Wrangling Report

UDACITY DATA WRANGLING PROJECT

1. Intro

Data Wrangling is an important process that makes the data ready and clean for analysis and visualization.

Data Wrangling consists of three important steps

1. **Gathering :** Gathering data from different sources as manual and programmtic download, web scraping, API querying or database quering to make the datasets.

2. **Assessment :** Inspecting the datasets to find errors and mistakes.

3. **Cleaning :** fixing the errors and mistakes from the assessment stage

2. Gathering

1. Enhanced twitter archive :

The WeRateDogs Twitter archive that has been manually downloaded from Udacity.

2. Image predictions file

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 following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Twitter API

Each tweet's retweet count, favorite count and followers count at minimum, Using the tweet IDs in the WeRateDogs Twitter archive, the Twitter API is queried for each tweet's JSON data using Python's Tweepy library and each tweet's entire set of JSON data is stored in a file called tweet_json.txt file.

Each tweet's JSON data is written to its own line. Tweet ID, retweet count, favorite count and followers count is written into a pandas DataFrame. Note: Twitter API keys, secrets, and tokens in your project submission is not included.

3. Data Assessment

Data Assessment has two types:

1. Visual Assessment : Open the three datasets using the notebook, MS Excel, or any text editor to detect problems and value mistakes visually

2. Programmatic Assessment : Using Pandas Functions as df.info(), df.nunique(), df.shape() to check issues in the datasets

4. Assessment Summary

Datasets issues has two types:

Quality issues: Content issues including inaccurate, corrupted, duplicated, inconsistent data, etc...

Tidiness issues: Structural issues of the datasets

4.1. Quality issues

4.1.1. archive_df

1. **inconsistency issues:**

- timestamp is object (it should be datetime).
- tweet_id is integer (it should be object).
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id are float (they should be object).
- retweeted_status_timestamp is object (it should be datetime).
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp will not be needed after removing all retweets and replies from dataframe.
- source is object (it should be categorical as it has 4 values only)

2. **Completeness issues:**

- Missing values in **expanded_url**
- a lot of missing values in **doggo, floofer, pupper, poppo, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp**
- some of the missing values are None that should be converted to Nan
- some tweets may have been deleted or unauthorized to get, Those should be deleted (note that number of tweets in api_df is less than archive_df)
- some tweets don't include images, Those should be deleted (note that number of tweets in image_prediction_df is less than archive_df)
- in **name** column, some names were not successfully extracted e.g. (index 37, 74, 151) (visual)
- some tweets are retweets and replies that have to be deleted as per the data wrangling scope mandated by project description (it must be deleted from all tables)
- **expanded_urls** has duplicated values (most of them are retweets by inspection)

3. **accuracy issues**

- a lot of errors in **name** column like 'a' or names less than 3 letters, so it should be extracted from the text or adding Null if not existing
- incorrect and weird values in **rating_numerator** like 0, 1776, 420
- incorrect and weird values in **rating_denominator** like 0, 2, 170
- **source** column should be more descriptive in its values instead of html format (twitter for iphone instead of its current value)
- **dog_stage** column has some errors and duplicated values

4.1.2. image_predictions_df

1. **inconsistency issues**

- **tweet_id** is int (it should be object)
- **img_num** is int (it should be categorical as it will be only 1,2,3 or 4)
- **p1_dog**, **p2_dog**, **p3_dog** have inconsistent captialization

2. **Completeness issues**

- some tweets may have been deleted or unauthorized to get, Those should be deleted (compare the dataframe with API_df)
- **img_url** has duplicated values (most of them are retweets by inspection)
- **p1, p2, p3** should be all combined in **algorithm precision** and all related columns also will be combined in one column
- non descriptive column names

3. **accuracy issues**

- weird values in **p1_conf**, **p2_conf**, **p3_conf** with their maximum values

4.1.3. api_df

1. **inconsistency issues**

- **tweet_id** is int (it should be object)

2. **Completeness issues**

- some tweets don't include images, Those should be deleted (note that number of tweets in image_prediction_df is less than api_df)

3. **accuracy issues**

- weird values in **favorite_count** lower than expected (most of them are retweets by inspection)
- weird values in **retweet_count** lower than expected (most of them are replies by inspection)

4.2. Tidiness issues

4.2.1. archive_df

- values are column names, **doggo, floofer, pupper and poppo** should be all combined in **dog_stage**

4.2.2. api_df

- it should be merged with archive_df as it doesn't have its own observational unit

5. Cleaning

5.1. Missing Data (completeness)

1. Some tweets in archive_df may have been deleted or unauthorized to get
2. Some tweets are retweets and replies that have to be deleted
3. Some tweets don't include images in archive_df and api_df
4. Remove retweets and replies columns after cleaning archive_df.
5. Some of the missing values are None in archive_df

5.2. Tidiness issues

1. values are column names, doggo, floofer, pupper and poppo in archive_df
2. api_df should be merged with archive_df

5.3. Quality issues

1. Incorrect Data types
2. Inconsistent values
3. Inaccurate Values