Wrangle report

Project: Wrangling and Analyze Data

Dmitry Zmienko

1. Introduction

The wrangle report is an internal document of the “Wrangle and Analyze Data” project. The project’s aim is together data from tweet archive of Twitter user WeRateDogs (@dog_rates), assess its quality and tidiness, clean it and then gather some useful insights.

2. Gathering data

Whole data is gathered from three pieces:

- The twitter_archive_enhanced.csv was given by Udacity platform and then uploaded to my Jupyter Notebook Workspace. This file contains information about 2000+ tweets (data, source, URLs, dog\s name, ratings, etc.)
- The image_prediction.tsv file with results of image predictions of breed of dog (top three only) from neural network. The file was downloaded programmatically from Udacity’s server.
- Additional data from the twitter API about tweet’s retweet and favorite counts. Using the tweet IDs I’ve tried to query the Twitter API for each tweet’s JSON data using Tweepy library. But have failed (my guess) due to problems with Twitter access from Russia (access is limited). So I used tweet_json.txt file from Udacity platform.

Files were open like DataFrames:

df_arch - twitter_archive_enhanced.csv

df_img - image_prediction.tsv

df_tw - tweet_json.txt

3. Assessing Data

After the data was gathered and stored in my Jupyter Notebook Workspace I assessed them visually and programmatically for quality and tidiness issues. The result is gathered in table:

№	Table	Column	Issue type	Method	Description
1	all tables	tweet_id	Quality	Programmatic	tweet_id is int64 format and has to be string
2	df_arch	name	Quality	Visual, Programmatic	Invalid dog’s names
3	df_arch	source	Quality	Programmatic	Has HTML tags, URL info and source info
4	df_arch	text	Quality	Programmatic	Has URL information in text description
5	df_img	p1, p2, p3	Quality	Programmatic	Dog’s breed in different standards

№	Table	Column	Issue type	Method	Description
6	df_arch	timestamp, retw_timestamp)	Quality	Programmatic	Date column not in date type
7	df_arch	-	Quality	Programmatic	The duplicates for one dog because of retweets
8	df_img	-	Quality	Programmatic	Duplicated images
9	df_arch	Rating_denominator, Ratings_numerator	Quality	Programmatic	Rating_denominator sometimes is not 10. Ratings numerator has several outliers (like 1776)
10	df_arch	doggo, floofer, pupper, puppo	Tidiness	Visual	4 last columns contain categorical variables and can be combined
11	all tables	-	Tidiness	Visual, Programmatic	Merging df_arch, df_tw and df_img into one table by tweet_id

4. Cleaning data

After the assessment the cleaning procedure (define – code – test) were performed. The three cleaned data frames were merged in one (also saved in twitter_archive_master.csv) for further analysis.

5. Analyzing and visualizing Data

To check the usefulness of cleaning data I produced 3 insights with 2 visualizations:

- Most common dog breed and dog name
- Most common rating
- Most popular dogs (most favorite & retweeted)
- Tweeting activity