DATA WRANGLE REPORT

## Introduction

Project has a variety of formats, quality and tidiness which are known as data wrangling.

The dataset is about the tweet archive of Twitter user @dog\_rates known as WeRateDogs. The twitter account that rates people's dogs with a humorous comment about the dog.

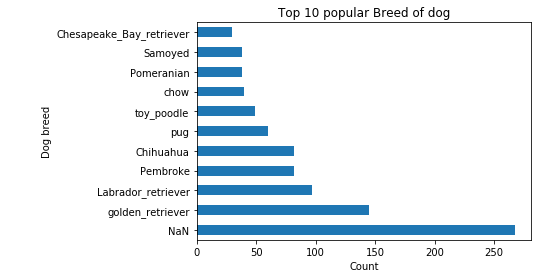
Mainly objectives are:

1. Data wrangling, which consists on:
   * Gathering data from the following sources:
     + The WeRateDogs Twitter archive called: **twitter\_archive\_enhanced.csv**
     + The tweet image predictions, i.e., what breed of dog and so on. The file is called: **image\_predictions.tsv** or we can access by [**https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad\_image-predictions/image-predictions.tsv**](https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)
     + To get access to tweet’s retweet count and “like” we need Twitter API and Python’s Tweepy library to gather whole useful information.
   * Assessing data consists in quality and tidiness in the dataset.
     + Quality data has four aspects:
       - Missing Value (Completeness)
       - Validity
       - Accuracy
       - Consistency
     + Tidiness has three aspects:
       - Each column is conforming with a unique variable
       - Each row has a unique observation
       - Each data (observation) has the own unit and form a table.
   * Cleaning data consists in three steps: define, code and test.
     + Files called df, images and twitter\_data were concated and saved as all\_data, this step was achieved reading <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.concat.html>
     + Some columns were dropped because It does not provide useful information to my analyses. Columns are: in\_reply\_to\_status\_id ,in\_reply\_to\_user\_id, retweeted\_status\_id, retweeted\_status\_user\_id, retweeted\_status\_timestamp, timestamp, source, tweet\_id and tweet\_id.
     + Some columns were modified its Dtype with the correct type.
     + Expanded\_urls and jpg\_url column have Duplicated values, so I dropped it.
     + The column name had some vowels and it was replaced by None.
     + Drop some row that do not have image in the jpg\_url column.
     + A column called diff\_dogs was created which contain: doggo, floofer, pupper and puppo values.
     + A column called image\_predic contains p1\_dog, p2\_dog and p3\_dog.
     + A column called confident has p1\_conf, p2\_conf and p3\_conf.
2. Storing whole data:
   * File called **twitter\_archive\_enhanced.csv** was save with the name **df.**
   * Using the url: [**https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad\_image-predictions/image-predictions.tsv**](https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)the file was saved as **images**.
3. Twitter API and Python’s Tweepy library were useful to gather information and it was saved as **twitter\_data.**
4. Files df, images and twitter\_data was merge in one file called: **all\_data**. At this point all data was cleaned according with our goals and saved it as **twitter\_archive\_master.csv**
5. Analyzing, and visualizing: Some columns were plotted which allowed to identify the favorite dogs, unfavorite, which has more retweets and other points.

Here, I will include some graphs but the whole analysis will be saved in **act\_report.pdf.**

**What is the 10 most popular dog breed?**

I plotted the column called **image\_predic**. These data are in the file called all\_data.



As we observed golden\_retriever is the favorite follow by Labrador\_retriever and pembroke. In addition, we have NAN that mean we do not have data.

**What is the 10 least popular dogs breed?**

There is a big difference between with favorite and least favorite dogs. In this case the number count by 2 and 1 as lest favorite.

