Welcome to CSC 276
Data Science

12/30/2022

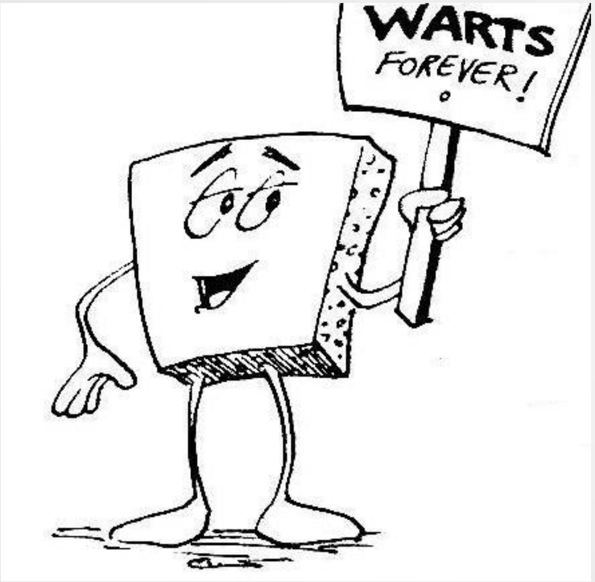
CSC 276: Data Science
Lecture #3
Introduction

Dr.Fatema Nafa

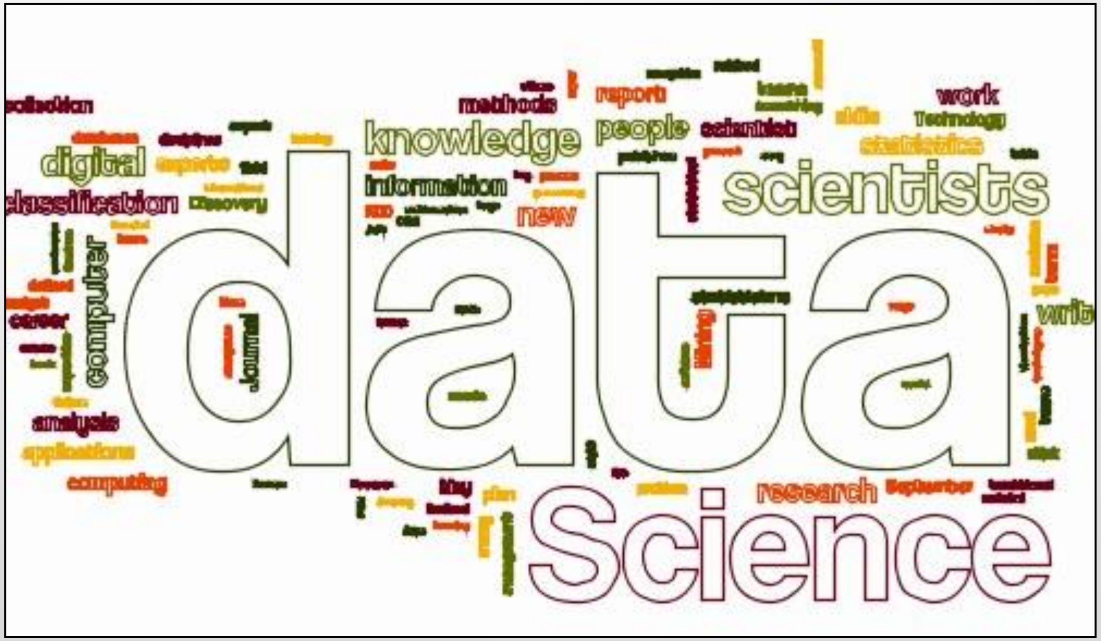
Fall 2022

12/30/2022

Welcome to CSC 276!



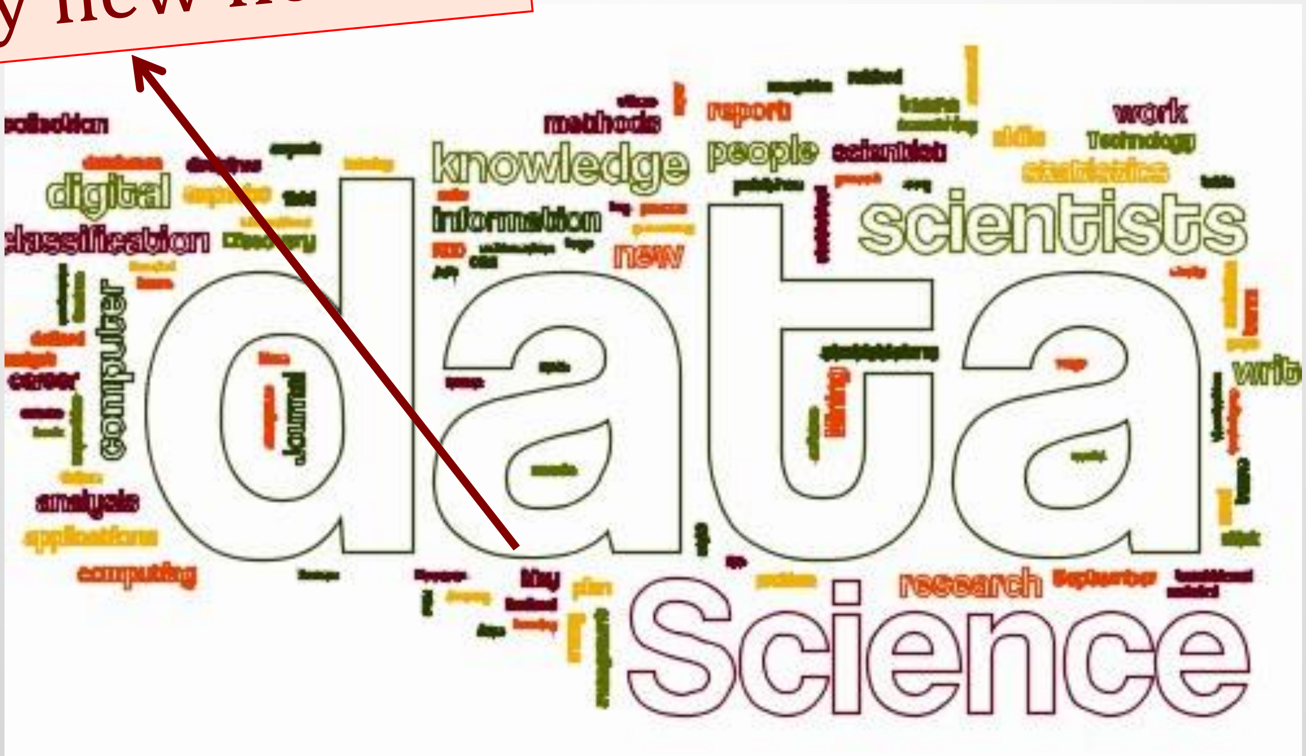
Data Science



Welcome to CSC 276!

Data Science

This class is truly seminar-style: I'm here, as you are, in order to gain insights into this very new field… .

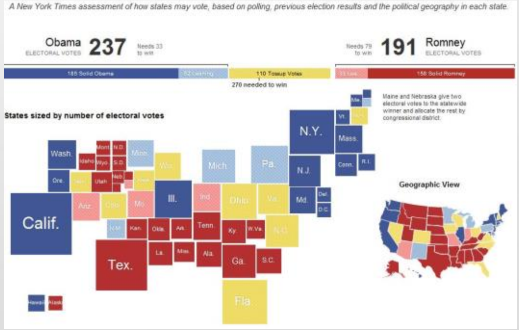


Lecture Outline

- **The Art of Data Science**
- Volume, Velocity, Variety
- The Logic of Data Science
- How to Be Agile
- Treating Data as Evidence
- **Python**
  - Fundamentals of Data Manipulation
  - Basic Data Processing with Pandas
  - Answering Questions with Messy Data

Any
Questions

1. Data sources
2. Collect data(**download**)
3. Prepare data (integrate, transform, clean, filter, aggregate)
4. Build model
5. Evaluate model
6. Visualize the results

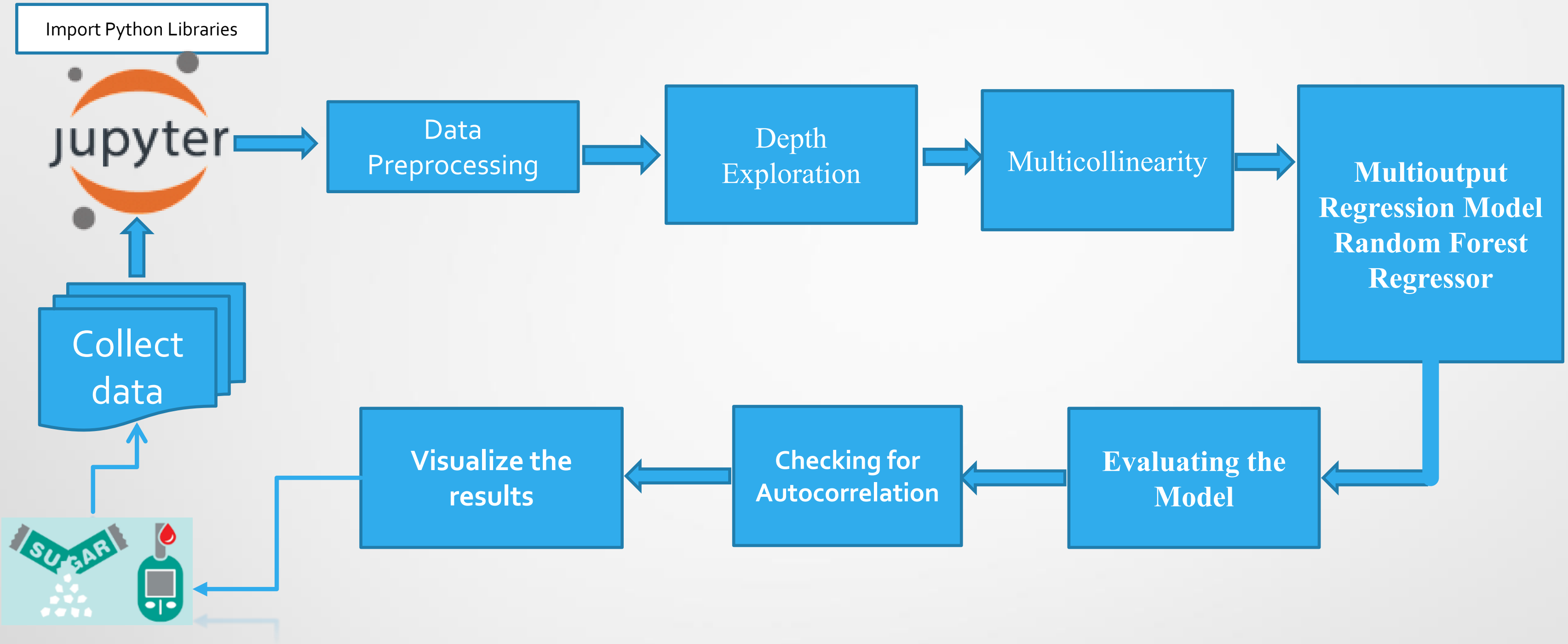


A New York Times assessment of how states may vote, based on polling, previous election results and the political geography in each state.
Obama ELECTORAL VOTES 237 Needs 33 to win
Needs 79 to win 191 Romney ELECTORAL VOTES
185 Solid Obama
110 Tossup Votes
158 Solid Romney
270 needed to win
States sized by number of electoral votes
Maine and Nebraska give two electoral votes to the statewide winner and allocate the rest by congressional district.
Geographic View
Wash.
Ore.
Calif.
Idaho
Mont.
N.D.
Wyo.
S.D.
Utah
Neb.
Minn.
Wis.
Mich.
Pa.
N.Y.
Vt.
Mass.
Conn.
R.I.
N.J.
Ill.
Ind.
Ohio
Va.
Md.
Del.
D.C.
Ariz.
Colo.
Mo.
N.M.
Kan.
Okla.
Ark.
Tenn.
Ky.
W.Va.
N.C.
Tex.
La.
Miss.
Ala.
Ga.
S.C.
Fla.
Hawaii
Alaska
Me.

Introduction

jupyter
SciPy
SPYDER
NumPy
DASK
bokeh
scikit learn
pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
Numba
TensorFlow
Panel
matplotlib
HoloViews
PYTORCH
CONDA®

Introduction



Import Python Libraries

Data Preprocessing

Depth Exploration

Multicollinearity

Multioutput Regression Model Random Forest Regressor

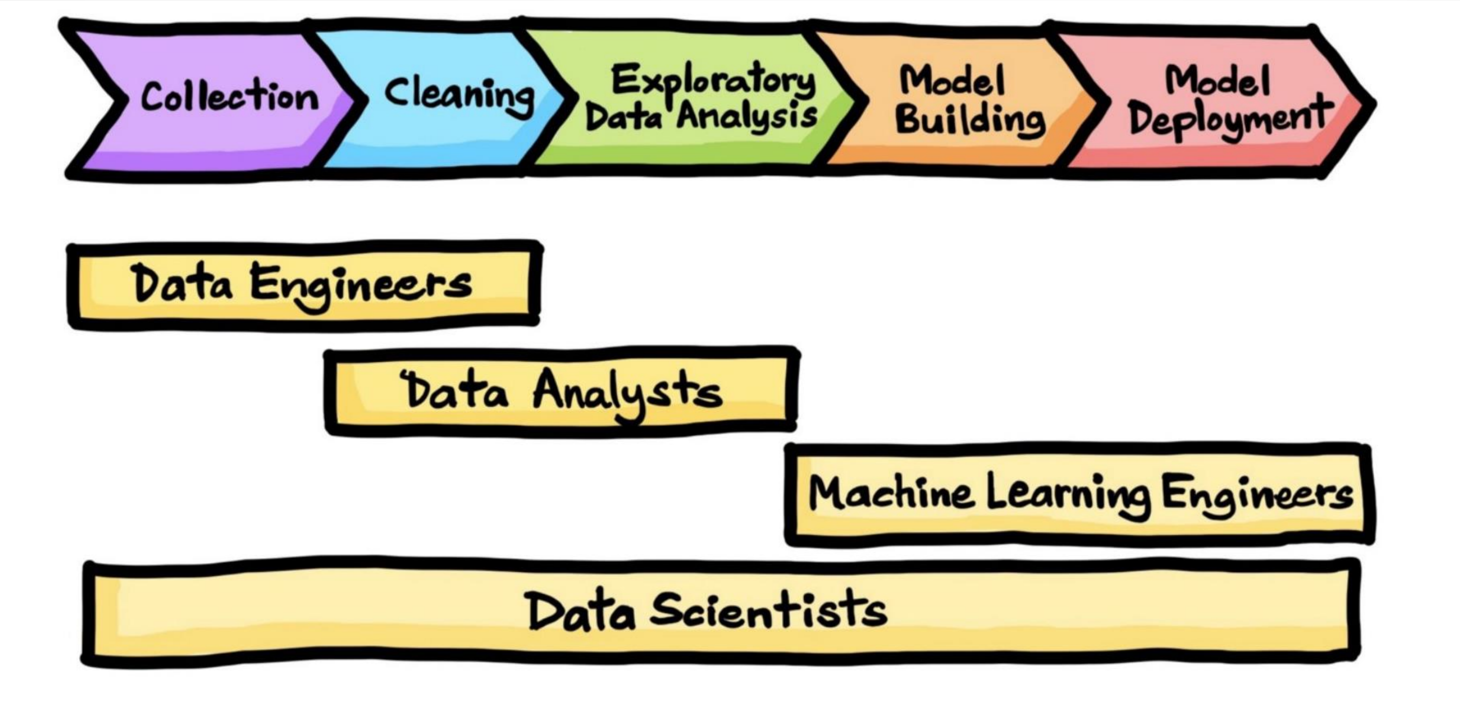
Evaluating the Model

Checking for Autocorrelation

Visualize the results

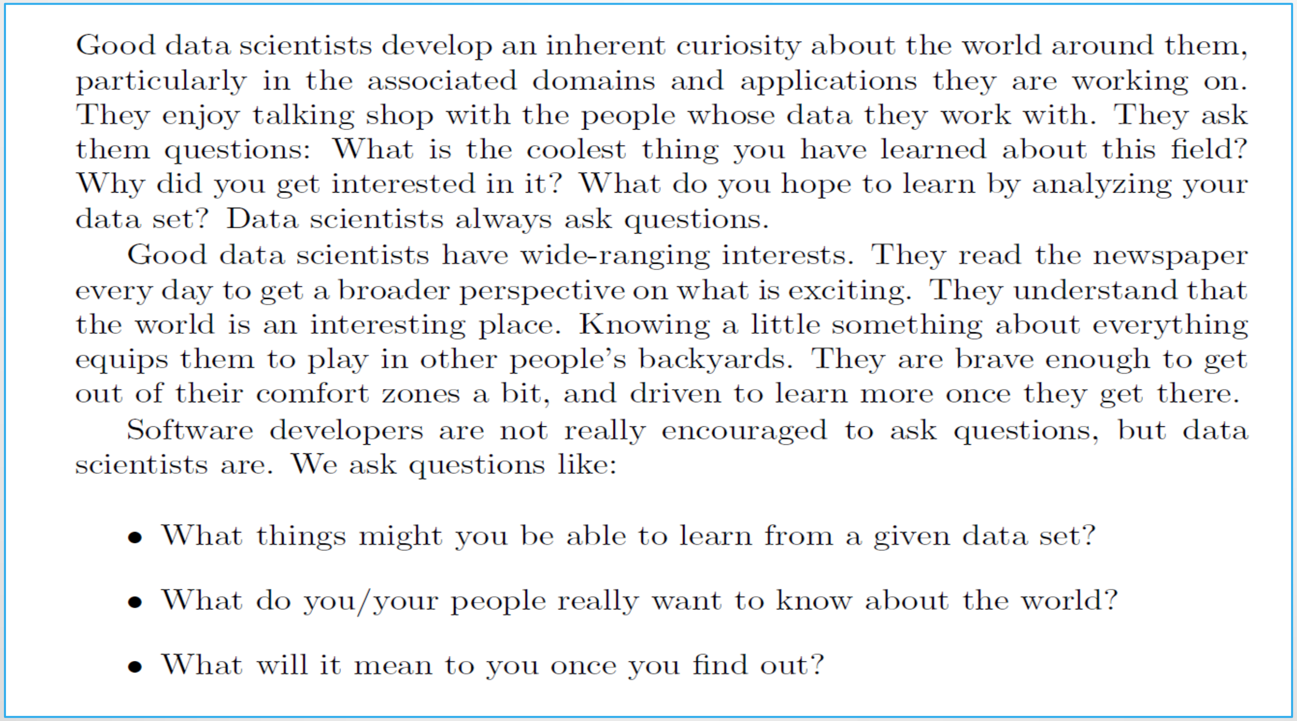
Collect data

12/30/2022


Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment
Data Engineers
Data Analysts
Machine Learning Engineers
Data Scientists

Asking Interesting Questions from Data

Introduction



Good data scientists develop an inherent curiosity about the world around them, particularly in the associated domains and applications they are working on. They enjoy talking shop with the people whose data they work with. They ask them questions: What is the coolest thing you have learned about this field? Why did you get interested in it? What do you hope to learn by analyzing your data set? Data scientists always ask questions.

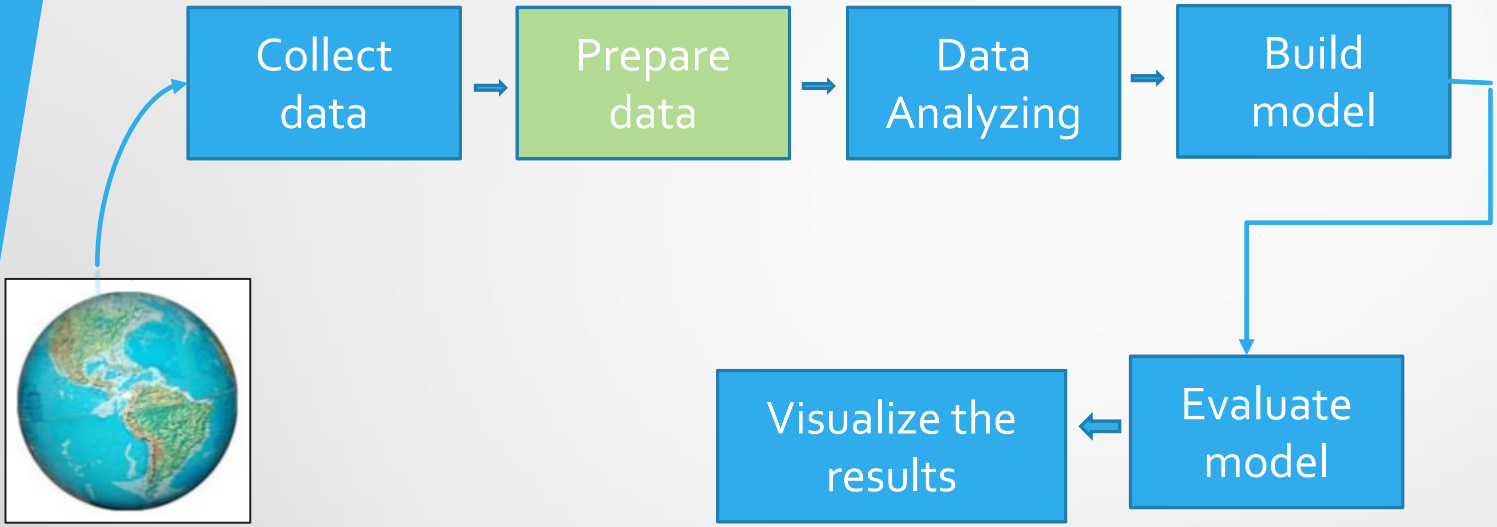
Good data scientists have wide-ranging interests. They read the newspaper every day to get a broader perspective on what is exciting. They understand that the world is an interesting place. Knowing a little something about everything equips them to play in other people's backyards. They are brave enough to get out of their comfort zones a bit, and driven to learn more once they get there.

Software developers are not really encouraged to ask questions, but data scientists are. We ask questions like:

- What things might you be able to learn from a given data set?
- What do you/your people really want to know about the world?
- What will it mean to you once you find out?

- After you understand what kind of information is available, try to come up with, say,
- 10 interesting questions you might explore/answer with access to the data set.

Introduction



Collect data

Prepare data

Data Analyzing

Build model

Evaluate model

Visualize the results

My DATA

Statistical information about my data

What’s Hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

Readings

Read next week readings and complete it before next class.

- Chapter Two: Python Language Basics, IPython, and Jupyter Notebooks
- Chapter Three: Built-In Data Structures, Functions, and Files

python for data analysis 2nd edition pdf

Any
Questions

Tutorial Content

Overview of Python Libraries for Data Scientists

Reading Data; Selecting and Filtering the Data;
Data manipulation, sorting, grouping, rearranging

Plotting the data

Descriptive statistics

Python Libraries for Data Science

Many popular Python toolboxes/libraries:

- NumPy
- SciPy
- Pandas
- SciKit-Learn

Visualization libraries

- matplotlib
- Seaborn

and many more …

Python Libraries for Data Science

NumPy:

- introduces objects for multidimensional arrays and matrices, as well as functions that allow to easily perform advanced mathematical and statistical operations on those objects

- provides vectorization of mathematical operations on arrays and matrices which significantly improves the performance

Link: http://www.numpy.org/

- many other python libraries are built on NumPy

Python Libraries for Data Science

SciPy:

- collection of algorithms for linear algebra, differential equations, numerical integration, optimization, statistics and more
- part of SciPy Stack
- built on NumPy

Link: https://www.scipy.org/scipylib/

Python Libraries for Data Science

Pandas:

- adds data structures and tools designed to work with table-like data (similar to Series and Data Frames in R)

- provides tools for data manipulation: reshaping, merging, sorting, slicing, aggregation etc.

Link: http://pandas.pydata.org/

- allows handling missing data

Python Libraries for Data Science

SciKit-Learn:

- provides machine learning algorithms: classification, regression, clustering, model validation etc.
- built on NumPy, SciPy and matplotlib

Link: http://scikit-learn.org/

Python Libraries for Data Science

matplotlib:

- python 2D plotting library which produces publication quality figures in a variety of hardcopy formats
- a set of functionalities similar to those of MATLAB
- line plots, scatter plots, barcharts, histograms, pie charts etc.

Link: https://matplotlib.org/

- relatively low-level; some effort needed to create advanced visualization

Python Libraries for Data Science

Seaborn:

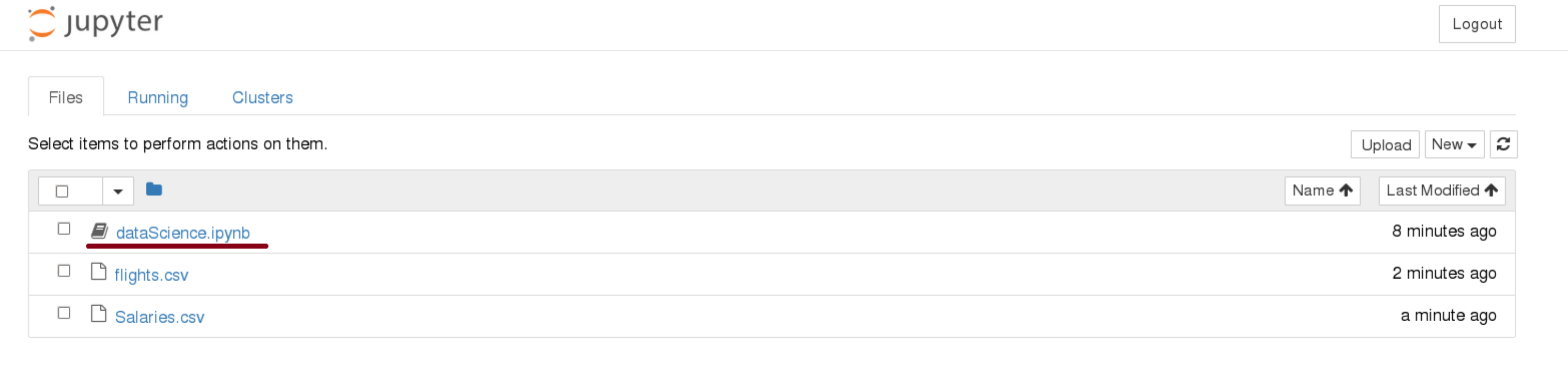
- based on matplotlib

- provides high level interface for drawing attractive statistical graphics

Link: https://seaborn.pydata.org/

- Similar (in style) to the popular ggplot2 library in R

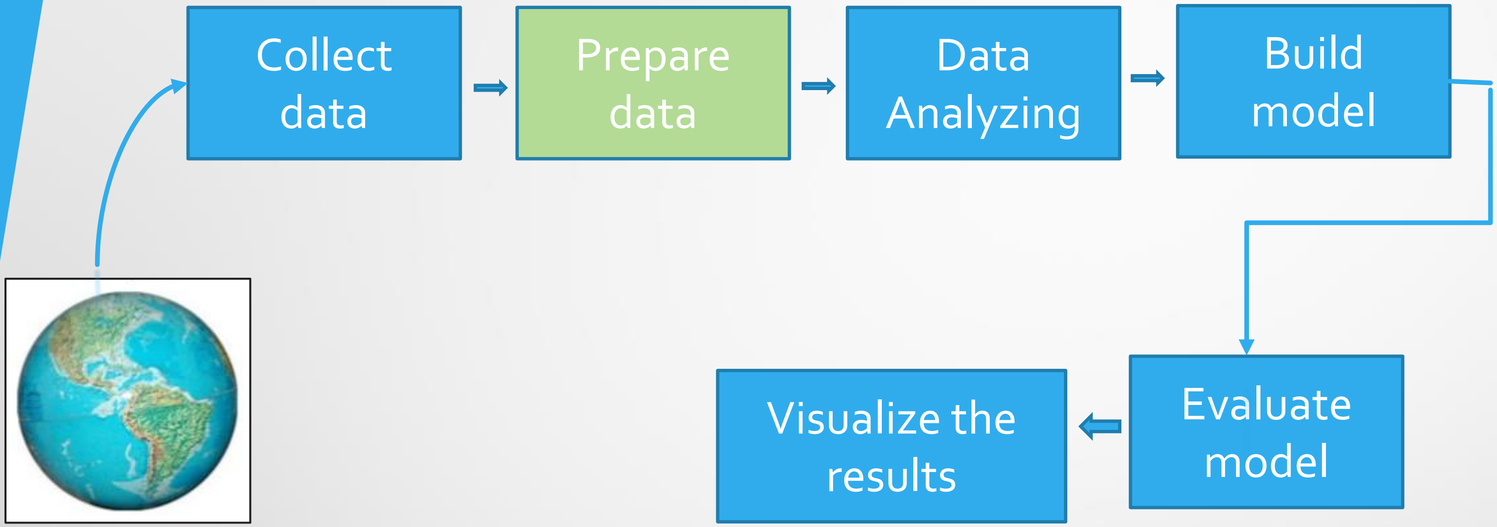
Start Jupyter nootebook



Select items to perform actions on them.

	Name	Last Modified
□	dataScience.ipynb	8 minutes ago
□	flights.csv	2 minutes ago
□	Salaries.csv	a minute ago

```
# On the Shared Computing Cluster

[scc1 ~] jupyter notebook
```



Collect data

Prepare data

Data Analyzing

Build model

Evaluate model

Visualize the results

1. Download your data
2. Provide the reference for your data.
3. Cite it

https://www.kaggle.com/secareanualin/football-events?select=events.csv

Loading Python Libraries

```python
#Import Python Libraries
import numpy as np
import scipy as sp
import pandas as pd
import matplotlib as mpl
import seaborn as sns
```

Press `Shift+Enter` to execute the *jupyter* cell

Reading data using pandas

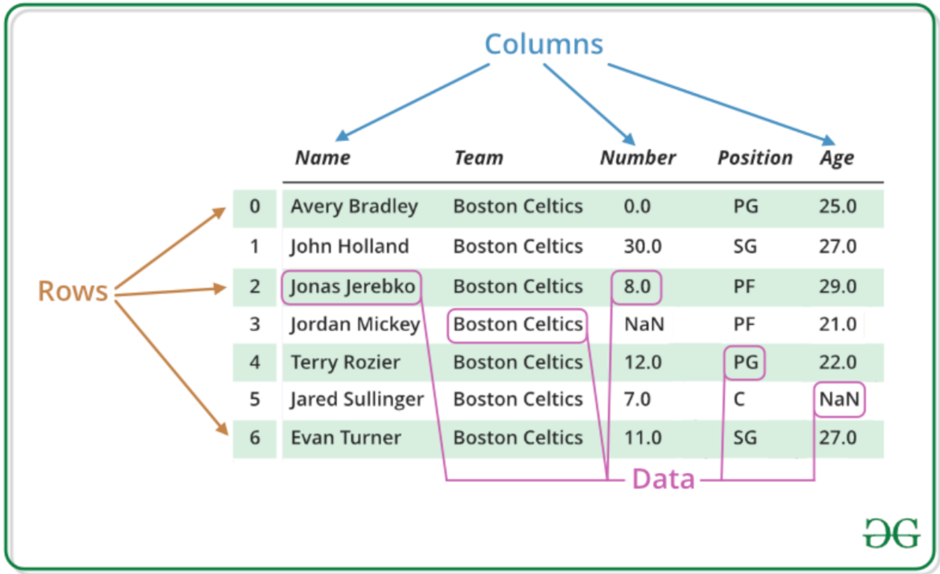
```
#Read csv file
Import pandas as pd
df = pd.read_csv("events.csv")
```

***Note:** The above command has many optional arguments to fine-tune the data import process.*

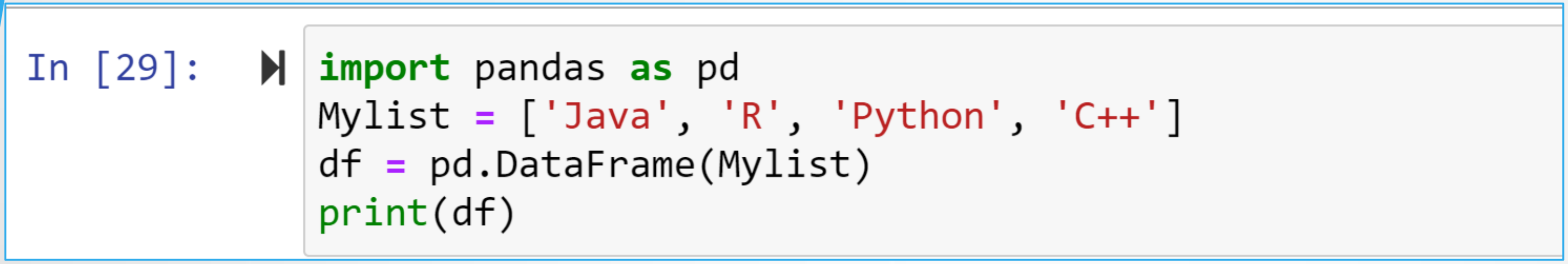
There is a number of pandas commands to read other data formats:

```
pd.read_excel('myfile.xlsx',sheet_name='Sheet1', index_col=None,
na_values=['NA'])
pd.read_stata('myfile.dta')
pd.read_sas('myfile.sas7bdat')
pd.read_hdf('myfile.h5','df')
```

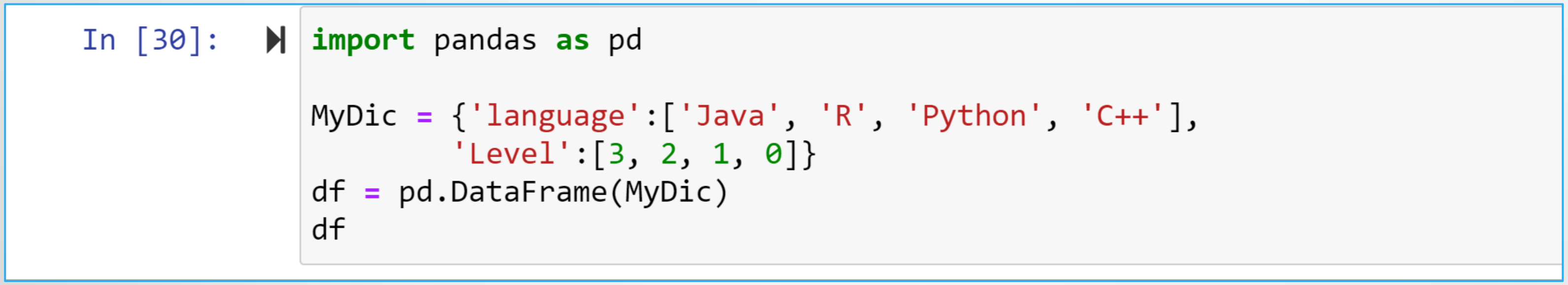
Pandas DataFrame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns). A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas DataFrame consists of three principal components, the **data**, **rows**, and **columns**.



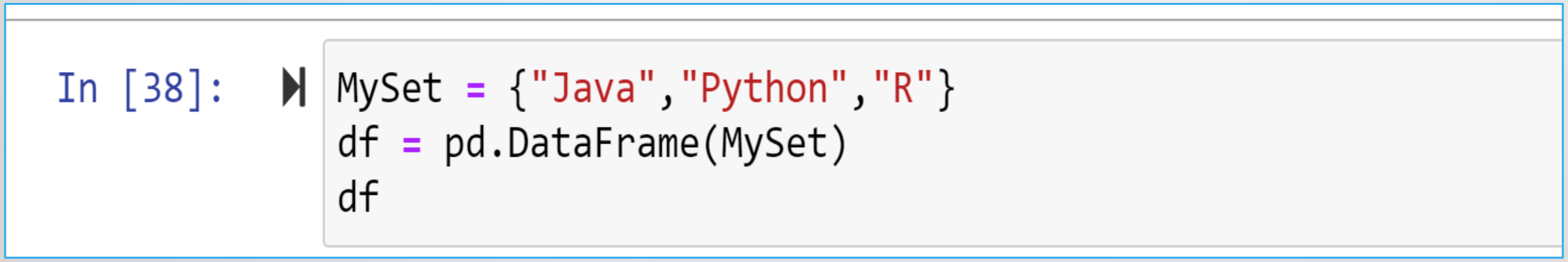
	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0



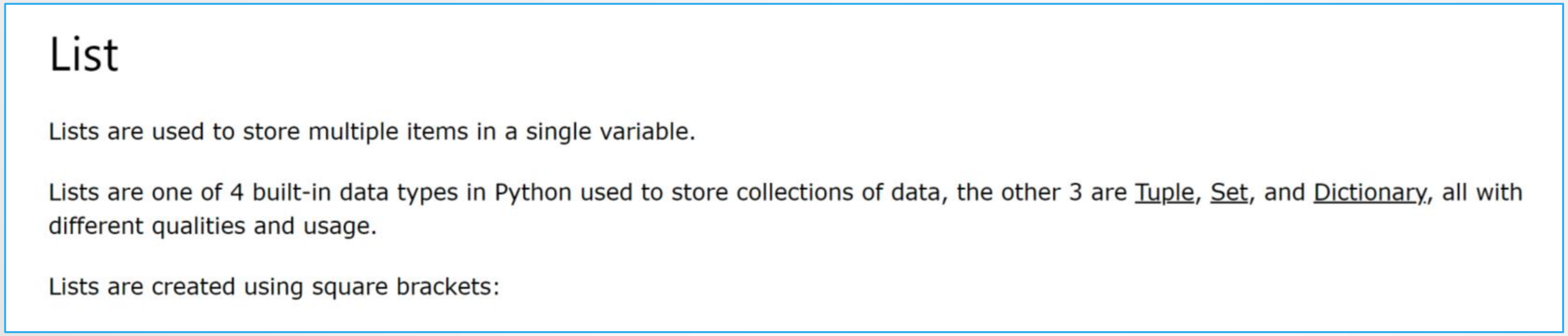
```
import pandas as pd
Mylist = ['Java', 'R', 'Python', 'C++']
df = pd.DataFrame(Mylist)
print(df)
```



```
import pandas as pd

MyDic = {'language':['Java', 'R', 'Python', 'C++'],
         'Level':[3, 2, 1, 0]}
df = pd.DataFrame(MyDic)
df
```



```
MySet = {"Java","Python","R"}
df = pd.DataFrame(MySet)
df
```

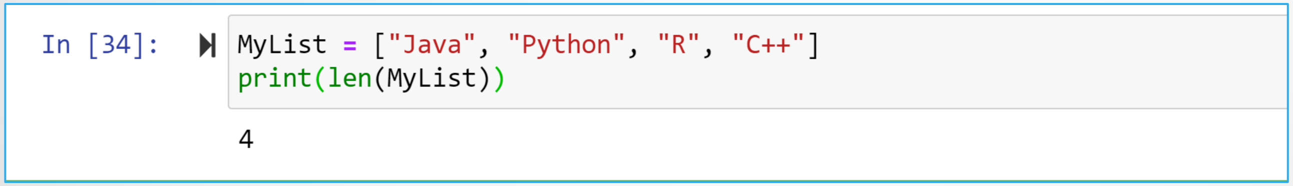


List

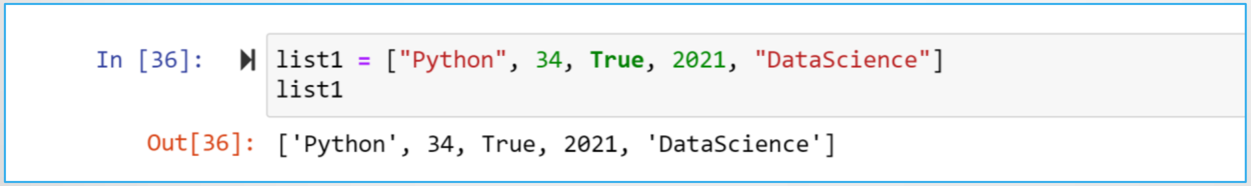
Lists are used to store multiple items in a single variable.

Lists are one of 4 built-in data types in Python used to store collections of data, the other 3 are Tuple, Set, and Dictionary, all with different qualities and usage.

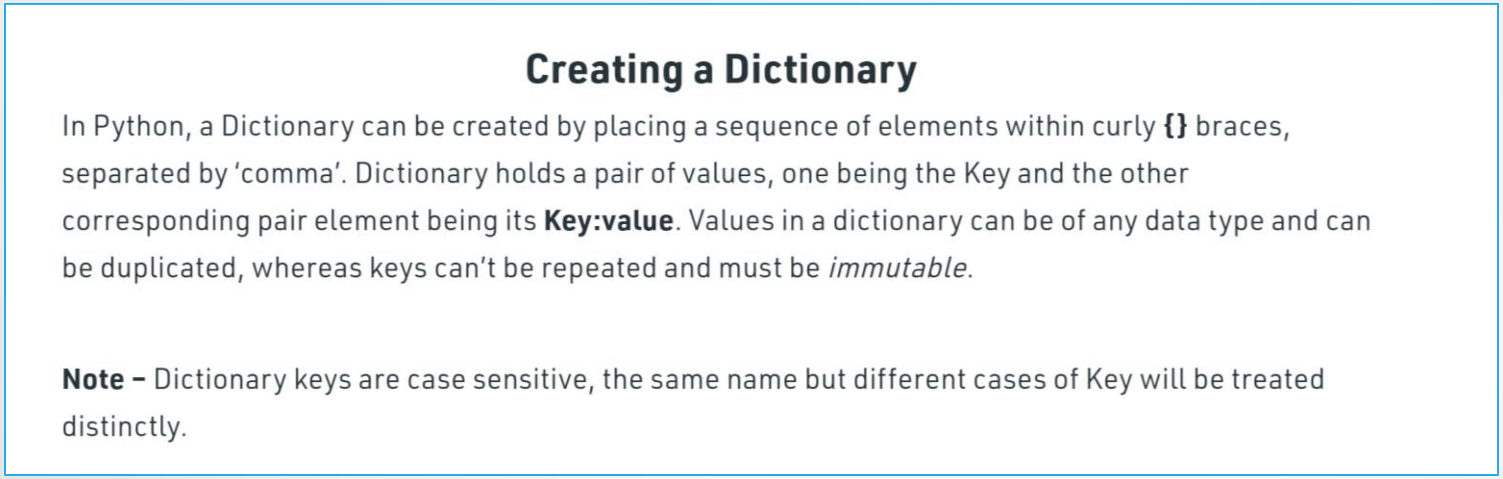
Lists are created using square brackets:



```
In [34]:  MyList = ["Java", "Python", "R", "C++"]
           print(len(MyList))

           4
```



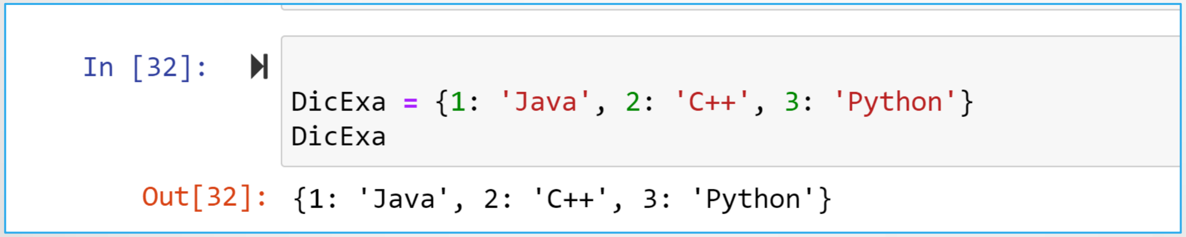
```
In [36]:  list1 = ["Python", 34, True, 2021, "DataScience"]
           list1

Out[36]:  ['Python', 34, True, 2021, 'DataScience']
```



Creating a Dictionary

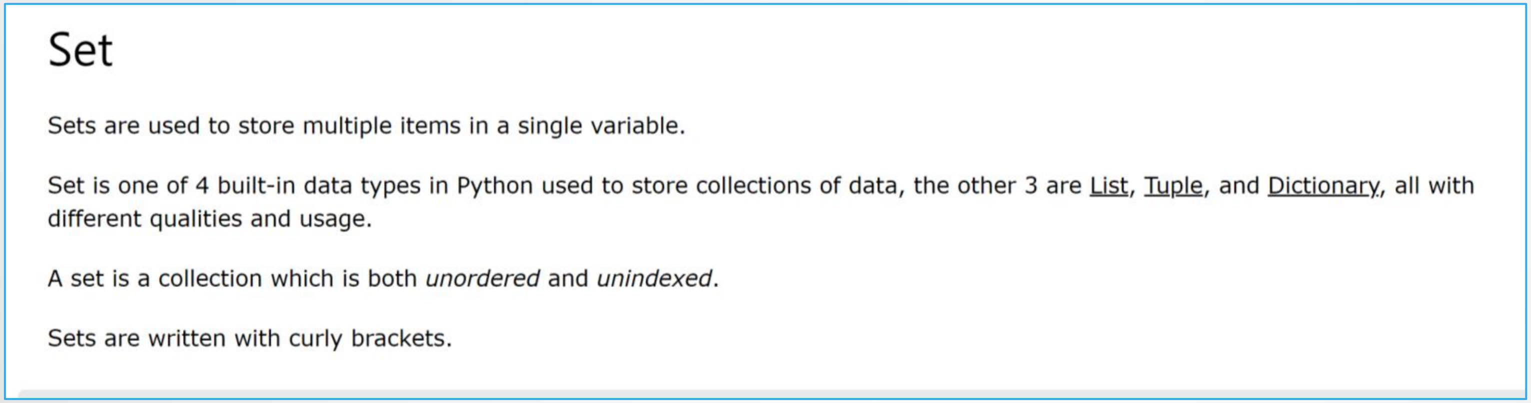
In Python, a Dictionary can be created by placing a sequence of elements within curly **{}** braces, separated by 'comma'. Dictionary holds a pair of values, one being the Key and the other corresponding pair element being its **Key:value**. Values in a dictionary can be of any data type and can be duplicated, whereas keys can't be repeated and must be *immutable*.

Note – Dictionary keys are case sensitive, the same name but different cases of Key will be treated distinctly.



```
In [32]:
DicExa = {1: 'Java', 2: 'C++', 3: 'Python'}
DicExa
```

```
Out[32]: {1: 'Java', 2: 'C++', 3: 'Python'}
```



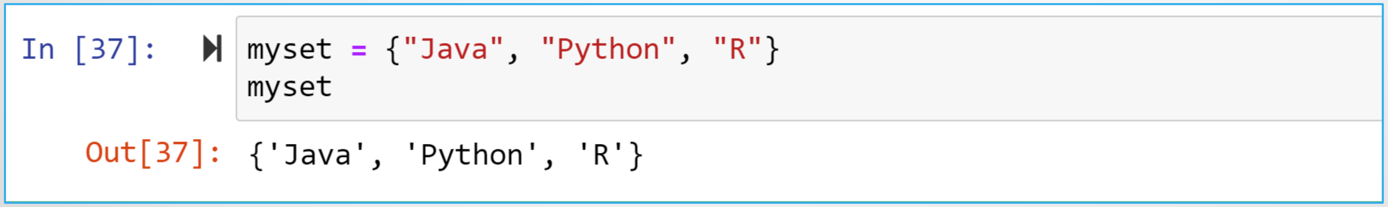
Set

Sets are used to store multiple items in a single variable.

Set is one of 4 built-in data types in Python used to store collections of data, the other 3 are List, Tuple, and Dictionary, all with different qualities and usage.

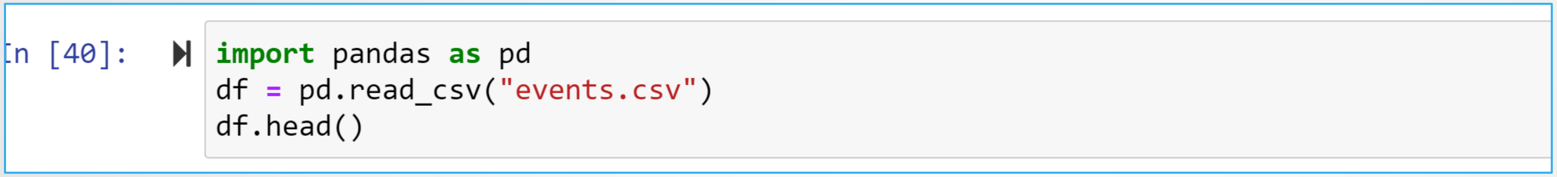
A set is a collection which is both *unordered* and *unindexed*.

Sets are written with curly brackets.



```
In [37]:  myset = {"Java", "Python", "R"}
          myset

Out[37]:  {'Java', 'Python', 'R'}
```

Exploring data frames



```
import pandas as pd
df = pd.read_csv("events.csv")
df.head()
```

Hands-on exercises

- ✓ Try to read the first 10, 20, 50 records;
- ✓ Can you guess how to view the last few records;

Data Frame data types

Pandas Type	Native Python Type	Description
object	string	The most general dtype. Will be assigned to your column if column has mixed types (numbers and strings).
int64	int	Numeric characters. 64 refers to the memory allocated to hold this character.
float64	float	Numeric characters with decimals. If a column contains numbers and NaNs(see below), pandas will default to float64, in case your missing value has a decimal.
datetime64, timedelta[ns]	N/A (but see the datetime module in Python's standard library)	Values meant to hold time data. Look into these for time series experiments.

Any
Questions