MSDS 7333 Quantifying the World

Example with Pandas: Baby Names

Lesson 1 (from Learn Pandas)

- Create Data
- Export Data to a text file
- Read Data from a text file. The data consist of baby names and the number of baby names born in the year 1880
- Prepare Data
  - Look for any missing data, inconsistencies in the data, or any other data that seems out of place
  - Make decisions on what to do with any anomalous records
- Analyze Data - Find the most popular name in a specific year
- Present Data – tables and graphs

Putting Names with Numbers

- *df['Names']* - This is the entire list of baby names, the entire Names column
  df['Births'] - This is the entire list of Births in the year 1880, the entire Births column
  df['Births'].max() - This is the maximum value found in the Births column
- [df['Births'] == df['Births'].max()] **IS EQUAL TO** [Find all of the records in the Births column where it is equal to 973]
  df['Names'][df['Births'] == df['Births'].max()] **IS EQUAL TO** Select all of the records in the Names column **WHERE** [The Births column is equal to 973]
- An alternative way could have been to use the ***Sorted*** dataframe:
  Sorted['Names'].head(1).value
- The ***str()*** function simply converts an object into a string

US Baby Names 1880-2010

- Visualize the proportion of babies given a particular name over time
- Determine the relative rank of a name
- Determine the most popular names in each year or the names with largest increases or decreases
- Analyze the trends in names: vowels, consonants, length, overall diversity, changes in spelling, first and last letters
- Analyze external sources of trends: Biblical names, celebrities, demographic changes

Examine the Files

- Get Data here
- Download national data file
- Unzip the names.zip file
  - Files by year from 1880 to 2010
  - Files contain the top 1000 names with at least 5 occurrences
- Import yob1880 into Python
- Import all years into a single Data Frame

Baby Names

- Group data by year and sex
- Obtain a column that gives the proportion of babies given each name relative to the total number of births
- In Python < 3.0, the births variable must be cast as a float to do the division
- Extract a subset of 1000 most popular names
- Analyze naming trends with some plots
- Analyze diversity of names with plots

Other References

- pandas tutorials
- Python for Data Science
- Python for Data Analysis