Wrangle and Analyze Data Report

Prepared by Sushanth Bobby Lloyds on 11/Sept/2019

1. Gather

Data is gathered from 3 sources

i. **twitter-archive-enhanced.csv**
This data source is provided by Udacity

ii. **image-predictions.tsv**
This data source is programmatically downloaded from
https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

iii. **tweet_json.txt**
This data source is download from Twitter,

a. Created twitter developer account and mails went back and forth as twitter was requesting justification for API access. Requested was accepted within a day and access was granted.
b. After providing the authentication details like consumer_key, consumer_secret, access_token and access_secret was able to run the tweepy API, which ran roughly 1hr 15mins to download the records.

2. Assess

i. Data Structures/Schema

a. eta_df : Enchanced Twitter Archive

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
tweet_id                      2356 non-null int64
in_reply_to_status_id         78 non-null float64
in_reply_to_user_id           78 non-null float64
timestamp                     2356 non-null object
source                        2356 non-null object
text                          2356 non-null object
retweeted_status_id           181 non-null float64
retweeted_status_user_id      181 non-null float64
retweeted_status_timestamp    181 non-null object
expanded_urls                 2297 non-null object
rating_numerator              2356 non-null int64
```

```
rating_denominator            2356 non-null int64
name                          2356 non-null object
doggo                         2356 non-null object
floofer                       2356 non-null object
pupper                        2356 non-null object
puppo                         2356 non-null object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

b. ip_df : Image Predictions

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
tweet_id    2075 non-null int64
jpg_url     2075 non-null object
img_num     2075 non-null int64
p1          2075 non-null object
p1_conf     2075 non-null float64
p1_dog      2075 non-null bool
p2          2075 non-null object
p2_conf     2075 non-null float64
p2_dog      2075 non-null bool
p3          2075 non-null object
p3_conf     2075 non-null float64
p3_dog      2075 non-null bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

c. tweet_api_df : Data downloaded via Twitter API

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2333 entries, 0 to 2332
Data columns (total 32 columns):
contributors                    0 non-null float64
coordinates                     0 non-null float64
created_at                      2333 non-null datetime64[n
s]
display_text_range              2333 non-null object
entities                        2333 non-null object
extended_entities               2061 non-null object
favorite_count                  2333 non-null int64
favorited                       2333 non-null bool
full_text                       2333 non-null object
geo                             0 non-null float64
id                              2333 non-null int64
id_str                          2333 non-null int64
in_reply_to_screen_name         77 non-null object
in_reply_to_status_id           77 non-null float64
in_reply_to_status_id_str       77 non-null float64
in_reply_to_user_id             77 non-null float64
in_reply_to_user_id_str         77 non-null float64
is_quote_status                 2333 non-null bool
```

```
lang                             2333 non-null object
place                            1 non-null object
possibly_sensitive               2199 non-null float64
possibly_sensitive_appealable    2199 non-null float64
quoted_status                    24 non-null object
quoted_status_id                 26 non-null float64
quoted_status_id_str             26 non-null float64
quoted_status_permalink          26 non-null object
retweet_count                    2333 non-null int64
retweeted                        2333 non-null bool
retweeted_status                 165 non-null object
source                           2333 non-null object
truncated                        2333 non-null bool
user                             2333 non-null object
dtypes: bool(4), datetime64[ns](1), float64(11), int64(4),
object(12)
memory usage: 519.5+ KB
```

ii. Data frame analysis

a. Reviewed all dataframes using below commands for datatype consistency across dataframes as they may be joined and checking to see whether they are having valid datatypes for the values stored.

   1. *dataframe.info()*
   2. *dataframe.head()*

b. Checked value_counts to see whether it’s a categorized data or you can find a specific value is commonly used. Like eta_df.name.value_counts( ) let you know commonly used dog name

   1. dataframe.column.value_counts()

c. Analyzed textual data to see if any data can be derived from it. For example, from eta_df.text below information can be derived

   1. Gender of the dog
   2. Name
   3. http:// twitter link for the image of the dog

iii. Quality

(Below points were ordered as quality issues were found)

1. eta_df : name : Dog names missing filled with incorrect names. None should be converted to NaN Eg., (None(745), a(55))
2. eta_df : Only original posts are required. (in_reply_to_status_id, retweeted_status_id) = NaN should be kept as they are the original tweet. 78 replies/retweets found

3. eta_df : retweeted columns not required(retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
4. eta_df : denominator : Max denominator is expected to be 10. But there are other values as well
5. eta_df : text : Some texts are not complete as they end with elipsis
6. eta_df : stage : if dog doesn't have a stage(doggo, floofer, pupper, puppo) it should NaN not None
7. eta_df : Some dogs are in multiple stages like below 733109485275860992 741067306818797568 751583847268179968 781308096455073793 785639753186217984 801115127852503040 808106460588765185 817777686764523521 854010172552949760 855851453814013952
8. ip_df : Some of data in image-predictions is not dogs (shopping_cart, box_turtle). 543 are not dogs.
9. ip_df : There are around 104 duplicate images having different tweet_ids. Probably user retweeted same images.
10. ip_df : Remove '_'(underscore) from dognames p1, p2, p3

iv. Tidiness

(Below points were ordered as tidiness issues were found)

1. eta_df : Wrong datatype : Need to be converted from float64 to int64 : in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id
2. eta_df : Wrong datatype : (timetamp, retweeted_status_timestamp) is defined as non-null object it should be datetime
3. eta_df : Columns not required as per requirement : in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
4. eta_df : Change four columns to one column 'stage' (doggo, floofer, pupper, puppo)
5. tweet_api_df: Datatype Conversion from float64 to int64 : in_reply_to_status_id, in_reply_to_user_id, quoted_status_id
6. tweet_api_df: Removing columns with no value/information to process(contributors, coordinates, display_text_range, favorited, geo, possibly_sensitive, possibly_sensitive_appealable, quoted_status_permalink, retweeted, truncated, place, extended_entities, user)
7. Mostly all information which entities have only extended_entities have so extended_entities can be removed
8. tweet_api_df - Mostly all rows have same values in tweet_api_df['user'] column can be dropped
9. eta_df : Create a new column to have tweet_id as string
10. ip_df : Create a new column to have tweet_id as string

11. tweet_api_df: move id & id_str column to the beginning of the dataframe & rename it was tweet_id & tweet_id_str
12. eta_clean : Find dog gender based on column 'text'

3. Clean (Quality & Tidiness)

- **Issue 1 : Missing dog names**

  Define

  eta_clean : name : None should be converted to NaN Eg., (None(745), a(55))

- **Issue 2 : Only original posts are required**

  Define

  eta_clean : (in_reply_to_status_id, retweeted_status_id) = NaN should be kept as they are the original tweet. 78 replies/retweets found

- **Issue 3 : Replace dog_stage from None to NaN**

  Define

  eta_clean : stage : Replace from None to NaN in columns (doggo, floofer, pupper, puppo)

- **Issue 4 : Remove rows which are not dogs in image predictions**

  Define

  ip_df : Some of data in image-predictions is not dogs (shopping_cart, box_turtle). 543 are not dogs

- **Issue 5 : Removing duplicate images**

  Define

  ip_df : There are around 104 duplicate images having different tweet_ids.

- **Issue 6 : Remove underscore from dog names**

  Define

  ip_df : Replace '_'(underscore) from dognames p1, p2, p3

- **Issue 7 : Change datatypes**

  Define

  1) eta_df : Wrong datatype : Need to be converted from float64 to int64 : in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id

  2) eta_df : Wrong datatype : (timetamp, retweeted_status_timestamp) is defined as non-null object it should be datetime

  3) eta_df : Columns not required as per requirement : in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp

- **Issue 8 : Merge four columns to one**

  Define

  4) eta_df : Change four columns to one column 'stage' (doggo, floofer, pupper, puppo)

- **Issue 9 : Some dogs have multiple stages**

  Define

  Below is quality point, handled in tidiness due to melt being part of tidiness(changes dataframe structure)

  7) eta_df : Some dogs are in multiple stages like below 733109485275860992 741067306818797568 751583847268179968 781308096455073793 785639753186217984 801115127852503040 808106460588765185 817777686764523521 854010172552949760 855851453814013952

  and remove duplicate rows introduced by melt.

- **Issue 10 : Creating new column(tweet_id_str) in eta_clean & ip_clean**

  Define

  7) eta_clean : Create a new column to have tweet_id as string

  8) ip_clean : Create a new column to have tweet_id as string

- **Issue 11 : Removing unused columns**

  Define

  6) tweet_api_df: Removing columns with no value/information to process(contributors, coordinates, display_text_range, favorited, geo, possibly_sensitive, possibly_sensitive_appealable, quoted_status_permalink, retweeted, truncated, place)

- **Issue 12 : Reorder & Rename id & id_str column to tweet_id & tweet_id_str**

  Define

  9) tweet_api_clean: move id & id_str column to the beginning of the dataframe & rename it was tweet_id & tweet_id_str

- **Issue 13 : Find dog gender**

  Define

  10) eta_clean : Find dog gender by text analyzing column 'text'

4. Store

Storing the cleaned dataframes to CSV files,

i. twitter_archive_master.csv
ii. image_predictions_master.csv
iii. tweet_api_master.csv

5. Analyze & Visualize data

i. Retweets & Favorites

ii. Dog stage statistics

   a. Counts
   b. Favorites & retweets counts
   c. Gender & Favorite counts
   d. Stage & Gender
   e. How users get to twitter
   f. Top 25 popular dog breeds

6. Learnings

i. Pandas int64 doesn't support NaNs, so its better to convert certain columsn to strings. No wonder certain key columns have _str counterpart. Initially thought it was a duplicate column and why would they do it and now i know.