We are seeing some large fluctuations in the main KPI of our site — something we call the “bad to good ratio” — that is, the ratio of users who have “quality” to users that don’t have it. We want to find the reasons for these fluctuations.

After some investigation, we think that the cause of the problem might have to do with changes in the sources of traffic. We have requested our development team to extract some data, and we are attaching a sample of it:

- bad.csv: CSV file containing the ones with no quality.

  An example of a row would be 104474, 3e14b0c730317aad79307c3c214fc838

  ,"2017-09-21 09:23:43", where ,3e14b0c730317aad79307c3c214fc838

  is the id of the user, 104474 is the id of the source of the traffic and "2017-09-21 09:23:43" is the date of the event.

- good.csv: CVS file containing the same information for quality ones.

In order to analyze this data, business would like to see several graphs:

- One graph representing how many no quality users we need in order to generate a quality one, for all the users of the site. For example, a value of 45 would indicate that for each 45 bad quality, one is becoming a quality one, on average.

  Ideally, we would like to have a point each 5 minutes and each point would be an average of the last hour, so bad joins for the last hour divided by the good joins in the last hour.

- Generate the same graph for the top 10 sources of traffic, so we can see the ratio for the biggest sources of traffic.

In addition to the graphs, business would really appreciate some conclusions about what we the possible influence of the traffic source on the ratio, so analyze them and add any valuable info. Any statistical analysis to try to find correlations between the source of the traffic and the main ratio will be highly valued.

Finally, we would like to consider the implementation of some Business Intelligence tool, where we could easily graph this data, in order to don’t repeat the analysis when a similar situation happens again. What would you suggest?.