WRANGLING REPORT
By Paul BEPE

INTRODUCTION

Our report is part of the Udacity Wrangle and Analyze Data project to document our work. We worked on the WeRateDogs Twitter account data, who is a page who rates people's dogs with humorous comment about the dog. This document content all the step of data wrangling known as gathering data, assessing data and cleaning data.

1. Gathering Data

We needed to gather data from several sources and different of formats.

a. **The WeRateDogs Twitter archive**: The file was provided by the project and could be also downloaded directly from Udacity website.
b. **The tweet image predictions**: The file was hosted on Udacity's servers. we downloaded this file programmatically by using the *Requests* library in Python.
c. **Twitter API & JSON:** normally this file should have been downloaded with the API, but because of some technical issues we used the given json file directly in our notebook.

2. Assessing data

After creating three data frame from the gathered data (***twitter-archive- enhanced data frame, retweet_count data frame, predict data frame***), we explore the data frames to found the issues in each data frame, after that we perform a visual assessing and we also did it programmatically to look for missing issues, quality issues and tidiness issues

a. Quality issues

twitter-archive-enhanced table

- Bad type of the timestamp columns (must be date time not string);
- Abnormal values for rating_numerator, rating_denominator columns, e.g., 170, 150, 130, etc.;
- Keep only date for timestamp columns;
- Rename timestamp column;
- Timestamp type is str, should be datetime, and we should remove +0000 in timestamp;
- The tweet_id columns must be string not integer;
- Drop None dog names;
- Drop useless columns.

Image prediction table

- many entries are not dogs, e.g., jaguar, mailbox, peacock, cloak, etc.
- bad capitalisation in first letters for breeds dogs;
- Underscore for many breed dogs names;
- jpg url duplicates;
- Convert the numbers to percentage format;
- The right predictions;
- No need for source of JPG url;
- Drop useless columns.

retweet table

- Tweet_id column type is a string.
- Some missing data

b. Tidiness issues

- Merging doggo, floofer, pupper, puppo to one column;
- Assembling many parts of 3 tables in same table.

3. Cleaning data

In this step, we tried to solve all issues found in the assessing step. Before starting the cleaning process, we make a copy of each table to avoid losing or changing our original. For the missed values issues detected after the visualization assessing, we tried to extract more values for numerator and denominator ratings. Then, the more important quality problems that needed more work where cleaning the abnormal values of numerator rating and get the year, month also day in **Twitter-archive-enhanced table**. On the other hand, in the **Predict table** the most complicate quality issues are get just the right predict with high percentage prediction, also extract the dog images names. Beside these complicated issues, we performed some formatting tasks like capital letters, type of columns, deleting useless columns...

Thereafter, they are two tidiness issues in this project. We merged all the three tables in one table after melting the dog stages columns in one column.

After fixing all this issues, we store clean data in a csv file.