**Open Source Tools Assignment 1**

**Due: Wednesday, October 7, 2015 at 11:59pm**

**Overview**

In this assignment, we'll use some of the UNIX tools we have seen so far to analyze some real world data.  Specifically, we will look at information about the New York City subway system.  Data from the [MTA developer website](http://web.mta.info/developers/sbwy_entrance.html) was extracted and put into the directory /home/unixtool/data/mta.  Inside this directory, there are a few scripts that were used to fetch and extract the data, as well as the raw web pages that provided the data (you are welcome to explore but the contents are not necessary for the assignment). We have extracted the data into subdirectories that you will be using to answer some of the questions. The structure of the data you will need is the following: there is a subdirectory for each subway *division*. There are three divisions: IRT, IND, BMT (historically New York had three transit agencies that have since been merged; see [here](http://www.nycsubway.org/wiki/Subway_FAQ:_Which_Lines_Were_Former_IRT,_IND,_BMT) for history). Within each division is a subdirectory for each subway *line*. For this data, the subway lines are not the letters/numbers you are used to seeing (those are actually called [*services*](https://en.wikipedia.org/wiki/New_York_City_Subway_nomenclature)); lines have names like "Astoria" or "Rockaway". Within each line directory is one file for each station on that line. The file name is sequence of letters and numbers that is guaranteed to be unique. The contents of each file is a list of the subway entrances for that station. For example, the N/R station on 8th St and Broadway has 8 entrances (2 by Waverly, 4 on the west side of 8th, and 2 on the east side of 8th). Each file is a comma-delimited set of data that has these fields:

|  |
| --- |
| * 1. **Station Name**   2. **Entrance Latitude**   3. **Entrance Longitude** |

**Before you begin**

Before answering the questions, explore the directories and experiment with UNIX commands. In particular, familiarize yourself with the commands: **cd**, **cat**, **head**, **tail**, **cut**, **sort**, **uniq**, **tr**, **wc**, **find**. **Do not use any other commands in answering the questions in this assignment**.

**Questions**

Each of these questions should be expressible as a single command or pipeline of commands. You should not need multiple lines or semicolons for any question.  Do not make assumptions that rely on the specific set of files that exist; keep your answers generally applicable (if new data were entered or existing data were changed, your answers should still be correct)  
  
**Part I**

Using **find** (and possibly other commands), write commands to do the following from /home/unixtool/data/mta.  Do not use shell wildcards with the **ls** command (patterns in quotes for the find command are ok).  Do not use any .csv files.

1. Print the total number of stations.
2. Print the number of stations in BMT.
3. Print the number of non-BMT stations.
4. Print the lines stop ids that are not readable by group.
5. Print the names of all the lines alphabetically (no duplicates).
6. Print the name of the division that has the most stations.
7. Print the station id that has the most entrances (or to impress, the station name).
8. Print the station with northenmost entrance (that will be the one with the highest latitude).

**Part II**

This part deals with the file *ridership.csv*, which has the information about how many riders use each subway station for a given year, over various years ([source](http://web.mta.info/nyct/facts/ridership/ridership_sub_annual.htm)). Each line contains the following fields:

|  |
| --- |
| 1. **Borough**  2.     **Station Name**  3-8. **Ridership** in 2009, 2010, 2011, 2012, 2013 and 2014  9.      **Change**: Increase or decrease of ridership between 2013 and 2014  10. **Services**: Which services are at this station, each separated by **:** (e.g. **N:R**) |

 With this file, answer each of the following:

1. Print the top 10 stations with the biggest increase in ridership.
2. Print the boroughs that have lowest 10 overall ridership (don't list any borough twice).
3. Print the names of all the stations that appear twice or more.
4. Print the borough with most stations.
5. Print the station with the least overall change of ridership (positive or negative).
6. Print out the number/letters of all subway services that appear in the data (each line should be a single character)
7. Print which subway service stops in the most stations.
8. Print stations that entered the top 10 highest ridership between 2009 and 2014 (you can use two commands for this).

**Turning in the assignment**

Copy the file /home/unixtool/data/assignment1 into your directory and fill in the answers in the appropriate places in this file.  It is important that you start with this file, which will help the graders partially automate testing.  Only submit the commands themselves, not the output of the commands.  Here is an example:

|  |
| --- |
| ## Assignment 1 ## Carmen Bianco N53943243  ############################################################### ## Question 1 ################################################# ###############################################################  find . -name myfile -print  ############################################################### ## Question 2 ################################################# ###############################################################  find . -name otherfile -print | wc |

Make sure each command has been tested!  You can use the following tool to verify that your homework is in a valid format:

   /home/unixtool/bin/check\_assignment1 *assignment1*

When you have finished, submit using the [homework submission system](https://nyu-csci.appspot.com/grader.cgi).