UDACITY

Udacity Data Analyst Nanodegree

Wrangle and Analyze Data:

Data Wrangling Report

Written by:

Anastasia Makarevich

Date: March, 5, 2019

Introduction

The goal of the project is to analyze twitter data gathered from WeRateDogs (@dog_rates) twitter account. This twitter account posts dogs pictures with ratings that typically range between 10/10 and 14/10. There are three sources we can get the data from:

- Twitter archive (dataset provided by Udacity) that contains some basic tweets data like id, publication date and time, tweet text as well as some extracted characteristics like dog's name, dog's rating and dog's stage (doggo, pupper, floofer or puppo).
- Predictions data set (available for download) that contains 3 most likely predictions of what's depicted on the photo for each tweet.
- Twitter API that allows us to download favorites counts and retweets counts for each tweet

The data wrangling process included three stages: **gathering**, **assessing** and **cleaning**. The main tools for this stage were Python and Jupyter notebook.

Gathering

The twitter archive was manually loaded from Udacity's website. The predictions dataset was downloaded programmatically using ***requests*** library.

Lastly, additional data was loaded using Twitter API and ***tweepy*** library. When fetching with Twitter API, I only queried for the original tweets excluding replies and retweets.

Assessing

This was the longest stage of the data wrangling process. I discovered several rather severe quality and tidiness issues. In terms of **quality**, I've identified the following major groups of issues (for full list of issues please see full report in wrangle_act.ipynb):

- Erroneous datatypes: *tweet_id* in all datasets is an integer, while conceptually it's not an integer and operations like adding/subtracting/multiplying don't make sense on it.
- Invalid ratings: there are numerous incorrect ratings (due to incorrect parsing (e.g., 11.75/10 → 75/10), puppies ratings scaled by the number of puppies (e.g., 88/80) or 'special dogs' (e.g. 1776/10 for Declaration of Independence dog).
- Invalid dog names: looks like the parser extracted anything that followed the phrase "This is … ", so "a", "an" and "the" are reported as names.
- Incorrect timestamp: the original dataset has all timestamp values converted to UTC while original tweets are in US/Pacific.

As for **tidiness** issues, most of them were contained in the twitter archive dataset:

- One variable spread out into 4 columns: doggo, floofer, pupper, puppo
- Tweets are mixed with retweets and replies (violates the principle of one table per observational unit)
- *text* column contains two other variables: unexpanded link and rating that were already extracted and that are not essential for text semantics
- *expanded_urls* column has multiple values in one cell most of which are duplicated

- *source* contains two variables
- all three datasets must be merged into one master dataset

Cleaning

The core steps performed during the cleaning stage include the following:

- *tweet_id* converted to str type
- *timestamp*: converted to datetime type and to US/Pacific timezone
- *text:* texts were cleaned from ratings, links and special symbols '\n\ and & using regexes
- *doggo*, *pupper*, *puppo* and *floofer* columns were converted to one categorical variable
- *name*: all lowercase values and "None" strings were converted to None type
- *source*: link and source name were extracted to separate file
- retweets and replies removed from dataset as well as 2 advertising tweets
- ratings were fixed where possible
- *expanded_urls*: duplicated removed, remaining links unpacked into two columns
- all three datasets were merged together by tweet_id and saved into **twitter_archive_master.csv** file