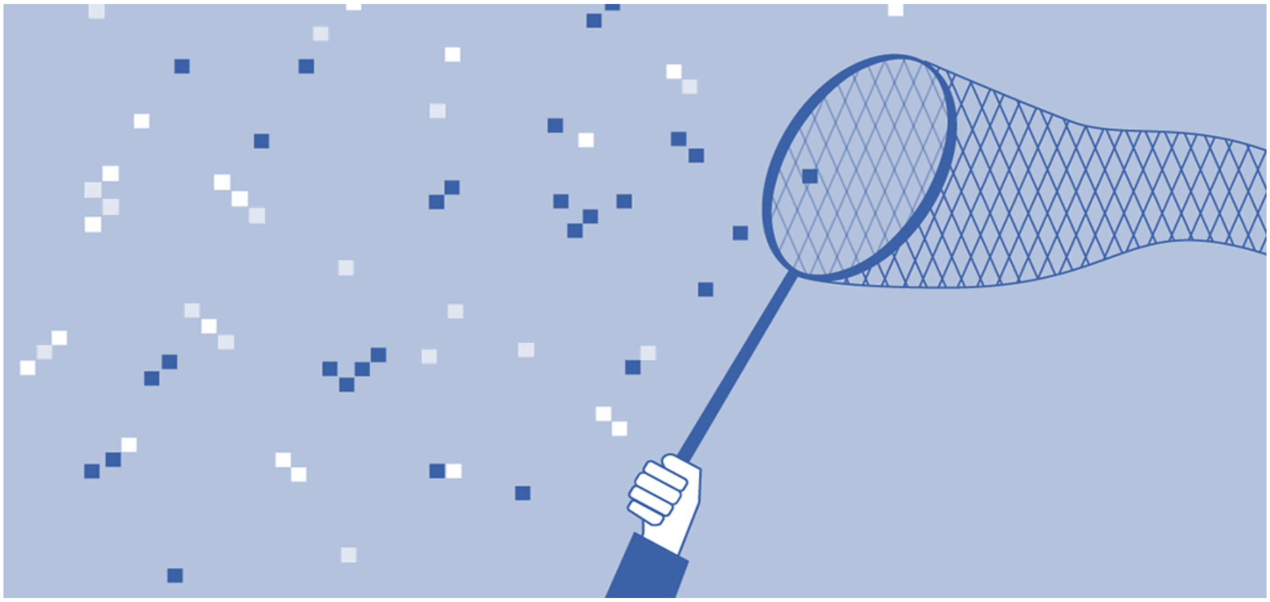
Report Data Wrangling

Gather, Assess, and Clean

Nada Alzahrani

Wrangle Report



Introduction:

Dataset in the project is tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a twitter account that rates people's dogs with humorous comment about the dog.

This report briefly describes my wrangling efforts.

Project details:

The WeRateDogs Twitter project goals (Wrangling the data through the following processes):

- Gathering Data
- Assessing Data
- Cleaning Data
- Storing, analyzing and visualizing data
- Reports

1. Gathering data:

The data for this project consists of three separate datasets which have been collected as follows:

- **Twitter archive file:** Udacity provided twitter archive enhanced.csv and manually downloaded it.

- **Image predictions file:** The predictions of the tweet picture, i.e. what breed of a neural network is present in each tweet. This file (image predictions.tsv) is hosted on the servers of Udacity and was programmatically downloaded using the Requests library and URL files.

- **Tweetjson file:** Twitter API & JSON: Using the tweet IDs in the WeRateDogs Twitter folder, I used the Python Tweepy library to query the Twitter API for each tweet's JSON data and stored the entire collection of JSON data of each tweet in a file named tweet_json.txt. With tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url, I read this .txt file line by line into a pandas dataframe.

2. Assessing Data:

I started to evaluate the data on both quality and tidiness issues once the data was collected.

Issue of Quality:

'twitter-archive-enhanced-2.csv':

- Remove unneeded columns
- Change timestamp from string to datetime
- Separate columns for date and time

'image_predictions.tsv':

- P1,p2 and p3 have unnessary underscore
- P1,p2 and p3 have inconsistent words after remove underscore
- Drop duplicate jpg_url.

'tweet_json':

- Rename id to tweet_id

'twitter-archive-enhanced-2.csv', **'image_predictions.tsv' and 'tweet_json':**

- Change tweet_id from number to string.

Issue of Tidiness:

- Combining dog stages
- Inner join between three data frames

3. Cleaning Data:

Define, Code and Test:

- Remove unneeded columns to be pure.
- Change timestamp from string to datetime.
- Separate columns for date and time to be clear.
- p1,p2 and p3 have unnecessary underscore, so I delete it.
- p1,p2 and p3 have inconsistent words after remove underscore, we should make all words lowercase to become consistent.
- Drop duplicate "jpg_url".
- Rename id to tweet_id to merge all datasets.
- Change tweet_id type, because tweet_id is string.
- Combining dog stages.
- Inner join between three data frames.