Course Outline

**Course Overview and Objectives**

This course is intended to teach data manipulation using Python, mainly using the Pandas library. Upon completion, students will be able to:

- Understand the fundamentals of data modeling and principles of data cleaning

- Read and write different file formats into Pandas dataframes

- Create and use Jupyter notebooks

- Convert unformatted or poorly formatted data into well-structured data

-Group and summarize data

-Work with text data effectively

-Work with dates and times effectively

-Broadly understand and be prepared to study more complex data science topics

**Prerequisites**

-Familiarity with basic Python syntax (Resources: Codecademy, learnpython the hard way, Coursera)

-How to create and when to use:

-Variables

-Lists, dictionaries, strings, floats, integers

-If/then/else

-For/while

-Functions

-Excel proficiency

**Course Materials**

**https://github.com/Robmattles/Python-For-Data-Analysis**

**Modules**

Introduction (10 minutes lecture):

* Motivations for this course
  + What Python can do that Excel can’t
    - Bigger files
    - More flexible
    - Easily repeatable
  + Introduction to more advanced data science
    - Importing data from dynamic sources
    - Natural Language Processing
    - Machine Learning
  + Installing Anaconda
  + Course File layout
  + Class format
  + Stack Overflow
  + Formal documentation

Module 1: Jupyter Notebooks (10 minutes lecture, 10 minutes guided exercise)

* Installing Anaconda
* Navigating files
* Jupyter notebooks
  + As final products
* Cells
  + Markdown
  + Code
  + Displaying Output
  + Running cells
    - Keyboard shortcut
  + Saving/checkpointing
  + %matplotlib inline

Module 2: Python Syntax overview and advanced Python syntax (20 minutes lecture, 2 hour guided exercises)

* Control flow
* Variable types
  + Set
  + Range
* Functions
* Strings
  + In
* Floats and integers
* Packages
* List Comprehensions
  + Using conditionals in list comprehensions

Module 3: Data Modeling, Terminology, and Formats (30 minute lecture with interactive components)

* Modeling
  + Appropriate data types for data types
  + Clean Data
  + Row vs Columns
  + Tables
  + 1-to-1 and relationships between tables
  + Keys
    - Foreign
    - Primary
* Formats
  + Flat
    - Excel
    - CSV
  + Complex
    - Multi-sheet Excel
    - Access
    - JSON
    - HTML
    - Completely unstructured

Module 4: Introduction to Pandas (30 minutes lecture, 1.5 hour guided exercise)

* Packages and how to use them
* Reading and writing data
  + Dataframe concept
  + Filepath slashes
  + From file
    - CSV and similar
      * Skiprows
      * Headers
      * Encoding
    - JSON
  + Manually
* Pandas Dataframe
  + Selecting row(s)
  + Selecting column(s)
  + Adding and removing columns
  + Renaming Columns
  + Pandas functions
    - Standard Syntax
    - Not in place
  + Getting basic summary statistics, value counts, data shapes, columns
  + Basic plots with pandas
  + To\_csv
* Pandas Series

Module 4.5: Parsing Data from Forms and Similar Text Documents

* Opening and viewing files in Python
* Working with strings
  + ‘in’ and slicing to identify key points in the document
  + Replace
  + Strip
* Toggle variables
* Extracting Data to python data types
* Converting extracted data to Dataframes
* Working with multiple files
  + Os.listdir
  + Looping over files
  + Collect data from multiple files into lists or dictionaries
  + Concatenating dataframes

Module 5: Manipulating Data Frames (45 minute lecture, 2 hours guided exercise)

* Subsetting
  + .loc
  + With basic operators
  + With other columns
  + With results of true/false functions like isin
  + Combining conditions with and/or
* Sorting
* Changing column data types
* Merging/joining/concatenating
* Removing Duplicates
* Melt
  + Pivot
* Wide\_to\_long
* Apply
  + Lambda functions

Module 6: Working with Missing Values (10 minutes lecture, 30 minute guided exercise):

* Nulls, Nas and blanks
* Filtering and counting missing values
* Removing rows with missing values
* Filling missing values
  + With constants
  + With bfill and ffill
  + With another column

Module 7: Grouping (15 minutes lecture, 1 hour guided exercises)

* Groupby
  + Single columns
  + Multiple columns
* Aggregation
  + Size
  + Built-ins
  + Custom functions
    - Must reduce
  + Using different functions for different columns
  + Converting to data frame
    - Reset\_index
    - unstack
* Transformations by group

Module 8: Debugging Common Problems (30 minute lecture, 1 hour guided exercise):

* Syntax errors
  + Missing close parens
* How to read an error message
  + Google!
* Try/except
* Using print and working backwards
* Key Error
* Data type errors
* Missing variable errors

Module 9: Dates and Times (20 minutes lecture, 1 hour guided exercises)

* Converting to date-time
* Grouping
* Time Deltas
  + Total\_seconds
* Converting back to strings
* Floor

Module 10: Strings and Text (30 minutes lecture, 3 hour guided exercises)

* Pandas string functions
  + Upper
  + Lower
  + Contains
  + Replace
  + Split
    - Expand
  + cat
* Regex
  + contains
  + Extract and extractall
  + Regex testers
  + Replace

Module 11: Html parsing (30 minutes lecture, 1 hour guided exercises)

* HTML structures
  + Tags
  + Properties
    - Class
    - ID
  + Inner text/HTML
* Saving as text
* Viewing in page inspector
* Beautiful Soup
  + Contents
  + get
* Moving data from Beautiful Soup to Pandas

(Expand into scraping)?

Module 12: Putting it All Together (30 minutes lecture, 2 hour final project/test)

* Combining individual functions into procedures
* Structuring your code
* Best practices for readability
  + Comments
  + Spacing
  + Variable naming
  + Do not repeat yourself
* Best practices for efficiency
  + Ordering your code properly
* Notebooks vs. Scripts
* Locating additional resources online
  + Stack Overflow
  + Official Documentation

Conclusion and Future Directions (15 minutes lecture)

* Excel will feel easier. Use Python anyway.
* Data Analyst position (minus SQL)
* Social Network Analysis
* Natural Language Processing
* Machine Learning
  + Supervised
  + Unsupervised
* Web scraping and APIs
* Advanced Visualizations
* How to seek help
* Where to find opportunities to practice