PROGRAMMING DATA SCIENCE FUNDAMENTALS

Matthew Hall

OVERVIEW

- General Workflows
- Tips & Tricks
- Breaking Down Future Training

GENERAL WORKFLOW

Learning The Routine Steps to What We Do.

STARTING A PROJECT

- **Identify (or be assigned) an objective**
  - Should align with business operations & goals
- **Figure out what data you need (or be given)**
- **Identify what the deliverable is going to be**
  - Daily reports, one-time analysis, tracking trends in data
- **Decide what tools you need**
  - Visualization, Analysis, and Storage of Data
- **Keep the big picture in mind!**
  - It'll guide your code and process from A to Z!

WHEN PROGRAMMING

1. Identify End Goal of Program & Requirements
2. Import Your Data & Preprocess if Needed
3. Figure Out Steps Needed to Run Your Analysis
4. Start Implementing Program, Running Tests at Each Step
5. When Done, Look Over Program, Run Tests!
6. Analyze Results, and Gauge Significance

WHEN STUCK

- **Take a Break!!**

- **Remember What You're Trying to Achieve at This Step**
  - And Look for Alternate Methods!

- **If It's a Technical Error, GOOGLE IT!**
  - Stack Overflow, Documentation, Towards Data Science, and Countless Other Sources Probably Have an Answer—Especially when you begin.

TIPS & TRICKS

Some of These I wish I knew, and still mess up.

TIPS TO DEBUG

- **Look For Basic Spelling Mistakes**
  - I've tried to debug something for hours to only realize I spelled a word wrong
- **Did You Reference the Right Variable?**
  - Turns out, referencing the wrong variable messes up for workflow!
- **Name Your Variables in a Meaningful Way!**
  - Don't have 50 variables named var1, var2, var3, … var50
  - If you want to store the current index, name it: current_idx

“It worked yesterday!”

- **It probably did not! If you’re using Jupyter, you may have something in memory that made it work but failed when not there.**
  - Reset your kernel! (The server storing and running everything)
  - Run everything as a Python script

“What did I do last year?”

- **Use the markdown cells and comments to your advantage.**
  - Something you write today might need to be updated in a year, or someone else may need to do that for you!
  - Comments should be in English, unless you’re certain it won’t be used by English speakers*
  - *System I saw in the workplace was developed in Costa Rica, and no one knew what it did because the end users didn’t speak English
- **Write good documentation!**
  - It makes sense now but might not in 2 years.

FUTURE TRAINING

Really Getting Into Data Science!

ARRAYS & DATAFRAMES

NumPy

- **Statistical Tests**
- **Random Number Gen**
- **Slicing & Sub-setting**

Pandas

- **Date & Time**
- **Column & Row-wise Operations**
- **Windows, Rolling, Shifts**
- **SQL-Like Functionality**
  - Group By, Summary Stats, Filters, Pivots
- **Multi-Indexed Data Sets**

DATA VISUALIZATION

MatPlot

- **Scatterplots**
- **Line Graphs**
- **Histograms**
- **Bar Charts**
- **And More!**
  - Making them look good, too

Seaborn

- **Heatmapping**
- **Better Looking Graphs**


MELIORA