W200 Week 10

Why pandas

- Pandas has been one of the most commonly used tools for Data Science and Machine learning, which is used for data cleaning and analysis.
- Pandas is the best tool for handling this real-world messy data. And pandas is one of the open-source python packages built on top of NumPy.
- Handling data using pandas is very fast and effective by using pandas Series and data frame, these two pandas data structures will help you to manipulate data in various ways.
- Based on the features available in pandas we can say pandas is best for handling data. It can handle missing data, cleaning up the data and it supports multiple file formats. This means it can read or load data in many formats like CSV, Excel, SQL, etc

My pandas go to functions

1. Read_csv - Read a csv
2. head() - quick review of what is in your DF
3. describe() - shape , size , data type
4. asType() - helps get your data to the right data type as needed for your exploration (or personal preference)
5. Loc[:] - Think select function
6. value_counts() - While checking null count likely - What have you used it for ?
7. drop_duplicates() - Cant get more straight forward
8. groupby() - Similar to SQL .. but be aware it changes the data frame setup - Helpful but be fully aware of what it does
9. merge()
10. sort_values()
11. fillna()

Some common starting points as you are ready to explore - EDA

Get to know your data first

1. What's the dimensions of our data? `my_dataframe.shape`
2. Get some info about your data: `my_dataframe.info()`
3. See the first 5 rows (the head) of your data: `my_dataframe.head()`
4. last rows: `my_dataframe.tail()`
5. remove duplicates `drop_duplicates()`
6. copy the frame to a new one, or overwrite it `my_dataframe.drop_duplicates(inplace = True)` (We can keep the first of the duplicates (the default) or drop all duplicates:`temp_df = my_dataframe.append(my_dataframe)` tomake a copy; `temp_df.drop_duplicates(inplace=True, keep=False)` would overwrite the df and remove all duplicates (keep=False).

Fixing Column Names, Extracting some data and counting ...

Consider stripping out and cleaning the titles of your columns (their metadata):

- *df*.columns.str.replace('\.*\s+', '_').str.strip('.').str.lower()
- Now try counting ... *df.column_name*.value_counts()
- Means? *df.column_name*.groupby(*df.col_name*).mean()
- Highest and lowest values?

Data Exploration

Preparing for Weeks 11-14 and our Project: Data Exploration & Analysis

Data **Exploration**:

- For data integrity
- To develop questions based on the variables
- To break your model - better now than in production

Data **Analysis**:

- Answer a research question or hypothesis
- Usually involves complex math, modeling statistics
- Likely to combine datasets
- Explore data by collapsing in groups in various ways
- Some functions are useful in exploration & analysis

Why use Pandas?

Pandas and other tools help in machine learning, scaling data, multiple regression, and more! Usually work together with numpy, scipy, matplotlib, scikit-le pysqlite3, psycopg2, and others. These two sites introduce a *lot* of features for python to read lots of data sources and more efficient problem-solving techniques: https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html and https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html#io-hdf5. For even more cool things, like Panels, see the documentation at https://www.tutorialspoint.com/python_pandas/python_pandas_panel.htm.

Data Exploration

Discussion Option: Work Processes and Systems Overview in Exploration & Analysis

Take time to get to know your data - the domain & range, parametric or non-parametric; explore the measures of central tendancy ... Here are some handy commands:

- `value_counts()`
- `describe()`
- `min(), max(), isnull()`
- Plot your data during exploration, too, not just during analysis.
- Consider the source(s) of your data and research the topic:
  - basic research methods require looking at threats to validity,
  - cross-validating the data,
  - issues of research "bias",
  - lack of precision in definitions (e.g., mismatched metadata when combining data)
  - look for any professional/industrial gold standards for measurement
- what might be confounding events in your data?

A standard text for new graduate students is learn about how research is expected to be conducted, the process, research questions, and more. You might want to read Booth, et al., *Craft of research*.

Activity - How would you do this - You can google it , but would be nice if you send me a screenshot of your notebook cell

Read a file or manually create a data frame - How would you execute these ?

- Drop column
- Fix header names
- Transpose axis
- Transform - add a new column at the end at the end of your data frame

Send me a bocurse email with your answer