# Wrangle & Analyze WeRateDogs Data

## Wrangle Report

### Introduction

The purpose of this project is to put into practice what I’ve learned in Data wrangling data course which is part of Udacity Data Analysis Nanodegree program.

The dataset that I will be wrangling is the tweet 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. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. (11/10, 12/10, 13/10).

### Project details

The tasks of this project are as follows:

- Gathering data
- Assessing data
- Cleaning data

### Gathering data

Data for the project was gathered from various sources:

- Twitter archive file:

This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017

The twitter-archive-enhanced.csv was provided by Udacity, downloaded manually then was loaded from the CSV file into a pandas data frame.

### • The tweet image predictions:

This file contains the top three predictions of dog breed for each dog image from the WeRateDogs Enhanced Twitter Archive. Data is downloaded programmatically using the Requests library from the URL address into a tsv file. The content of image-predictions.tsv file is then loaded into the pandas' data frame, with the size of 2075 rows and 11 columns.

The table contains the top three predictions, tweet ID, image URL, and the image number that corresponded to the most confident prediction.

### • Twitter API File:

Twitter API file contains tweet id, favorite count and retweet count. Data was provided by Udacity, downloaded manually then was loaded from the tweet-json.txt  file into a pandas data frame. Dataframe size is 2354 rows and 2 columns. The tweeter ID column is used as an index.

## Assessing data

After gathering, the data is assessed for tidiness and quality as follows:

### • Enhanced Twitter Archive

- As a first step, a sample of data is assessed visually and a summary of data types and non-null values is displayed. This allows to identify columns with the incorrect data type and/or null values. Then, IDs are checked for duplicates. Next, the number of tweets which are replies and retweets is calculated.

- Name of dog column is assessed programmatically checked for the number of values. And all tweets were checked for dogs with more than one dog category (stage) assigned.

- Rating denominator is assessed visually by displaying a sample of data, and then ratings with denominator greater than 10 are printed out for further investigation. Rating numerator is also assessed visually. Based on the visual assessment of rating columns, we check programmatically text column for any float ratings.

- Expanded URLs are firstly assessed visually and then checked programmatically for the existence of two or more URLs in one cell

### • Image Predictions

- As a first step, a sample of data is assessed visually and a summary of data types and non-null values is displayed. This allows to identify columns with the incorrect data type and/or null values. Then, the jpg_url column was checked for duplicates, also it was checked to confirm if it contains only jpg and png images. As the last step, the 1st prediction is checked to see how many images have been classified as dog images.

### • Twitter API Data

- As a first step, a sample of data is assessed visually and a summary of data types and non-null values is displayed. This allows to identify columns with the incorrect data type and/or null values. Then, IDs are checked for duplicates.

## Cleaning data

The quality and tidiness issues identified in the Assessing Data section are cleaned using pandas:

### ▲ Enhanced Twitter Archive

- As a first step, a copy of dataset is created for use throughout the cleaning exercise. As some of the gathered tweets are replies and retweets, we remove them together with other unnecessary columns.

- Dog 'stage' classification (doggo, floofer, pupper or puppo) which was broken into four separate columns, is merged into one column.
- Next, we fix the timestamp which has an incorrect data type - is an object - by converting it to DateTime.
- Float ratings, which have been incorrectly read from the text of tweet are gathered again, this time correctly. The denominator of some ratings is not 10, while numerator of some ratings is greater than 10 - the fact that the rating numerators are greater than the denominators does not need to be cleaned, however, we introduce a normalized rating which will be used for plots.
- We have 639 expanded URLs which contain more than one URL address, therefore we build correct links by using the tweet id field.

### ▲ Image Predictions -

- As a first step, a copy of dataset is created for use throughout the cleaning exercise. As some of the column names are confusing and do not give much information about the content, so we rename columns.

- Then we clean dog breeds - we replace underscores with whitespace and capitalize the first letter to have consistent and clean formatting.

- 66 image_url duplicates were removed.

- since only 2075 images have been classified as dog images for the top prediction (1st prediction), we use the dog breed predicted in the 2nd or 3rd predictions for the remaining rows.

### ▲Twitter API Data -

There is no need to perform cleaning tasks in this data set.

As a last step of the cleaning process, we merge all datasets into one and export to twitter_archive_master.csv file.