1 . Gathering Data

Gather Twitter archive CSV file

Using the link provided by Udacity (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter archiveenhanced/twitter-archive-enhanced.csv) I downloaded the WeRtwitter_archive_enhanced.csv

file and imported this file into a dataframe(df_1)

Gather tweet image predictions

I downloaded the tweet image predictions file hosted on Udacity's servers programmatically using Python's Requests library and saved it locally to image_predictions.tsv file. Then, I imported this file into a Python Pandas dataframe (df_2).

Gather data from Twitter API

first, I tried to set up a developer account and flow instruction in course they send me first email to ask me about more details then I send all details about codes and how I use but they inform that my app is not accepted

second, I used the tweet_json.txt provided in resources then i read the file line by line and use pandas to extract 'tweet_id', 'favorite_count', 'retweet_count to wrangle it.

2 Assessing Data

Visual Assessment

I opened the twitter_archive_enhanced.csv and image_predictions.tsv in spreadsheet and

scrolled through them, looking for quality and tidiness issues. I was able to spot a problem about rating and names.

Programmatic Assessment

I used pandas' info method on df_1 to spot wrong datatypes and other quality issues. Then I

used value_counts method on rating_numerator, rating_denominator and name columns to look up the range of their values and its distribution I also used google sheet and this helps me to identify the most import points I work on it.

Through this I wrote the following issue

Quality

wrong data types (ex:tweet_id,timestamp)

missing some expanded_urls(url+tweet_id)

there is unnecessary columns(in_reply_to_status_id','in_reply_to_user_id,retweeted_status_id)

some names is not actual name

unnecessary html tags in source column

p1, p2, p3 inconsistent capitalization

inaccurate rating numerator

inaccurate rating denominator

Tidiness

doggo, floofer, pupper and puppo columns in df_1 table should be merged into one column

named "stage" .merge all 3 files.

3. Cleaning Data

I created a copy of all 3 files and named it df_clean_1,2,3. For each quality/tidiness issue, I performed the programmatic data cleaning and write script my codes and test it

4.Storing Data

After the completion of the cleaning process, I stored the archive_clean DataFrame in twitter_archive_master.csv file.