

# (Intro to ) Data Analysis in Python

John Serences, [jserences@ucsd.edu](mailto:jserences@ucsd.edu)

Feb 27th, 2019

Class 07

# Reminder:

- All in-class code and lecture slides can be found on GitHub
  - [https://github.com/JohnSerences/PSYC193 IntroPython W2019](https://github.com/JohnSerences/PSYC193_IntroPython_W2019)

# Course Schedule (approximate)

- Week00, January 9: What is Python?, Jupyter Environment (Google Colab), First Program, Intro to object types and methods
- Week01, January 16: More on object types, lists, for loops, list comprehensions, slicing lists
- Week02, January 23: If...elif...else statements, dictionaries
- Week03, January 30: User input, while statements, try/except statements
- Week04, February 6: NO CLASS
- Week05, February 13: Midterm, writing functions
- Week06, February 20: Classes, object-oriented programming
- **Week07, February 27: File Input/Output, data formats for files (e.g. JSON)**
- Week08, March 6: NumPy (numerical computing), Plotting (Matplotlib)
- Week09, March 13: Pandas (data frames and Seaborn)
- Final: Room/Time TBD

# In-class work

- To launch a new notebook and do basic setup
  - Log in to: <https://colab.research.google.com/notebooks/welcome.ipynb>
    - Can just google: “google colab”
    - Use your AD credentials to login
  - File menu -> “New Python 3 Notebook”
  - File menu -> “Rename” (or just type in the name field in upper left corner)
    - File name should be: Lastname\_PSYC193\_Class00.ipynb
    - Use exactly this convention EVERY time, only updating the 00 counter (so next class is 01, etc).
  - File menu -> “Locate in drive”
    - This will launch a new window with a file list
    - Right click, Move, navigate to our shared folder and move file
    - Now all your work will be saved in our shared folder...

# In class quiz on material from last week

- Questions from last week...

# File I/O and File Types

- File I/O (input/output) refers to reading and writing files of various types
- Very important skill because if you are just running code in the notebook, all of your data lives in volatile memory – when you close the notebook or crash your computer, its gone!
- Can write out data at various stages of analysis to different types of file format
  - Text
  - Binary

# Text Files

- Contain characters/strings
- Human readable using common editors (WordPad, Preview, etc)
- Common types:
  - CSV (comma separated values) - .csv file
  - Space delimited text - .txt file
  - Tab delimited text - .txt file
  - JSON (Java Script Object Notation) - .json file

# Binary Files

- Computer readable (not human readable)
- Written in specified format (e.g. int8, uint8, int16, float64 etc).
  - These formats set a limit on the encoded numbers
  - For example, uint8 files have the values 0:255 (00000000 to 11111111)
  - Whereas int8 files have the values -127:128
  - More generally, its  $2^{*N}$ , where N is 8,16,64, etc
- Often more efficient than txt files: can pack in more data and computer can make faster use of info (no conversion needed)
- Examples: .exe, .dmg, .png, .bmp, .jpg etc



# Reading and Writing files

- Open the file to get a **file object**
- The file object refers to the file and grants access to related **methods**
- On open, you need to specify a mode: read, write, append, or read/write
- You also specify a format: text (default) or binary
- After your read/write to the file, you must **close** the file!

# Structure of one common file type: jpeg (jpg)

- Joint Photographic Experts Group
- Lossy compressed images
  - Side note: FFT the image and discard information at spatial frequencies that we're not very sensitive too
- Data format:
  - 3D image matrix –  $[X, Y, 4]$ , where X and Y are the dimensions in pixels
  - Each entry in the 3<sup>rd</sup> dimension contains information about the amount of red, green, blue in each pixel, as well as the alpha value for each pixel (which indicates transparency)
  - The values in the 3<sup>rd</sup> dimension span 0:255 to indicate the relative intensity of each color

## In class exercises

- Please finalize by Friday at midnight so that I can grade in time to give feedback.
- We'll reconvene today at 11:30 (or so) to go over answers.
- Does anyone want to go over the exam more? If so, please let me know.