Class 13 - Advanced Pandas

[w200] MIDS Python Course Spring 2018

Agenda

Election Data Discoveries

Analysis Design

MatPlotLib

Project 2

Final Exam Preparation

Schedule

Class 10 - Working with Text and Binary Data

Class 11 - NumPy

Class 12 - Data Analysis with Pandas

Class 13 - More Data Analysis with Pandas

Class 14 - Group Work, Code Testing and Final Project Showcase

https://docs.google.com/spreadsheets/d/1Skg_b0rM5jPcVg0ixGrPnK5-QCGrHaFVr1afgchUN5c/edit?usp=sharing

Schedule | Projects/exams

Live Session 11 - Discuss Final Project

Live Session 12 - Proposal Finalized

Live Session 13 - Final Exam Distributed

Live Session 14 - Final Project Showcase

https://docs.google.com/spreadsheets/d/1Skg_b0rM5jPcVg0ixGrPnK5-QCGrHaFVr1afgchUN5c/edit?usp=sharing

Schedule | Due Dates

Final Exam - Before last class (Tues: April 17 Thurs: April 19)

Final Projects - 11:59 PM PST, day after last class

(Tues: April 18 Thurs: April 20)

If you need an extension please email all four of us ASAP with your request!

https://docs.google.com/spreadsheets/d/1Skg_b0rM5jPcVg0ixGrPnK5-QCGrHaFVr1afgchUN5c/edit?usp=sharing

Assignment Review | Week 12

Discussion: What did you learn from the Election Data?

Agenda

Election Data Discoveries

Analysis Design

MatPlotLib

Project 2

Final Exam Preparation

Pandas | Analysis Design

You can think about an analysis as a **series of dataset transformations**

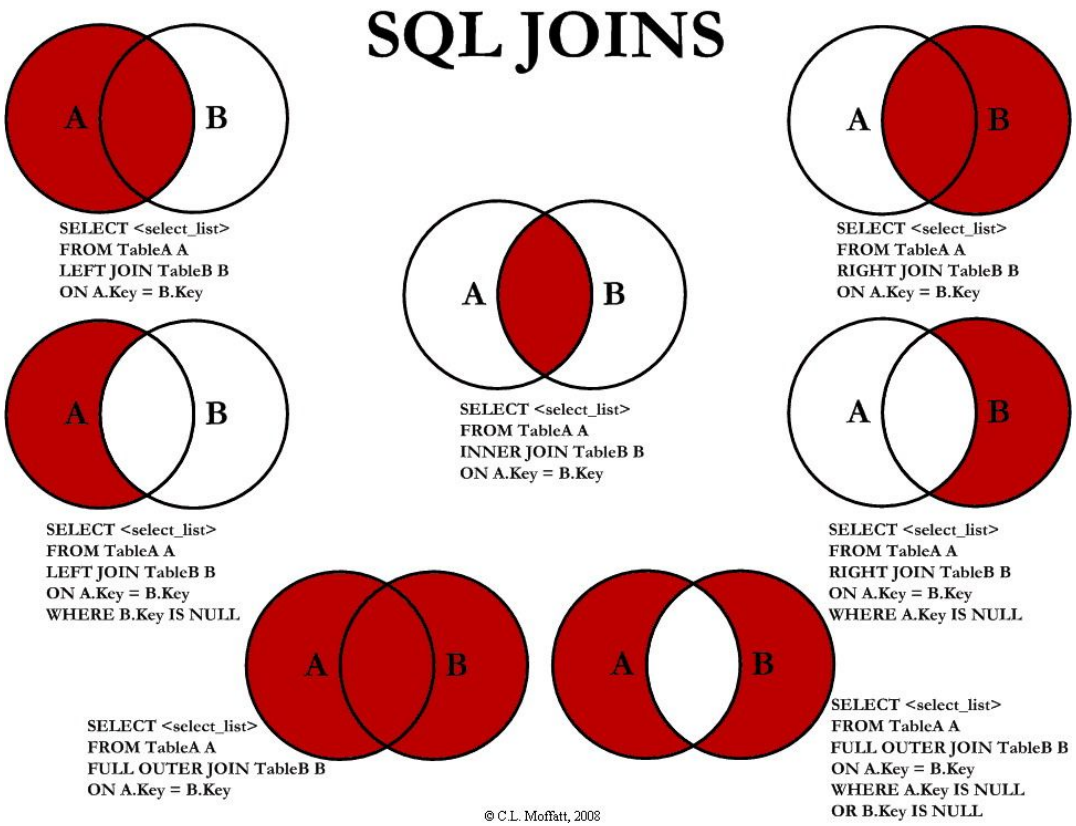
You might filter **out rows based on conditions**

You might **create new columns**

You might **aggregate** or **collapse by groups**

You might **join two datasets together**

Pandas | Join Types - Discuss



Pandas | Some Functions

- groupby()
- cut()
- agg()
- apply()
- reset_index()
- pivot()

Agenda

Election Data Discoveries

Analysis Design

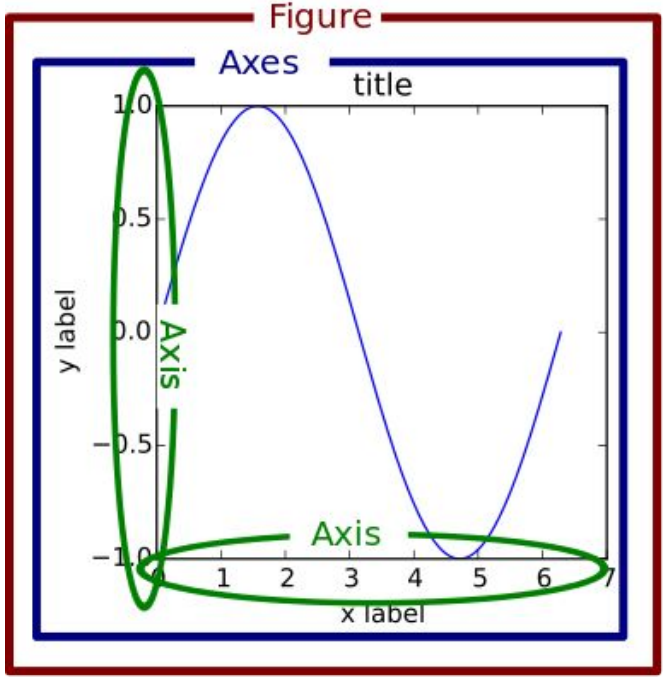
MatPlotLib

Project 2

Final Exam Preparation

MatPlotLib | Overview

```
fig = plt.figure()  # an empty figure with no axes
fig, ax = plt.subplot()  # a figure and axes
```



Pulling it All Together | Demo

Jupyter Lab

Agenda

Election Data Discoveries

Analysis Design

MatPlotLib

Project 2

Final Exam Preparation

Grading | Reminder of Breakdown

1. Homework (30%)
2. Midterm (10%)
3. Project 1 (20%)
4. Final (10%)
5. Project 2 (20%)
6. Participation (10%)

Project 2 | Preliminary presentation

This will be mostly a “working session” for project teams. We will balance time between your project group and breakouts where you can discuss challenges with others.

Lets work up to a 2-minute “elevator pitch” on their project to the full room, followed by 2-minutes of Q&A. Please pick who you would like to present.

Project 2 | Grading

- Proposal **(10%)**
- 10-15 Minute Final Class Presentation **(20%)**
- Report **(70% as follows)**
  - Lay out the question and describe the data set clearly. That includes defining columns and the source of the data **(10 pts)**
  - Check the data for internal inconsistencies and convince us that you know your dataset **(20 pts)**
  - Tell a story that shows significant exploration of the data set in text and appropriately figures **(40 pts)**
    - Roughly **20 pts** will relate to your text, and **20 pts** to your figure -- but we may be flexible on this if you have particularly compelling stories or figures

Project 2 |The Team Review

- Contributions in the paper (e.g Jim Bond: data cleaning -70%, writing 10%)
- You will also be asked to answer a quick survey about your team to ensure everyone has contributed.

Project 2 | Team Feedback

We'll take 30 minutes now to let you work as a group.

For the first **10 minutes**, you will be with your group to plan and discuss your project.

For the second **15 minutes**, I will combine groups together. Discuss your projects, and give each other feedback.

For the last **5 minutes**, you will be back with your own team to recap and close out.

Agenda

Election Data Discoveries

Analysis Design

MatPlotLib

Project 2

Final Exam Preparation

Final Exam | Logistics

Final Exam (10%) - Due by Class 14.

You will have 24 hours to complete the exam. It will cover:

1. Object Oriented Programming (briefly)
2. Data Analysis

Much of the exam will be short answer or discussion format

However, there will be some short problems that require you to code.

Final Exam | Content Reminder

Unit 7 - Classes
Unit 8 - Object-Oriented Programming
Unit 10 - NumPy and Functional Programming
Unit 11 - Data Analysis with Pandas
Unit 12 - More Data Analysis with Pandas

Final Exam | Review

Please answer the following questions...

Final Exam | Review

- What is inheritance?
- What is polymorphism?
- Why might you use either?
- What are the products in the PyData Ecosystem?
- When should you use NumPy? What about Pandas?
- Let’s talk about how to explore a dataset… what do you do?
- Why is data exploration important? Make up a horror story.
- What is a good process for designing an analysis?
- What are two methods of accessing variables in a dataset?
- What is the difference between “groupby” and “agg”?