Handling Data with Pandas

Review different ways to pull data into pandas and the link between objects in python and pandas
Understand the differences btwn a DataFrame and a Series
Practice part of the ACES data exploration model
Learn imputation strategies

Panadas will work across a variety of data inputs, including csv, excel, JSON, and using additional python libraries to connect to databases. For today, we will focus on using the csv input. We'll use data about heart disease from the UCI Machine Learning data repository.

http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

there are a few diff data types, some processed, some not, and a "names" file. The names file will expose for us the columns for the processed data:

```
7. Attribute Information:    -- Only 14 used
-- 1. #3  (age)
-- 2. #4  (sex)
 -- 3. #9  (cp)
-- 4. #10 (trestbps)
-- 5. #12 (chol)
-- 6. #16 (fbs)
-- 7. #19 (restecg)
-- 8. #32 (thalach)
-- 9. #38 (exang)
-- 10. #40 (oldpeak)
-- 11. #41 (slope)
-- 12. #44 (ca)
-- 13. #51 (thal)
-- 14. #58 (num)       (the predicted attribute)
```

lets grab those fields as headers, and the processed Cleveland data to work with in pandas (the .names file refers that this is the primart file used in research). Pandas io tools [2] handles http requests to grab files from the internet, though reminder that when doing so, it only saves the file in memory (in python), and not as a file on your machine.

Columns are attributes, number of attributes is m
Number of observations is n
Cancer is our target variable, should be dependent on inputs but be independent
Just because a variable is dependent doesn't mean it's a target variable
Can have multiple target variables

Int + Float -> Float
Float / Int -> Float
(Int + Float) / Int => Float

u means Unicode in python
terminal comands
echo #PATH path veriable
./ will run a local program
which –a python will tell you what is running python
import sys
print sys.path shows what files and directories python is using

Pandas
pd.DataFrame? documentation
also see all pd. Then press tab
also shift-tab gets documentation
put ?? gives you docString and sourcecode

ipython
dd deletes cells

Pandas is a tool that allows us to represent and summarize data in tabular form
UCI dataset has that form

Pandas is really just python
Pandas is a library for python, built heavily around the task of manipulating and presenting data. If you're writing pandas code, you're writing python code! Pandas contains primarily two new python objects:
DataFrame: a wrapper around a 2 dimensional numpy ndarray (in math we call this a matrix)
Series: a wrapper around a 1 dimensional numpy ndarray (in math, we call this a vector)

Math
When working through data matrices and vectors, we'll also often use the words feature and observation.
Feature: a feature is represented by a column. It is a segmentation of your data. Features are usually either continuous values (representing –ing to ing, and 1<2) or categorical values (each values represents its own space; 1 1!< 2).
Observation: an observation is a row of your data. It should represent a single entity (for example, a survey responder).
Target Variable: often we'll be working with a column called a target variable, or predicted value. In data analysis, it is often the goal to be able to statistically explain this variable using the observations and features.

Archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/

1. Bring the data into a dataframe with pandas. This file is auto-mpg.data. Since it is space separated, you'll have to tell read_csv to use spaces ('\s+') and not commas (',') as the delimiter. Likewise, you have to name the columns.

2. Compare the data for cars from the year 1970 and the year 1982. In general in this data set, have cars changed in terms of mpg, horsepower, etc? (any of the continuous values)
3. Horsepower is missing several values. What are some basic techniques to fill in the missing data?

More advanced

1. Consider splitting the data by year and filling in horsepower that way. What would the python code look like to handle this?
2. Read doc for split, apply, combine.
3. To understand the choices in data storage ready about tidy data
4. Matplot lib and seaborn

DataFrames behave like lists
DataFrames support many of the functionalities of lists, like slicing.

DataFrames also behave like dictionaries

DaraFrames sipport returning by column in a similar way a dictionary returns by key. Note that passing a string for a key will return a pandas Series, while a list of keys will return a Data Frame

Python we have the list type
Numpy took lists and put them on steroids- inspired by matlab – matrix laboratory
Numpy handles numerical analysis – solving computations representing systems of equations – linear algebra- optimized to make quick calculations
On top of numpy there is pandas, pandas extends numpy
Pandas is a tool for representing and performing calculations on tabular data
Tabular data is called a dataFrame
Also handles things like plotting, statistical tools, also analysis
Pandas is representing a numpy nd array, and numpy is using python lists

Numpy.ndarray
Pandas is wrapping the dataframe in indices, the column and rows that make up our dataframe

Numpy stores data at a low level written in c – very fast represents data – an average or statistical analysis

Pandas is just an interface to interact at numpy, all the stuff it is doing is through numpy
For example hear_data['age'].mean() the mean calculated is being done by numpy on a pandas object

Pandas – datafram- there are a bunch of methods- these methods are numpy
It is calling numpy internally in source code

It is very difficult to program those matricies calculations
Numpy Calling low level functions that are implemented in c but in python environment
Numpy is using a bunch of tools that are written in c called basic linear algebra support, takes the c implantations and wraps them in a familiar c interface and conforms to how we interact with lists in python

You can also represent a dataframe as a list of dictionaries

Selecting data: is filtering data that fills required criteria
When you are selecting data you are reducing or 'filtering' the number of rows or a (subset of all rows)
Ex. All patents older than 50, patients who don't have cancer

Projection- filtering/removing a subset of columns.
Ex. Height is irrelevant, so exclude that column

List is an array
The following tables of code shows similar code, dependent on your object type: