In [1]:
```
import pandas as pd
```

WeRateDogs Project

Author: Eduardo Kaneko

Date: 10.04.2019

Process

Purpose: Gather and clean data from multiple sources for analysis

Language: Python

Skills:

- gathering data: programmatic file download, downloading API data
- assessing data: documenting cleanliness and tidiness issues
- cleaning data: cleaning for completeness, tidiness, validity, accuracy, consistency
- storing data: CSVs
- analyzing data
- visualizing data
- written report

Libraries: matplotlib.pyplot, missigno, seaborn, pandas, numpy, resquests, warnings, tweetpy, json, re, os.
Software: Jupyter Notebook. </p>

Introduction

This project is about a twitter's account named as WeRateDogs. WeRateDogs is an account that classify people's dog with a really funny comment. It was born in 2015, by a college student, Matt Nelson, that received an international attention.

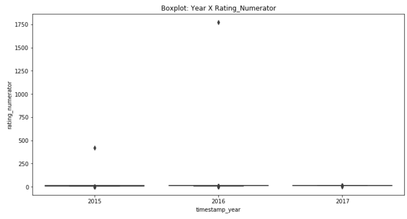
By October 2017 it had an additional 3.7 million followers. These classifications almost always have a denominator of 10. But what about numerators? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Because? Because "they're good dogs, Brent."

Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to **analyzing data sets** to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

Ref: Wikipedia

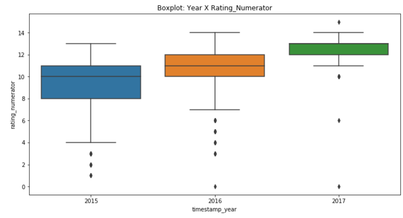
First of all, after I got the dataset clean, I've had the focus to find any possible outliers. So, I've started by the variable `rating_numerator`. I found two discrepant values, that did not make sense for this variable.



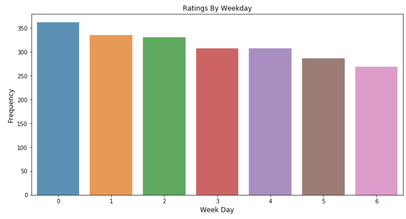
As this aroused my curiosity, I went after the two tweets that were considered outliers:

In this specific case, to get back the analysis, I dropped both outliers to see the results without it. So, I could have seen that, over the years, the rating get better and better. I do not know if the dogs were better or the person who evaluated was less strict:

- The lower quartile increase over the years. Just like the median, the upper quartile and the end values (min and max).



The second point is with regard to the days of weeks. If we look below, in the histogram Ratings By Weekday, we notice that there is an asymmetric distribution on the right. So, we can conclude that sunday is the day the dogs get the most ratings (the mode).



Scatter plots serve to compare two quantitative variables. A summary statistic that relates to the scatter plot is the correlation coefficient commonly denoted by a r. To the below plot, we can see a positive and strong relation between retweets and favorites. So, when one variable increase, the other variable increase too.