datapolitan

# Data Analysis with Python

Follow along at: http://bit.ly/data-analysis-python

See the code at: http://bit.ly/data-analysis-python-code

## Open your browser to:

## http://student___.datapolitan.com/julia

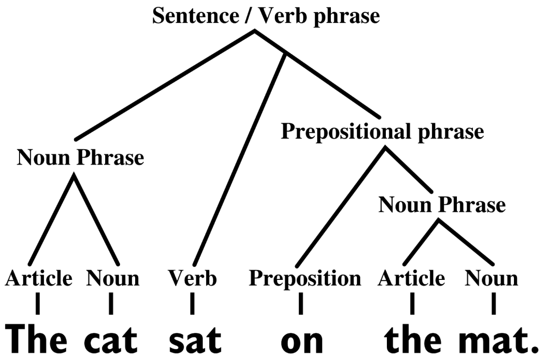
Username: **rstudio**    Password: **rstudio**

## Key Questions for the Morning

| What is python? | How is it useful in analysis? |
|---|---|
| | |

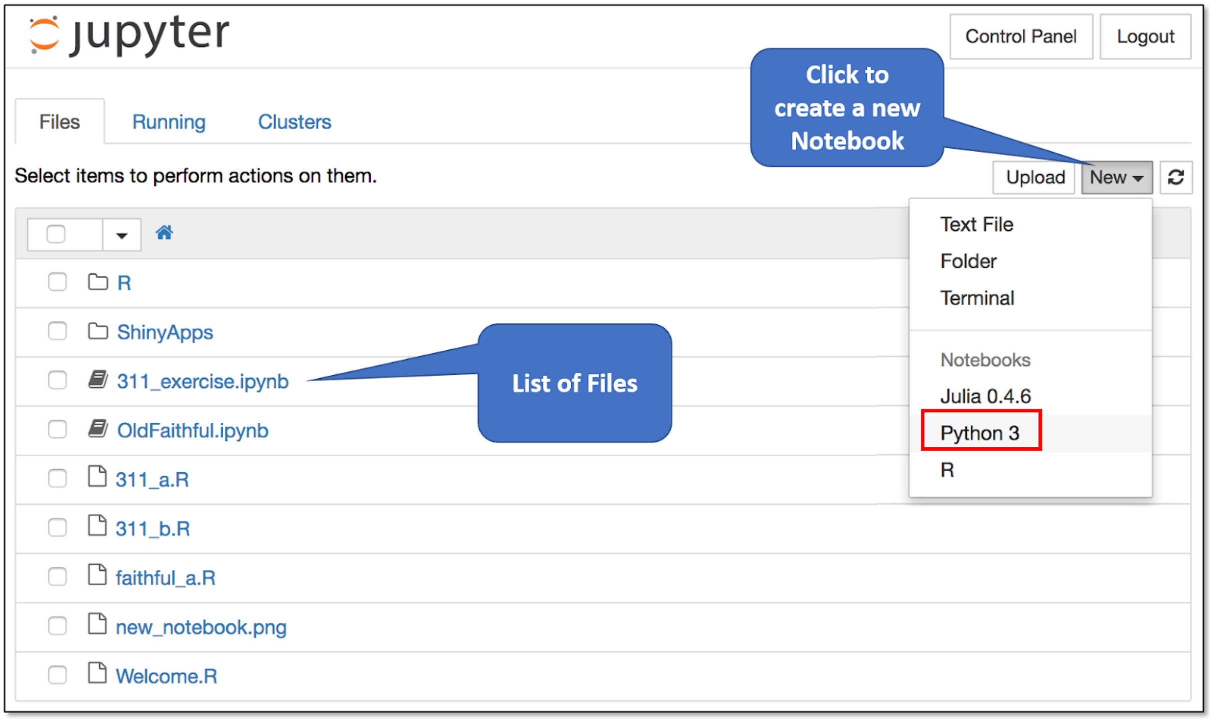
## What is Syntax?

Basic constituent structure analysis of a sentence:



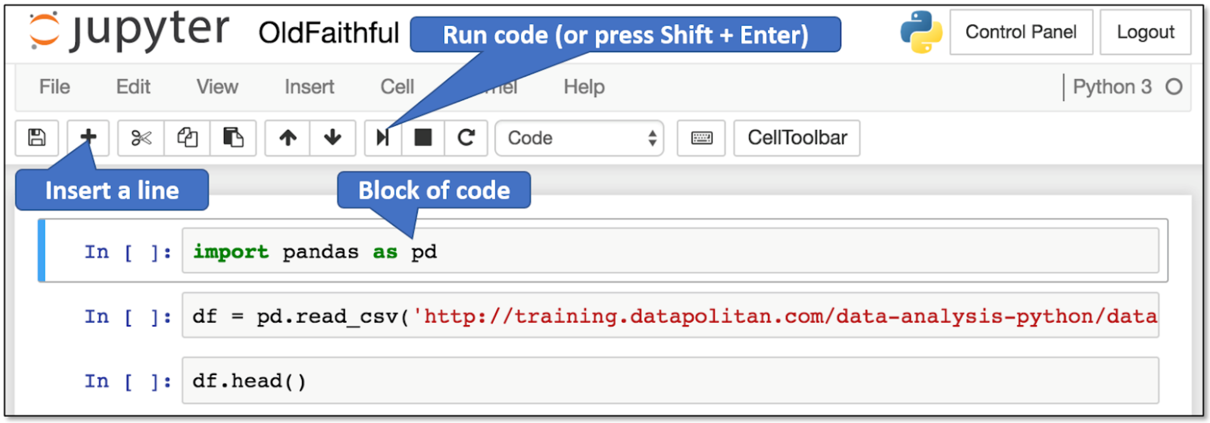
## Python Syntax

- Variables hold some value
- We create variables and assign a value using the = sign
- We can perform operations with mathematical operators
- We can use built-in functions for operations
- Reference a particular column like df['Column Name']
- Use a dot (.) to call a function on an object



Click to create a new Notebook

List of Files



Run code (or press Shift + Enter)

Insert a line

Block of code

```
In [ ]: import pandas as pd

In [ ]: df = pd.read_csv('http://training.datapolitan.com/data-analysis-python/data

In [ ]: df.head()
```

Key pandas Functions

- **read_csv()** - import data from CSV into a DataFrame
- **read_excel()** - import data from .xls and .xlsx files into a DataFrame
- **head()** & **tail()** - first (head) and last (tail) 5 rows of DataFrame
- **count()** - count of all rows in a DataFrame column
- **max() & min()** - maximum and minimum values in a DataFrame column
- **mean() & median()** - mean and median values of numbers in a DataFrame column
- **describe()** - summary statistics for DataFrame
- **plot()** - plot data from a DataFrame
- **hist()** - create a histogram of values
- **groupby()** - group values together in data frame
- **sort_values()** - sort by values

## Key Questions for the Afternoon

| How is Python/pandas helpful for me? | What key tasks do I need to learn/practice? |
|---|---|
|  |  |

### Key Links

- Download Python 3 (Anaconda distribution): https://www.anaconda.com/distribution/
- Download exercise files from this class: http://bit.ly/data-analysis-python-code
- Jupyter Quick Start Guide: https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/

### Learning and Practicing More with Python

- **Beginner's Python Tutorial** (https://en.wikibooks.org/wiki/A_Beginner%27s_Python_Tutorial)- A good way to get started with basic tasks
- **Python for Data Analysis** (https://www.oreilly.com/library/view/python-for-data/9781491957653/) - The textbook on using pandas for data analysis
- **Real Python Tutorials** (https://realpython.com/) - Great tutorials, videos, and mini-courses on Python (mix free and paid content)
- **Whirlwind Tour of Python** (http://nbviewer.jupyter.org/github/jakevdp/WhirlwindTourOfPython/blob/master/Index.ipynb) - An in-depth, fast-paced introduction to Python with code hosted online

### Other Useful Resources

- Peter Gleeson "**An A-Z of useful Python tricks**" (https://medium.freecodecamp.org/an-a-z-of-useful-python-tricks-b467524ee747)
- George Seif "**23 great Pandas codes for Data Scientists**" (https://towardsdatascience.com/23-great-pandas-codes-for-data-scientists-cca5ed9d8a38)
- **Stack Overflow** (https://stackoverflow.com/questions)- One of the best Q&A sites for technology
- **Datapolitan training classes** https://www.datapolitan.com/ - The online home of our training materials

Contact Us

training@datapolitan.com @datapolitan https://www.datapolitan.com

About Datapolitan

For over 5 years, Datapolitan has worked to empower public sector clients with the skills, techniques, concepts, and mindsets necessary to make data meaningful and actionable to effectively manage their resources and realize operational value from the information they collect. We do this through a range of consulting and training services for local government agencies and nonprofit organizations, customized to their strategic vision and operational needs.

# Key Code Examples From Today

## Exploring a Dataset

```
import pandas as pd
```

```
df.head() # Show the first 5 rows of data
df.count() # Count the number of non-null values in each column
```

```
df.max() # Find the maximum value
df.min() # Find the minimum value
```

```
df.mean() # Find the mean value of all non-null columns
```

```
df.describe() # Provide summary statistics for all columns
```

## Filtering Data

```
df['Column Name'] # Example of the syntax for referencing a single column
df[['column1','column2','column3']] # select columns 1, 2, and 3
df[df['Column'] == 'Value'] # filter rows for "Value" in specified "Column"
df[(df['Column 1'] == 'Value 1') & (df['Column 2'] == 'Value 2')]
df[df['Complaint Type'].str.contains('Noise')] # fuzzy match on "Noise"
```

## Aggregating Data

```
df.groupby(['Column to group'])['Column to count'].count()
```

```
df.groupby(['Borough'])['Unique Key'].count()
```

## Sorting Data

```
df.groupby(['Borough'])['Unique Key'].count().sort_values(ascending=False)
```

## Visualizing Data

```
import matplotlib.pyplot as plt # Imports the visualization package
%matplotlib inline
# This tells Jupyter to show images inside notebook cells
# instead of in a separate window
```

```
df.hist() # Create a histogram
```

```
df.plot(kind='scatter',x='waiting',y='eruptions') # Create a scatter plot
```

```
df.groupby(['Borough'])['Unique Key']\
    .count()\
    .sort_values(ascending=False)\
    .plot(kind='bar',ylim=(0,75000),\
    title='Count of NYC 311 Service Requests by Borough')
```

# Your Notes