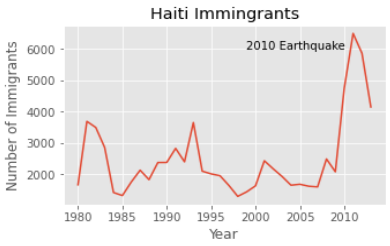
Nama : Ghina Khoerunnisa
Program : Data Scientist Intern

Basic Visualization

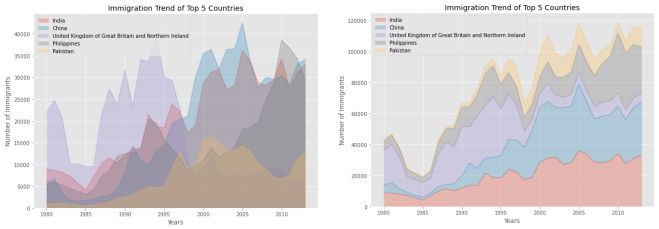
Visualisasi data merupakan suatu representasi visual dari data. Visualisasi data digunakan di data science untuk memperkuat storytelling, proses EDA, dan dapat digunakan untuk membuat keputusan pada saat pemilihan model. Sebelum memulai visualisasi data kita perlu melakukan import library yang dibutuhkan yaitu pandas dan matplotlib. Berikut merupakan jenis-jenis visualisasi dasar pada course ini:

- **Line Plot,** digunakan ketika kita ingin menampilkan perkembangan suatu variabel biasanya digunakan pada data dengan periode waktu tertentu. Untuk menampilkannya dapat menggunakan .plot(kind='line'). Contohnya: Kita akan menampilkan perkembangan masuknya imigran Haiti ke Ka

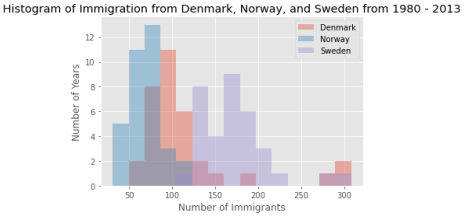
```
df_can.loc['Haiti', years].plot(kind='line', figsize=(5,3))
plt.xlabel('Year')
plt.ylabel('Number of Immigrants')
plt.title('Haiti Immingrants')
plt.text(1999,6000,'2010 Earthquake')
plt.show()
```



- **Area Plot,** digunakan ketika kita ingin menekankan volume atau kuantitasnya. Untuk menampilkannya dapat menggunakan .plot(kind='area'). Pada area plot terdapat atribut stacked yang jika False akan menampilkan plot seperti line plot tetapi terdapat area di bawahnya, sedangkan jika True akan menampilkan area plot yang bertumpuk sesuai dengan kuantitas variabelnya. Contohnya: Kita akan menampilkan top 5 penyumbang imigran terbanyak di Kanada. Sebelah kiri merupakan area plot dengan stacked False dan yang di sebelah kanan dengan stacked True.

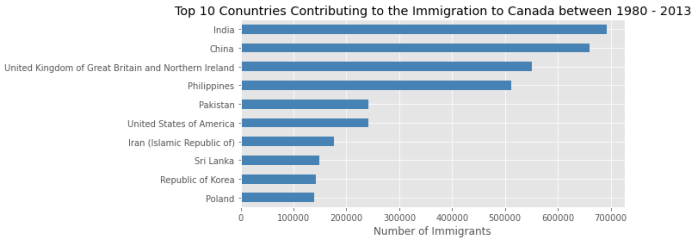


- **Histogram,** digunakan ketika kita ingin menampilkan distribusi frekuensi. Untuk menampilkannya dapat menggunakan .plot(kind='hist') dan kita dapat mengatur binsnya serta stacked atau tidak sesuai dengan kebutuhan. Contohnya: Kita akan menampilkan distribusi frekuensi imigran dari Denmark, Norway, dan Sweden dengan binsnya 15 dan stacked False.



- **Bar Chart,** biasanya digunakan untuk menampilkan data kategorikal. Untuk menampilkannya dapat menggunakan .plot(kind='bar') atau .plot(kind='barh') untuk bar yang horizontal. Contohnya:

```
df_top10.plot(kind='barh', figsize=(8, 4), color='steelblue')
```

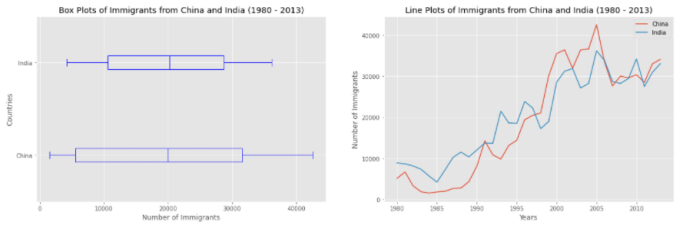


- **Pie Chart** merupakan diagram berbentuk lingkaran yang digunakan untuk menampilkan persentase yang totalnya harus mencapai 100%. Untuk menampilkannya dapat menggunakan .plot(kind='pie').
- **Box Plot** merepresentasikan distribusi data melalui lima dimensi utama. Pada data science, biasanya box plot digunakan untuk mendeteksi outlier. Untuk menampilkannya dapat menggunakan .plot(kind='box').
- **Subplot** digunakan pada saat kita akan menampilkan grafik yang berisi beberapa grafik. Untuk menggunakannya pertama kita membuat figure dengan plt.figure(). Kemudian menambahkan subplot ke figure dengan .add_subplot(row, column, plot ke berapa). Misalnya kita akan menampilkan dua plot yaitu box plot dan line plot,

```
fig = plt.figure() # create figure
ax0 = fig.add_subplot(1, 2, 1) # add subplot 1 (1 row, 2 columns, first plot)
ax1 = fig.add_subplot(1, 2, 2) # add subplot 2 (1 row, 2 columns, second plot).
```

```
China_India.plot(kind='box', color='blue', vert=False, figsize=(20, 6), ax=ax0) # add to subplot 1
```

```
China_India.plot(kind='line', figsize=(20, 6), ax=ax1) # add to subplot 2
```



- **Scatter Plot** digunakan pada saat kita ingin melihat pola hubungan antar dua variabel yang berbeda. Misalnya korelasi antara harga rumah dengan luas area. Untuk menampilkannya dapat menggunakan .plot(kind='scatter', x=var_x, y=var_y).
- **Bubble Plot** merupakan plot yang mirip dengan scatter plot tetapi pada bubble plot terdapat dimensinya. Cocok untuk memvisualisasikan hubungan 3 variabel misalnya ada luas tanah sebagai variabel x, luas bangunan sebagai variabel y, dan price sebagai ukuran dari bubblenya.