CS5665 Project 1: Who can you trust?

Austin Derbique A01967241

Objective

Determine if Google Analytics Data is consistent with the data found in the weblogs over a year or longer.

Background Info

Provided for this assignment are four data sets. This consists of: Desktop-Referral, Mobile-Referral, Desktop-Sessions, and Mobile-Sessions. The two referral datasets measure how many sessions are recorded for each website between the period of Jan 1, 2016 to September 8, 2017. The session datasets measure how many sessions were recorded for each day between that same time period.

Hypothesis

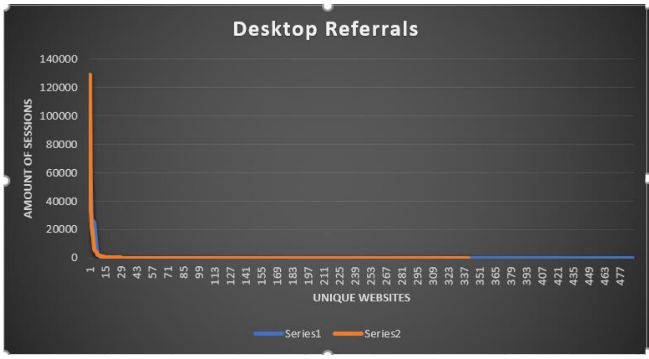
I believe that there will be inconsistencies between the web logs and google analytics data. This will be due to error in recording and categorizing of data. Maybe some things are logged, others are not. Or it may be possible data is not logged properly. Generally, I expect the data to be more or less similar between the files.

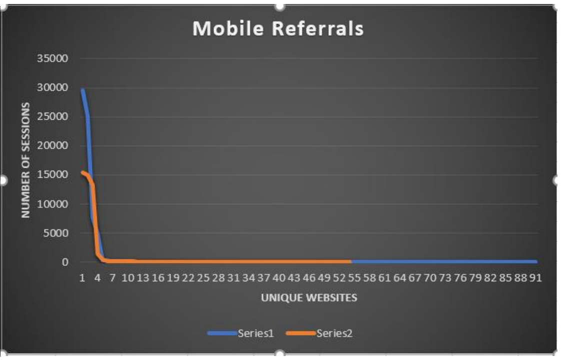
Findings

To answer this question, we must break down this complex data set into smaller pieces and analyze them separately. Notably, there are four different measurements that will be used to arrive at a conclusion.

Mobile & Desktop Referrals

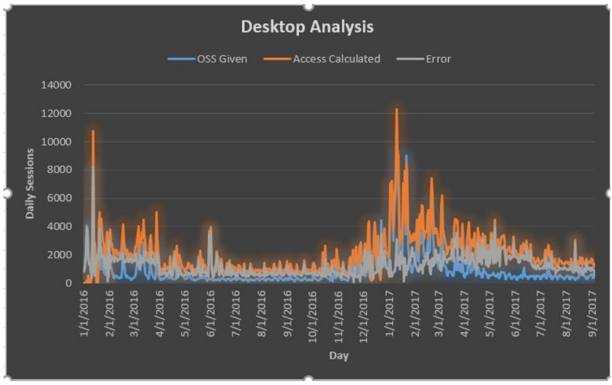
In this case, the orange line is the given referral data showing how many sessions there are in descending order for each unique website. The blue is my calculated dataset from the access.log file. Notice how many more websites are discovered in my calculated dataset.

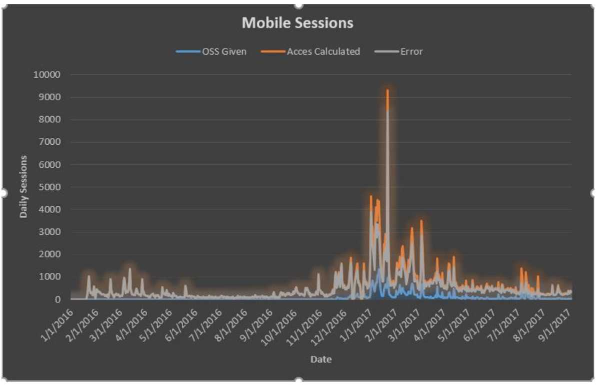




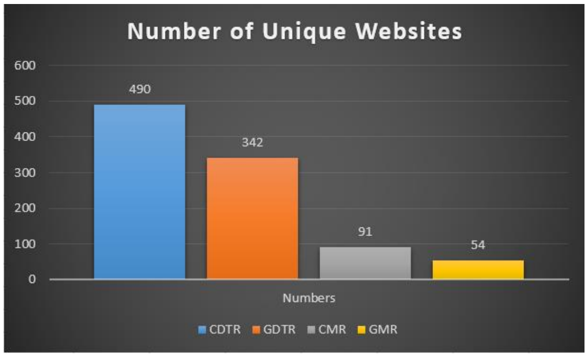
Mobile & Desktop Sessions Per Day

Shown below is how many sessions are recorded per day over the span of almost two years. The orange line is the dataset generated from the access.log file and the blue line is given by the weblogs. The gray line is an absolute error of the two measurements.





Number of Unique Websites



In this graph, the blue bar is calculated desktop referral sites, the orange is given desktop referral sites, the gray is calculated mobile referral sites, and yellow is given mobile referral sites. In this context, calculated comes from access.log and given comes from the weblogs.

What to make of the Data

Referrals: There appear to be many more calculated mobile referral numbers than what is given in the web logs. On the contrary, there are more peak sessions for certain websites in the desktop graph according to the weblogs than calculated access log data. This could be due to the approximation of what is considered a “session.” Knowing that mobile browsers and desktop browsers refresh differently, this could give indication for why there’s a difference.

Sessions: Although there are differences between what is calculated and what is given, there appears to be a relationship between the data that is nearly linear. This means that even though the data is off by a certain margin, it is consistently off, therefore showing that the two forms of data collection are in some way consistent with one another.

Number of Unique Websites: The calculated dataset shows more unique websites than what is given by the weblogs. Once again the sets appear to be proportional to one another.

Conclusion

While these two datasets are not exact, it can be proven by looking at the graphs that they are at least consistent with one another by analyzing referral, session, and website data.

Summary of Approach

This appendix page aims shed some light onto the inner works of the processing for how I came to my solution. Below are listed the scripts written for this assignment and what they do:

- date_session_count.py – counts the sessions for each day
- filter_data.py – leaves behind only outside requests
- mobile_desktop_split.py – splits data into mobile and desktop files
- range2file.py – saves a given range of lines to a file
- referral_session_count.py – counts the amount of sessions per website
- tabulate.py – parses the log file and saves to csv file

For full documentation, please view the data and scripts directories inside the project folder. In step by step order, I will attempt to walk through my thought process for this assignment.

1. The first thing I did was run filter_data.py which took the 28 million line access.log file and turned it into 1.9 million lines. This script throws out the line if it contains the expression "oss.weathershare.com". Only 7% of the data set appears to be useful.

2. Next I ran the mobile_desktop_split.py which takes each of the lines in the newly leaned access file and writes it to either the desktop log file or the mobile log file. This case statement depends on if "Mobile" or "mobile" are found in the line. This appears to work fairly well as it caught more mobile websites than the web logs did.

3. Now that the logs are split into mobile and desktop versions, I ran date_session_count.py to run through and count the amount of sessions for each day there were. I attempted to clump sessions from the same user within a 20 minute period together to reduce the amount of false sessions. This was a simple sum for each day. This outputs as a csv file.

4. Using referral_session_count.py, I used a dictionary of unique websites and incremented the count each time it found the same website again. I called this script on both the mobile and desktop version. This outputs as a csv file.

5. I sorted the values of the sessions by date and compared web log sessions directly with the google analytics access log dataset. The third column is the magnitude of error between the two values. I then graphed this to form what is shown in this report.

6. Next, I sorted the values of referral data in descending order with the two datasets placed next to each other. Graphing this shows the sessions in descending order in addition to how many unique websites there are for both data sets.

7. A simple excel file used to hold the values of unique websites between datasets is used to graph.