Construction of User Interfaces (SE/ComS 319)

Ali Jannesari

Jinu Susan Kabala

Department of Computer Science

Iowa State University, Spring 2021

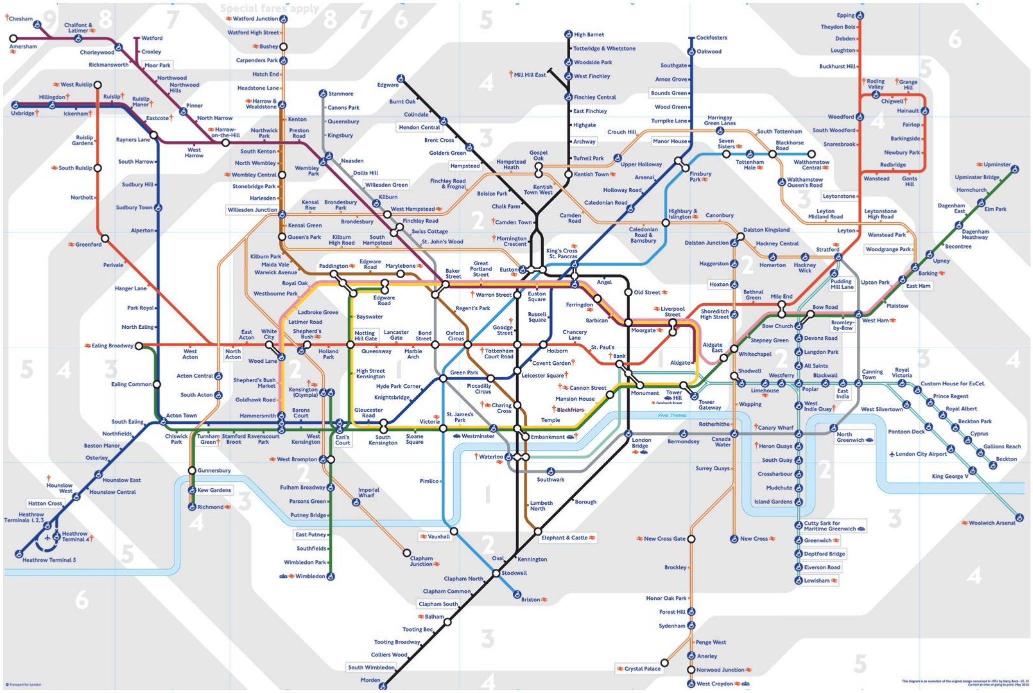
DATA VISUALIZATION

Outline

- Data Visualization

Data visualization

- Techniques for displaying large amounts of information
- "Use a picture. It's worth a thousand words."
  - Visual representation (e.g. Map of London Subway: too awkward to communicate in words)

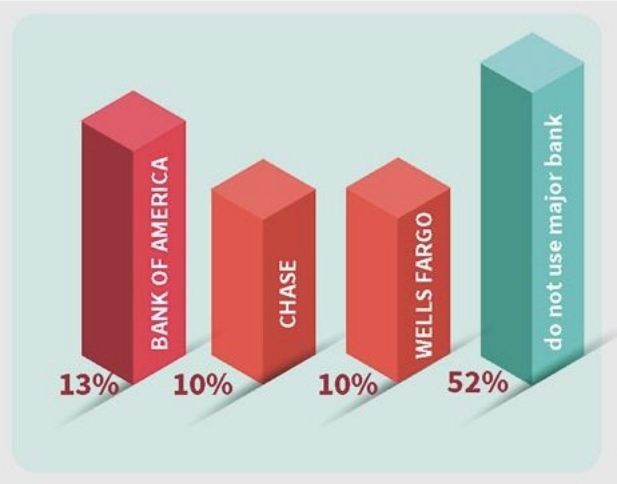


As the volume of data available to us increases **exponentially** in every field of endeavor, information visualization is becoming increasingly **important**!

Data visualization (2)

- May reveal relationships and/or trends of data
  - Could improve human problem-solving performance
  - Could influence business decisions
  - **Wrong inferences!!** (**Risks** of Visualization)
    - By choosing what information to represent and what information to leave out, there are now "lies, damned lies and information visualizations".

For example, a viewer looking at this chart might incorrectly assume that the majority of people use one of the top three US banking chains, instead of the other way around (neglecting the scale).

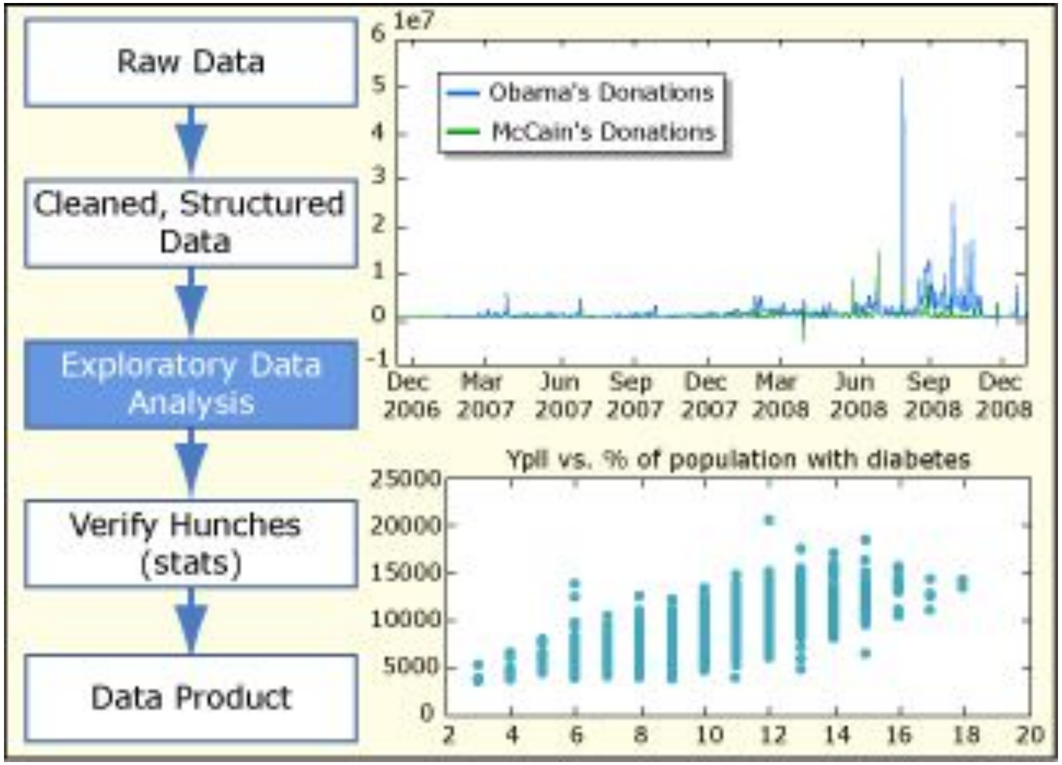


Why data visualization?

- Data exploration is easy
  - Use visualization to see where relationships in data may exist (**explorative analysis**)
- Easy to communicate and present idea
  - **Presentation** for Understanding or Persuasion
- Easy to share data and results with audience
- Easy to share findings with stakeholders
  - Use visualization to help confirm your understanding and analysis of data (**confirmation analysis**)

Explorative analysis

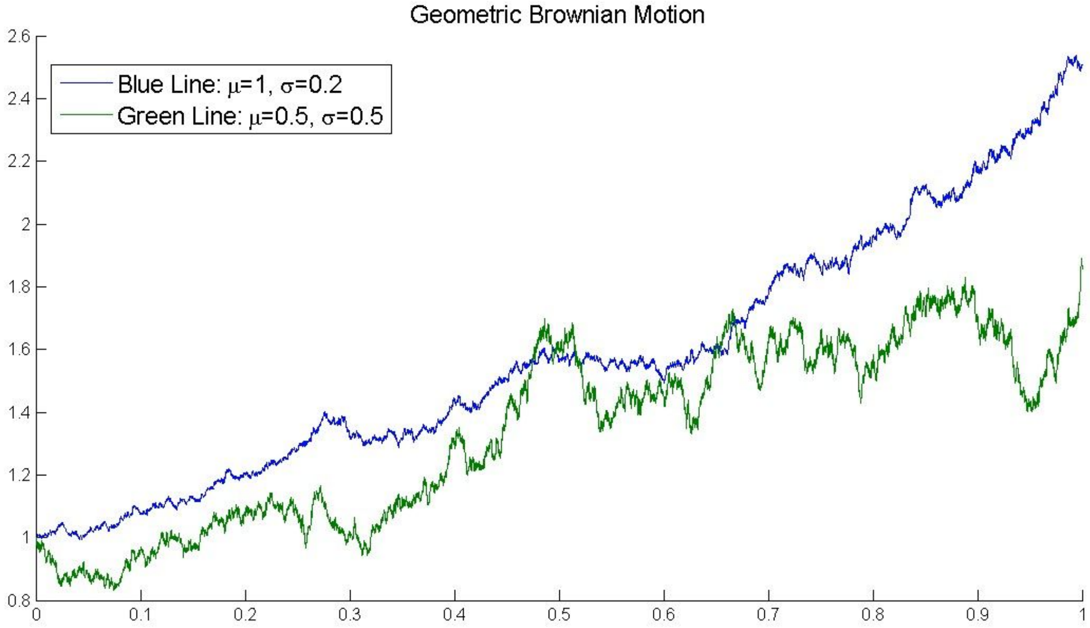
- Visualize data
- Calculate main characteristics
- Understand data and explore relationships and new hypothesis
  - Example: Campaign contributions and county health rankings



https://ocw.mit.edu/resources/res-6-009-how-to-process-analyze-and-visualize-data-january-iap-2012/

Confirmation analysis

- Confirmation analysis helps confirm our understanding and analysis of data
- Example: Brownian motion between sets of particles
- Confirm the break in the relationship towards the end of the graph



Basic visualization techniques

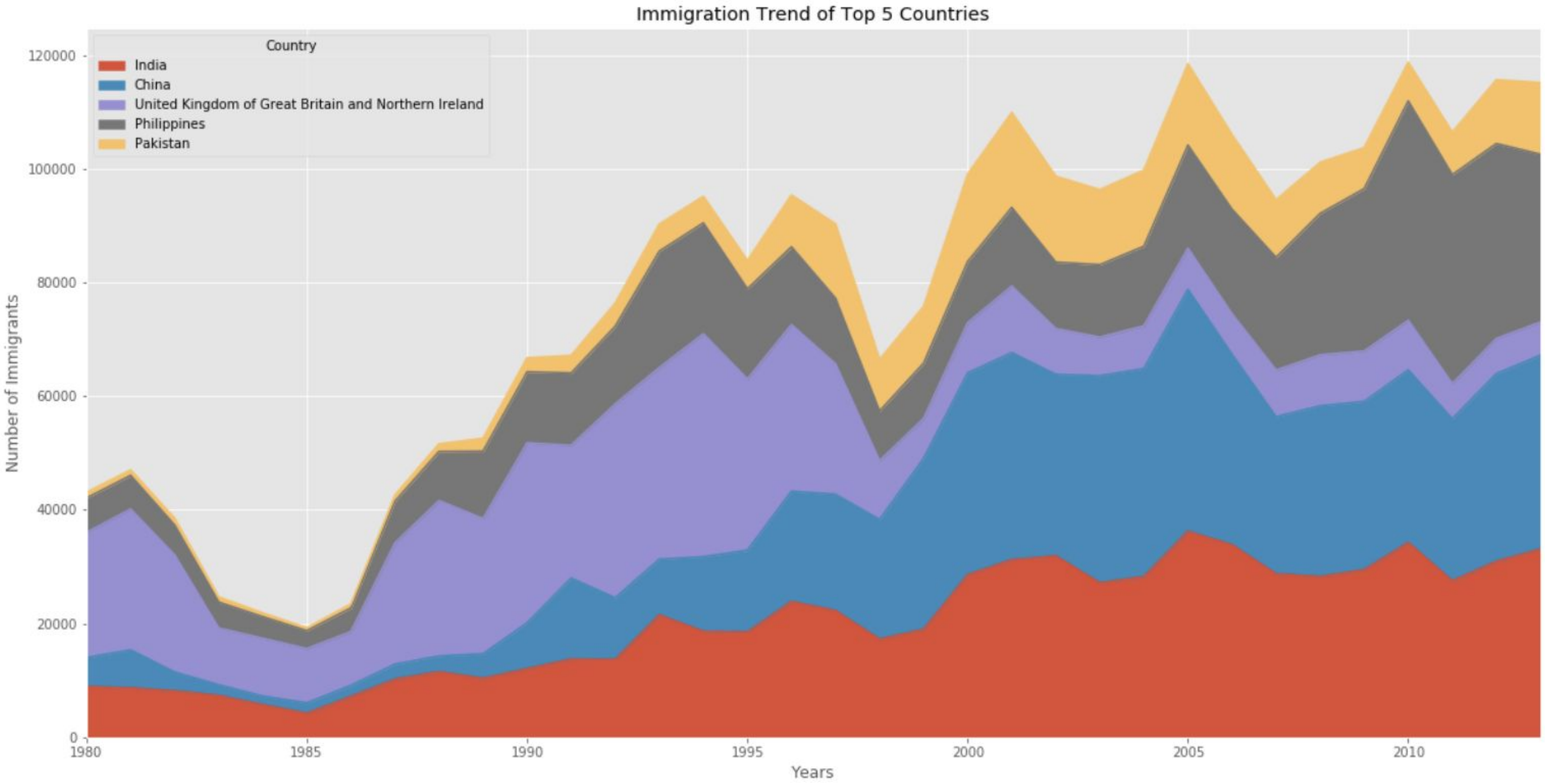
- Area plots
- Histograms
- Bar charts
- Pie charts
- Box plots
- Scatter plots

Area plots

- Also known as area charts or area graphs or stacked line plots
- Commonly used to represent cumulated totals using numbers or percentages over time
- Is based on the line plots
- More often used to compare two or more quantities

Area plots – Example

- **Example**: Area plots of countries with highest number of immigrants to Canada

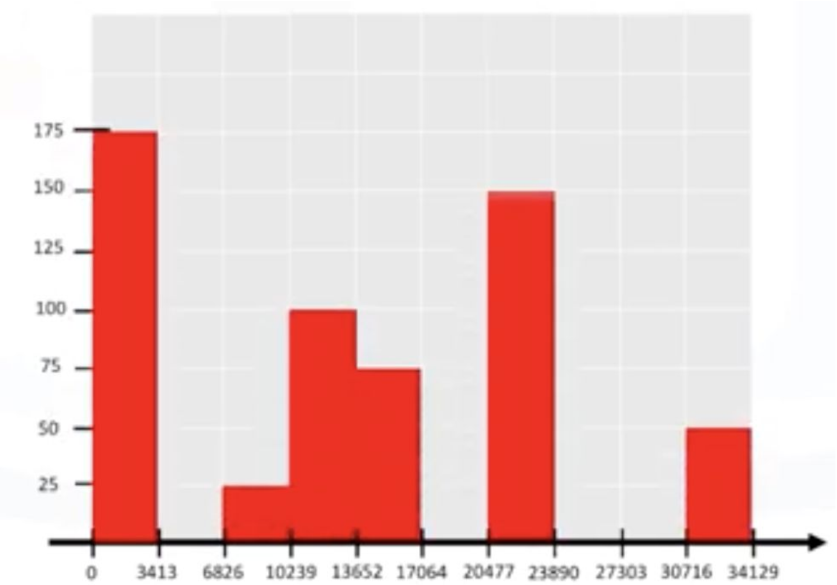


Histograms

- A histogram is a way of representing the frequency distribution of a variable
- It partitions the spread of the numeric data into bins
- Assign each data point in dataset to a bin
- Counts # of data points that have been assigned to a bin
- Y-Axis is frequency or # of data points in each bin

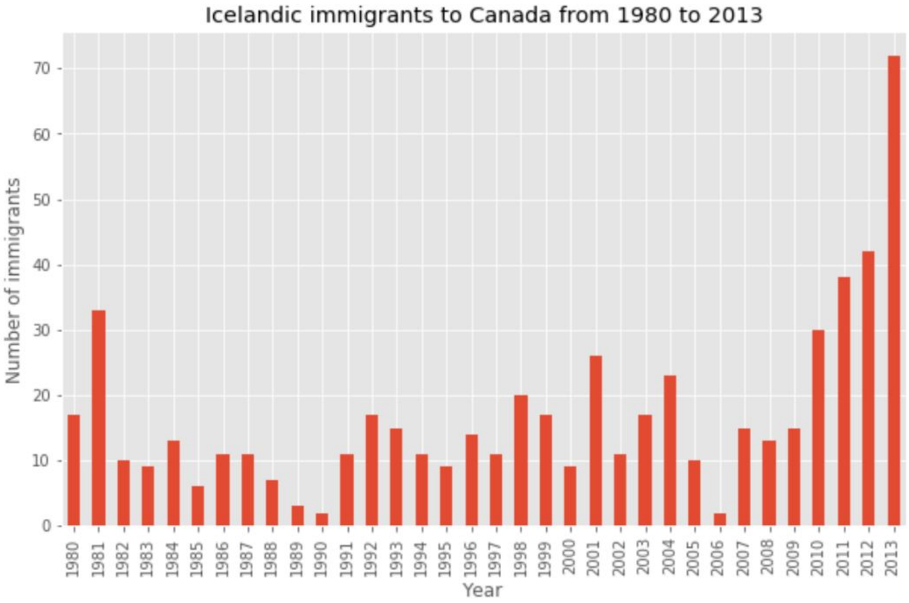
Histograms – Example

- Let us suppose a range of numeric value is 34,129
- Partition the horizontal axis of equal size
- Histogram is built on # of data points belong to these bins
- If no data point belongs to a bin, its height is 0



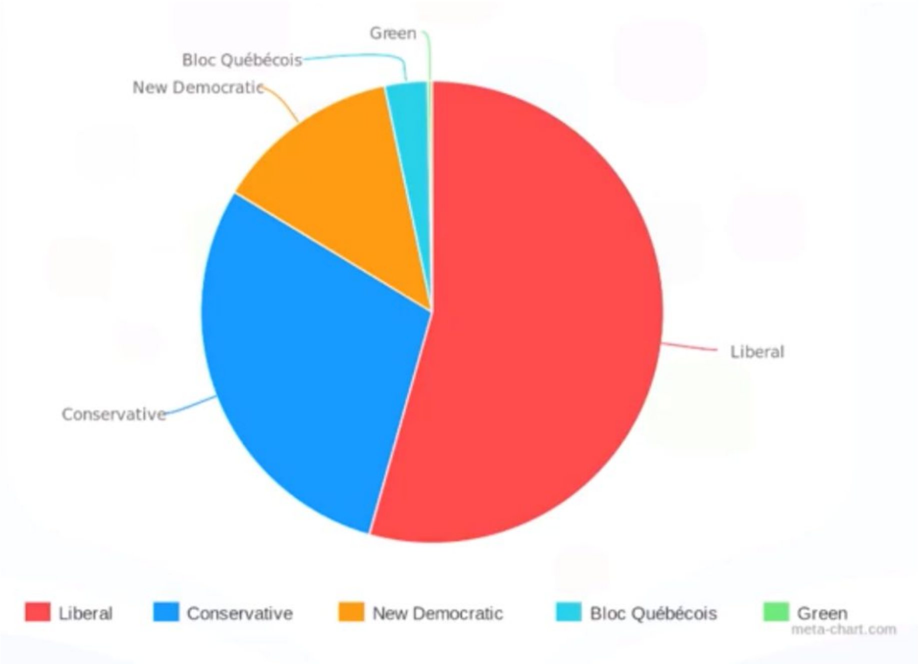
Bar charts

- A bar chart is commonly used to compare the values of a variable at a given point of time
- Each bar is proportional to the value of a variable at a **given point in time**
- **Example**: Immigration from Iceland to Canada from 1980 to 2013



Pie charts

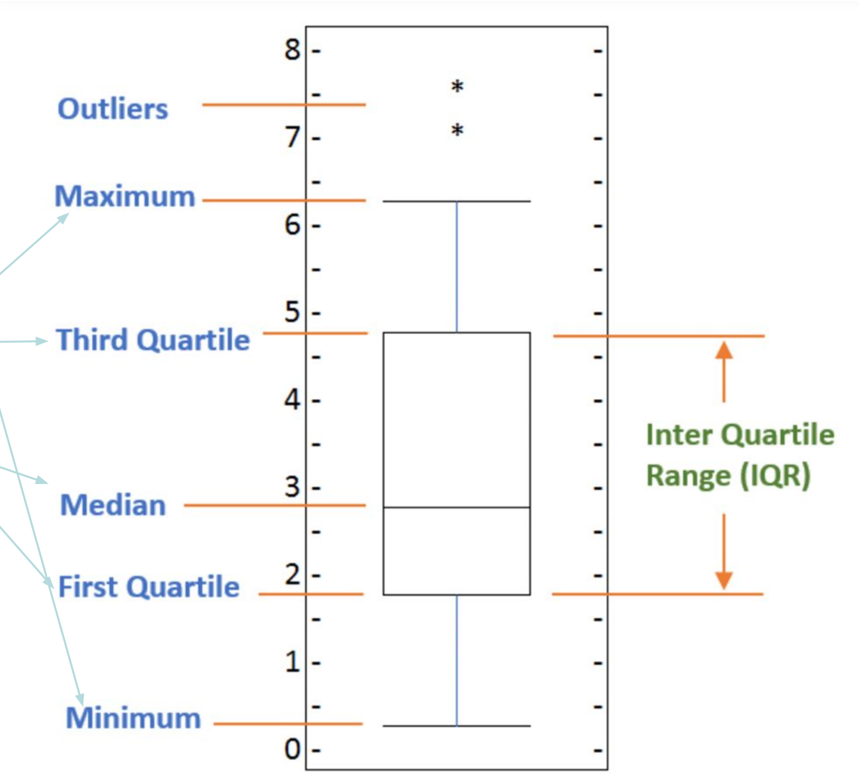
- A pie chart is a circular statistical graphic divided into slices to illustrate numerical proportion.
- **Example**: A pie chart of the Canadian federal election back in 2015.



Box plots

- A boxplot is a way of statistically representing the distribution of given data through five main dimensions.

- **Minimum:** Smallest number in the dataset.
- **First quartile:** Middle number between the minimum and the median.
- **Second quartile (Median):** Middle number of the (sorted) dataset.
- **Third quartile:** Middle number between median and maximum.
- **Maximum:** Highest number in the dataset.

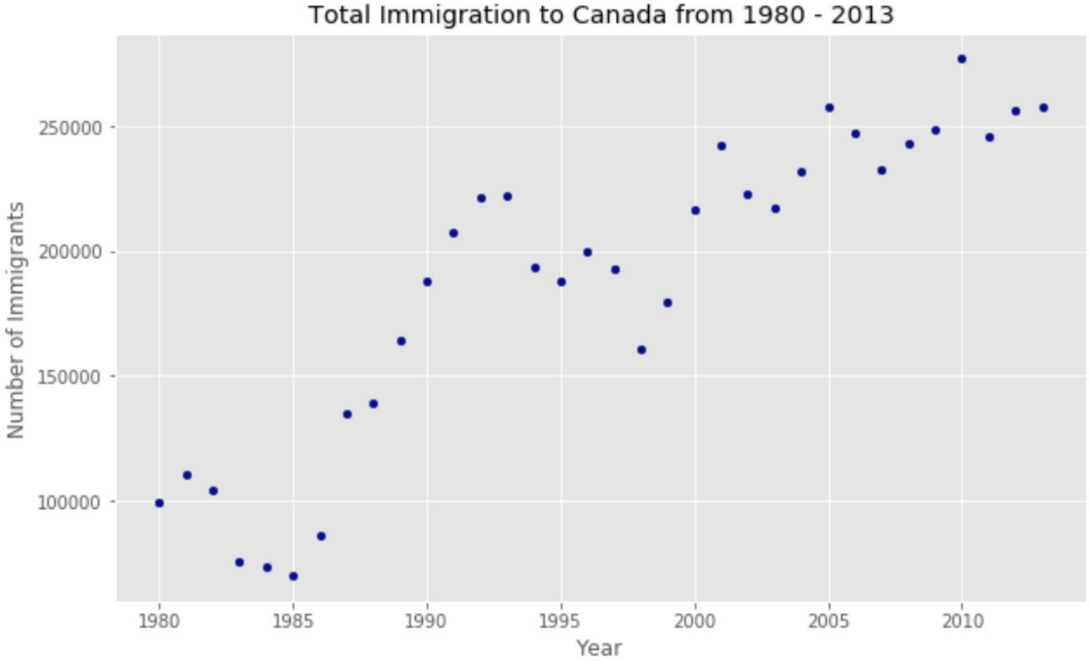


Scatter plots

- A scatter plot is a type of plot that displays values pertaining to typically two variables against each other.
- Usually it is a dependent variable to be plotted against an independent variable in order to determine if any correlation between the two variables exists.

Scatter plots – Example

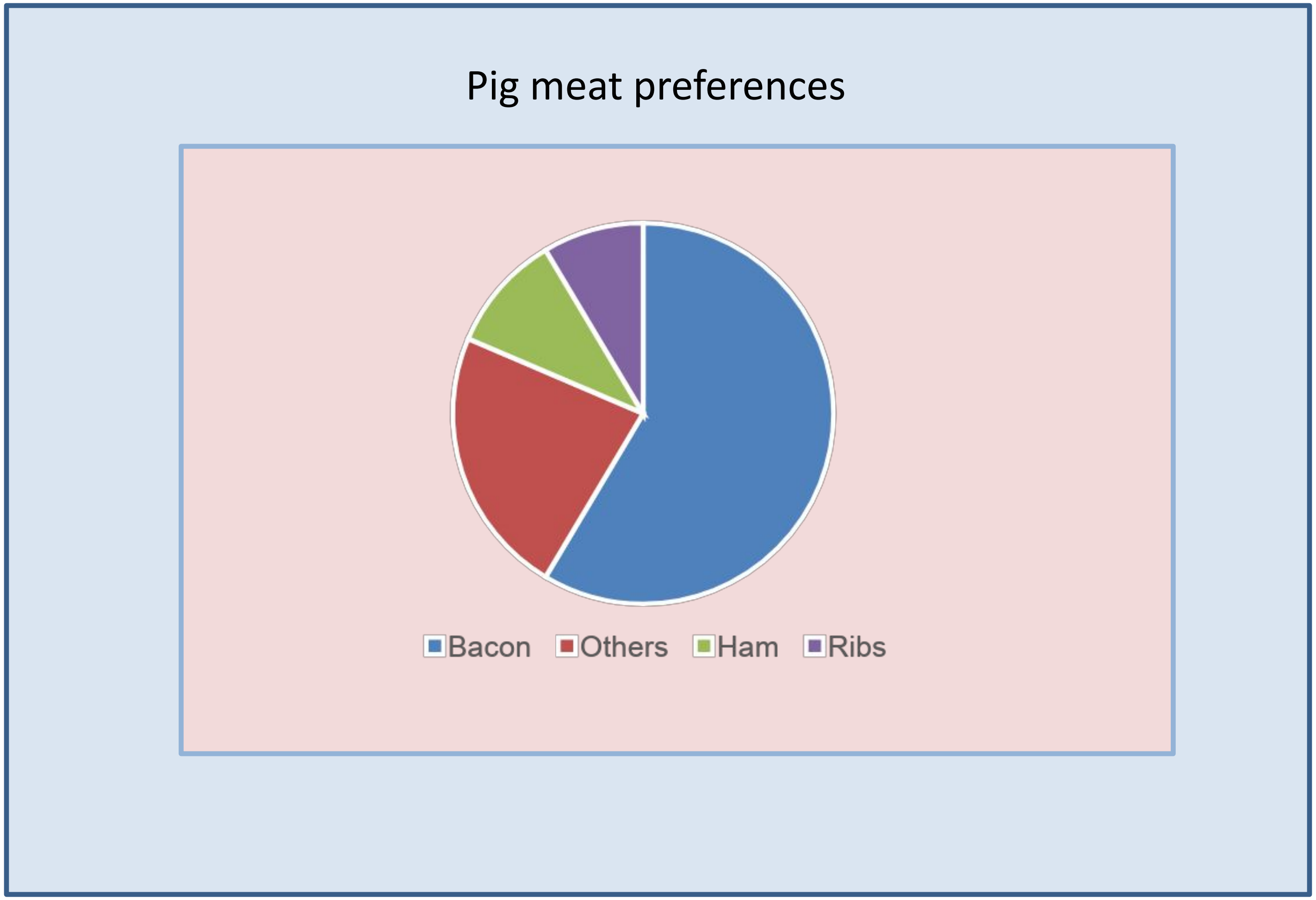
- **Example**: a scatter plot of the total annual immigration to Canada from 1980 to 2013.



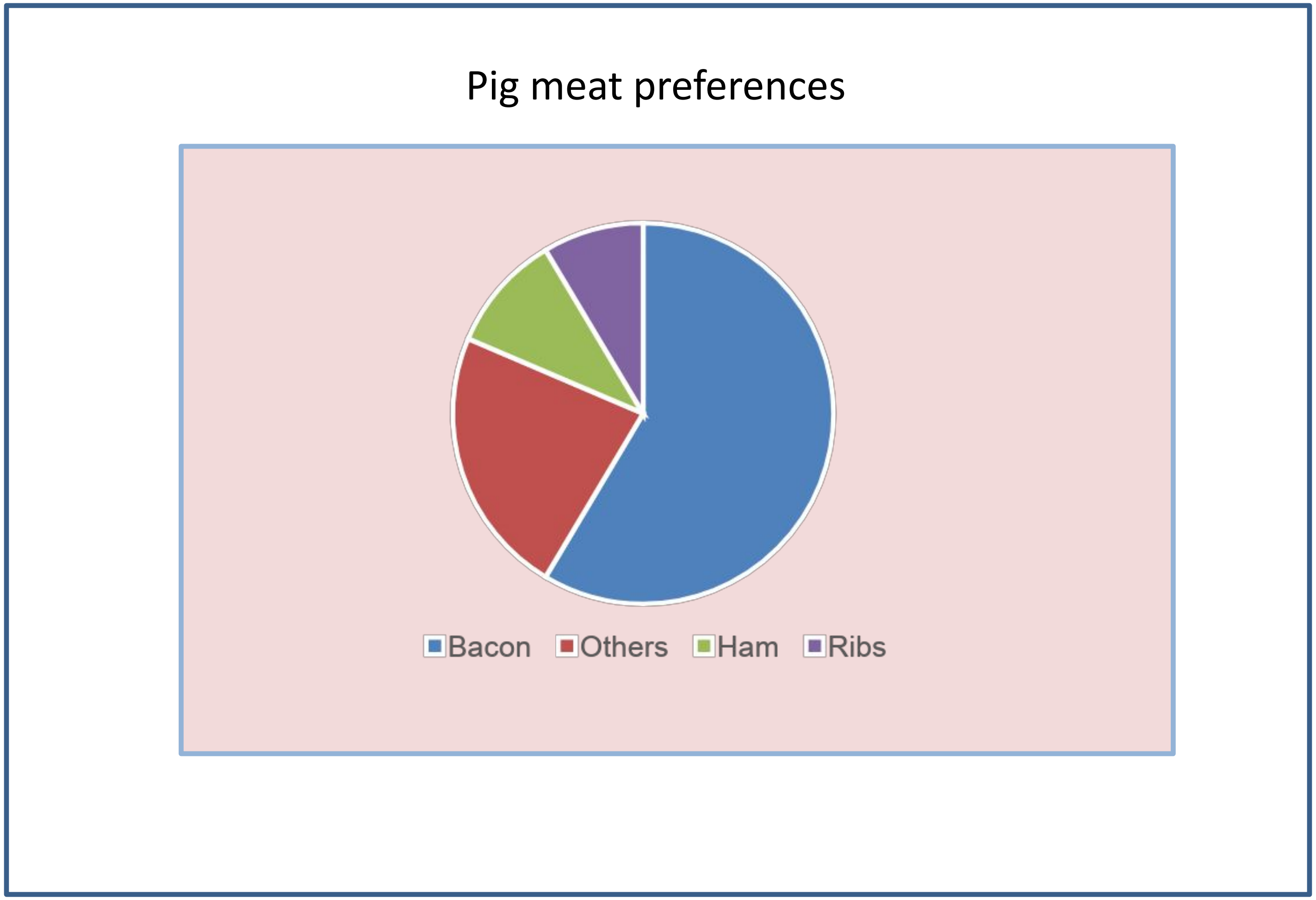
Best practices for data visualization

- Less is more effective
- Less is more attractive
- Less is more impactive

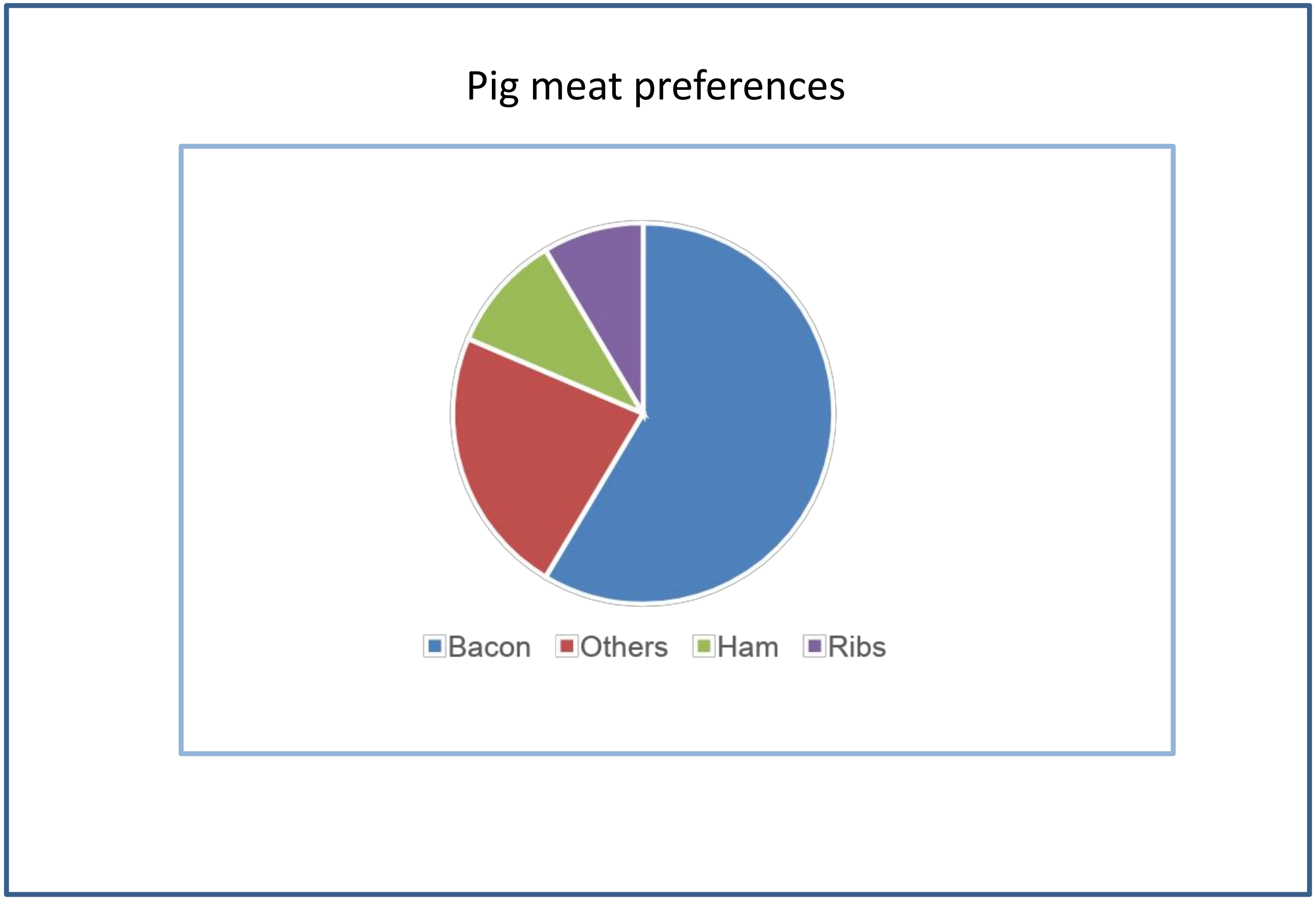
Best practices – Example (1)



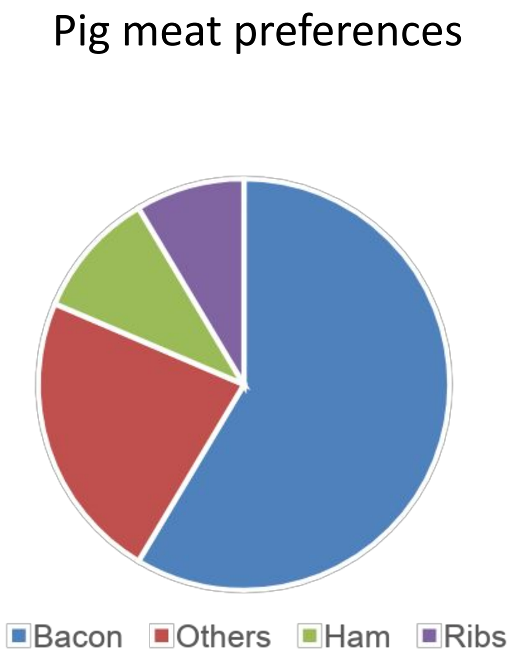
Best practices – Example (2)



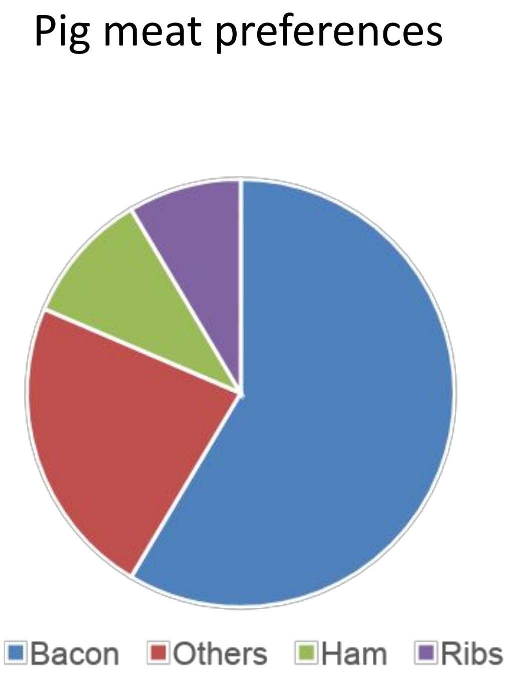
Best practices – Example (3)



Best practices – Example (4)

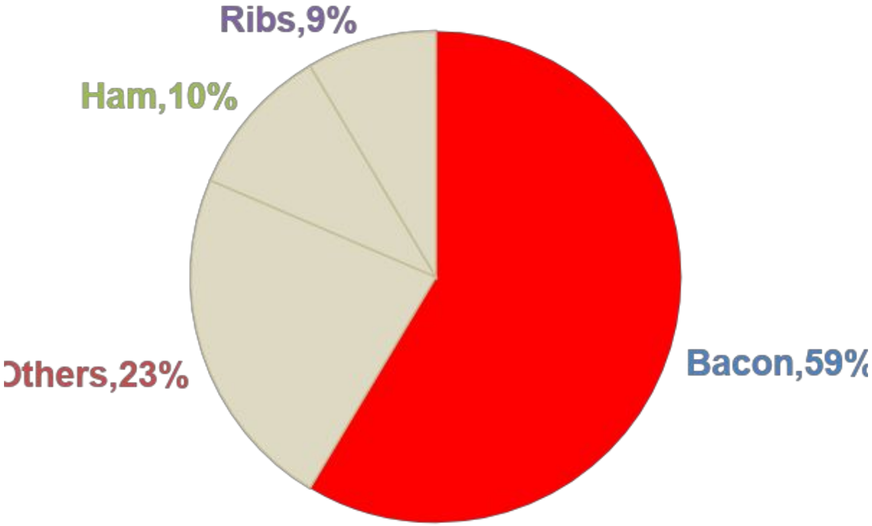


Best practices – Example (5)



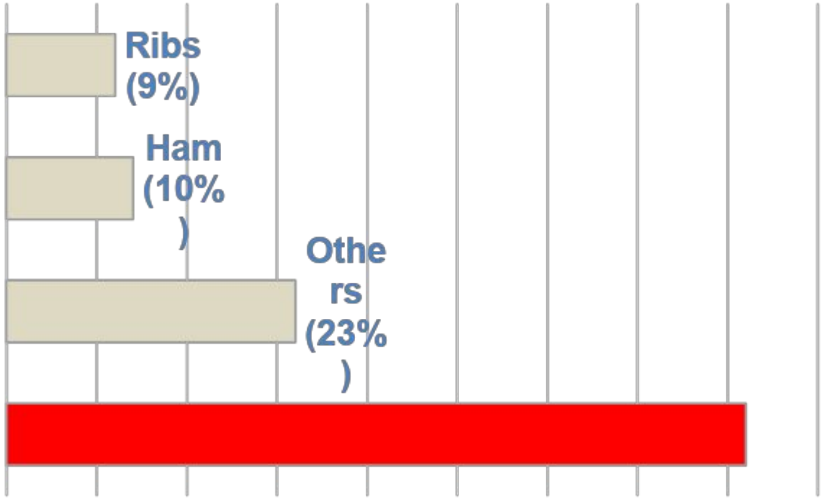
Best practices – Example (6)

Pig meat preferences



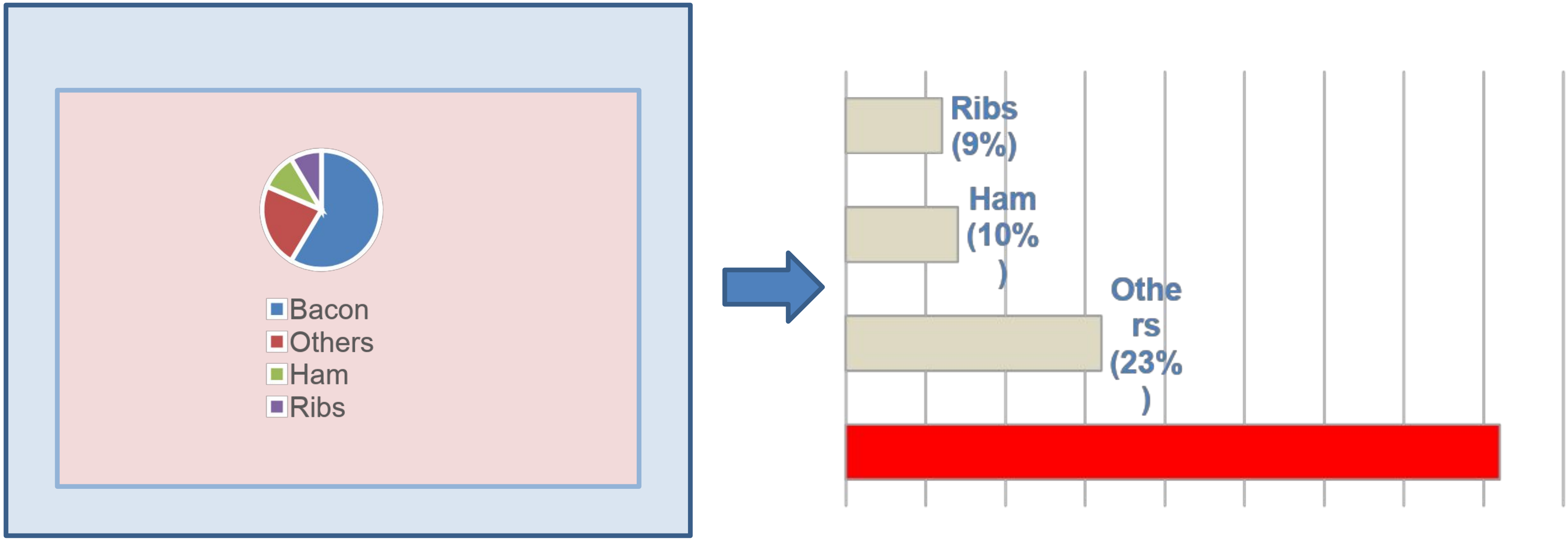
Best practices – Example (7)

Pig meat preferences



Best practices – Example (8)

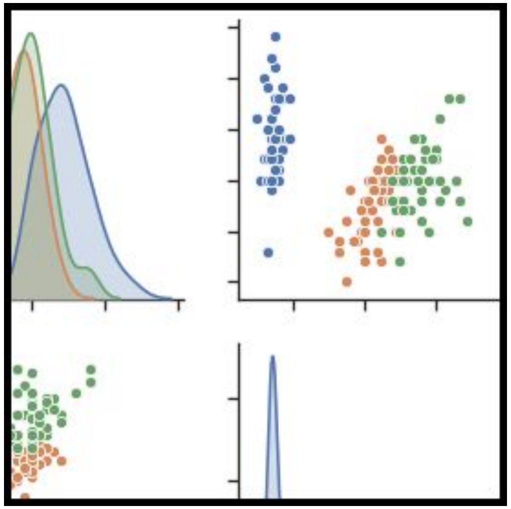
Pig meat preferences



Pie vs. bar and charts

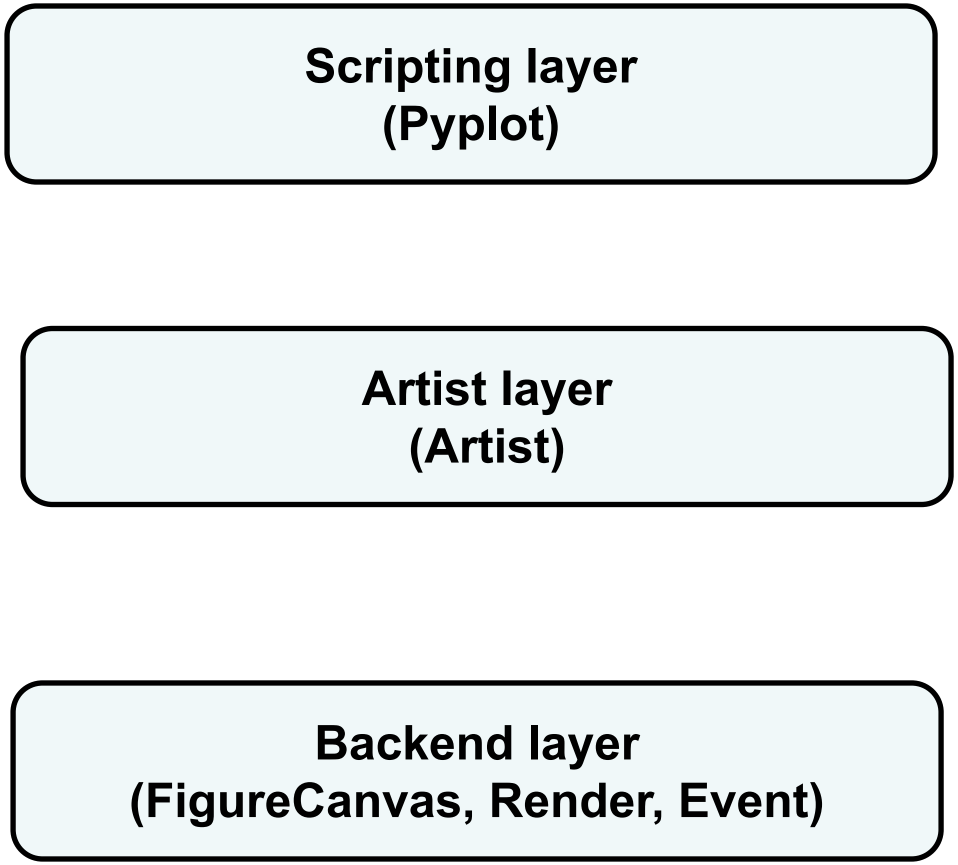
- Pie charts are less crunchy
- Bar and Charts are more crunchy and effective

Python libraries for visualization



- Seaborn
  - A Python data visualization library based on Matplotlib
  - It provides a high-level interface for drawing attractive and informative statistical graphics
  - https://seaborn.pydata.org
- Matplotlib
  - One of the most widely used, visualization library in Python
  - http://aosabook.org/en/matplotlib.html

Matplotlib architecture



Scripting layer
(Pyplot)

Automates the process of defining a canvas and defining a figure artist instance and connecting them

Artist layer
(Artist)

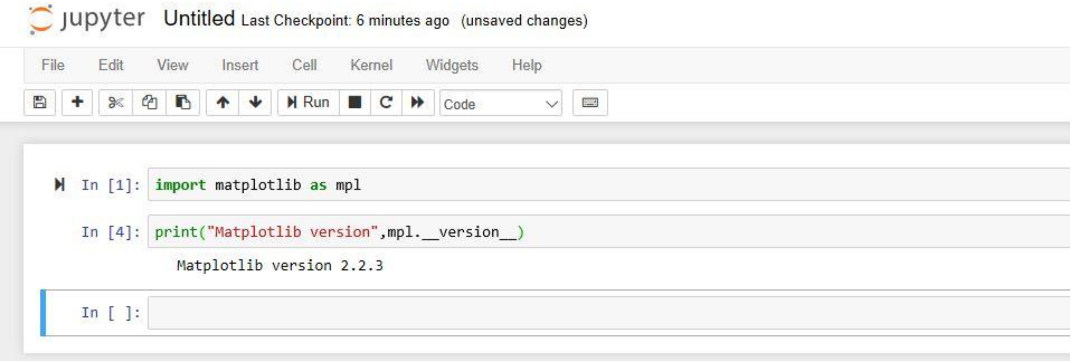
Knows how to use renderer to draw on canvas

Backend layer
(FigureCanvas, Render, Event)

Defines a canvas and knows how to draw figure on canvas

Jupyter notebook

- Open source web application that allows live code visualizations
- Numerical simulation, statistical modelling, data visualization, machine learning
- https://jupyter.org
- Jupyter has some specialized support for Matplotlib

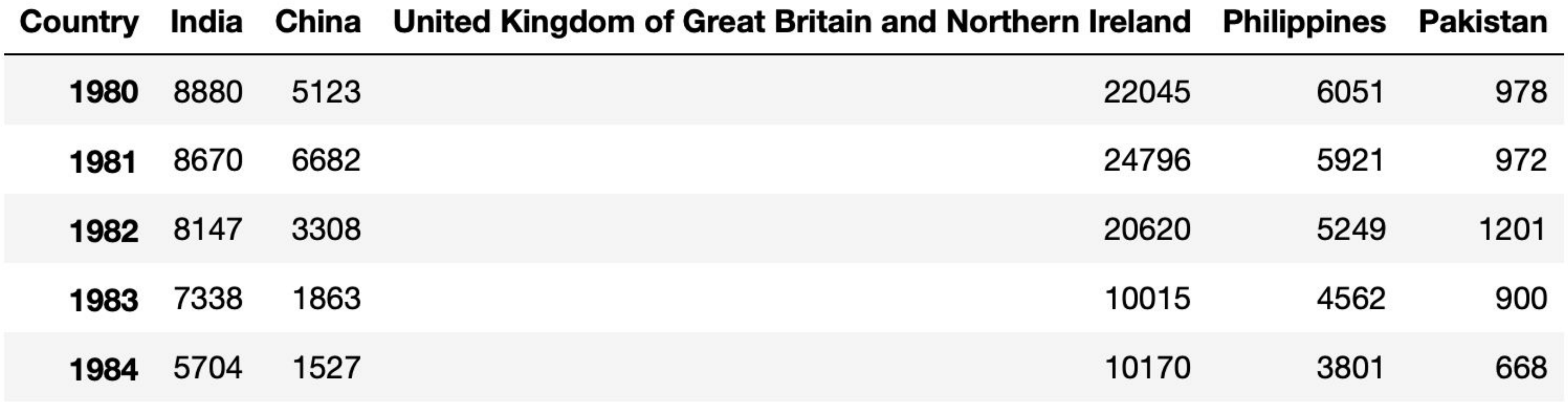


jupyter Untitled Last Checkpoint: 6 minutes ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run
Code

```
In [1]: import matplotlib as mpl

In [4]: print("Matplotlib version",mpl.__version__)
           Matplotlib version 2.2.3

In [ ]:
```

Self-study: Matplotlib example for area plot (1)

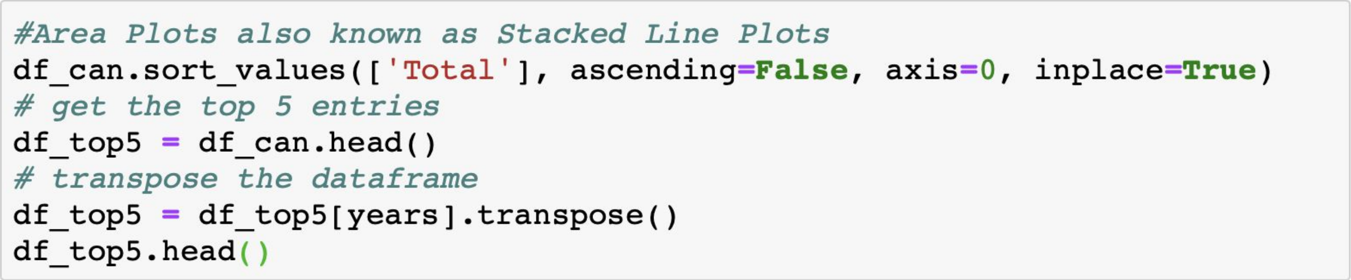
- **Dataset**: Total number of immigrants from all over the world to each of the 45 countries as well as other metadata pertaining to the immigrants countries of origin
- https://ibm.box.com/shared/static/lw190pt9zpy5bd1ptyg2aw15awomz9pu.xlsx



Country	India	China	United Kingdom of Great Britain and Northern Ireland	Philippines	Pakistan
1980	8880	5123	22045	6051	978
1981	8670	6682	24796	5921	972
1982	8147	3308	20620	5249	1201
1983	7338	1863	10015	4562	900
1984	5704	1527	10170	3801	668

Self-study: Matplotlib example for area plot (2)

- Generate the area plots of countries with highest number of immigrants to Canada
- Sort the data in descending order (sort_values)



```
#Area Plots also known as Stacked Line Plots
df_can.sort_values(['Total'], ascending=False, axis=0, inplace=True)
# get the top 5 entries
df_top5 = df_can.head()
# transpose the dataframe
df_top5 = df_top5[years].transpose()
df_top5.head()
```

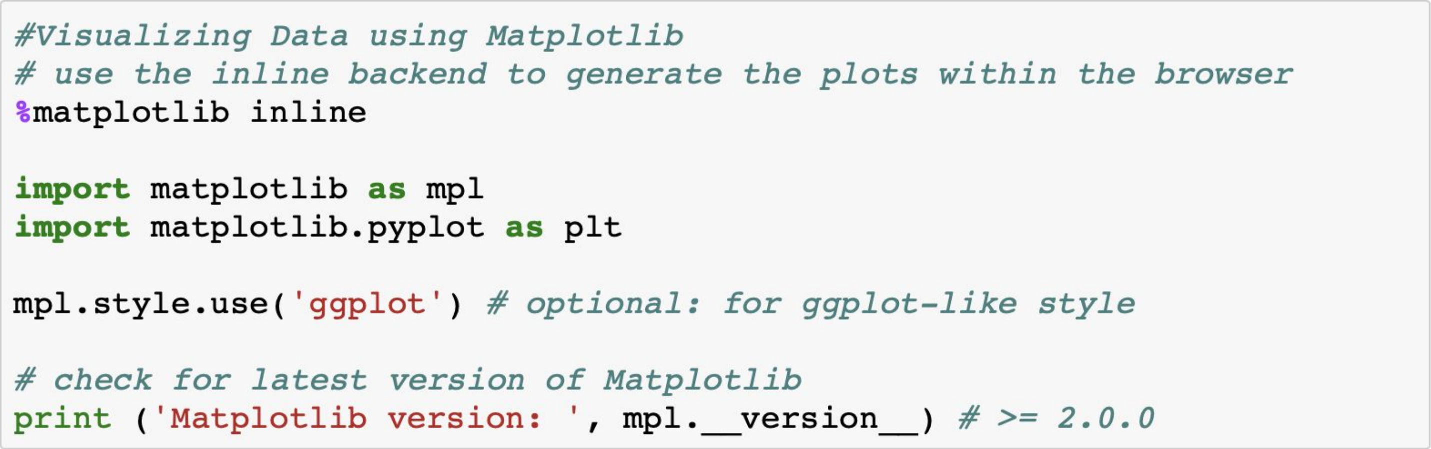
Self-study: Matplotlib example for area plot (3)

Generating area plots

- Need to create a new data frame and exclude rest of countries
- Years should be plotted on horizontal axis
- # of immigrants should be plotted on vertical axis
- Matplotlib plot indices on horizontal axis i.e. Countries
- Transpose function

Self-study: Matplotlib example for area plot (4)

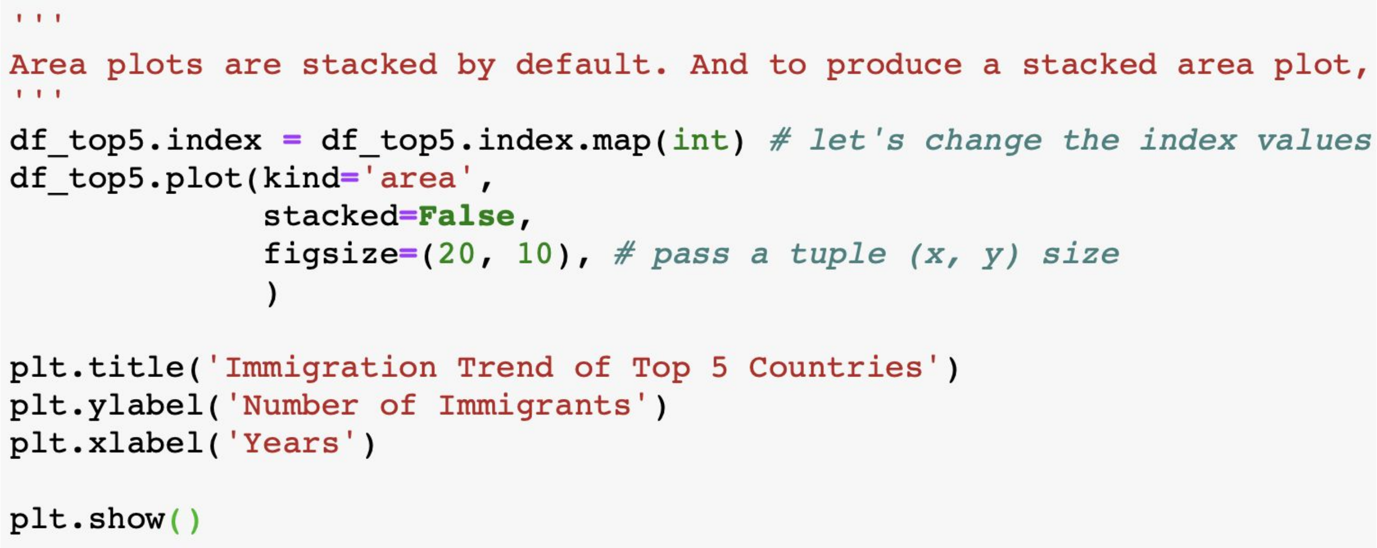
- Generate the required data frame



```
#Visualizing Data using Matplotlib
# use the inline backend to generate the plots within the browser
%matplotlib inline

import matplotlib as mpl
import matplotlib.pyplot as plt

mpl.style.use('ggplot') # optional: for ggplot-like style

# check for latest version of Matplotlib
print ('Matplotlib version: ', mpl.__version__) # >= 2.0.0
```

Self-study: Matplotlib example for area plot (5)

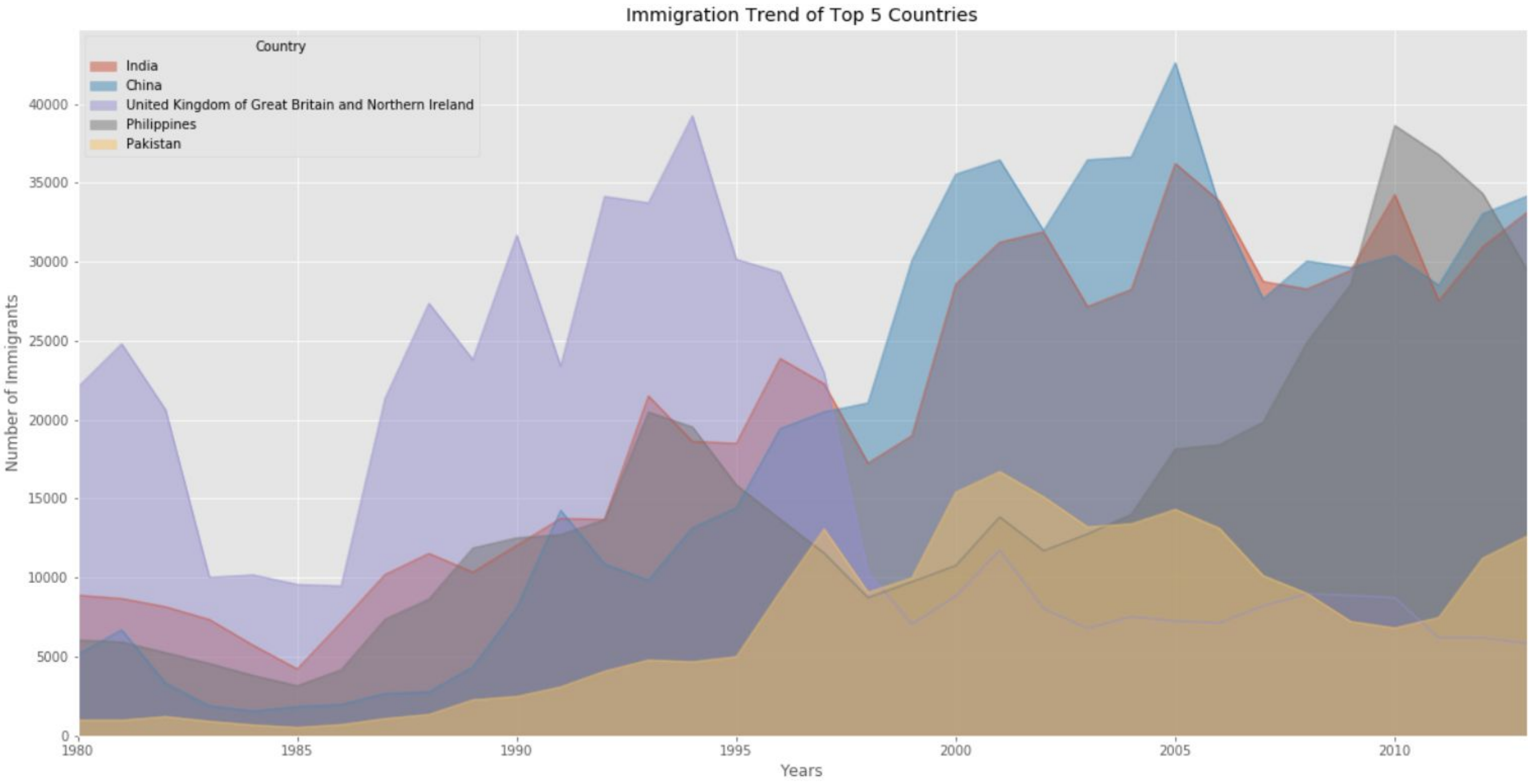
- Generate the area plot



```
'''
Area plots are stacked by default. And to produce a stacked area plot,
'''
df_top5.index = df_top5.index.map(int) # let's change the index values
df_top5.plot(kind='area',
             stacked=False,
             figsize=(20, 10), # pass a tuple (x, y) size
             )

plt.title('Immigration Trend of Top 5 Countries')
plt.ylabel('Number of Immigrants')
plt.xlabel('Years')

plt.show()
```

Self-study: Matplotlib example for area plot (6)

- Result of area plot, stacked=False



Data driven user Experience

- Allow users to analyze large amounts of data interactively using a very intuitive visual experience which help make decisions and further explorations
- Tools that are popular: Tableau, Power-BI
- Embed in web pages to offer a seamless user experience to present results or data in an interactive fashion
  - https://help.tableau.com/current/pro/desktop/en-us/embed.htm
  - https://powerbi.microsoft.com/en-us/blog/easily-embed-secure-power-bi-reports-in-your-internal-portals-or-websites/
- Some examples:
  - https://public.tableau.com/

References

- https://powerbi.microsoft.com/
- https://www.tableau.com