Wrange Report

❖ Introduction

- ➢ In the purpose of applying data analysis techniques we will use twitter user @dogs_rate tweets to do a wrangling process to discover some insights and analytics regarding dogs rates

❖ Process

- ➢ Gather
  - ■ We need to gather data in order to do our process on it so there swerve 3 suggested datasets that we will use in this process
    - ● Twitter archive file proved by @dogs_rate
    - ● Tweet image predictions provided by Udacity using neural networks on the Twitter archive dataset
    - ● Querying dataset from Tweepy API based on the Ids provided by Twitter archive dataset to get the complete informations about the tweets
- ➢ Assess
  - ■ Assessing the data visually and programmatically to define the issues that needs to be fixed in the data
- ➢ Cleaning the data
  - ■ Clean each issue discover by the assessing step in order to get the right results

❖ Quality issues :

- ➢ 'twitter-archive-enhanced.csv'
  - ■ retweeted_user_id & retweeted_status_id:
    - ● There are retweets we must drop them
  - ■ expanded_urls
    - ● Tweets without images we must drop them
  - ■ Timestamp:
    - ● Object format instead of datetime
  - ■ name:
    - ● some names are false (a, by,0,my..)
  - ■ text & rating_numerator:
    - ● Some tweets include more than one rating or decimal numbers causing wrong or missing data in the rating_numerator and rating_denominator
  - ■ pupper, puppo, floofer and doggo column:
    - ● missing data
    - ● There are some IDs with more than one dog "stage" information (two dogs are rated).

        - We should add a column for the fraction of rating_numerator and rating_denominator
        -
- Predictions¶
        - p1,p2,p3 columns: dog breeds needs to be converted to lower case all of them

❖ Tidiness Issues

- twitter_archive:
    - 4 columns (dogger, floofer, pupper and puppo) they should be only one column dog_stage
    - Other 2 datasets should be joined to twitter_archive
- Predictions
    - The dog breed prediction should be one column breed_prediction
    - The prediction confidence should be one column one column prediction_confidence