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

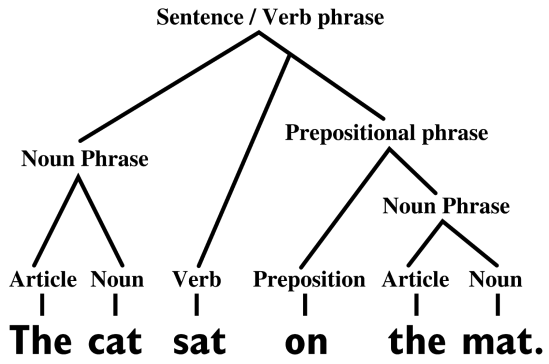
Username: **rstudio** Password: **rstudio**

Key Questions for the Morning

What is python?	How is it useful in analysis?

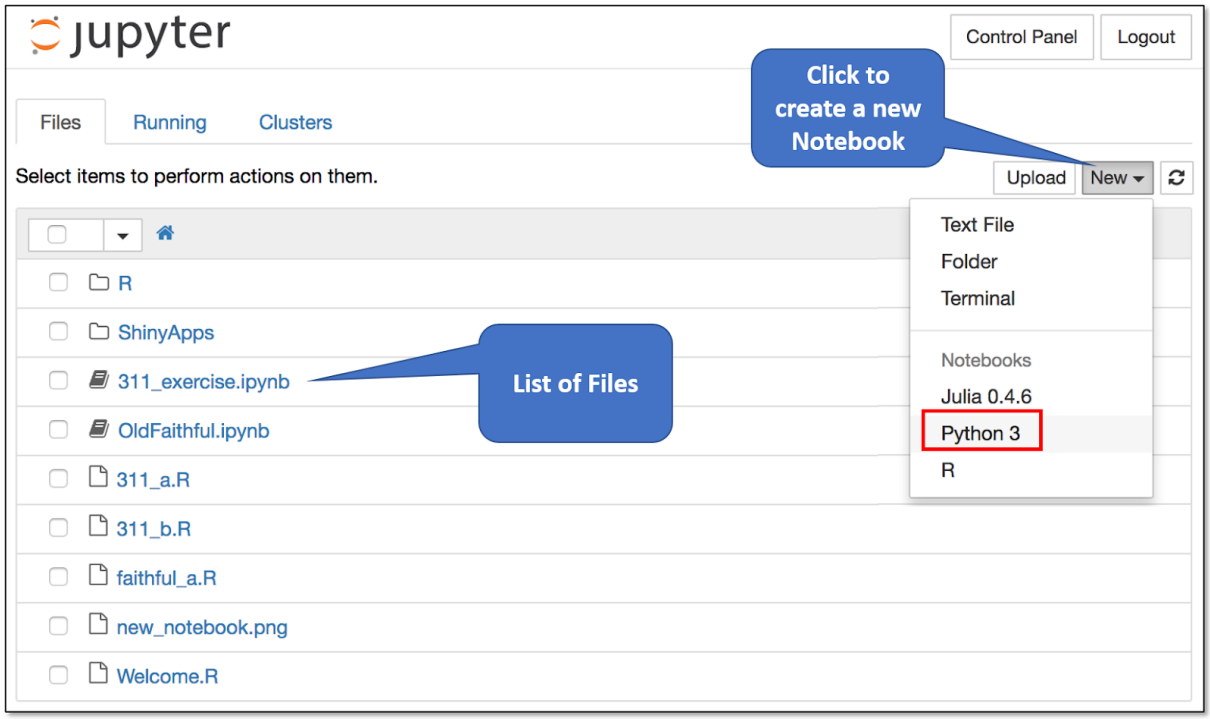
What is Syntax?

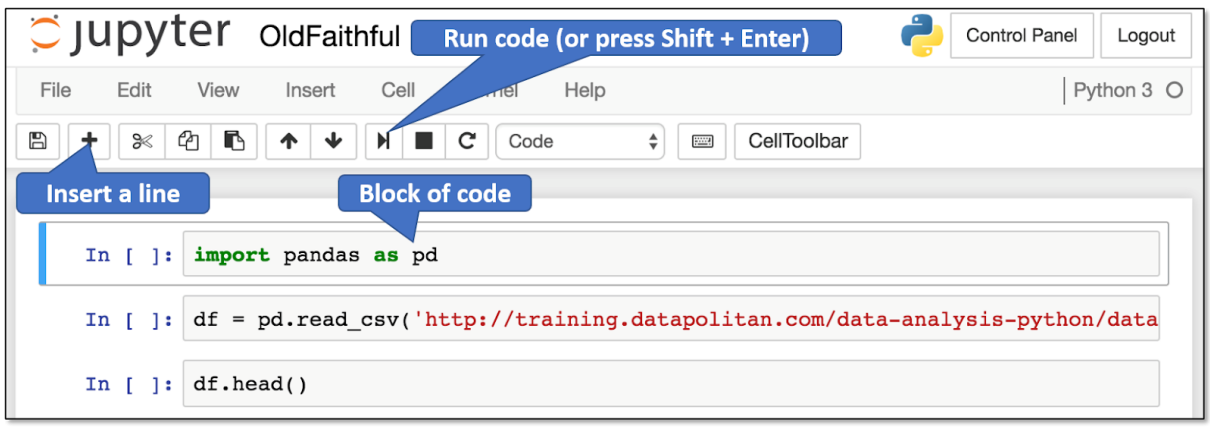
Basic constituent structure analysis of a sentence:



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





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?

Learning More

- **Python for Data Analysis** (http://shop.oreilly.com/product/0636920023784.do) - The textbook on using pandas for data analysis (2nd edition coming soon)
- **Beginner's Python Tutorial** (https://en.wikibooks.org/wiki/A_Beginner%27s_Python_Tutorial)- A good way to get started with basic tasks
- **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
- **Style Guide for Python Code**- https://www.python.org/dev/peps/pep-0008/
- George Seif "**23 great Pandas codes for Data Scientists**" (https://towardsdatascience.com/23-great-pandas-codes-for-data-scientists-cca5ed9d8a38)
- Peter Gleeson "**An A-Z of useful Python tricks**" (https://medium.freecodecamp.org/an-a-z-of-useful-python-tricks-b467524ee747)

Other Resources

- Stack Overflow (https://stackoverflow.com/questions)- One of the best Q&A sites for various technical questions
- Datapolitan training classes https://www.datapolitan.com/

Contact Us

training@datapolitan.com @datapolitan http://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.median() # Find the median value
```

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