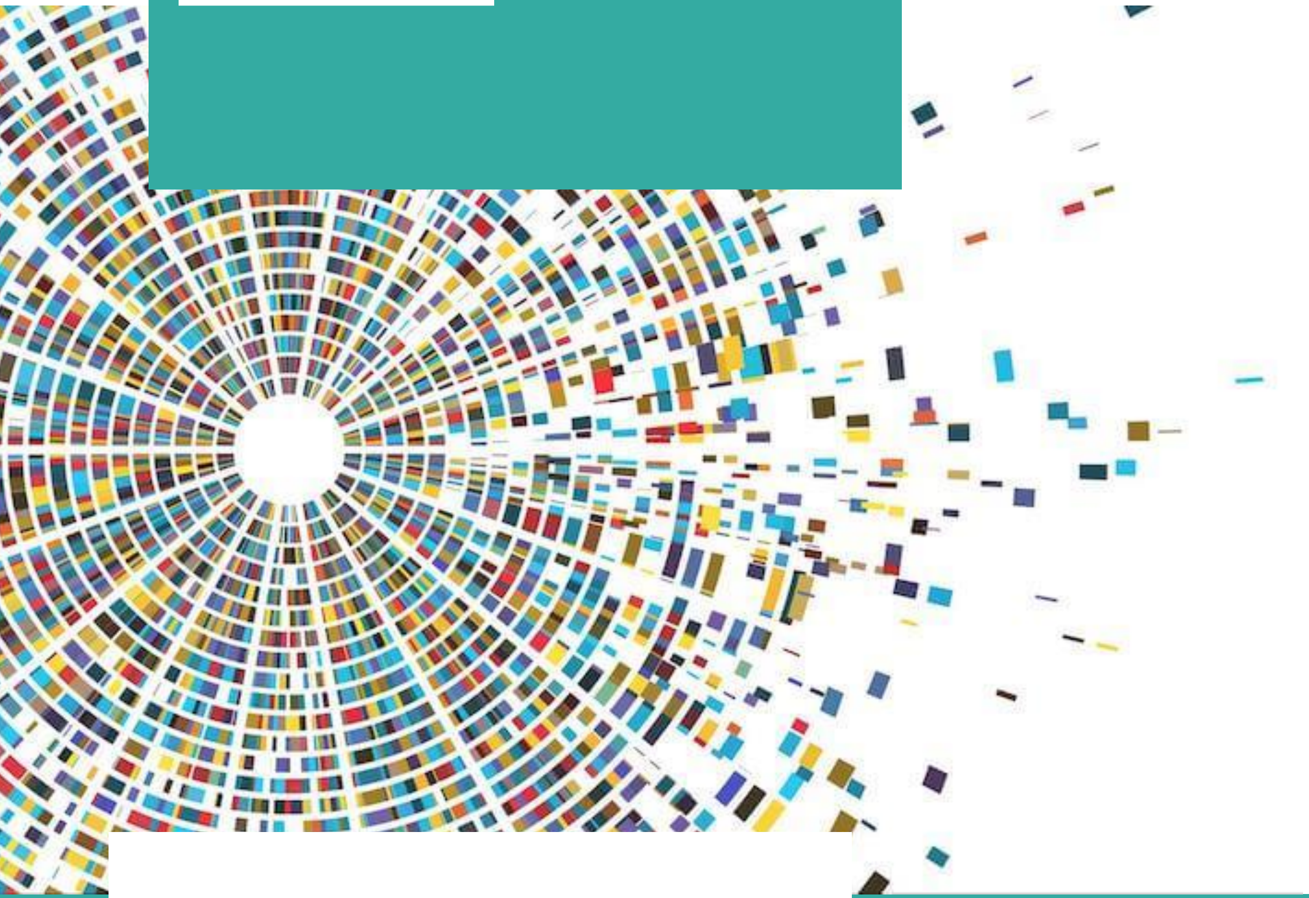
Data Wrangling Project



DECEMBER 15

Authored by: Abdelrahman Mohsen

Overview

The main goal for this project is to practice data wrangling processes on a real dataset to master the required skills, dataset came from WeRateDogs.

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The rating system in this account is special as dogs can take rate higher than the rating scale (0 to 10) so its fine to find that the nominator of the rating exceeds the dominator.

Data wrangling process:

There are three processes to be done on the dataset to be able to extract useful information from it.

Gathering data: in this step you build the dataset from all possible sources (if needed) to get a reliable data (e.g. web scrapping, using APIs etc.)

Assessment: look at the data visually and programmatically to detect quality and tidiness issues.

Cleaning: fix all issues appeared in assessment process which can be done programmatically or manually.

In order to work on the project I used python notebook with pandas library as the main tool, I also used useful libraries like numpy, re, requests, tweepy and json.

All code is included in wrangle_act.ipynb.

Gathering Process

In this process we gathered data from three sources (csv file - tsv file - twitter API) creating data frames that are ready for assessing and cleaning steps.
I go through four steps:

1. Read the csv file and make dataframe for it.
2. Download the image predictions file programmatically through requests then read its content to pandas dataframe.
3. Retrieve more data from account using tweepy API, store them in txt file.
4. Read text file (generated from the previous step) line by line to create dataframe.

Input:

- twitter_archive_enhanced.csv
- image_predictions.tsv
- Data from twitter API

Output:

- DataFrame for twitter archive data
- DataFrame for image prediction data
- DataFrame for twitter API data
- Tweet_json.txt
- image-predictions.tsv

Assessment Process

In this process we will assess the dataframes we created above visually and programmatically for quality and tidiness issues. We will extract eight (8) quality issues and two (2) tidiness issues.
I used useful methods for pandas dataframe like head, info, sum and value_counts

After visual and programtic assessment we can summary the issues in our data:

Quality issues

archive dataframe

- source column needs to be cleaned from tags and url that not related
- date and time need to be datetime object
- drop rows with no images
- drop replies tweets
- drop retweets
- change None values in doggo, floofer, pupper and puppo columns to empty string
- rating numerator column must be float
- correct wrong values of rating_numerator and rating_denominator

Tidiness issues

archive dataframe

- doggo, floofer, pupper and puppo need to be one column

image predictions dataframe

- merging image prediction dataframe with archive dataframe

Cleaning Process

We do cleaning into three steps:

1- Define: clearly describing how you will solve the issue.
2- Code: write codes the fix the issue.
3- Test: test the output to make sure of the solution.