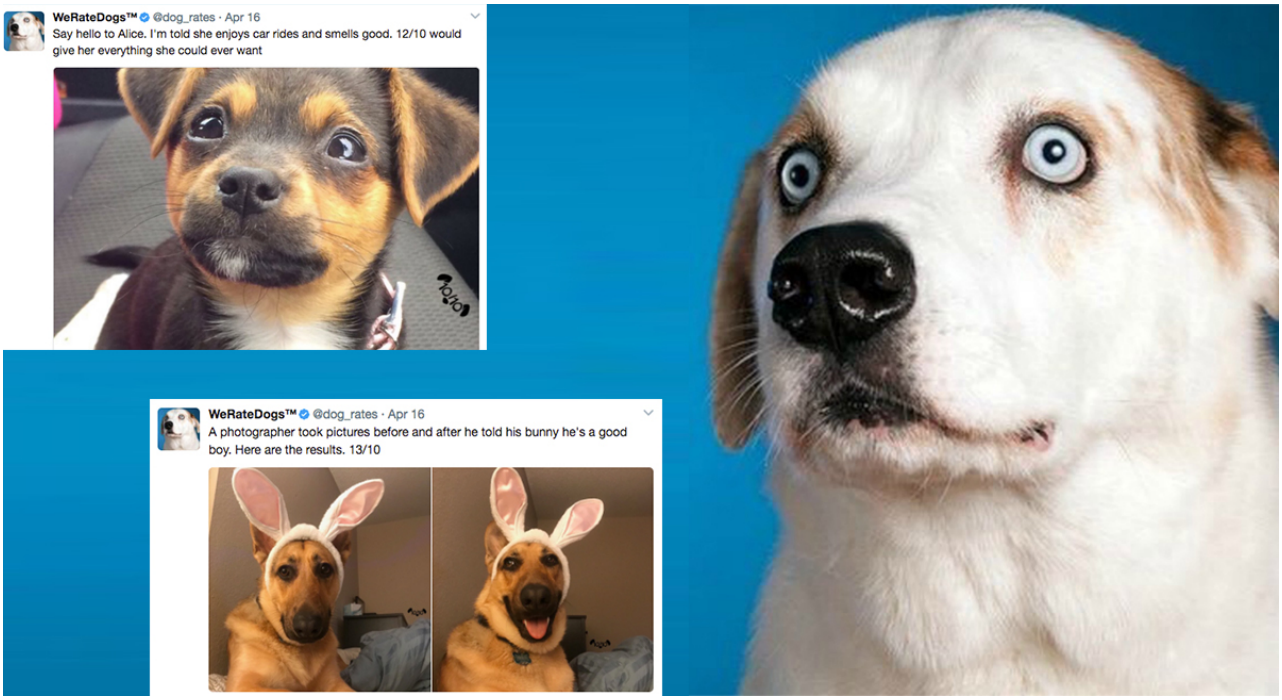
Udacity Data Analyst Nanodegree

Wrangle and Analyze Data



Project Overview

The project is about the complete end to end data wrangling process, as:

- Data Wrangling (not part of this document)
- Data Analysis

The tool which is used is called Jupyter Notebook.

The exercise is to analyze a harmonized three dataset. The dataset is sourced from an archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Insight 1: Favourite Sources

I'm interested in what are the favourite twitter sources, let's get an overview:

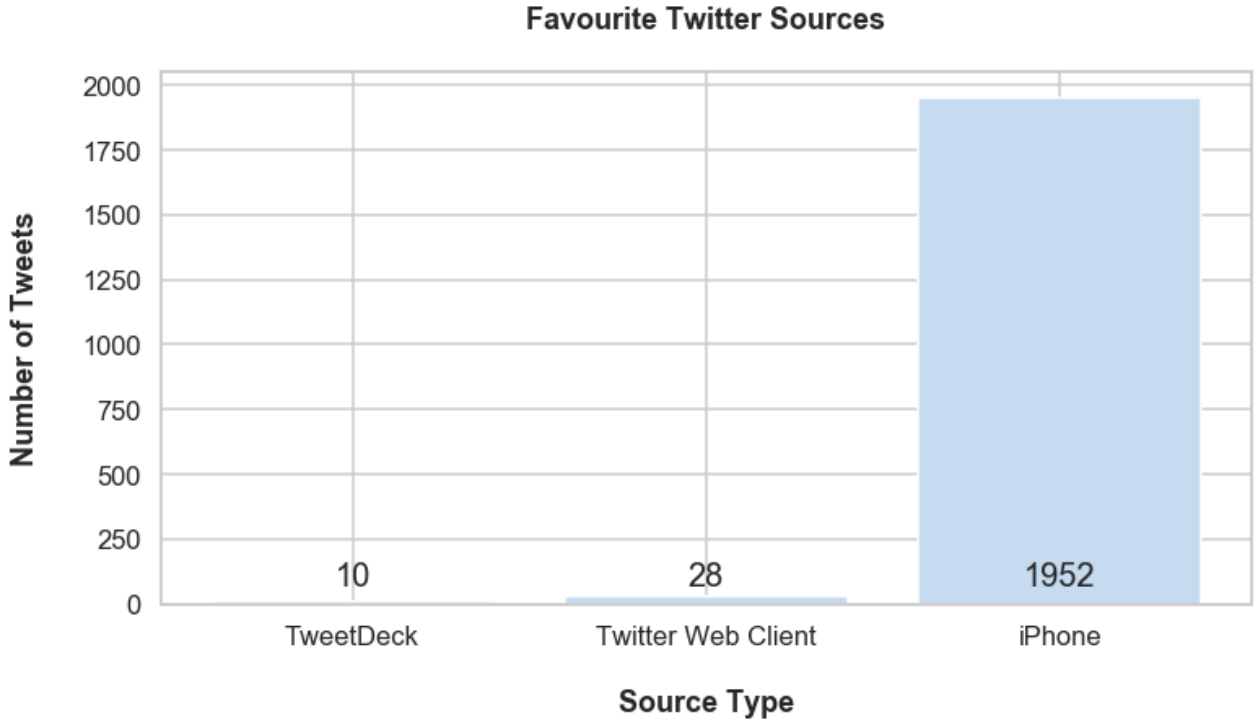


Figure 1 Twitter Sources

Ok, iPhone seems to dominate the twitter sources, this is also visible looking at the proportions:

source	Number of Tweets	%
TweetDeck	10	0,50
Twitter Web Client	28	1,41
iPhone	1952	98,09

The iPhone as source is extraordinarily high, pretty much exactly 98%. With a hypothesis test we could look if that is also true for the entire population of "we rate dog" users. A definition would look like that.
For "We rate dogs":

- The Null Hypothesis (H_0) states that the proportion of iPhone sources is <= the proportion of all other sources

- The Alternative Hypothesis (H_1) states that the proportion of iPhone sources is > the proportion of all other sources

Statistical Hypothesis Notation:

- $H0: P_{iPhone} <= P_{all\_other}$

- $H1: P_{iPhone} > P_{all\_other}$

Maybe one should look deeper in the relation of tweets to users, as it could be that some few users using the iPhone twitter app are extensively sending tweets, this would bias the hypothesis. Furthermore, the number of overall tweets is 10200 at the moment (just looked it up). So, the sample size is very promising.

Insight 2: Retweet vs. Favourites Counts

10200 tweets more than 8 million followers and 142000 Likes, the likelihood that the count of favourites and the count of retweets are positively correlated is very high.

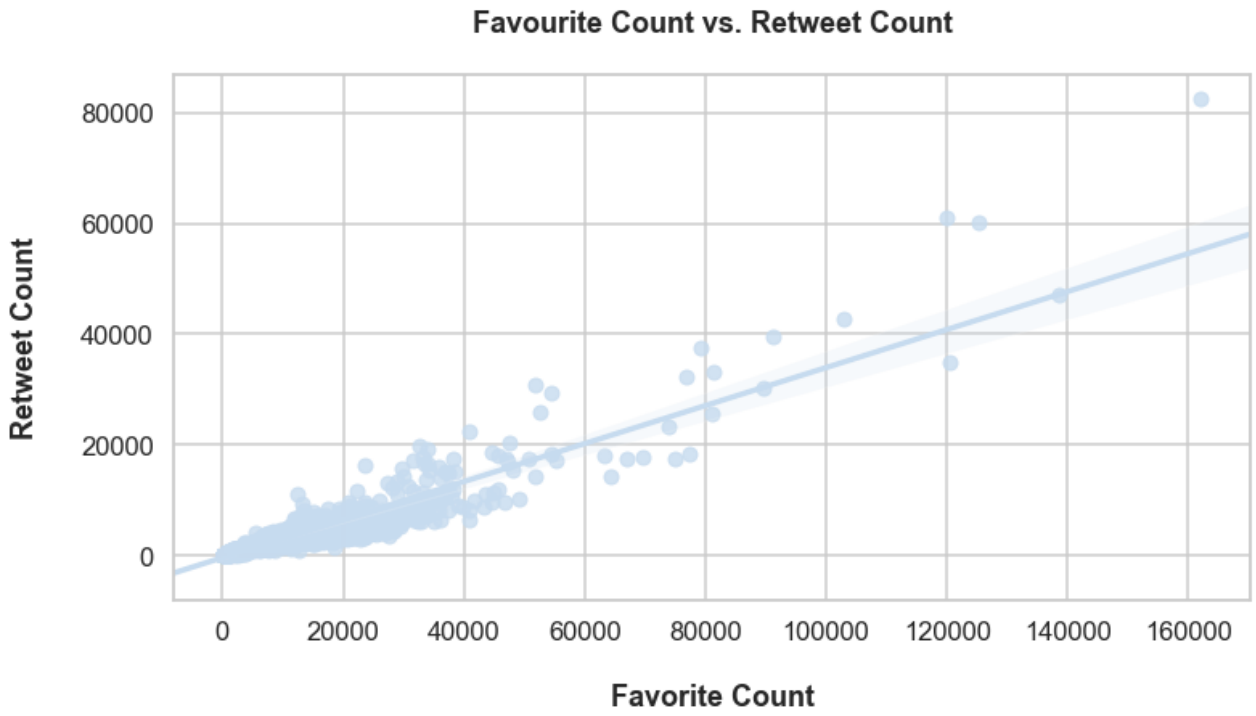


Figure 2 Favourite Count vs. Retweet Count

The scatterplot clearly shows the strong positive correlation relationship of retweets and favourites counts. Means if a tweet is retweeted more often it is more likely that somebody likes it.

Insight 3: Number of Tweets over Time

Finally look to the evolution of “we rate dogs” .

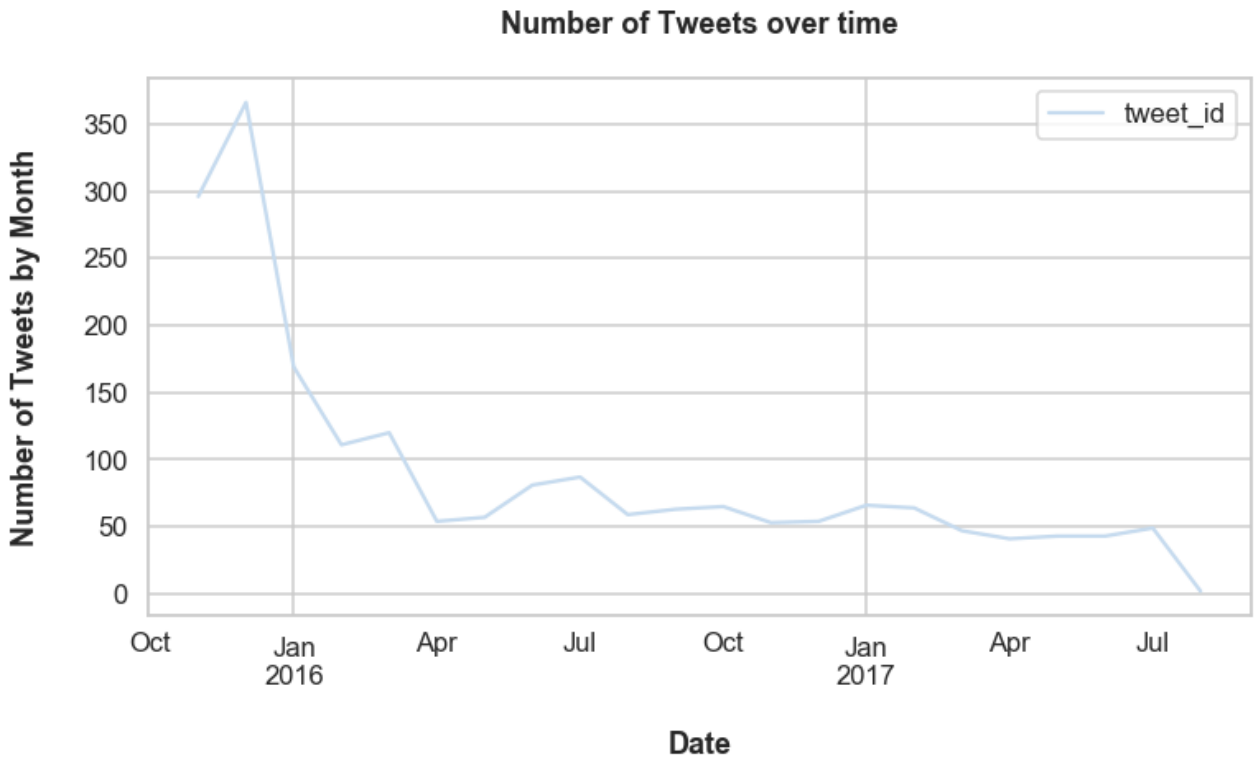


Figure 3 Number of tweets over time

The chart above is showing the aggregated count of tweets by month.
Ok, I believe the hype seems to be dampened. There was huge activity late 2015 and then constantly declining till April, then it remained kind of constant till appr. July, 2017 and then it dropped. To examine that further we would need to gather more recent data, but definitely the hype is over.