**36 Hour Study Sessions?**

Raunak Pednekar

December 15, 2016

Data accuracy is incredibly important when collecting and storing data. Downstream analyses done on data is meaningless unless data is accurate. In this data set, there are concerns about the veracity of certain features of the data. The data explored in this section of the report comes from the library files for 2015.

Some connection times were recorded as lasting for several days. The validity of this connection time is suspect. Hence, all connection times over 24 hours have been shortened to one day. It is possible that the devices are not actually connected, and so are not using any traffic. If this is true, then we suspect there would be disproportionately lowtraffic concentrationwhere traffic concentration is defined as the ratio of total traffic to connection time. Traffic concentration is the number of Megabytes used per minute.

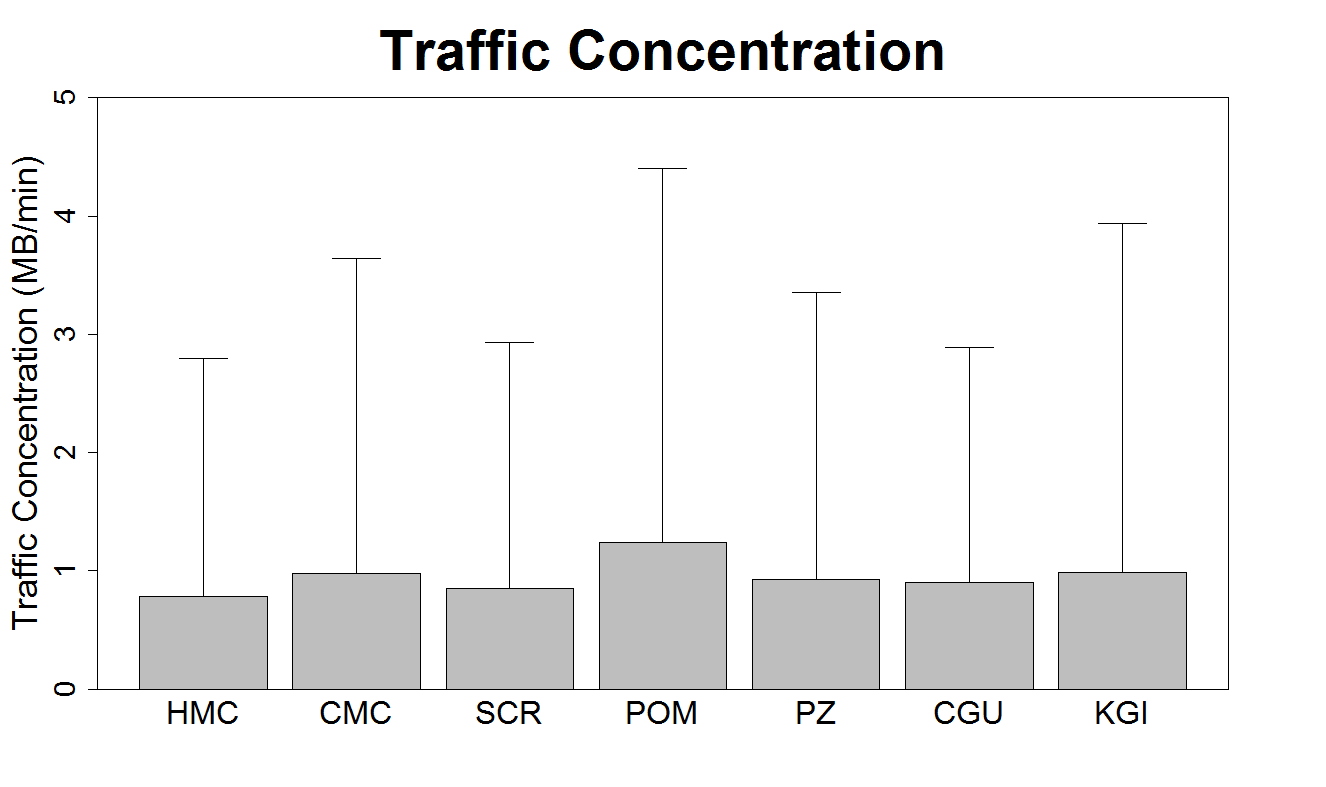


Figure 1. The traffic concentration plot reveals that the average connection consumes data at a rate of 1 MB/min. The standard deviation for this metric, as one might expect, is very large (Some people stream Netflix, others are chatting on Messenger). There aren’t differences between the traffic concentration for different schools (especially when taking the large standard deviations into account)

If certain connections had innacurate connection times- if they were indeed artificially inflated- then traffic concentration for these artificially inflated connections should be lower than for other connections.

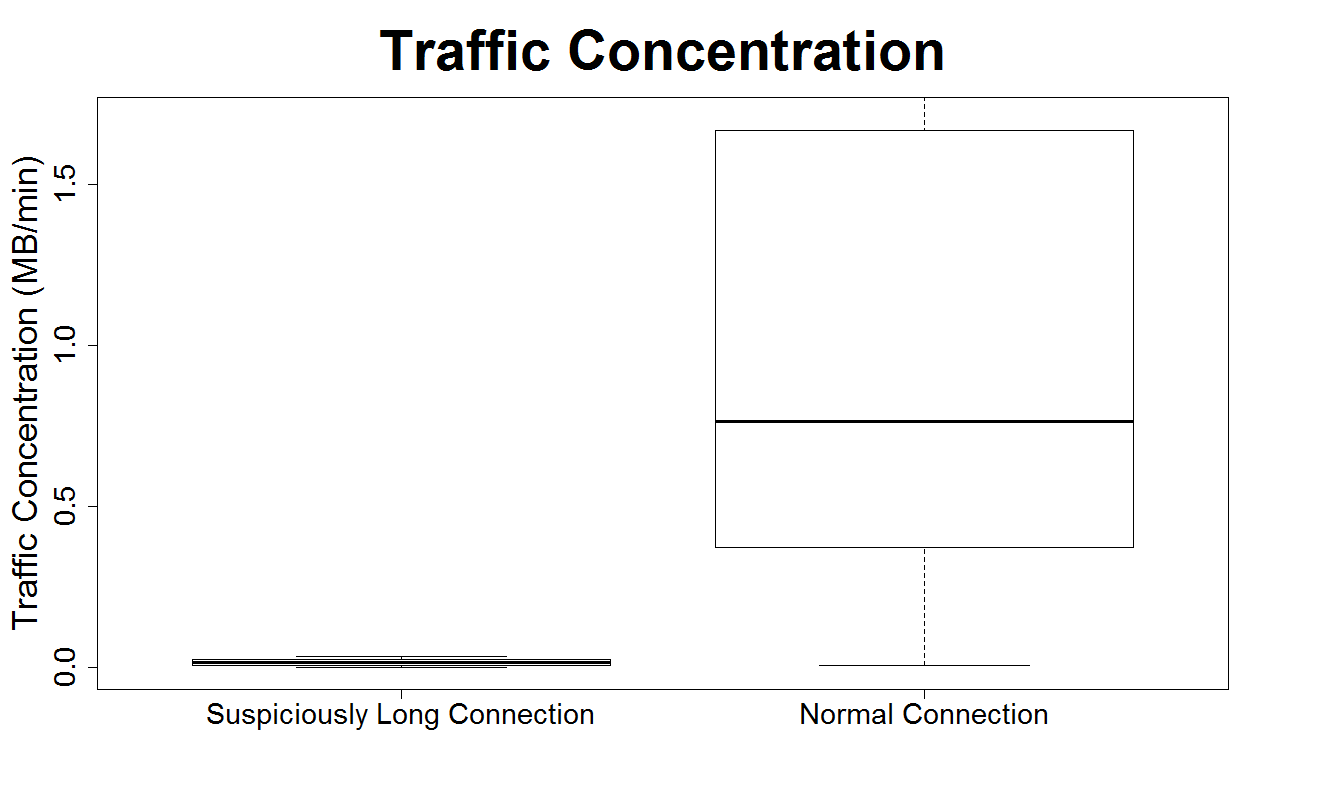


Figure 2. As can be seen, the traffic concentration for suspiciously long connections is lower than that for regular connections.

Now that we’ve determined that certain data points are inaccurate, the next question to explore is whether these inaccurate data points have any distinguishing markers. For instance, does a certain WiFi Network result in more false data than another? In other words, is the proportion of false data higher for a certain WiFi network?

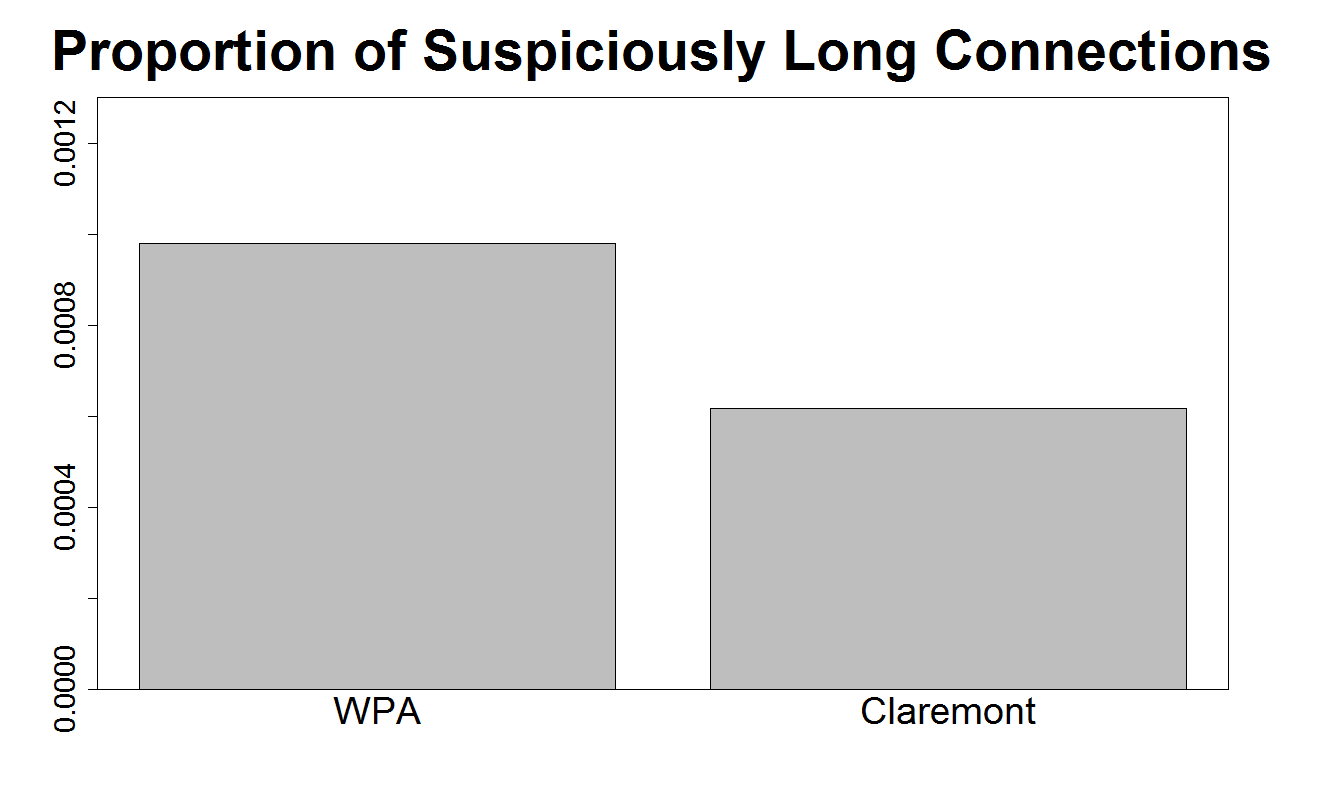


Figure 3. As Claremont-WPA and Claremont encompass over 97% of all WiFi connections, we compared the proportion of false data for these two WiFi networks. WPA had a slightly higher proportion of false data than Claremont. Similar results were obtained when comparing the proportion of suspiciously long connections across (1) Manufacturer (2) Type of Device (phone/tablet/laptop) (3) Operating System (Android/Windows 8/OSX/ iOS).

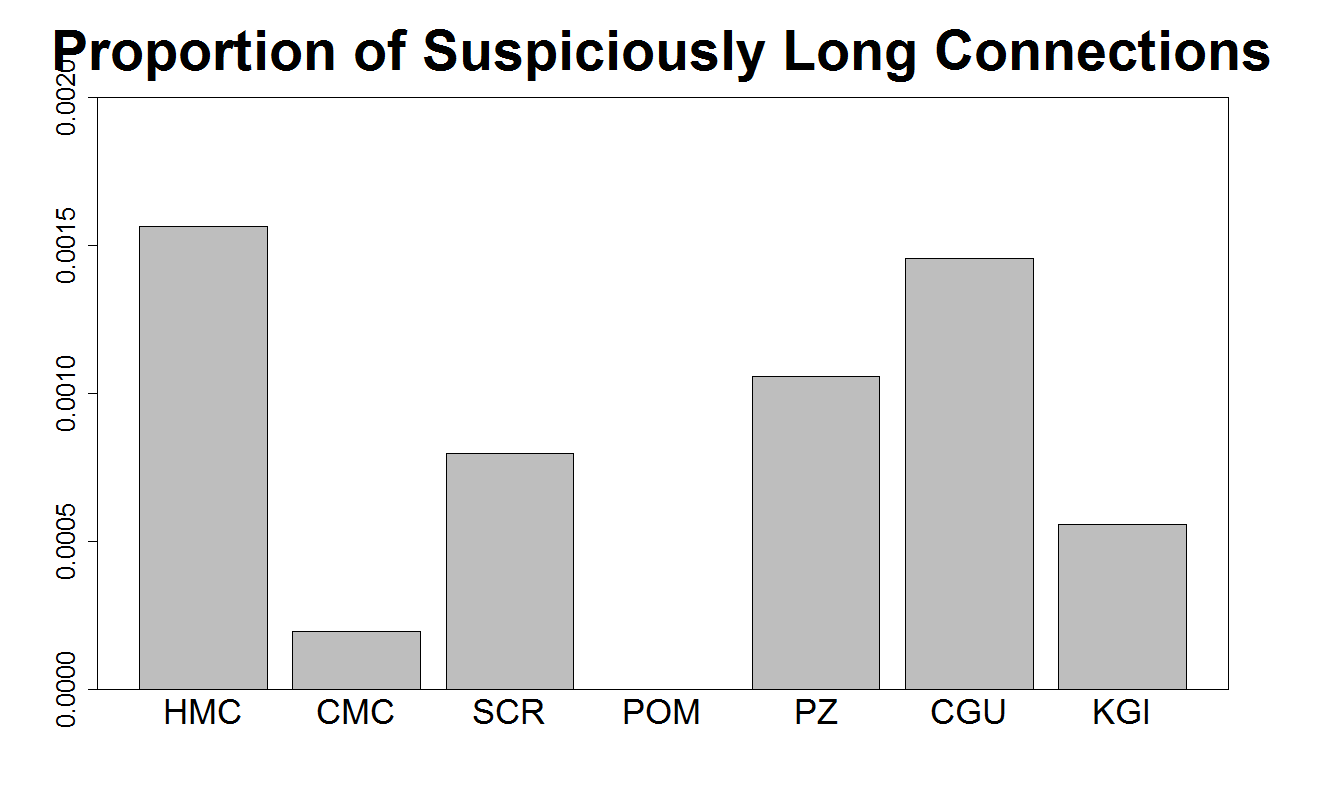


Figure 4. While no school contributes much more than all others to the false data, there is one school that has no false data*:* Pomona.