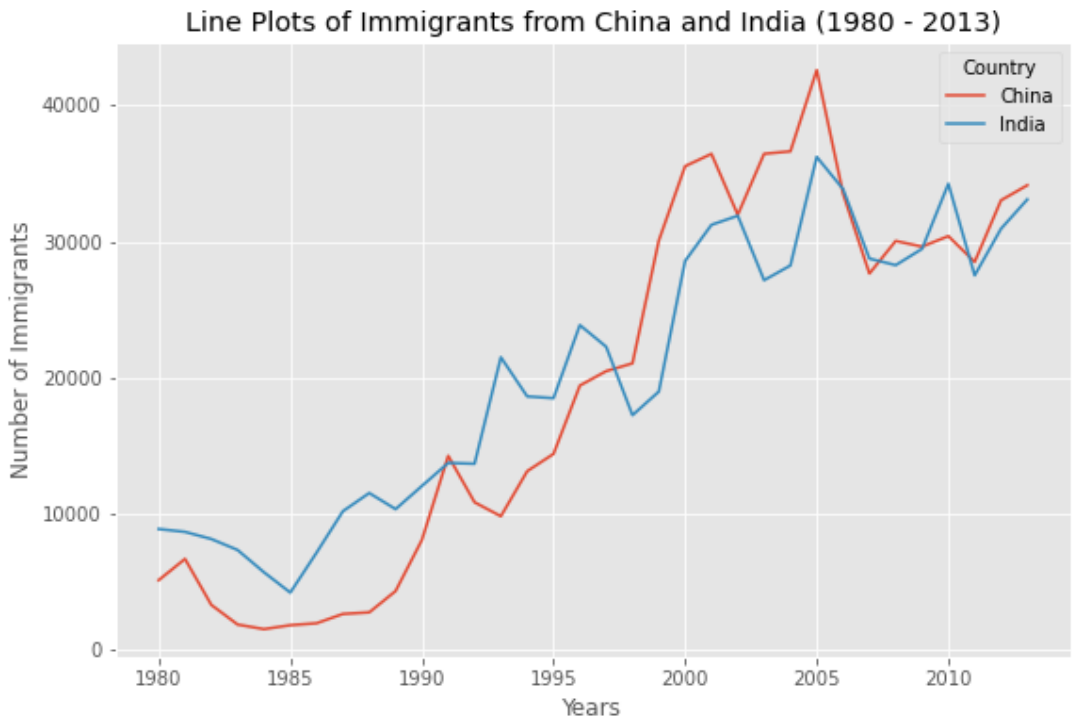
File: Part 3

Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

Subplots



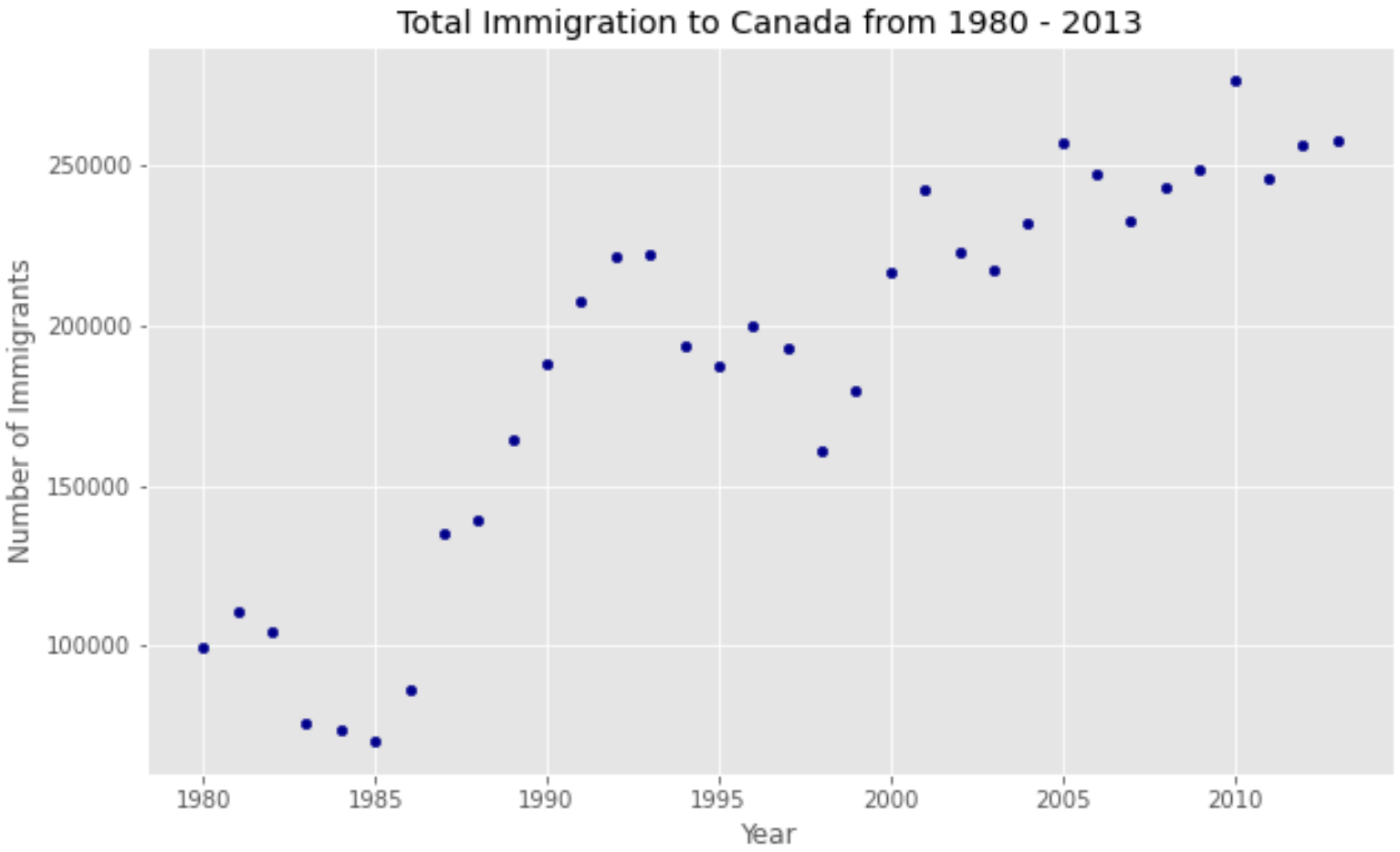


Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

File: Part 3

Scatter Plot

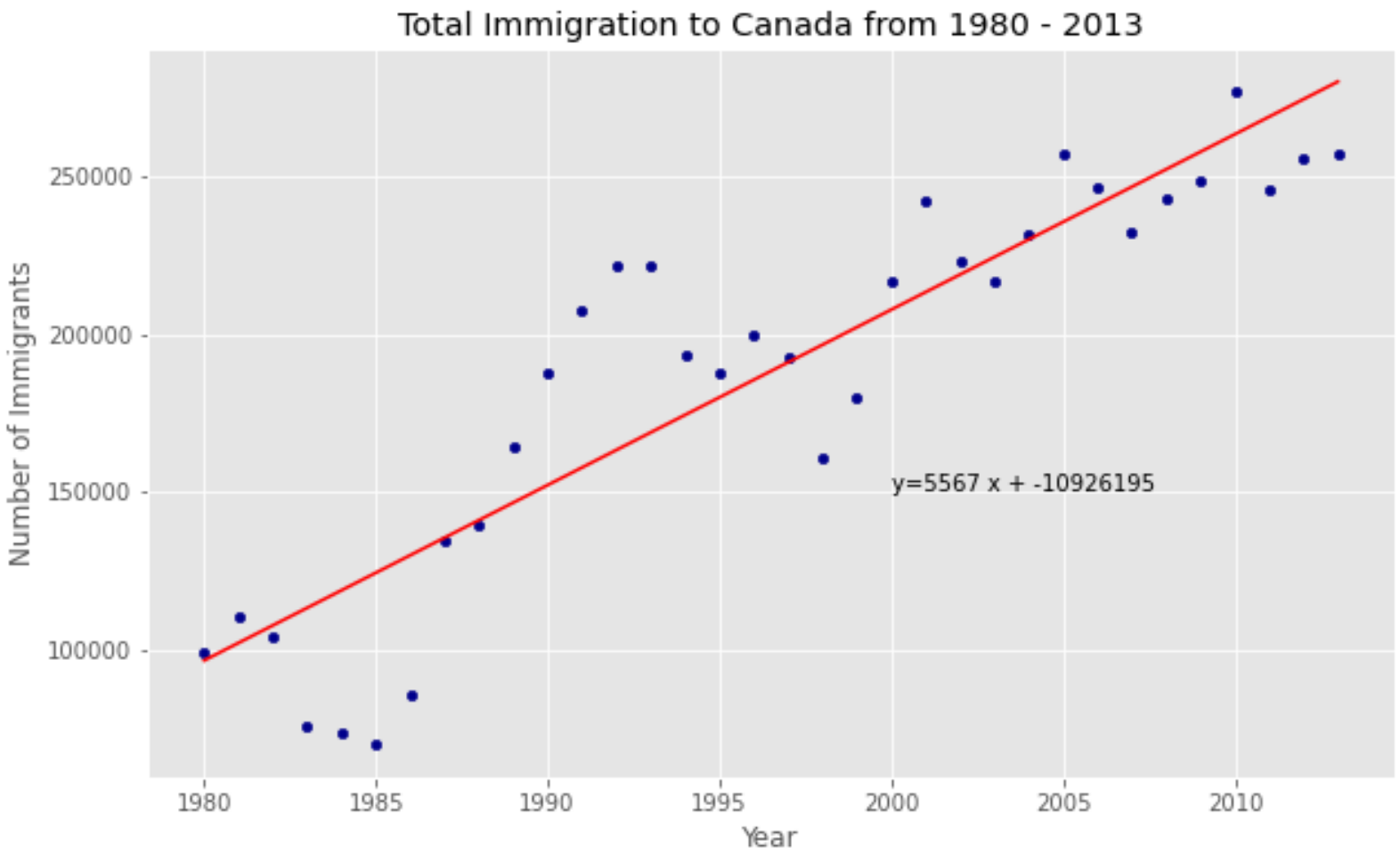


File: Part 3

Library: Matplotlib

Data Source: Canada dataset
International migration flows to Canada

Regression Plot (Matplotlib)

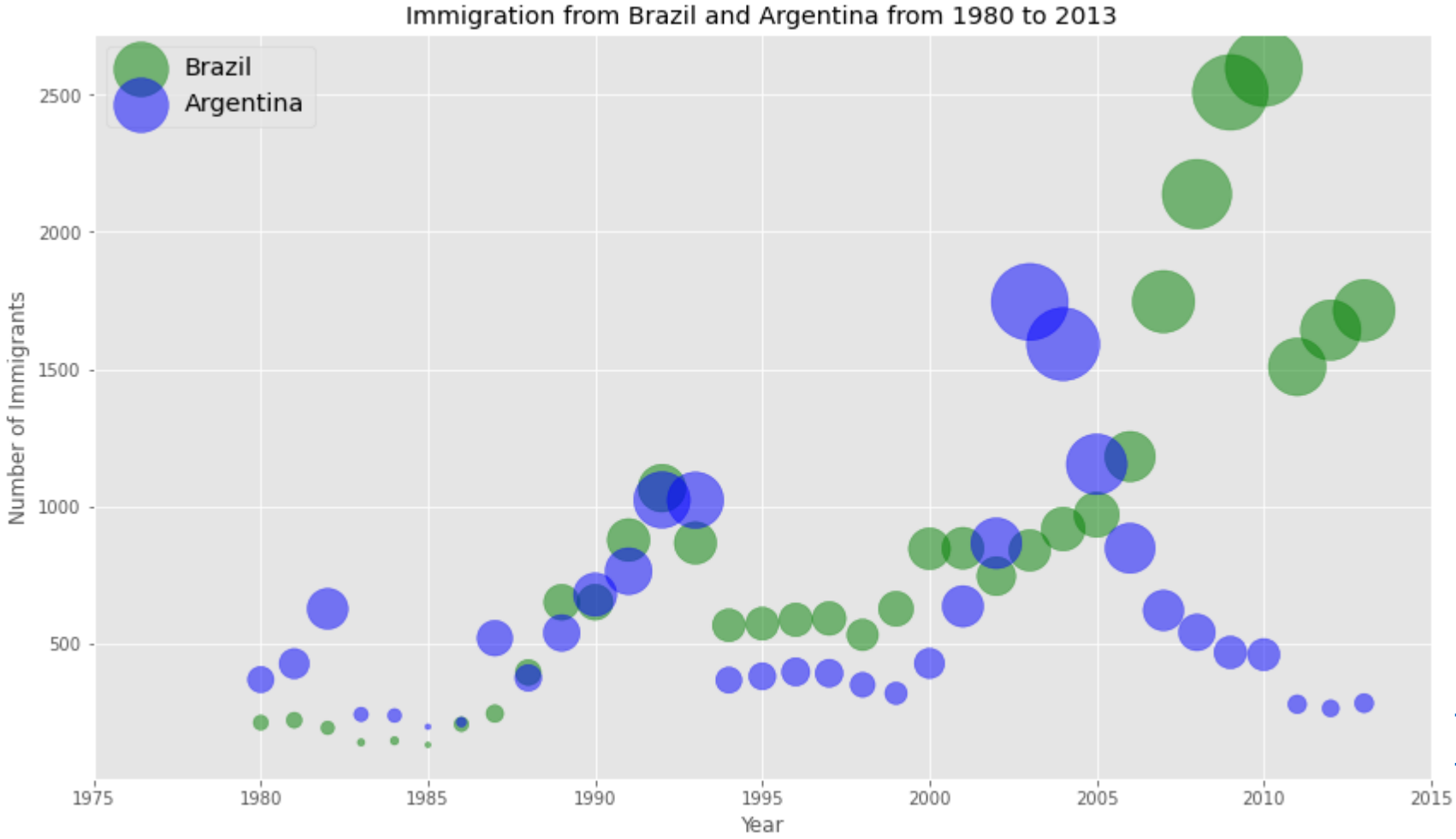


File: Part 3

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 Specialized Charts

- ***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 form China.
- ***Subplots*** are used when data on different types of charts need to be considered together. Example: Both box plot and line chart are used together to compare immigrants from India and China to provide more insights into the trend and pattern of immigration from these two countries.

Discussion Specialized Charts cont'd

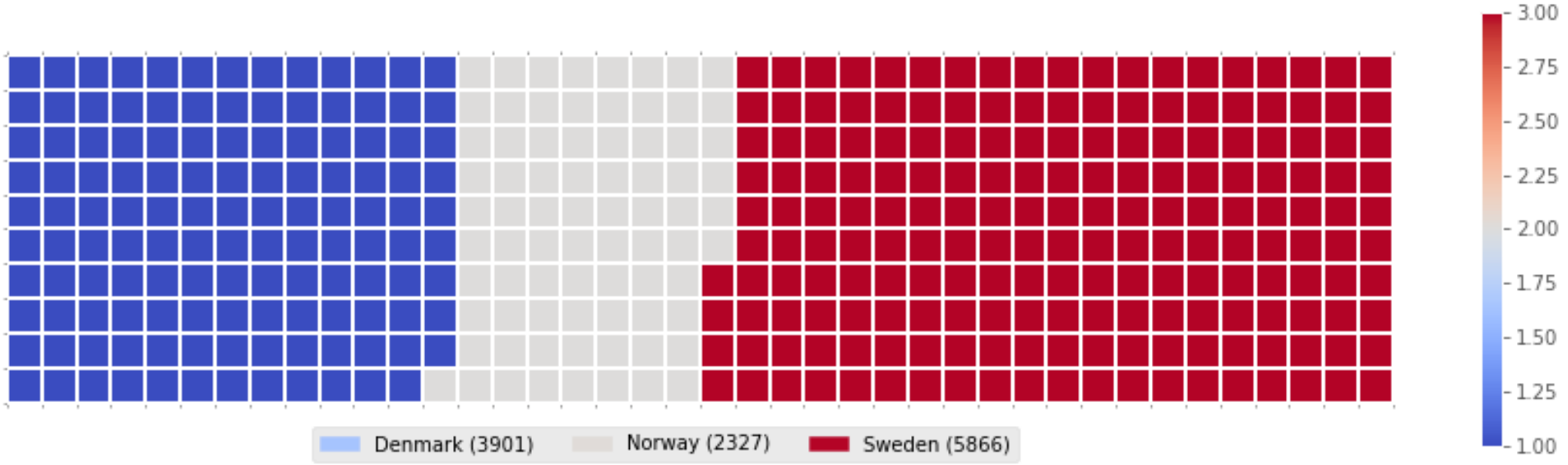
- ***Scatter chart*** is best for revealing whether the relationship between two variables are linear or non-linear and how the data fluctuate. Example: The number of immigrants to Canada generally continued to rise as time went by. However, a temporary peak appeared between 1992 and 1993. Then, the 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 might be appropriate for forecasting the number of immigrants for several more decades to come.
- ***Bubble plot*** presents 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. Meanwhile, Canada had kept the door wide open for immigrants. These caused a large inflow of immigrants from Brazil.

Advanced Charts

(Files: Part 4 and Part 5**)**

Waffle Chart

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



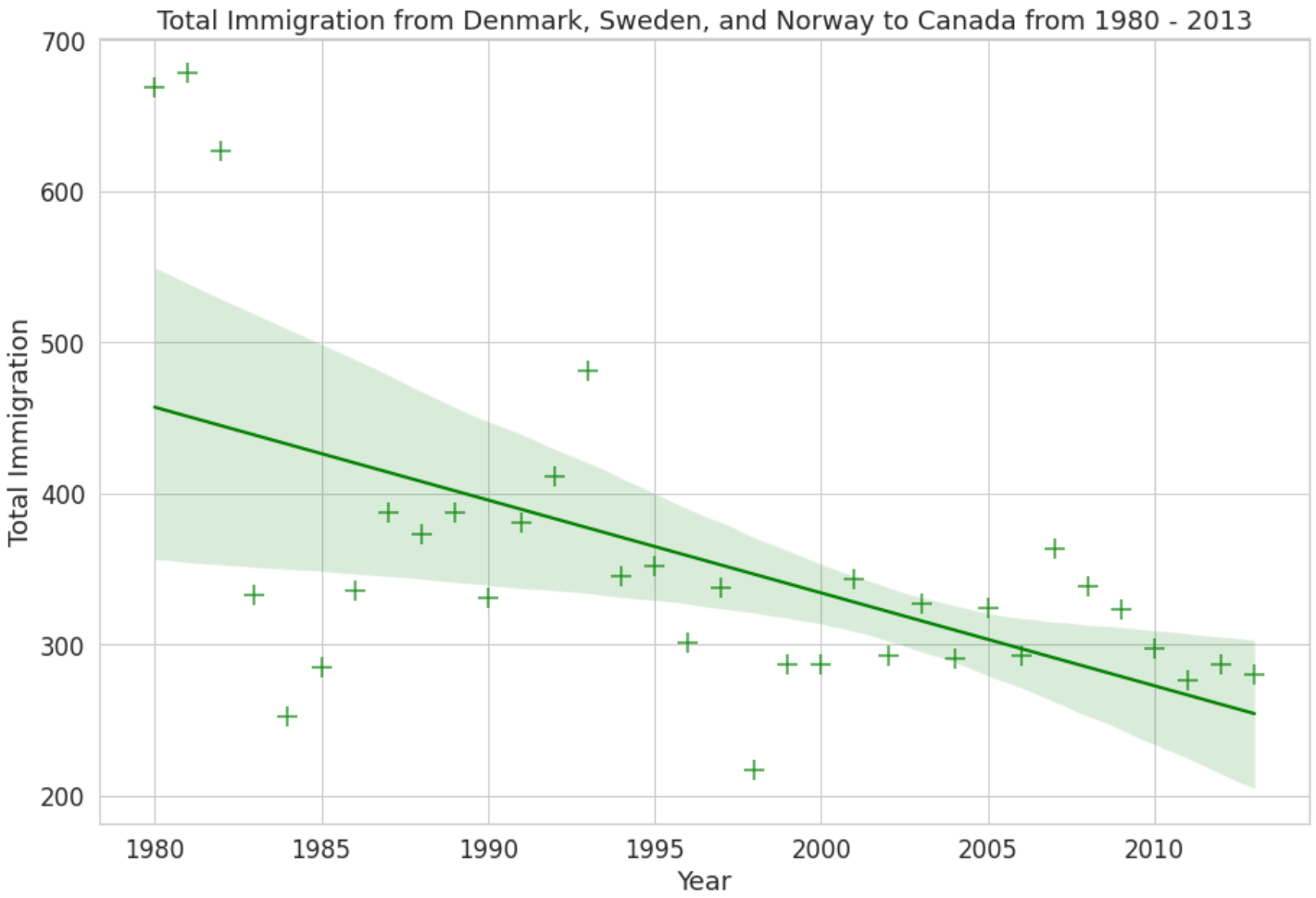
Library: Matplotlib

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

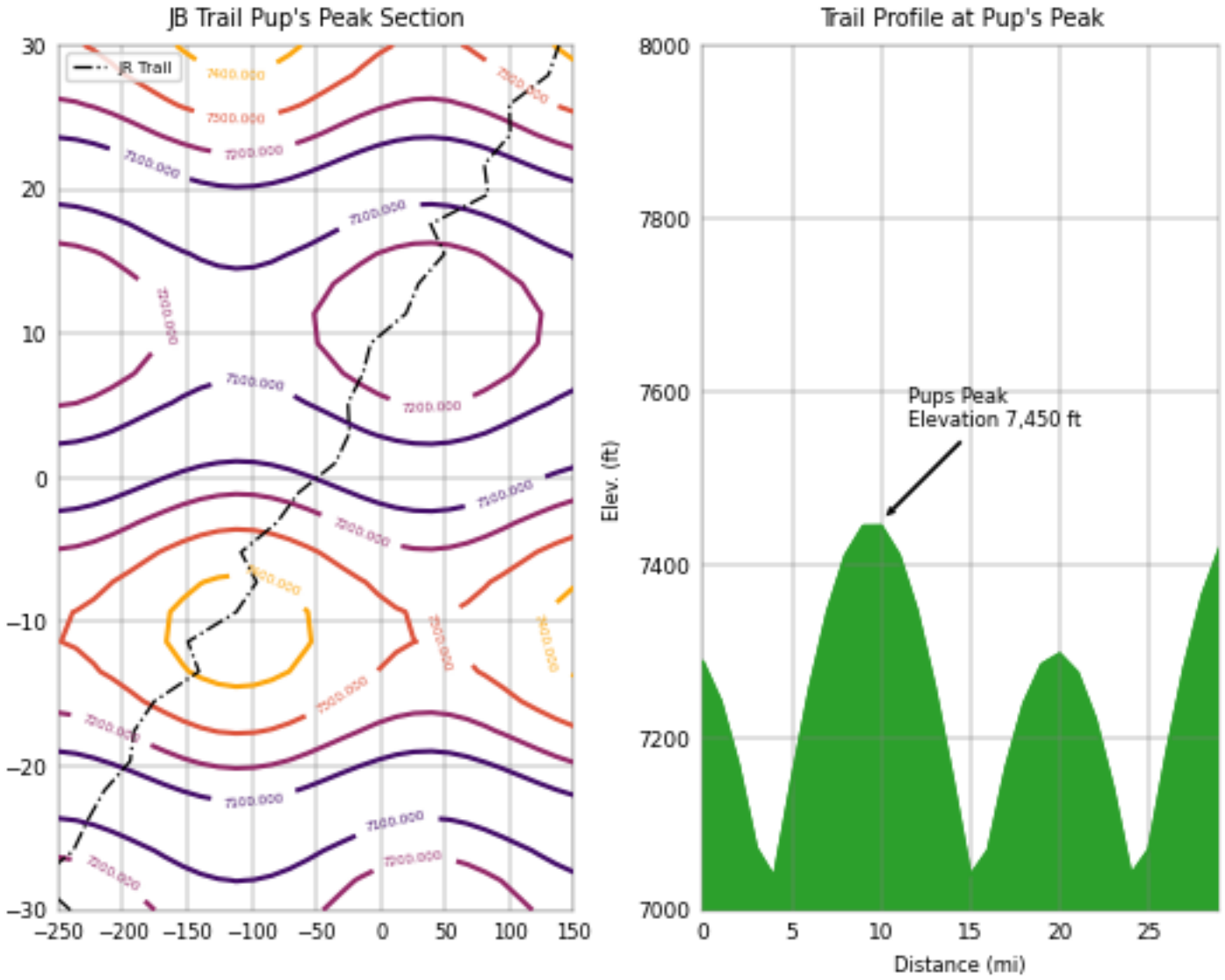


File: Part 4

Library: Seaborn

Data Source: Canada dataset
International migration flows to Canada

Complex Subplots



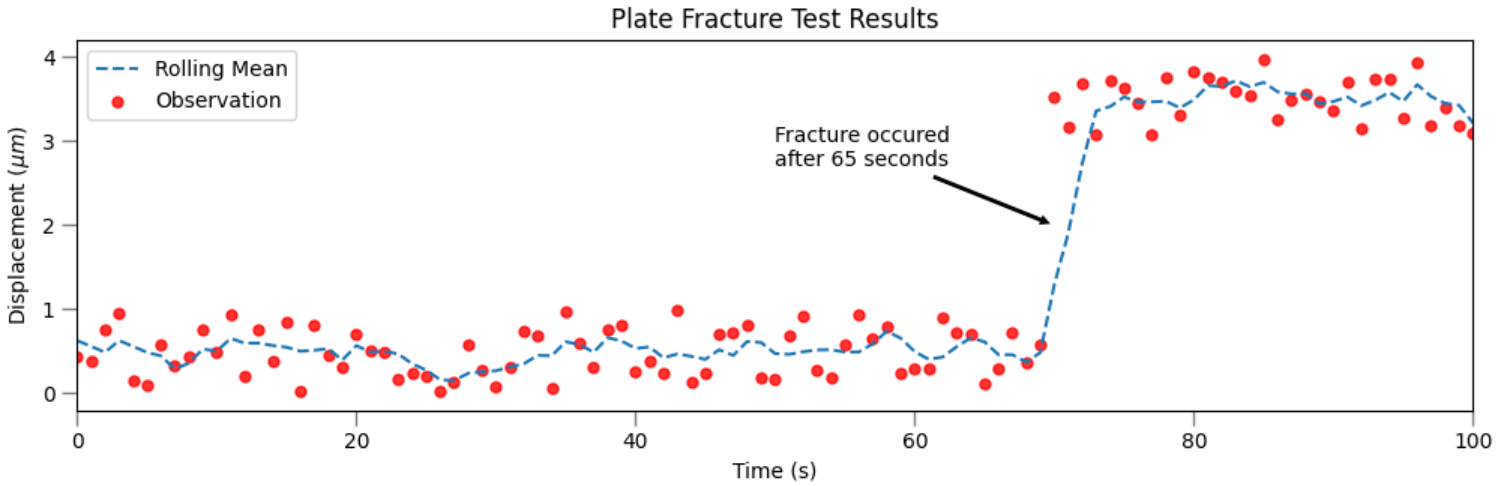
File: Part 5_A

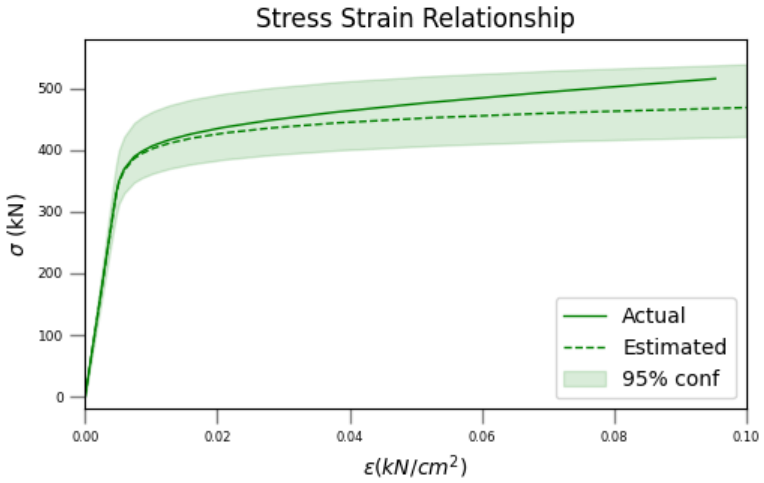
Library: Matplotlib

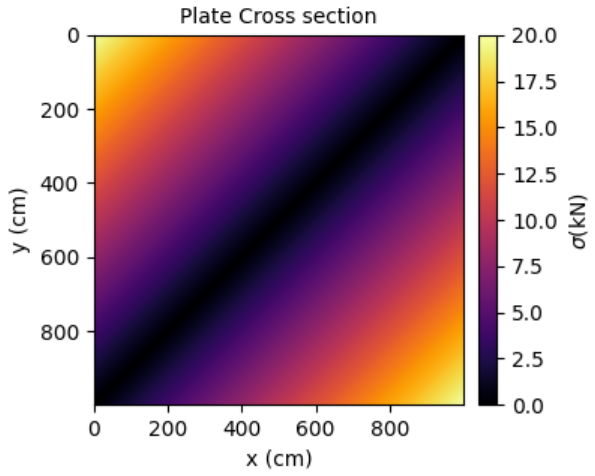
Data Source:

Generated in the code

Complex Subplots







File: Part 5_B

Library: Matplotlib

Data Source:

stress_strain.csv (in folder)

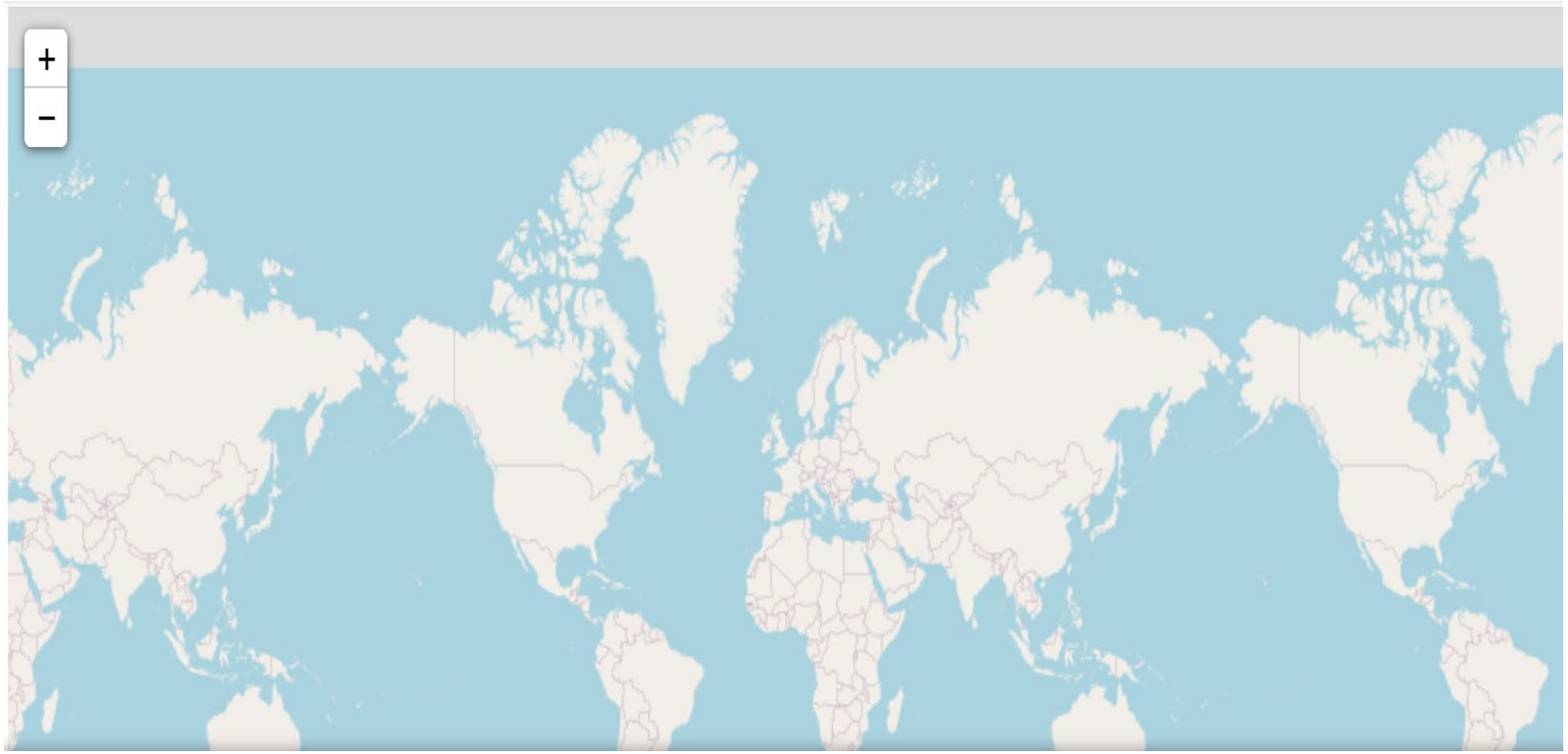
Discussion Advanced Charts

- ***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: One 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 while Matplotlib accomplishes these separately. In addition, Seaborn includes the confidence level (the shaded area) along the regression line, which Matplotlib doesn't do.
- ***Complex Subplots*** are helpful when multiple aspects of the data that need to be considered 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).

Geospatial Charts

(File: Part 6**)**

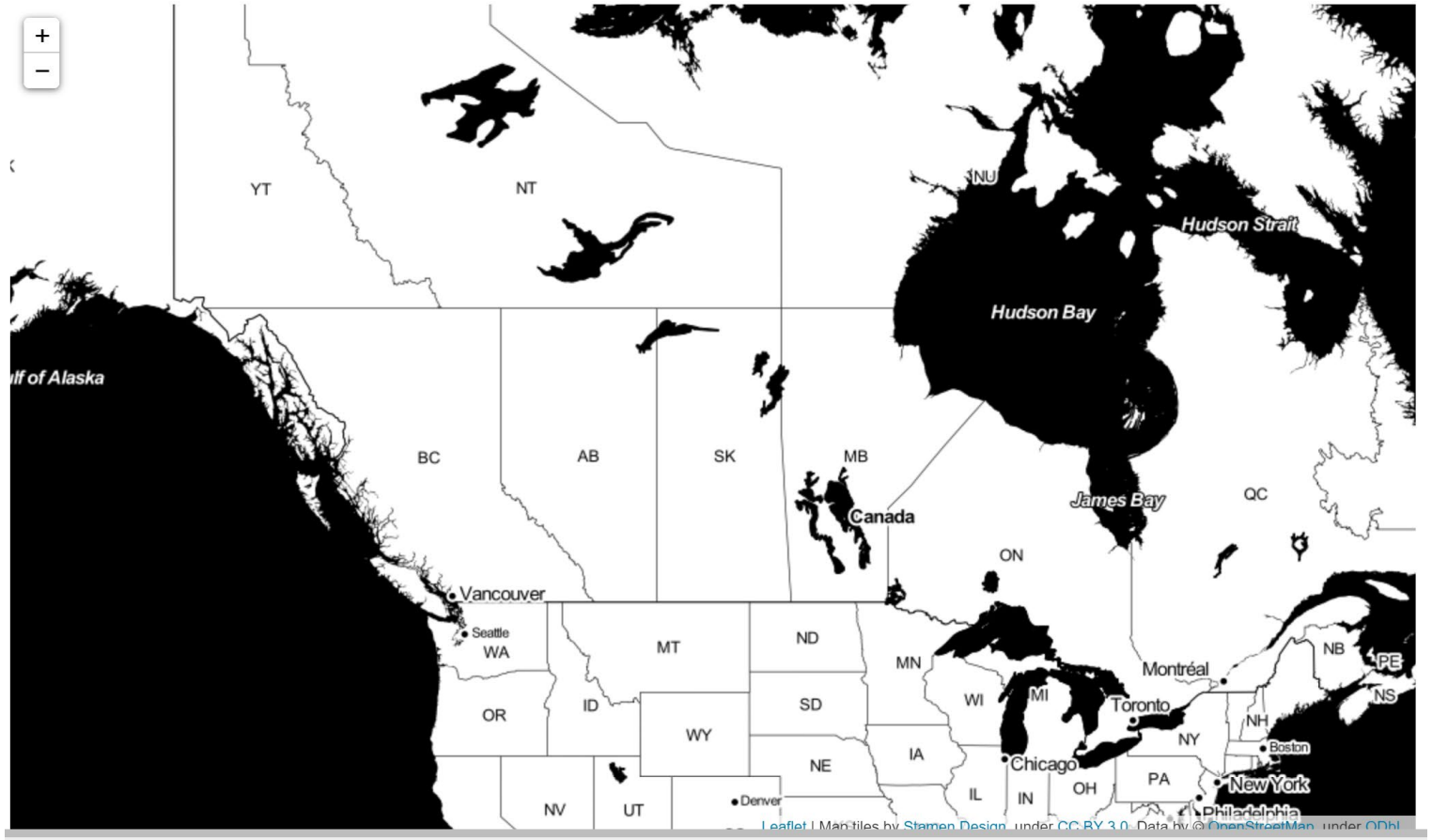
World Map



File: Part 6

Library: Folium

Stamen Toner Map



File: Part 6

Library: Folium

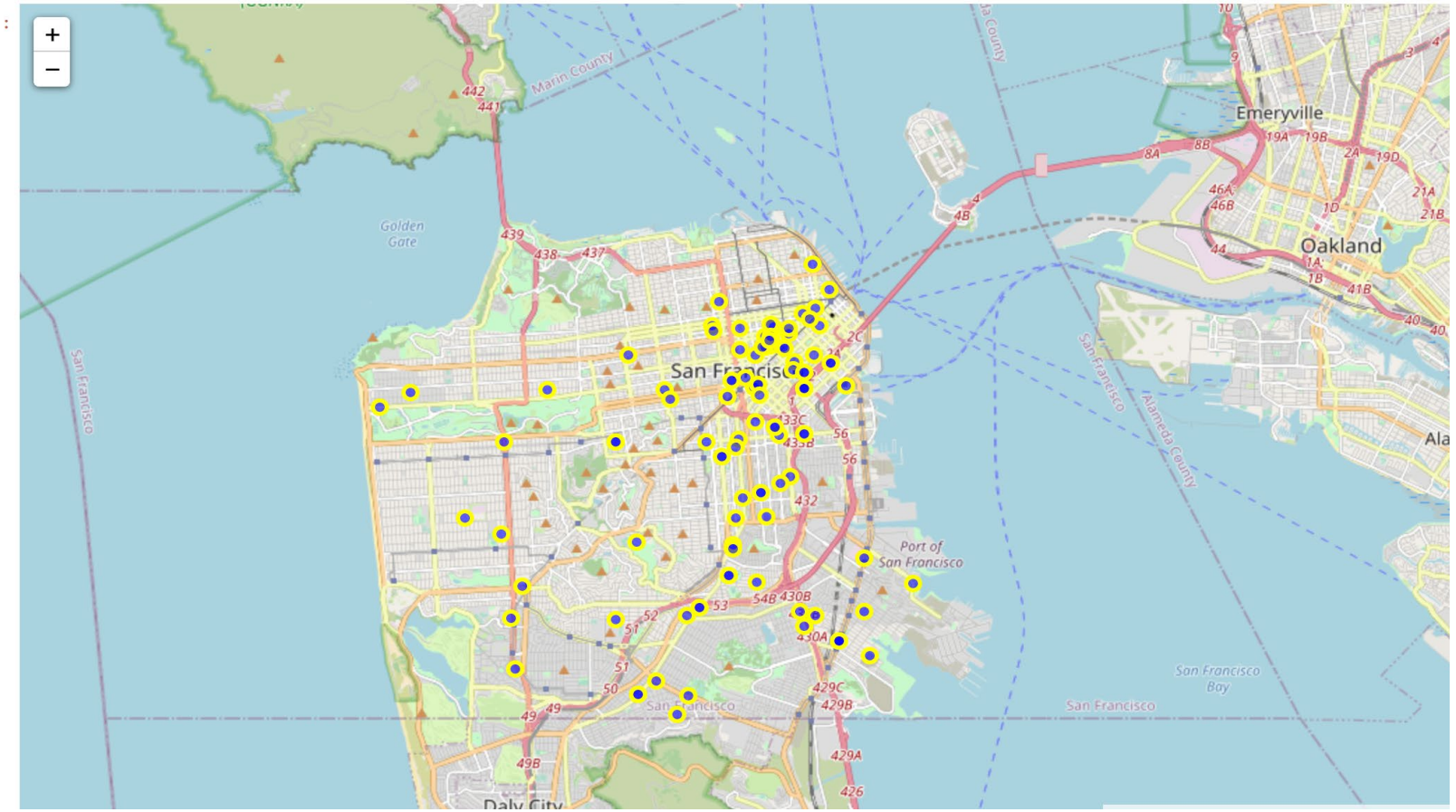
Stamen Terrain Map

File: Part 6

Library: Folium

Map with Markers

Crimes near San Francisco, 2016



*Information like crime categories can be added directly to the circle markers.

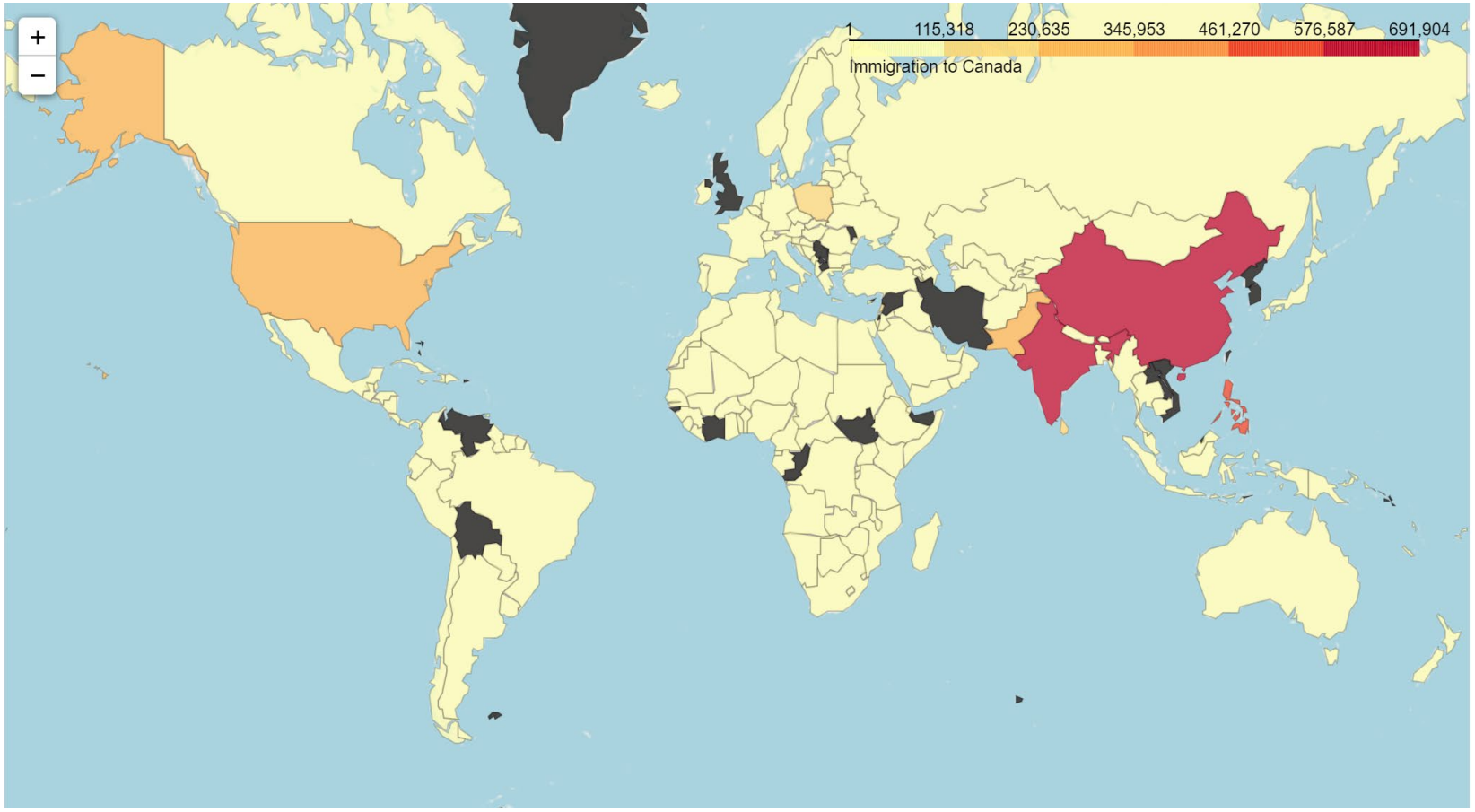
File: Part 6

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 Geospatial Charts

- 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 can be easily spotted as two major countries contributed the immigrant inflow to Canada.