Lecture 6 Notebook

September 11, 2018

1 ER190C Lecture 6 Notebook

Data Cleaning and Exploratory Data Analysis

Duncan Callaway

September 11 2018

Today we'll work with PurpleAir data to explore the concepts of Structure, Granularity, Scope, Temporality and Faithfulness. Along the way we'll talk about data cleaning as well.

Here's PurpleAir's website -- They have really cool maps!

The way I developed this lecture was by pulling the data down and exploring it. You'll see my (edited) process of examining the data.

This began by me visiting this website to look for data. I used the Chrome browser to pull data (other browsers didn't work).

The folks are PurpleAir also sent me this pdf describing their data.

```
In [1]: import numpy as np
        import pandas as pd
        import os
```

1.1 Structure.

First let's look at what's in the data directory using `os.listdir` (remember this is a set of command line-style commands that work across platforms, i.e. mac, linux, windows)

```
In [2]: os.listdir('data')
```

```
Out[2]: ['.DS_Store',
         'Ecole Bilingue de Berkeley (37.854830799999995 -122.28937169999999) Primary 08_05_201
         'Ecole Bilingue de Berkeley (37.854830799999995 -122.28937169999999) Secondary 08_05_1
```

What can we learn from these file names? * the sensor location is probably the French School in Berkeley. * Looks like lat / lon coordinates in parens * the date range is listed * there is a secondary / primary distinction.

Before proceeding let's find the size of some of these files:

```
In [3]: os.path.getsize('data/Ecole Bilingue de Berkeley (37.854830799999995 -122.289371699999
```

```
Out[3]: 2381187
```

What are the units? Let's shift tab in to `getsize` to find out.

```
In [4]: os.path.getsize

Out[4]: <function genericpath.getsize>
```

Not much information. Google search reveals this information page, which says the units are bytes.

```
In [5]: os.path.getsize('data/Ecole Bilingue de Berkeley (37.854830799999995 -122.289371699999

Out[5]: 2.381187
```

SO 2.4 Mb.

```
In [6]: os.path.getsize('data/Ecole Bilingue de Berkeley (37.854830799999995 -122.289371699999

Out[6]: 2.497975
```

Before we go furhter, what's the primary vs secondary data file?

Checking out the "Using Purple Air data" pdf, provided to my by them, it looks like the two files contain different data. We'll focus on PM2.5, which is in the primary file.

In this directory there is a python file (`utils.py`) that has some useful utilities -- we'll pull some in over the course of the lecture. First to use is `line_count`

```
In [7]: from utils import line_count

In [8]: help(line_count)

Help on function line_count in module utils:

line_count(file)
    Computes the number of lines in a file.

    file: the file in which to count the lines.
    return: The number of lines in the file



In [9]: line_count('data/Ecole Bilingue de Berkeley (37.854830799999995 -122.28937169999999) Pr

Out[9]: 29894

In [10]: from utils import head

In [11]: head('data/Ecole Bilingue de Berkeley (37.854830799999995 -122.28937169999999) Primary

Out[11]: ['created_at,entry_id,PM1.0_CF_ATM_ug/m3,PM2.5_CF_ATM_ug/m3,PM10.0_CF_ATM_ug/m3,Uptime
          '2018-08-05 00:00:31 UTC,111170,1.96,4.34,4.96,135.00,-67.00,84.00,33.00,4.34\n',
          '2018-08-05 00:01:51 UTC,111171,2.13,3.89,6.83,136.00,-67.00,84.00,33.00,3.89\n',
          '2018-08-05 00:03:11 UTC,111172,3.04,4.93,6.18,137.00,-68.00,84.00,34.00,4.93\n',
          '2018-08-05 00:04:31 UTC,111173,2.17,4.26,6.83,139.00,-65.00,84.00,33.00,4.26\n']
```