Class 13 - Advanced Pandas

[w200] MIDS Python Course

Today's Agenda

1. Schedule | Project/Exams
2. Election Data Discoveries
3. Analysis Design
4. Data Vis & Plotting Review (e.g., MatPlotLib)
5. Project 2 – schedules for presentations, expectations
6. Final Exam Preparation

1 Schedule | Projects/exams

This week … we review some of our last new material with Panda …

The final exam will be released shortly.

As usual, you have a full week in which to complete the exam. Once starting the exam you have a 24-hour clock.

Please complete the "student evaluation" site when it is convenient for you (before 12/14).

Good luck!

If you need an extension please email your request **asap** all instructors & TAs!

2 Assignment Review

Discussion: What did you learn from the previous home-works, such as the election data?

3 Pandas | Analysis Design

Think about an analysis as a **series of dataset transformations … why?**

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

You might **create new columns**

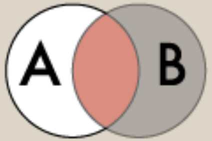
You might **aggregate** or **collapse by groups**

You might **join two datasets together**

The next slide is an example of how our commands can include/exclude data, using SQL and Venn diagrams to demonstrate.

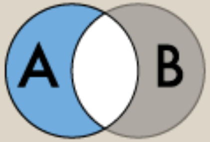
3 Pandas | Join Types - Discuss

SELECT <list> FROM tableA.A
LEFT JOIN tableB.B
ON A.key = B.key

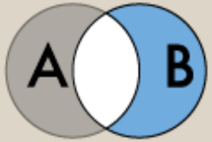


SELECT <list> FROM tableA.A
INNER JOIN tableB.B
ON A.key = B.key

SELECT <list> FROM tableA.A
RIGHT JOIN tableB.B
ON A.key = B.key



SELECT <list> FROM tableA.A
LEFT JOIN tableB.B
ON A.key = B.key
WHERE B.key IS NULL



SELECT <list> FROM tableA.A
LEFT JOIN tableB.B
ON A.key = B.key
WHERE A.key IS NULL

SELECT <list> FROM tableA.A
FULL OUTER JOIN Table B.B
ON A.key = B.key

SELECT <list> FROM tableA.A
FULL OUTER JOIN Table B.B
ON A.key = B.key
WHERE A.key IS NULL OR B.key IS NULL

3 Pandas | Some Functions

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

There's an example of "groupby" on the next slide to suggest how we create the data, group those data, and then start to review/analyze them.

```
import pandas as pd

raw_data = {'Group': ['Cats','Cats','Cats','Dogs','Dogs','Dogs'],
    'company': ['1st','1st','2nd','1st','2nd','2nd','1st'],
    'name': ['Tom', 'Jane', 'Abdul', 'Ming', 'Jeff', 'Toby'],
    'preTestScore': [5, 5, 4, 3, 4, 5],
    'postTestScore': [6, 6, 7, 7, 6, 6]}
df = pd.DataFrame(raw_data, columns='Group','company','name','preTestScore',
'postTestScore'])
```

Create some data and a frame …

```
groupby_group = df['preTestScore'].groupby(df['group'])
# use list to see 'em…
list(df['preTestScore'].groupby(df['Group']))

# descriptive stats by group
df['preTestScore'].groupby(df['Group']).describe()
```

Group and list these data to examine.

Why not some desc stats, too?

```
for name, group in df.groupby('Group'):
    #print group name
    print(name)
    # print the data of that Group
    print(group)
```

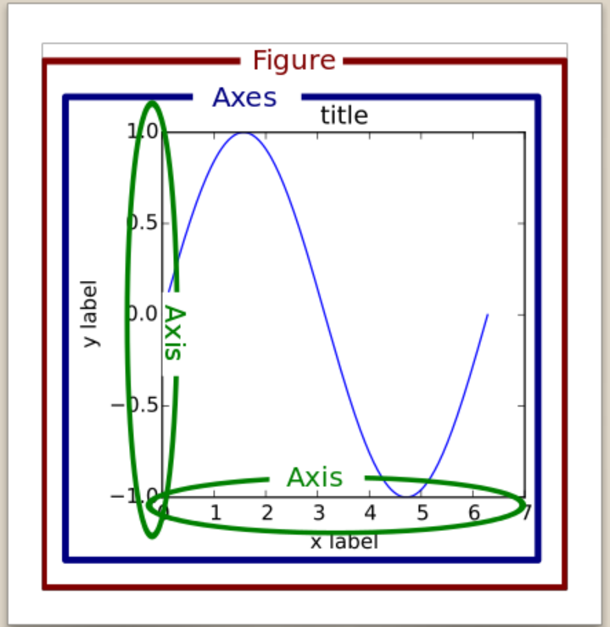
Let's view and perhaps share these results.

4 Data Vis & Plotting (Overview)

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

Recall: We need a "container" to hold the drawing area. In matplotlib we might use a "figure" object; in HTML5 there's the "Container" object; in Java we have JFrame; in Adobe Illustrator you'll encounter the "bounding box." They're all functionally the same - they designate a Graphic drawing area for "painting" your data.

Note, btw, that most of the drawing area commands may have a built-in "erase - then paint" command. Just FYI.



4 Pulling it All Together | Demo

We've seen how matplotlib + seaborn are popular and with practice become useful tools in your DataSci arsenal. In practice people often combine Stata, R, NumPy, Pandas and other tools for data cleaning, analysis, statistical work, and presentation of data.

5 **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.

Let's work up to a 2-minute "elevator pitch" on their project to the full room, followed by 2-minutes of Q&A. Please pick whom you would like to present.

5 **Grading** | Reminder of Breakdown

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

5 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

5 **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.

6 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

There will be some short problems that require you to code.

6 Final Exam | Review

We know and you know you can program in Python!
Let's give you a chance to write about your knowledge, too.

- 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"?

Thanks!

Before we progress to your presentations, we are all glad to have had each of you in our classes and the DataScience program.

Thanks for all your hard work and persistence. Best wishes for great success in the exam and your presentations and, of course, in all your MIDS courses.