Datapolitan

Data Solutions for the Modern Metropolis

tiny PANTHER

Overview of Data Analysis with Python

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

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

What to Expect Today

- 9:15 – Welcome and Introductions
- 9:25 – What is Python?
- 10:40 – 10 min break
- 10:50 - Exercise
- 12:30 – Dismissal

What Not To Expect Today

- Becoming a Python expert
- Becoming a data analytics pro
- Becoming a visualization wizard

What is Python?

- Open-source programming langage
- Useful in standalone scripts or powering fully-featured applications
- Strong support for data analysis and visualization, as well as other programming tasks

Python vs. Excel

- Language vs. program
- Big data
- Different structures and data types
- More potential

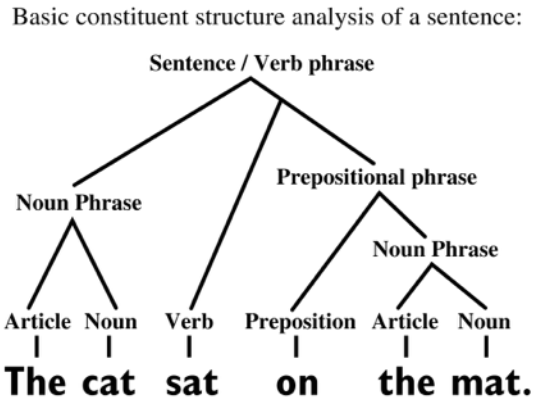
Your student number is:_____

Your weblink is: **http://student__.datapolitan.com**

Username: **rstudio**

Password: **rstudio**

What is Syntax?



Python Syntax

- Variables hold some value
- We assign value using the = sign
- We can perform operations with mathematical operators
- We can use built-in functions for operations

Analyzing the Old Faithful Data

- Import the data
- Inspect the data
- Count the number of rows
- Find the range of values
- Find the mean (average)
- Find the median (middle)

Function Chaining

- We can string operations together using the dot method
- This means we can chain operations using a dot between operations
- Python executes these from left to right (like we read)
- This is a paradigm called object-oriented programming
- You don't need to fully understand this to program in Python but it helps

Your Turn

- How many columns are in the data?
- How many rows are in the data?
- What is the time range of the data?
- Which borough has the most complaints?
- Which Complaint Type has the most service requests?
- And why might that be a little misleading?
- Bonus Question: Find the Location Type that has the most rodent complaints

Key pandas Functions

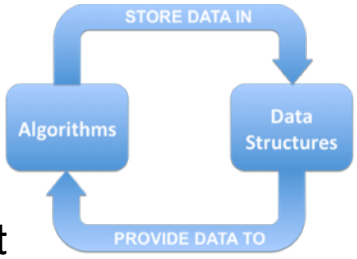
- **read_csv()** - import file from CSV (**read_excel()**)
- **head()** & **tail()** - first and last 5 rows of data frame
- **count()** - count of all rows in column
- **max()** & **min()** - max and min values in column
- **mean()** & **median()** - mean and median values of numbers in column
- **describe()**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

What we've covered

- Basic Python syntax
- Working in Jupyter
- Opening a dataset
- Exploring a dataset
- Visualizing a dataset

What we haven't covered

- Data Structures
- Algorithms
- More Packages and there are a lot of packages
- How to be Pythonic
- How to use APIs
- So much more...



Final Thoughts

- Python is a powerful tool for cleaning, analyzing, and visualizing data
- Integrating it into your workflow takes practice and a commitment to not giving up (Google is your friend)
- Distributions like Anaconda make it easy to get started (and you should be able to install it on your work computer)
- It's best if you just start off with Python 3 (what we've been using)

Richard Dunks

- Email: richard@datapolitan.com
- Website: http://www.datapolitan.com
- Twitter: @datapolitan

Julia Marden

- Email: julia@tinypanther.com
- Website: http://tinypanther.com
- Twitter: @juliaem

Resources

- 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
- Stack Overflow (http://stackoverflow.com/)- One of the best Q&A sites for technical questions of all kinds