Code Chalange Tenchi Security

João L M Resende

30 january 2022

Report for Abin

This is a report to show about cyber breach information.

Summary of the Data

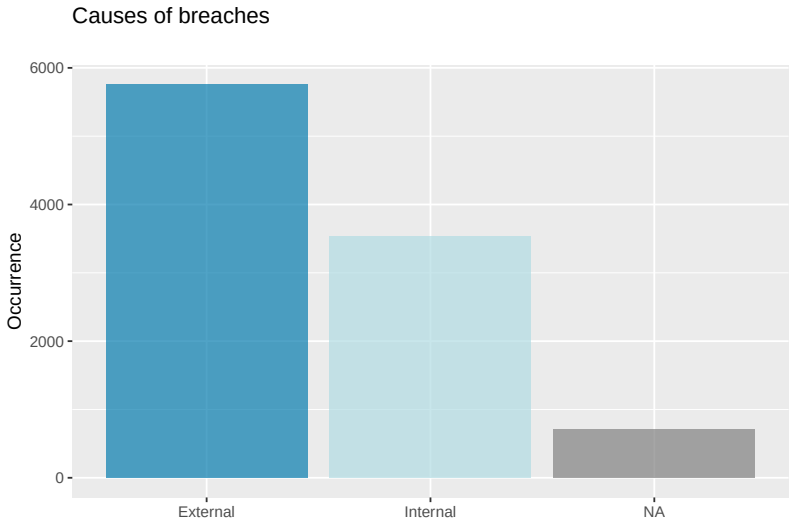
This data shows that the mean value of each breach is **$4,661,739** . Te minimum value of the breach is **$0** , and he maximum value is **$7,915,369,031**

The 5 sectors which more occurrence of breach of information are shown in the table below:

Table 1: Ocurrence per secto

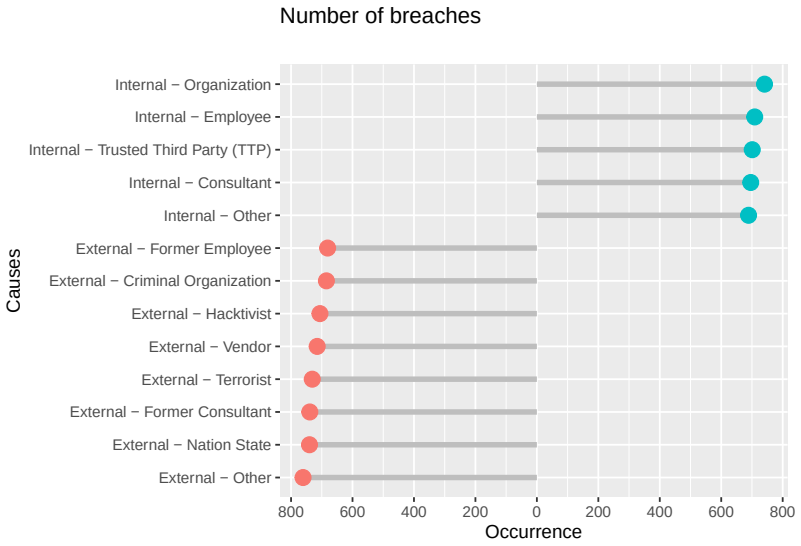
sector	Occurrence
Financial	1258
Professional	1232
Healthcare	1158
Administrative	1089
Education	719

As we can see, the **Financial** is the leader of breach of information, with **1258** occurrence. Now, in the plot below we can see the division between external and internal causes.



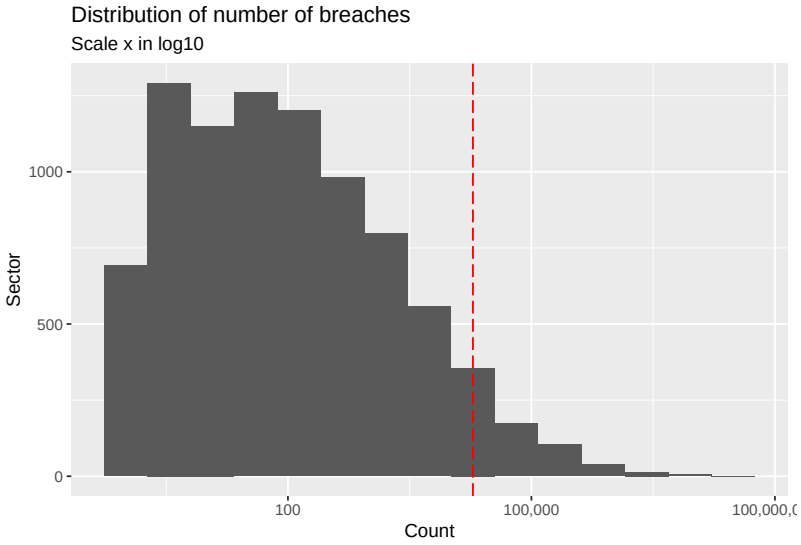
As we can see, the majority of breaches are caused from external causes, with almost double of the internal ones.

In the next plot we gonna see occurrence of breaches, divide in external and internal causes, for this analysis the events without causes are going to be excluded.

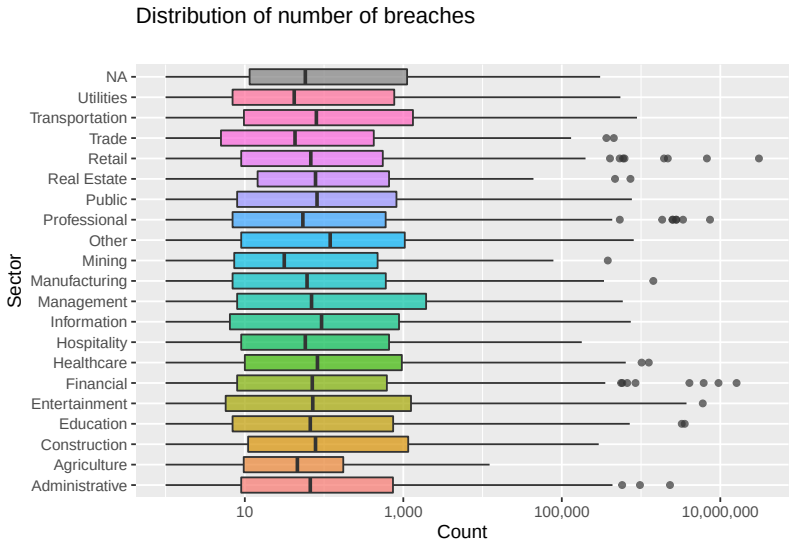


As we expected, the variety of external causes is bigger, as seen in the previous plot. Another fact to be in mind is that in both groups external and internal, the difference between the biggest reason for breaches and the smallest is approximately 100.

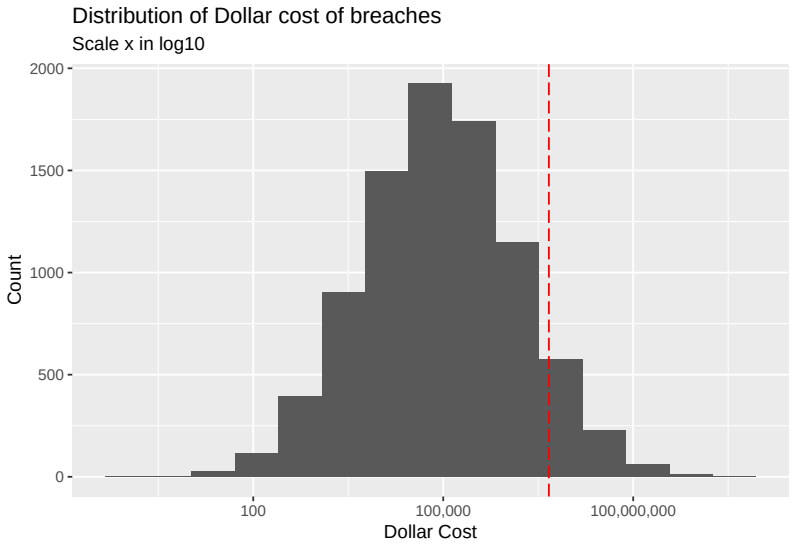
In the next analysis we can see the distribution of affected_count, that is, how the count of data records involved in the breach are distributed.



Because of the big difference between the number, we choose to use a logarithmic scale (log10). That is, for each same size distance, the value of x increases by 10 times. As we can see in the plot, the distribution of the number of data records involved in the breaches is concentrated at the begging of the plot, that is, the majority of the number of records is smaller than the mean value, represented by the red dashed line.

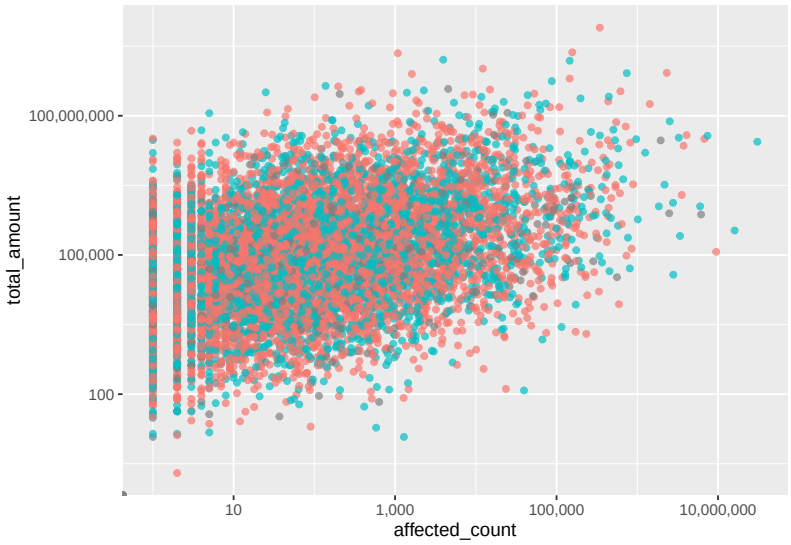


We can see in the plot above that there is no statistical difference between the means of breaches in each sector. We also can see that the outliers, that is the big numbers that do not follow the distribution are more present in the Retail, Financial and Professional sectors.



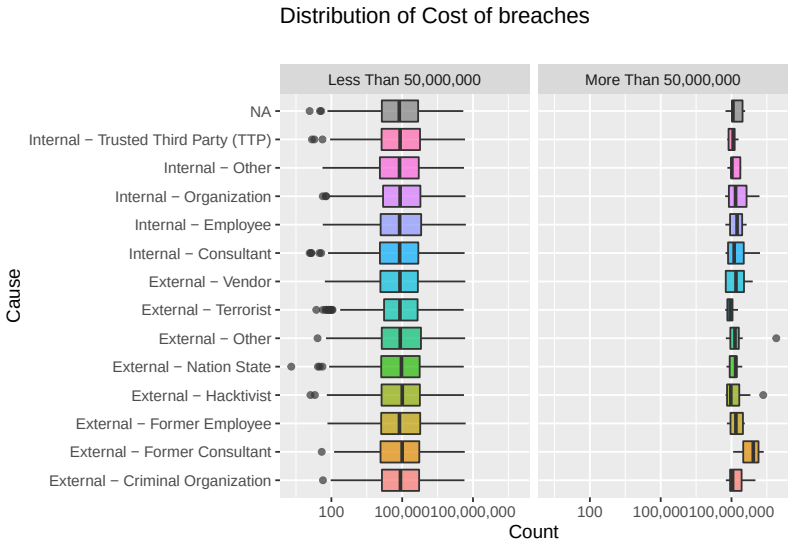
On this distribution, the distribution is similar to the normal distribution, that is, the distribution is symmetric around the center value, however, these values are not the mean, in fact, the vast majority of the values are smaller than the mean, showing that the in the distribution there is a lot of big outliers. The median, or the most typical value is **US$ 49,926**.

Next let's see the relationship between total_amounts and affected_count.



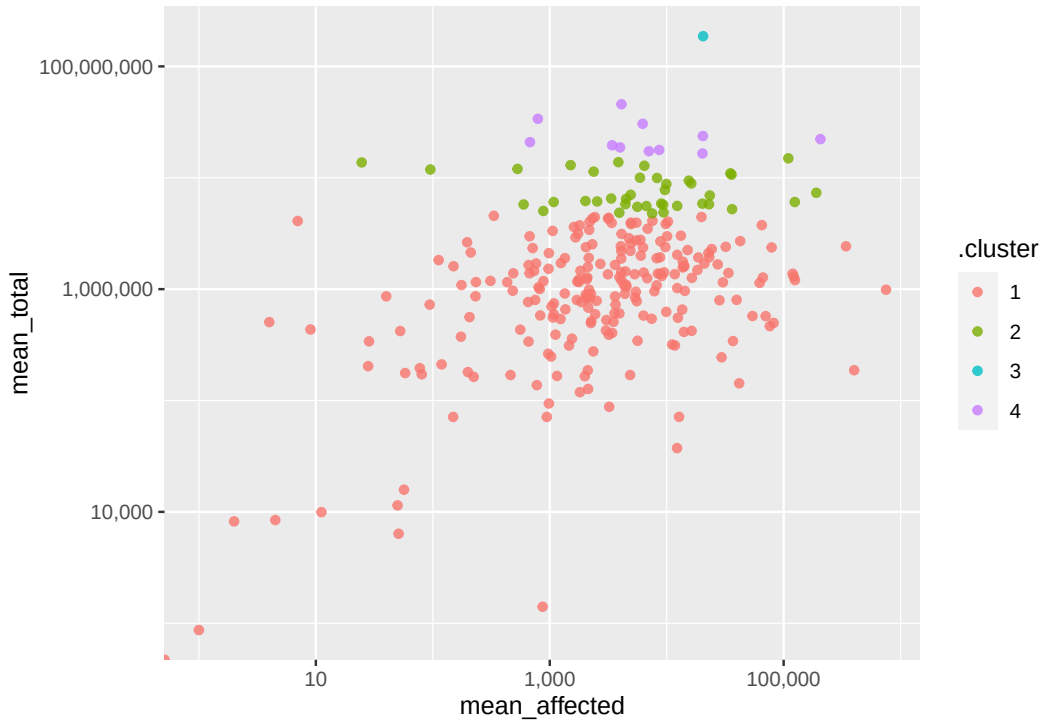
Since we are on a log10 scale, we cannot affirm the size of the relationship, however, we can say that there is a positive relationship between the number of affected counts and the total cost of the breach.

Since we have big outliers, next we gonna divide our data into two samples, one with the Cost of Breaches bigger than US$50,000,000, and the other with the Cost of Breaches smaller than US$50,000,000.



We can see, that the distribution is very similar for the Cost of Breaches smaller than US$50,000,000, however, for the other sample, we can see that the External - Formal Consultant have a significantly greater mean cost of breach than the other Causes.

The last analysis is the k-means, in the plot below, points thar are in the same klustes, have similarities betweenn than. We cann see that the moijority of breaches of the same sector are together, howesver the cause beeing external or internat does not put the observation togeher, hance this is not a factor to determiny if there are similiarities between than.


100,000,000
1,000,000
10,000
mean_total
10
1,000
100,000
mean_affected
.cluster
1
2
3
4