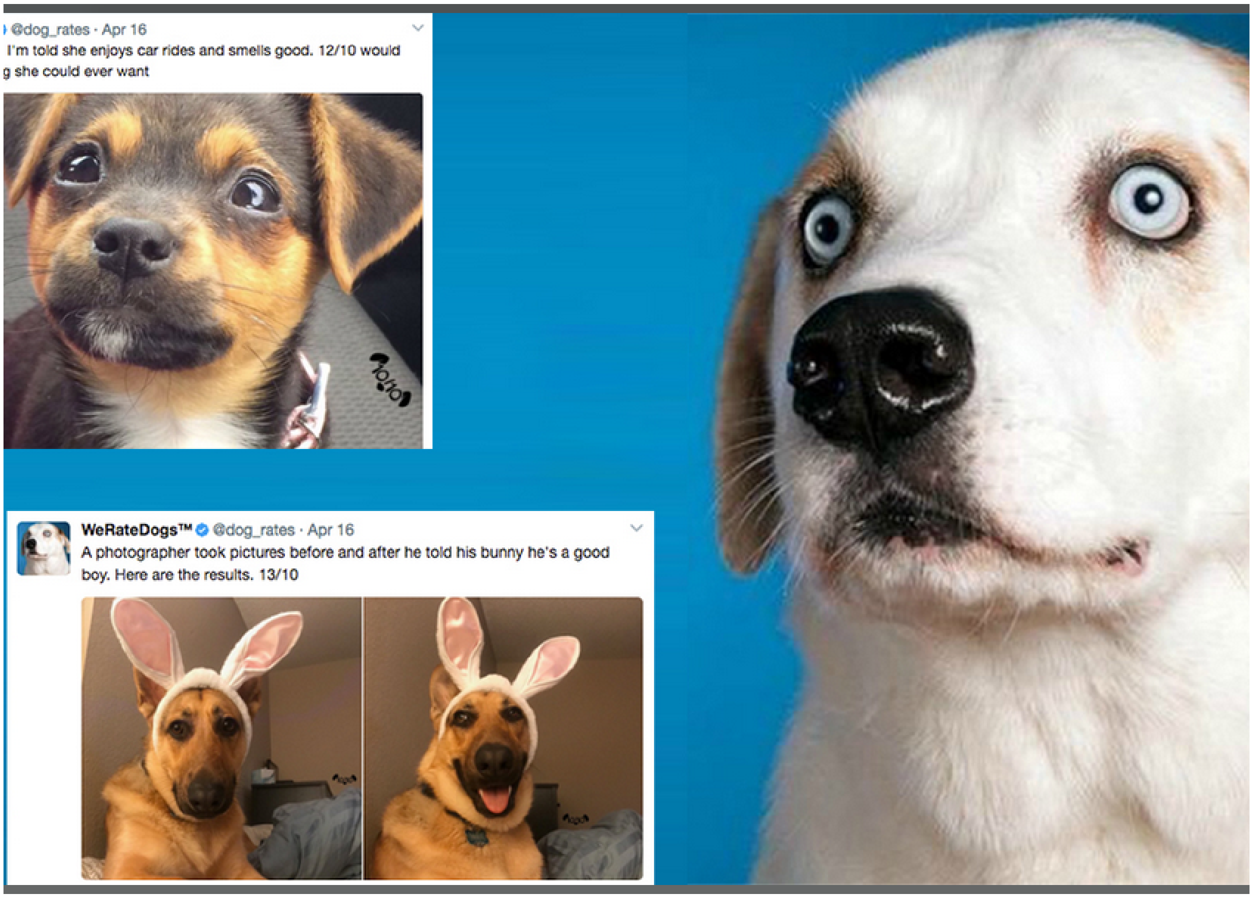
ALX-T Data Analyst Nanodegree Program

Project 2: Data Wrangling

WeRateDogs twitter Archive



@dog_rates · Apr 16
I'm told she enjoys car rides and smells good. 12/10 would
g she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

BY: Joshua Regan Lenge

Project: https://github.com/Reganmatics/Udacity_WeRateDogs_project

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

The tasks in this project are as follows:

- Step 1: Gathering data
- Step 2: Assessing data
- Step 3: Cleaning data
- Step 4: Storing data
- Step 5: Analyzing, and visualizing data
- Step 6: Reporting

Gathering Data

At this stage, the data tables(DataFrames) are created from;

1. Twitter_archive_enhanced.csv
2. Image_predictions.tsv
3. tweet-json.txt

Then using the pandas library we have;

df_1, df_2 = pd.read_csv(file_path) for the **csv** and **tsv** file

And df_3 -> tweet-json.txt

Assessing Data

At this stage, we perform exploratory analysis on the datasets to get the defects latent in the data such as dirty data (quality issues) and untidy data (structural issues)

Though the datasets have more than 10 issues combined, for this project we only identify 8 quality issues and 2 tidy issues.

The issues identified are;

Quality issues

df_1

1. Invalid name entries in name column e.g None, a, *et̀al*..
2. Tweet_id should be of string Dtype not int64. (df_1 and df_3).
3. The columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_id, retweeted_status_timestamp have lower than 25% of non-null entries, all less representative.
4. timestamp should be of Dtype datetime instead of object.
5. Dog ratings are not standardized
6. Expanded_urls has Nan values.

df_2

7. Remove duplicate entries in the jpg_url column.
8. df_2: Remove entries with p1_dog == False, p2_dog == False AND p3_dog == False. They are not dogs.

Additional Issues

9. dog_breed should have a separate column

Tidiness issues

1. df_1: melt the four dog stages into one column.
2. merge df_1_clean, AND df_3_clean into one dataFrame.

Cleaning Data

At this stage, we clean all the first eight **quality issues** identified and the two **tidiness issues.** The cleaning is carried out in the order; Define Code and Test but first we create new copies of each DataFrame then perform our cleaning operation on the cleaned DataFrame.

#**Define**

Write out in issues using Verbs (action statements) this is more like a pseudocode for the testing stage.

#**Code**

Translate the action statement to Code.

#**Test**

Write the code to check if the issue has been resolved.

Storing Data

At this stage, we merge our datasets as we deem fit based on the specifications of our cleaning protocol. This is to enable anyone carry out the same research and track your

Analyzing, and visualizing data

At this stage, we perform exploratory analysis on the cleaned data to identify 3 insights and then make at least one visualization

Some insights from cleaned data

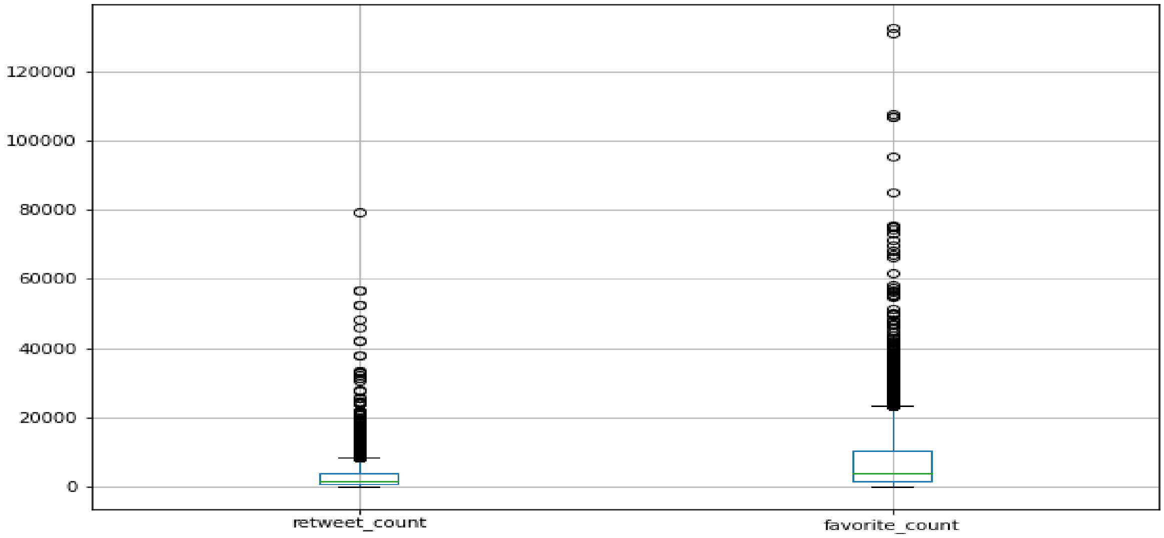
1. From the barplot we confirm that like are far more than retweets.
2. For the retweets;
   - The dog with the highest `retweet_count` is `Stephan` with `56625` retweets.
   - The dog with the lowest `retweet_count` is `Scout` with `23` retweets.
   - observe that the favorite_count for `Stephan` is 0 which isn't ideal for the topology of tweets in reality and backed by the +ve correlation earlier shown in the visualisation.
   - **Note:** this is just an observation, not an expression of doubt.
3. For the ratings;
   - The dog with the highest rating is `Atticus` with 177.6.
   - The dog with the lowest rating is `Crystal` with 0.2.

Check for dogs with min and max retweet_count (retweets)

	173	**1499**
tweet_id	842892208864923648	666447344410484738
timestamp	2017-03-18 00:15:37+00:00	2015-11-17 02:46:43+00:00
source	<a href="http://twitter.com/download/iphone" r...	<a href="http://twitter.com/download/iphone" r...
text	RT @dog_rates: This is Stephan. He just wants ...	This is Scout. She is a black Downton Abbey. I...
expanded_urls	https://twitter.com/dog_rates/status/807106840...	https://twitter.com/dog_rates/status/666447344...
rating_numerator	13	9
rating_denominator	10	10
name	Stephan	Scout
rating	1.3	0.9
dog_stage	None, None, None, None	None, None, None, None
retweet_count	56625	23
favorite_count	0	107

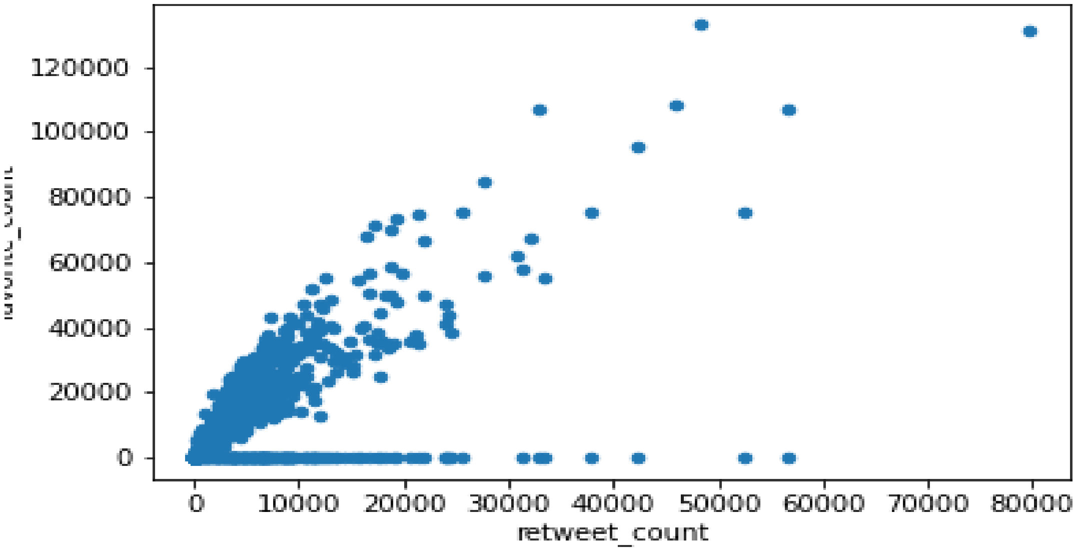
Check for dogs with min and max rating

	696	1186
tweet_id	749981277374128128	678424312106393600
timestamp	2016-07-04 15:00:45+00:00	2015-12-20 03:58:55+00:00
source	<a href="https://about.twitter.com/products/tw...	<a href="http://twitter.com/download/iphone" r...
text	This is Atticus. He's quite simply America af....	This is Crystal. She's a shitty fireman. No se...
expanded_urls	https://twitter.com/dog_rates/status/749981277...	https://twitter.com/dog_rates/status/678424312...
rating_numerator	1776	2
rating_denominator	10	10
name	Atticus	Crystal
rating	177.6	0.2
dog_stage	None, None, None, None	None, None, None, None
retweet_count	2772	2880
favorite_count	5569	5916



Favorite_count

The number of likes are far more than the number of retweets, this could be because its easier to click the like button than to retweet with a text.(a possibility :) -> :| -> :) -> :|)



Likes Vs retweets

Clearly likes and retweets are positively correlated with 1.5 times more likes than retweets.

Reporting

At this stage, we document our wrangling and project reports in **act_report.pdf** and **wrangle_report.pdf**

Side notes: This project can be used by anyone to practice thier wrangling skills for free