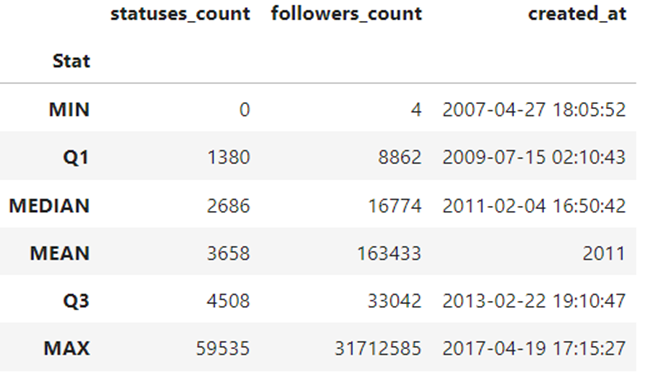
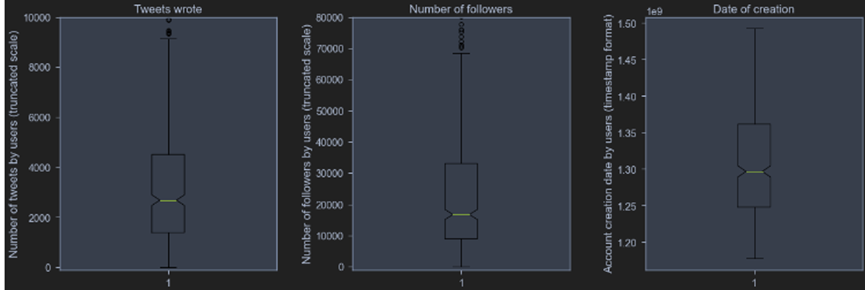
1. First, I want to see some characteristics of the data to get an overview, so l will use the five number repartition and mean and finally make some boxplots on three variables (statouses\_count, followers\_count, created\_at) and see how they are distributed.



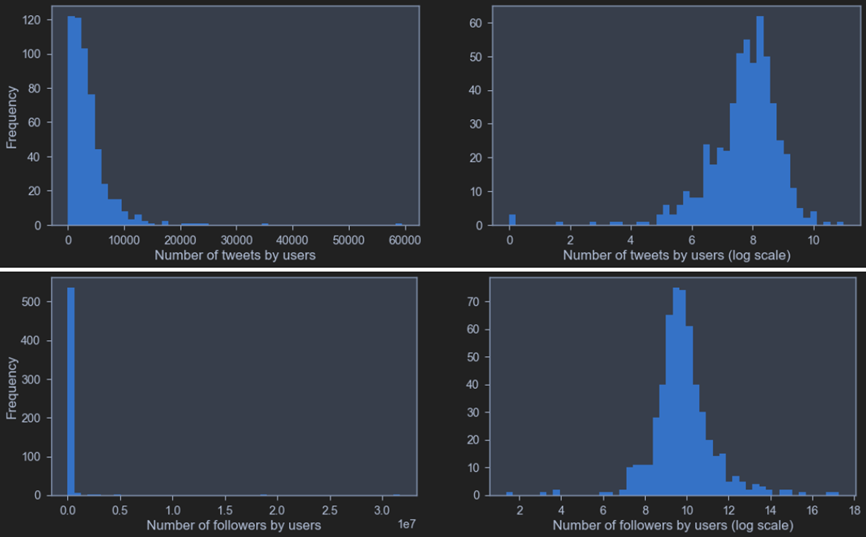
1. All features show big outlier(s) on highest values. Indeed, there is a rough linear growth from Q0 to Q3, and probably still nearQ4, but some high-value outlier(s) extend Q4 value. Under boxplot representation, the lower whisker and two halves of the box (IQR) should have similar lengths.

2. Except created at, the two other variables have mean values very different from associated median values. Under symmetrical distribution, the values would be equal. Since it's not the case, both features distribution must be highly right skewed (i.e. same conclusion from first point: high-value outliers)



一張含有 螢幕擷取畫面, 繪圖軟體, 3D 模型 的圖片

自動產生的描述

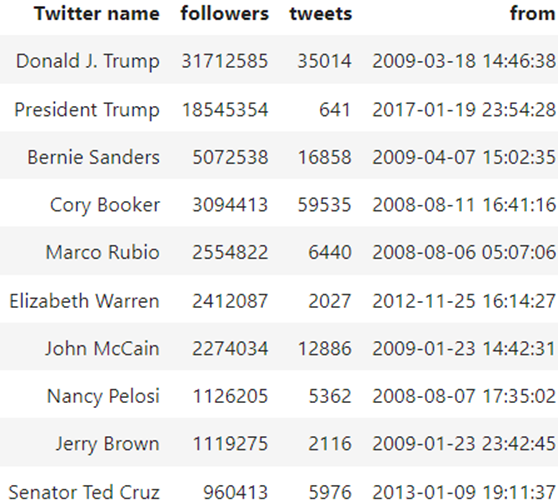


一張含有 螢幕擷取畫面 的圖片

自動產生的描述

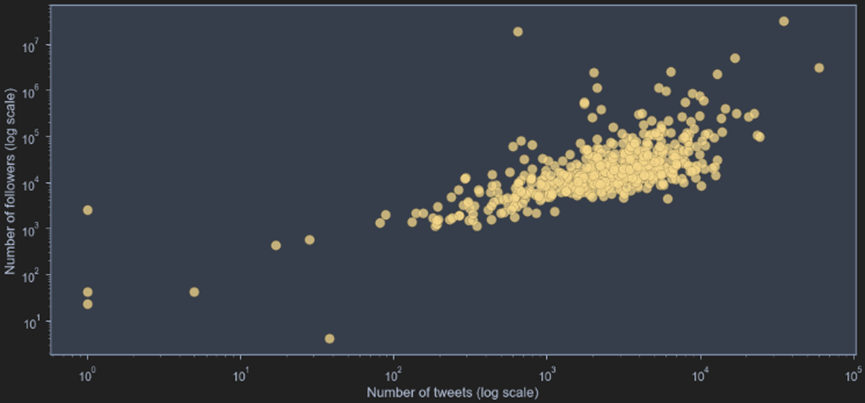
Indeed, most of previous assumptions of the boxplots and distribution representations have been confirmed. However, the dates of account creation are surprisingly distributed very erratically. There have been strong spikes in the appearance of members of Congress on Twitter every two years in January since 2009.

1. Twitter Popularity investigation



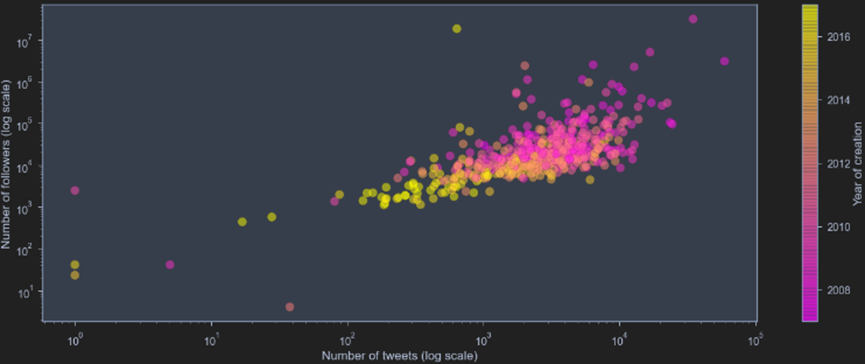
Here is top 10 accounts ordered by their followers.

Let's see now the relationship between popularity and activityon Twitter as well as the relation with states population.



It seems that the more number of tweets the more number of followers. This is relevant with our initial hypothesis but it has to be ever more investigated with p-values.

Also if we add a filter and see the date of creatin of accounts wecan see that the older the account the more followers it has. Thisalso supports our initial hypotheses but this has also to be more specifically searched with p-values.



1. What about colors?



It is difficult to see a valuable distinction between the colors because the is a big variety of range on each color. Although we can observe that at profile link color blue and red is dominant which verifies our initial hypothesis about colors.