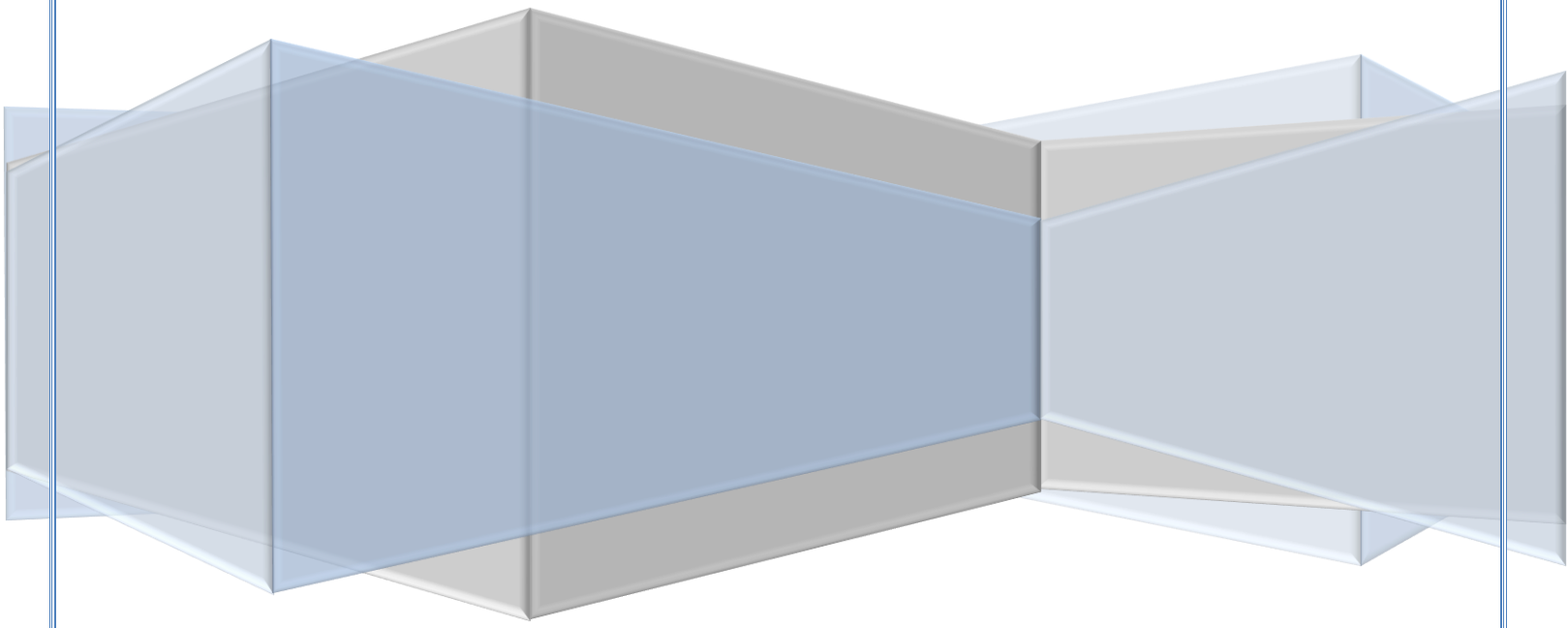
UDACITY PROJECT

WARANGLE REPORT

Mohamed Najm Mohamed

Mohamednajm250@gmail.com



Overview

this is a data wrangling project which means that gather data from a many sources and with a variety of formats, assess them (quality and tidiness) and define the problem in them clearly, then clean it and doing all of that by using python with its libraries.

Data wrangling project

First libraries:

I used those libraries and import them

- pandas
- numpy
- json
- tweepy
- requests
- matplotlib.pyplot

Data wrangling has three phases:

• Gathering data
• Assessing data
• Cleaning data

Phase one, Gathering data:

As the description of the project that what i was been asked to do:

Gather each of the three pieces of data as described below in a Jupyter Notebook titled (wrangle_act.ipynb):

1. The (WeRateDogs)Twitter archive. I am giving this file to you, so imagine it as a file on hand. Download this file manually by clicking the following link: twitter_archive_enhanced.csv

2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the followingURL:https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_mage predictions/image-predictions.tsv

3. Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Phase two, Assessing data:

There I define some problems in the data (manually and programmally) like:

First: Quality issues

twitter_archive_data:

- Drop retweets of retweeted_status_user_id by filtering the NaN
- alot of nan values in many columns:in in_reply_to_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_user_id and retweeted_status_id.
- split time column into 3 after changing its type from string
- cleaning the source text column form tage of html
- drop exetrem values and then Correct any Invalid values at rating_numerator
    - Correct numerators with decimals
- drop exetrem values and then Correct any Invalid values at rating_denominator
    - Manually (assessed individual by print text).
    - programarlly(any Tweets without denominator equal to 10 is a multiple dogs).
- convert float numerator_rating

json_tweet:

- Drop any retweets

image_dogs_predictions:

- Removing any retweets which will turn retweeted_status column into valueless column.
- Drop duplicated jpg_url
- Create 2 columns:one for confidence level and another for dog image prediction

Second: Tidiness issues

- turn type of tweet_id into int64 to be ready to marge with the other two tables
- create column for dog stage which out of merging four columns: floofer, puppo, doggo and pupper
- finally, Merge the three datasets into one.

Phase 3: Data cleaning:

The last phase and in it I tried to solve all the problems which was defined before at assessing phase using python and for every problem statement there was a more detail define and the code part in which I solved it and then by the end the test part to test if it was solved or not.

Phase four, data analysis:

Some analysis and visualization to have some insight about the data.

And that it so I tried to go through what was defined clearly in the discretion of the project and achieve every single demand.