Gather

1. Image Pridections Data is downloaded from here

2. Twitter Archive Enhanced table is downloaded from Udacity resources and is uploaded with this file.

3. In order to extract data from Twitter API you should:

   1. First, if you do not already have one, you need to sign up for a Twitter account.
   2. Next, to set up a developer account, follow the directions on Twitter’s Developer Portal, in the “How to Apply” section.
   3. If you can't set up a Twitter developer account, or you prefer not to create a Twitter account for some reason, you may instead download it tweet_json.txt.

Assess

Quality

`archive` table

1. 109 erroneous name values (assigned as 'a', 'an', 'actually', 'his', ...)
2. tweet_id is an integer not a string
3. Inconsistent datatypes (timestamp, retweeted_status_timestamp)
4. Tweets without photos exist
5. Retweets & replies exist
6. Unuseful data (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
7. None values in dog stage columns

`image_pridections` table

1. Undescriptive columns' headers for image_predictions
2. tweet_id is an integer not a string

`twitter_API` table

1. tweet_id is an integer not a string

Tidiness

1. Dog stage are values represented as column names in archive table
2. Information about one type of observational unit (tweets) is spread across three different files/dataframes

---

Clean

Missing Data

1. 109 erroneous name values (assigned as 'a', 'an', 'actually', 'his', ...)

Define

archive: Iterate text columns trying to extract dog names, and putting NaN if name is not found.

Quality

2. tweet_id wrong data type

Define

Converting tweet_id to string in the three tables using astype

3. Inconsistent datatypes (timestamp, retweeted_status_timestamp)

Define

archive: Convert datatypes of 'timestamp' & 'retweeted_status_timestamp' to date using to_datetime.

4. Tweets without photos exist

Define

Use the image_predictions table to guide the selection and removal of tweets without photos in the archive table

5. Retweets & replies exist

Define

Using the following columns (in_reply_to_status_id, in_reply_to_user_id, 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'), we will shed the retweets and replies from our datasets and then will drop them.

6. Unuseful data (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)

Define

Drop all unuseful data from archive table

7. None values in dog stage columns

Define

Replace the “None” string with empty string ""

8. Undescriptive columns' headers for image_predictions

Define

Convert columns names into a more descriptive names.

Tidiness

1. Dog stage are values represented as column names in archive table

Define

doggo, floofer, pupper, puppo columns in twitter_archive_enhanced.csv should be combined into a single column as this is one variable that identify stage of dog.

2. Information about one type of observational unit (tweets) is spread across three different files/dataframes

Define

Using pd.merge function we will put all tables into one master table as they are part of the same observational unit.

Storing Data to csv file

Define

Using to_csv function we will store our final data to a csv file