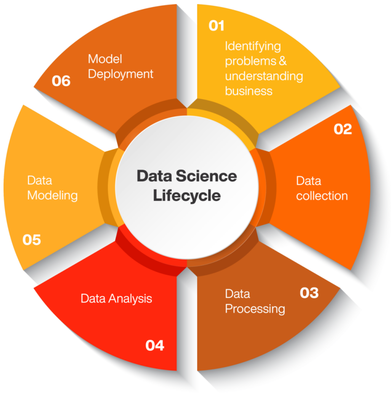
Data Science Life Cycle



What is Data Science?

The Study of data in order to ro make more meaningful conclusions

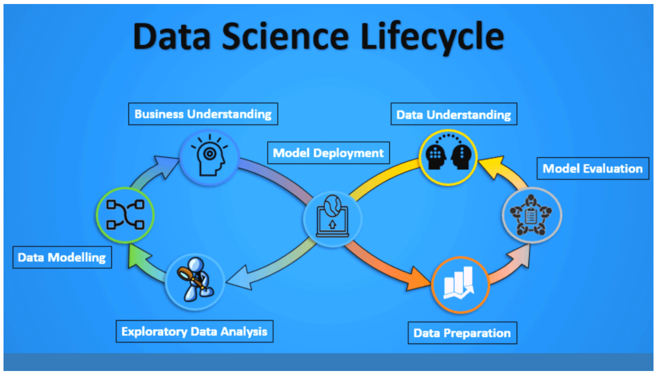
The difference between Data and Information

Data: Is unprocessed raw input

Information: Is processed data that can be understood

Data science Life Cycle steps

1. Question
2. Collection
3. Wrangling Data
4. Analyze
5. Visualize
6. Communicate



Python Fundamentals

Collection Types

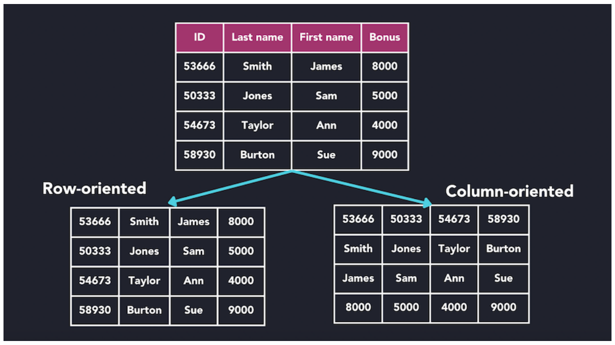
Very useful in data science because datasets are the collection of data

Python collection types:

- Dictionary
  - Ordered
  - Changeable
  - Duplicates not allowed
- List
  - Ordered
  - Changeable
  - Duplicates allowed
- Tuple
  - Unordered
  - Unchangeable
  - duplicates not allowed
- Set
  - Unordered
  - Unchangeable
  - duplicates allowed

Code Representation Datasets

- Column- Oriented:
  - Grouping by features
- Row-oriented:
  - Grouping by a single observations



ID	Last name	First name	Bonus
53666	Smith	James	8000
50333	Jones	Sam	5000
54673	Taylor	Ann	4000
58930	Burton	Sue	9000

Row-oriented

53666	Smith	James	8000
50333	Jones	Sam	5000
54673	Taylor	Ann	4000
58930	Burton	Sue	9000

Column-oriented

53666	50333	54673	58930
Smith	Jones	Taylor	Burton
James	Sam	Ann	Sue
8000	5000	4000	9000

Indexing

- To access values, we need to INDEX

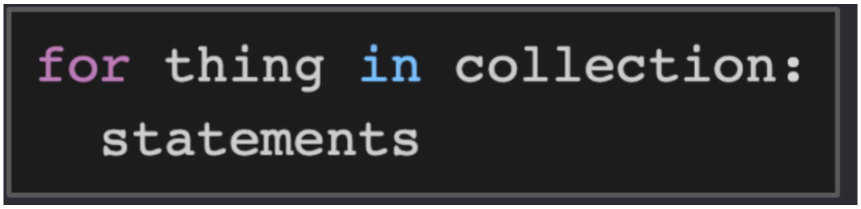
Type	Indexing Pattern
List	name[index]
Dictionary	name[key]
Set	For loop (next slide)
Tuple	Name [index}

Iteration

- You can repeat processes with loops or recursion in python

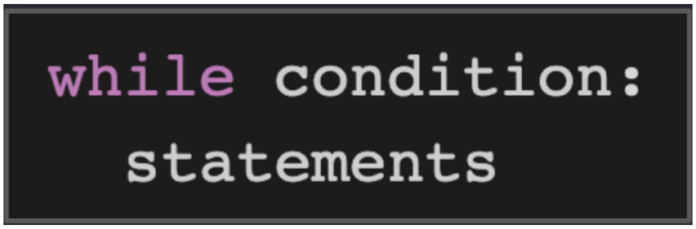
Python loop types: (https://www.w3schools.com/python/python_ref_dictionary.asp)

❖ For loop

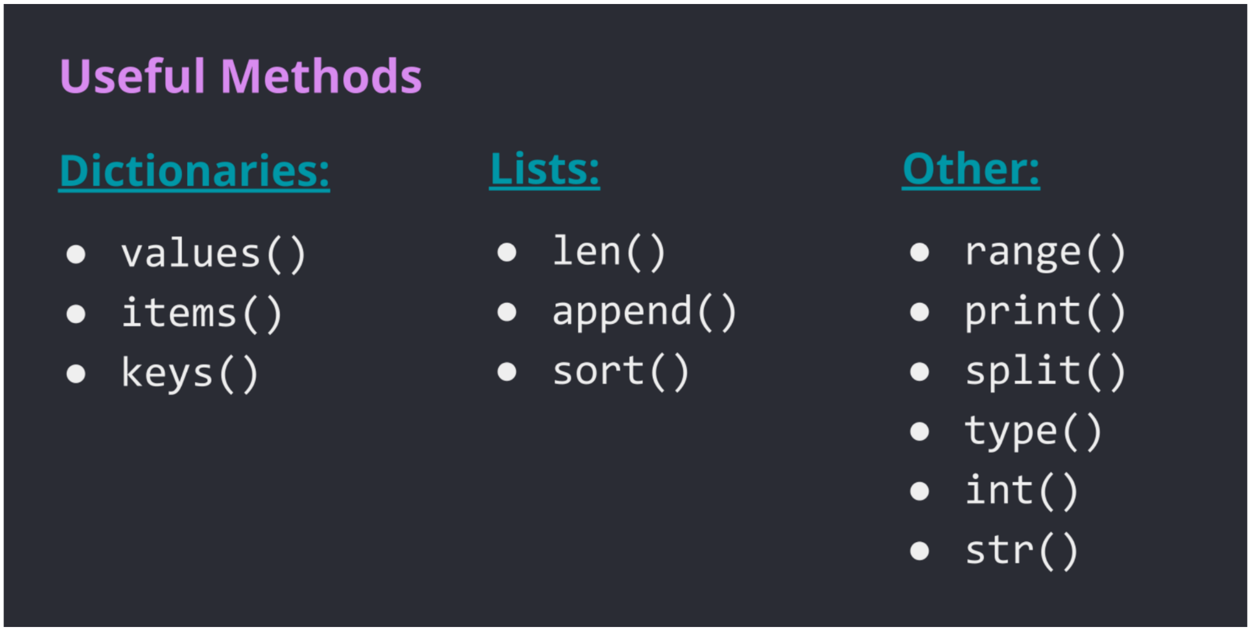


```
for thing in collection:
   statements
```

❖ While loop



```
while condition:
   statements
```



Useful Methods

Dictionaries:

- values()
- items()
- keys()

Lists:

- len()
- append()
- sort()

Other:

- range()
- print()
- split()
- type()
- int()
- str()

Central Tendency

(Reference crash course statistics)

Measures of central tendency:

Mean, Median, Mode are statistical measures that help us describe the behavior of a **collection of data points**

Definitions

- Mean
  - Weight tendency for things to occur in a data set
  - Affected by outliers
- Median
  - The middle of sorted data
  - Does Not use all data points
  - Less affected by outliers
- Mode
  - The most frequently/ popular occurring value in data set
  - Most helpful with moderately large data sets
- Central tendency
  - The summarized version of data set that is based in the middle of the data
- Normal distribution
  - Not Skewed
  - Symmetrical
- Skewed Distribution
  - The median is nearly identical but the mean is pulled toward the skewed data