Wrangle Report (Udacity - Data Analyst Nanodegree Program)

Alen Mrsic

In this document, I'll describe how I gathered, analyzed and cleaned the data. The goal of the project is to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

1. Gathering the data

I gathered each of the three pieces of data as described below:

- from HTTP server, I downloaded manually by clicking the following link: **twitter_archive_enhanced.csv.**
- **image_predictions.tsv** file which is hosted on Udacity's servers I downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv.
- Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called a **tweet_json.txt** file

The first step of the data wrangling is to gather data. In our case, we have three different methods to make that. First, I manually downloaded a file from the server, then I gathered data via python Request library and the last I scraped the data via python and Twitter API. From the Twitter API, I received an error "No status found with that ID" with an error code "144" for 11 tweets. From https://developer.twitter.com/en/docs/basics/response-codes I found out that the requested Tweet ID is not found (if it existed, it was probably deleted).

2. Assessing the data

After gathering a data from different sources and importing them into data frames (memory) I investigated each data separately (visually and programmatically) to detect data quality and tidiness issue. For this project, it was a requirement to find at least eight data quality and two tidiness issues.

Data quality issues were:

- there are retweets in the dataset, remove the data and columns
- different number of entries in image_predict (2075) from twitter archive (2356), so that means that some tweets have no image

- display full content of text column
- remove HTML part from source column to be easier to read
- correct the dog names, find out a name from tweet text (some cases like 'a', 'an')
- rename columns to have a more meaningful name (to be easier understand what that column represent)
- change the data type for columns which have a wrong data type.
- from tweet text remover URL

Tidiness issues were:

- create one variable from four columns: doggo, floofer, pupper, puppo
- create one data set, join image_predict and tweet_json with twitter_archive_enhanced

3. Cleaning the data

Before I started with data cleaning, I created a copy of each dataframe to persist original data. I divided a cleaning process into three parts. For every issue were Defined tasks, Coded logic and Tested results. I merged three data sets into one and passed through all data quality and tidiness issue. Cleaned data were saved to a csv file.