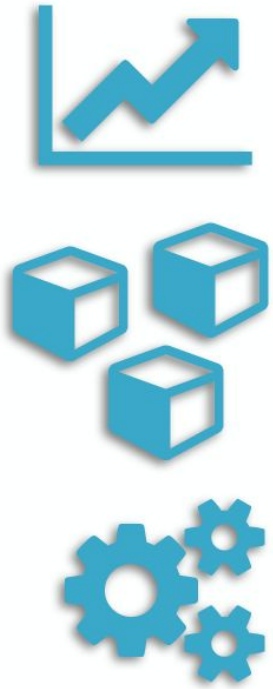
Big Data - Foundations and Applications

Lesson #3 - Intermediate Python for Data Science

Ivanovitch Silva

July, 2017

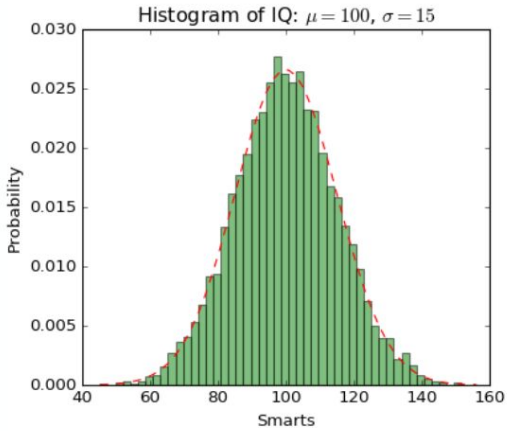
Agenda

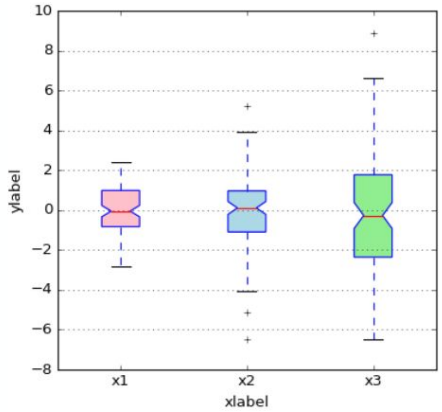


- Visualization
- Data Structures
- Control Structures

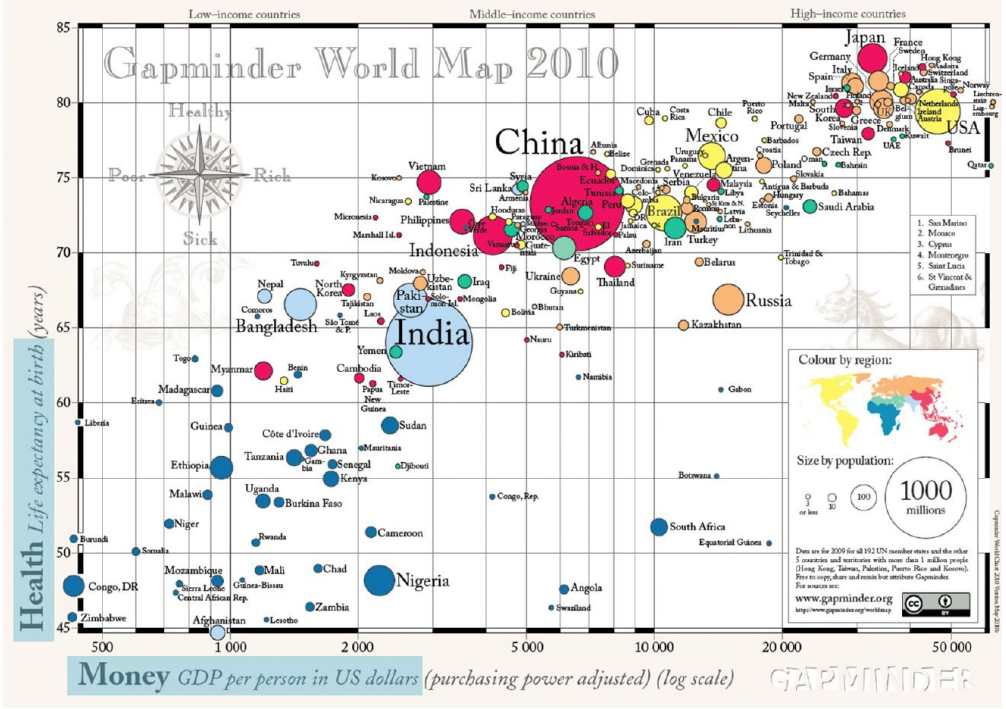
Data Visualization

- Very important in data analysis
  - Explore data
  - Report insights





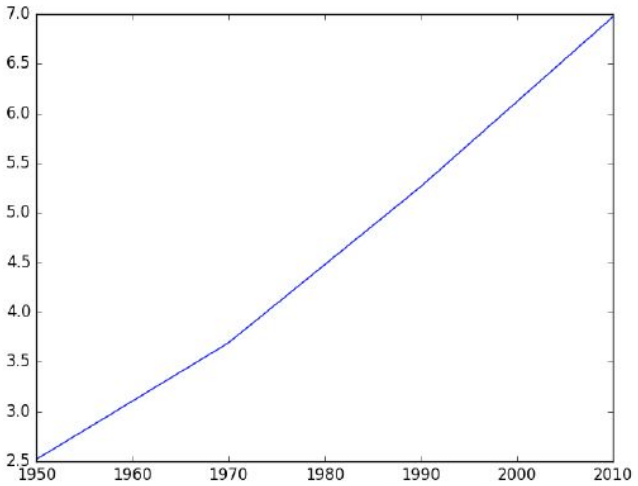
Data Visualization



http://www.gapminder.org/data/

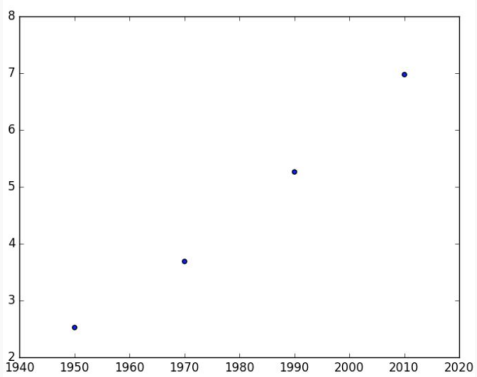
Matplotlib

```
In [1]: import matplotlib.pyplot as plt

In [2]: year = [1950, 1970, 1990, 2010]

In [3]: pop = [2.519, 3.692, 5.263, 6.972]

In [4]: plt.plot(year, pop)
                 x     y

In [5]: plt.show()
```



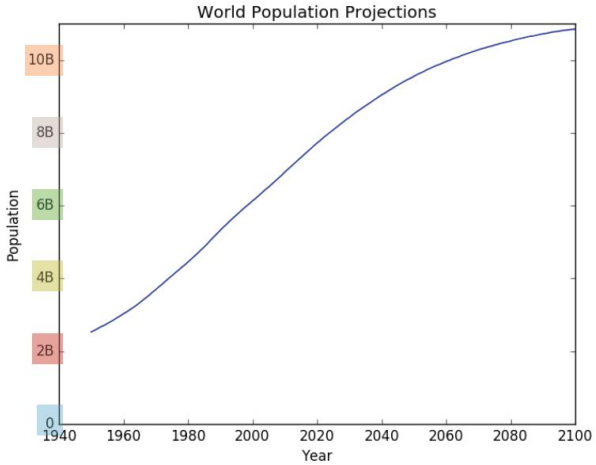
Scatter plot

```
In [1]: import matplotlib.pyplot as plt
In [2]: year = [1950, 1970, 1990, 2010]
In [3]: pop = [2.519, 3.692, 5.263, 6.972]
In [4]: plt.scatter(year, pop)
In [5]: plt.show()
```



Ticks

```
import matplotlib.pyplot as plt
year = [1950, 1951, 1952, ..., 2100]
pop = [2.538, 2.57, 2.62, ..., 10.85]

plt.plot(year, pop)

plt.xlabel('Year')
plt.ylabel('Population')
plt.title('World Population Projections')
plt.yticks([0, 2, 4, 6, 8, 10],
           ['0', '2B', '4B', '6B', '8B', '10B'])

plt.show()
```



List

```
In [1]: pop = [30.55, 2.77, 39.21]

In [2]: countries = ["afghanistan", "albania", "algeria"]

In [3]: ind_alb = countries.index("albania")

In [4]: ind_alb
Out[4]: 1

In [5]: pop[ind_alb]
Out[5]: 2.77
```

Not convenient
Not intuitive

Tabular Dataset

temperature	measured_at	location
76	2016-01-01 14:00:01	valve
86	2016-01-01 14:00:01	compressor
72	2016-01-01 15:00:01	valve
88	2016-01-01 15:00:01	compressor
68	2016-01-01 16:00:01	valve
78	2016-01-01 16:00:01	compressor

row = observations
column = variable

country	capital	area	population
Brazil	Brasilia	8.516	200.4
Russia	Moscow	17.10	143.5
India	New Delhi	3.286	1252
China	Beijing	9.597	1357
South Africa	Pretoria	1.221	52.98

Pandas

- 2D Numpy array?
  - One data type
- Pandas!
  - High level data manipulation tool
  - Built on Numpy
  - DataFrame or Serie

country	capital	area	population
Brazil	Brasilia	8.516	200.4
Russia	Moscow	17.10	143.5
India	New Delhi	3.286	1252
China	Beijing	9.597	1357
South Africa	Pretoria	1.221	52.98
`str`	`str`	`float`	`float`

Dataframe from Dictionary

```
In [2]:  dict = {
    "country":["Brazil", "Russia", "India", "China", "South Africa"],
    "capital":["Brasilia", "Moscow", "New Delhi", "Beijing", "Pretoria"],
       "area":[8.516, 17.10, 3.286, 9.597, 1.221]
 "population":[200.4, 143.5, 1252, 1357, 52.98] }
```

keys (column labels) **values (data, column by column)**

```
In [3]: import pandas as pd

In [4]: brics = pd.DataFrame(dict)
```

Dataframe from Dictionary

```
In [5]: brics
Out[5]:
     area    capital       country  population
0   8.516   Brasilia        Brazil      200.40
1  17.100     Moscow        Russia      143.50
2   3.286  New Delhi         India     1252.00
3   9.597    Beijing         China     1357.00
4   1.221   Pretoria  South Africa       52.98

In [6]: brics.index = ["BR", "RU", "IN", "CH", "SA"]
In [7]: brics
Out[7]:
      area    capital       country  population
BR   8.516   Brasilia        Brazil      200.40
RU  17.100     Moscow        Russia      143.50
IN   3.286  New Delhi         India     1252.00
CH   9.597    Beijing         China     1357.00
SA   1.221   Pretoria  South Africa       52.98
```

DataFrame from CSV file

brics.csv

```
,country,capital,area,population
BR,Brazil,Brasilia,8.516,200.4
RU,Russia,Moscow,17.10,143.5
IN,India,New Delhi,3.286,1252
CH,China,Beijing,9.597,1357
SA,South Africa,Pretoria,1.221,52.98
```

```
In [8]: brics = pd.read_csv("path/to/brics.csv")
In [9]: brics
Out[9]:
  Unnamed: 0       country    capital    area  population
0         BR        Brazil   Brasilia   8.516      200.40
1         RU        Russia     Moscow  17.100      143.50
2         IN         India  New Delhi   3.286     1252.00
3         CH         China    Beijing   9.597     1357.00
4         SA  South Africa   Pretoria   1.221       52.98

In [6]: brics = pd.read_csv("path/to/brics.csv", index_col = 0)
In [7]: brics
Out[7]:
        country  population      area    capital
BR       Brazil         200   8515767   Brasilia
RU       Russia         144  17098242     Moscow
IN        India        1252   3287590  New Delhi
CH        China        1357   9596961    Beijing
SA  South Africa          55   1221037   Pretoria
```

Column Access [ ]

	country	capital	area	population
BR	Brazil	Brasilia	8.516	200.40
RU	Russia	Moscow	17.100	143.50
IN	India	New Delhi	3.286	1252.00
CH	China	Beijing	9.597	1357.00
SA	South Africa	Pretoria	1.221	52.98

```
In [4]: brics["country"]
Out[4]:
BR           Brazil
RU           Russia
IN            India
CH            China
SA     South Africa
Name: country, dtype: object

In [5]: type(brics["country"])
Out[5]: pandas.core.series.Series
```

1D labelled array

Column Access [ ]

```
In [6]: brics[["country"]]
Out[6]:
         country
BR        Brazil
RU        Russia
IN         India
CH         China
SA  South Africa

In [7]: type(brics[["country"]])
Out[7]: pandas.core.frame.DataFrame
```

	country	capital	area	population
BR	Brazil	Brasilia	8.516	200.40
RU	Russia	Moscow	17.100	143.50
IN	India	New Delhi	3.286	1252.00
CH	China	Beijing	9.597	1357.00
SA	South Africa	Pretoria	1.221	52.98

Column Access [ ]

```
In [8]: brics[["country", "capital"]]
Out[8]:
         country    capital
BR        Brazil   Brasilia
RU        Russia     Moscow
IN         India  New Delhi
CH         China    Beijing
SA  South Africa   Pretoria
```

	country	capital	area	population
BR	Brazil	Brasilia	8.516	200.40
RU	Russia	Moscow	17.100	143.50
IN	India	New Delhi	3.286	1252.00
CH	China	Beijing	9.597	1357.00
SA	South Africa	Pretoria	1.221	52.98

Row Access [ ]

indexes

	country	capital	area	population
0 BR	Brazil	Brasilia	8.516	200.40
1 RU	Russia	Moscow	17.100	143.50
2 IN	India	New Delhi	3.286	1252.00
3 CH	China	Beijing	9.597	1357.00
4 SA	South Africa	Pretoria	1.221	52.98

```
In [9]: brics[1:4]
Out[9]:
   country    capital    area  population
RU  Russia     Moscow  17.100       143.5
IN   India  New Delhi   3.286      1252.0
CH   China    Beijing   9.597      1357.0
```

Row Access `loc`

	country	capital	area	population
BR	Brazil	Brasilia	8.516	200.40
RU	Russia	Moscow	17.100	143.50
IN	India	New Delhi	3.286	1252.00
CH	China	Beijing	9.597	1357.00
SA	South Africa	Pretoria	1.221	52.98

```
In [10]: brics.loc["RU"]
Out[10]:
country       Russia
capital       Moscow
area            17.1
population     143.5
Name: RU, dtype: object
```

Row as Pandas Series

```
In [11]: brics.loc[["RU"]]
Out[11]:
   country capital  area  population
RU  Russia  Moscow  17.1       143.5
```

DataFrame

Row & Column `loc`

	country	capital	area	population
BR	Brazil	Brasilia	8.516	200.40
RU	Russia	Moscow	17.100	143.50
IN	India	New Delhi	3.286	1252.00
CH	China	Beijing	9.597	1357.00
SA	South Africa	Pretoria	1.221	52.98

```
In [13]: brics.loc[["RU", "IN", "CH"], ["country", "capital"]]
Out[13]:
   country    capital
RU  Russia     Moscow
IN   India  New Delhi
CH   China    Beijing
```

Row Access iloc

```
In [17]: brics.loc[["RU", "IN", "CH"]]
Out[17]:
   country    capital    area  population
RU  Russia     Moscow  17.100       143.5
IN   India  New Delhi   3.286      1252.0
CH   China    Beijing   9.597      1357.0

In [18]: brics.iloc[[1,2,3]]
Out[18]:
   country    capital    area  population
RU  Russia     Moscow  17.100       143.5
IN   India  New Delhi   3.286      1252.0
CH   China    Beijing   9.597      1357.0
```

	country	capital	area	population
0	BR	Brazil	Brasilia	8.516	200.40
1	RU	Russia	Moscow	17.100	143.50
2	IN	India	New Delhi	3.286	1252.00
3	CH	China	Beijing	9.597	1357.00
4	SA	South Africa	Pretoria	1.221	52.98

Row & Column iloc

```
In [21]: brics.loc[:, ["country", "capital"]]
Out[21]:
         country    capital
BR        Brazil   Brasilia
RU        Russia     Moscow
IN         India  New Delhi
CH         China    Beijing
SA  South Africa   Pretoria

In [22]: brics.iloc[:, [0,1]]
Out[22]:
         country    capital
BR        Brazil   Brasilia
RU        Russia     Moscow
IN         India  New Delhi
CH         China    Beijing
SA  South Africa   Pretoria
```

	0	1	2	3
	country	capital	area	population
0	BR	Brazil	Brasilia	8.516	200.40
1	RU	Russia	Moscow	17.100	143.50
2	IN	India	New Delhi	3.286	1252.00
3	CH	China	Beijing	9.597	1357.00
4	SA	South Africa	Pretoria	1.221	52.98

References

- Notebook
- Dataset