Assignment 3 Mistakes
Austin Derbique A01967241
9/12/17

1. The definition of Data Science
    a. Data Science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms. https://en.wikipedia.org/wiki/Data_science
    b. Data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them. http://datascience.nyu.edu/what-is-data-science/
    c. A field of Big Data which seeks t provide meaningful information from large amounts of complex data. Data Science combines different files of work in statistics and computation in order to interpret data for the purpose of decision making. http://www.investopedia.com/terms/d/data-science.asp

2. Computer Science is an essential part to Data Science as it enables the scientist a given tool set to effectively and accurately determine a solution that would otherwise be monotonous or potentially impossible.

3. python3 egrep.py "\[(.*)]\s\[(.*)]\s\[client.(.*)].(.*)"
Look at egrep.png to see how it's used.
[Thu Nov 13 14:32:01 2014] [crit] [client 153.90.200.206] (13)Permission denied: /home/www/public_html/OSS/.htaccess pcfg_openfile: unable to check htaccess file, ensure it is readable

This regular expression separates the line information into groups in the following order:
Group 1: Time Stamp
Group 2: Error Type
Group 3: IP Address
Group 4: Error Message

4. ehead.py complete
5. etail.py complete

6. The ehead.py script will run faster than the etail.py script as the ehead script grabs the first N lines whereas the etail script has to count how many lines are in the file before pulling the last N lines. This will likely lead to a slow down in large files. With small lines <10,000 there should not be a noticeable difference.

7. datecount.py complete. Look at datacleangraph.png for regex usage
"\[\w\w\w\s(\w\w\w)\s(\d\d).*.(\d\d\d\d)]\s\[(.*)]\s\[client.(.*)].(.*)"
8. graph & chart file complete. Look at graph.png

9. On January 1st, 2018, there will be 1,123 unique days for log information. This is by taking the amount of log days and adding how many more days there will be until the new year. Currently, there are 1194962 entries in the log file. We will take an average error per day of 1064 and account for the continual increased growth, rounding that number to 1100. Knowing that there are approximately 100 days left in the year, we will multiply these numbers to achieve an estimated number of errors added to

the log file. Adding approximately 110,000 errors will result in a log file of roughly 1304962 errors on January 1st, 2018.