1.Data munging

In this challenge, you'll practice the command line concepts you've learned so far by munging datasets using just the command line. [Data munging](https://en.wikipedia.org/wiki/Data_wrangling) involves transforming datasets to make them easier to work with. Some datasets are too large to load into Python, so looking at them or transforming them beforehand can be useful. Even for smaller datasets, simple exploration on the command line is faster than exploration in Python, and file-based tasks like unifying datasets can be faster on the command line.

You'll be interacting with datasets on U.S. housing affordability from the [U.S. Department of Housing & Urban Development](http://www.huduser.org/portal/datasets/hads/hads.html) in this challenge. To start things off, let's explore the datasets in the first few steps.

home/dq$ ls -la

total 6740

drwxr-xr-x 1 dq dq 4096 Oct 30 17:14 .

drwxr-xr-x 1 root root 4096 Jun 14 13:43 ..

-rw-rw-rw- 1 dq dq 4496 Oct 30 17:14 .bashrc

drwxr-xr-x 1 dq dq 4096 Jun 14 13:43 .byobu

drwxr-xr-x 1 dq dq 4096 Jun 14 13:43 .cache

drwxr-xr-x 1 dq dq 4096 Jun 14 13:43 .config

-rw-r--r-- 1 dq dq 25 Jun 14 13:43 .gitconfig

-rwxr-xr-x 1 dq dq 2051577 Oct 30 17:14 Hud\_2005.csv

-rwxr-xr-x 1 dq dq 1874334 Oct 30 17:14 Hud\_2007.csv

-rwxr-xr-x 1 dq dq 2902856 Oct 30 17:14 Hud\_2013.csv

drwxr-xr-x 1 dq dq 4096 Oct 30 17:14 .ipython

drwxr-xr-x 2 dq dq 4096 Jun 14 13:43 .jupyter

drwx------ 3 dq dq 4096 Jun 14 13:43 .local

-rw-rw-rw- 1 dq dq 17 Jun 14 13:03 .tmux.conf

**Data exploration**

It looks like there are 3 different CSV files, each corresponding to a separate year.

You learned about the tail command to display the last n rows in a file. To display the first nrows (10 by default), you can instead use [the head command](http://bit.ly/22Ia4gh).

Instructions

* Use the head command to display the first 10 rows of each of the 3 CSV files.

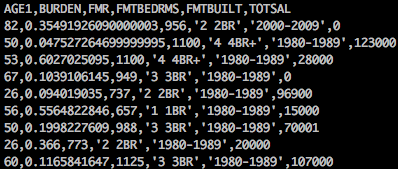
head –n 10 Hud\_2013.csv

tail –n 10 Hud\_2013.csv

**Filtering**

The goal is to eventually get this into a Pandas Dataframe so let's combine the datasets into one file so it can be read in easily. Since each dataset contains the same columns, you need to combine the datasets into one file. You can't, however, just use append the full contents of each file to one final file since each dataset contains the header row. The consolidated file should only contain the header row once (in the first row). You need to instead append the header row to the consolidated file once, then append only the non-header rows from the 3 datasets to the consolidated file.

Here's a reminder of how the first 10 rows of Hud\_2013.csv looks like:



Since the header row is always the first row in each of the datasets, you can just select all rows after the header row. You can use [the command wc](http://bit.ly/1ZqAXDh) along with the l flag to return the number of lines for a specified file. You can use each file's line count combined with the tail command to return the last n lines of a file.

Instructions

* Create the file combined\_hud.csv and append the header row from one of the datasets.
* Select all non-header rows from Hud\_2005.csv and append to combined\_hud.csv.
* Display the first 10 rows in combined\_hud.csv to verify your work.

/home/dq$ wc -l Hud\_2005.csv

46854 Hud\_2005.csv

/home/dq$ wc -l Hud\_2007.csv

42730 Hud\_2007.csv

/home/dq$ wc -l Hud\_2013.csv

64536 Hud\_2013.csv

/home/dq$ touch combined\_hud.csv

/home/dq$ head 1 Hud\_2005.csv > combined\_hud.csv

head: cannot open ‘1’ for reading: No such file or directory

/home/dq$ head -n 1 Hud\_2005.csv > combined\_hud.csv

/home/dq$ head combined\_hud.csv

AGE1,BURDEN,FMR,FMTBEDRMS,FMTBUILT,TOTSAL

/home/dq$ tail -n 46853 Hud\_2005.csv >> combined\_hud.csv

/home/dq$ head -n 10 Hud\_2005.csv

AGE1,BURDEN,FMR,FMTBEDRMS,FMTBUILT,TOTSAL

43,0.513,680,'3 3BR','1980-1989',20000

44,0.2225915493,760,'4 4BR+','1980-1989',71000

58,0.2178312657,680,'3 3BR','1980-1989',63000

22,0.21745562129999998,519,'1 1BR','1980-1989',27040

48,0.28285714289999997,600,'1 1BR','1980-1989',14000

42,0.2922857143,788,'3 3BR','1980-1989',42000

-9,-9.0,702,'2 2BR','1980-1989',-9

23,0.14475,546,'2 2BR','1980-1989',48000

51,0.2962,680,'3 3BR','1980-1989',58000

**Consolidating datasets**

Looks good! Now finish the job by adding in the data from the other datasets.

Instructions

* Append the remaining datasets in the order of the years they describe.
  + Select all non-header rows from Hud\_2007.csv and append to combined\_hud.csv.
  + Select all non-header rows from Hud\_2013.csv and append to combined\_hud.csv.
* Display the last 10 rows of combined\_hud.csv and verify that they match the last 10 rows of Hud\_2013.csv.

**/home/dq$ tail -n 42729 Hud\_2007.csv >> combined\_hud.csv**

**/home/dq$ tail -n 64535 Hud\_2013.csv >> combined\_hud.csv**

**/home/dq$ head -n 10 combined\_hud.csv**

**AGE1,BURDEN,FMR,FMTBEDRMS,FMTBUILT,TOTSAL**

**43,0.513,680,'3 3BR','1980-1989',20000**

**44,0.2225915493,760,'4 4BR+','1980-1989',71000**

**58,0.2178312657,680,'3 3BR','1980-1989',63000**

**22,0.21745562129999998,519,'1 1BR','1980-1989',27040**

**48,0.28285714289999997,600,'1 1BR','1980-1989',14000**

**42,0.2922857143,788,'3 3BR','1980-1989',42000**

**-9,-9.0,702,'2 2BR','1980-1989',-9**

**23,0.14475,546,'2 2BR','1980-1989',48000**

**51,0.2962,680,'3 3BR','1980-1989',58000**

**Counting**

Now that you have a consolidated dataset, you can start to answer basic questions on the entire dataset.

Instructions

* Count and display the number of lines in combined\_hud.csv containing 1980-1989.

**/home/dq$ grep "1980-1989" combined\_hud.csv | wc -l**

**19711**

In this challenge, you learned about a few useful commands for exploring files and practiced data munging from the command line. Next in this course is a guided project where you'll explore how to create Python scripts from the command line for more robust and reusable logic.