Data Wrangling Report

by : Nourhan Yousry

Objective:

The goal of this project is to gather enough data from weRateDogs twitter account's archive in order to be able to draw conclusions about this data. This was done in various steps:

1. Gathering data from three different sources
2. Assessing the data
3. Cleaning the data
4. Saving the new cleaned data
5. Extracting information from this new data

Every point will be discussed in detail.

Gathering Data:

The data was gathered from three different sources.

1. Twitter Archive:

This was a file containing weRateDogs enhanced tweets archive with ones that have ratings only giving a total of 2356 tweets. This is the file that contains most of the information where it includes data about the tweet itself such as:

- Tweet id
- Tweet source
- Tweet text
- Timestamp
- Information about the user and tweet it replied to (if any)
- Information about the user and tweet it was retweeted from (if any)

Then there was information about the dog(s) in the tweet such as their ratings, stage(or type as it will be called later in this document) and name.

2. Additional Data via the Twitter API:

This data was extracted using twitter API (tweepy) and the tweet ids available in the twitter archive file. The end result was a text file containing json objects that has information about the tweets. However, some tweets were not found. In order to achieve this project's goal, tweet id, retweet and favorite counts were extracted from the json objects and saved in a dataframe.

3. Image Predictions File:

   Image predictions file was received via requests library from a specific url, it contains the 3 highest predictions ,from a neural network, of the objects in the image in order to know the dog breeds. It also contains three boolean columns for the three predictions that are true if the prediction is a dog breed and false if else.

After extracting information from the three different files, three different pandas dataframes were created for the three files. These data frames were then assessed.

Assessing the data:

Each of the three dataframes was assessed both visually and programatically. Many pandas functions were used to help figure out the data and reach the issues easily. Many quality issues were pointed out as well as tidiness issues.

The quality related points were:

twitter_archive_df

- Some tweets are retweets (not original ratings)
- some tweets are replies (not original ratings)
- some dog names are incorrect (a or the instead of a name)
- timestamp has wrong datatype
- some rating_numerators are too large
- missing records in expanded_urls
- incorrect rating_denominator (0)
- very large rating_denominator
- incorrect urls in the expanded_urls column (gofund me website)
- the same url repeated multiple times and separated by comma in expanded_urls
- incorrect numerator and denominator for Sam (tweet id:810984652412424192)
- incorrect numerator and denominator for tweet: 666287406224695296
- incorrect numerator and denominator for tweet: 682962037429899265
- incorrect numerator and denominator for tweet: 740373189193256964

Image_pred_df

- some rows do not have dog breeds but other animals/ objects instead (invalid data)

While the Tidiness issues where:

Twitter_archive_df

- dog types should be merged into one column

Different dataframes:

- tweet_df and twitter_archive_df should be merged
- image_pred_df and twitter_archive_df should be merged

Most of these points were fixed in the cleaning part.

Cleaning The Data:

At first, some quality points were addressed followed by tidiness points followed by most of the quality points. Here are the points and how they were solved:

Quality Points	How they were cleaned
twiter_archive_df: Some tweets are retweets (not original ratings)	At first, any tweet that had information in the retweet related columns was removed then the retweet related columns were dropped
twiter_archive_df: Some tweets are replies (not original ratings)	At first, any tweet that had information in the reply related columns was removed then the reply related columns were dropped
twiter_archive_df: missing records in expanded_urls	These three issues had the same solution which is to re-insert the expanded url data. Since the expanded url data is the original link to the tweet which consists of twitter address followed by the account name then the tweet id so that was how it was re-constructed

incorrect urls in the expanded_urls column (gofund me website)¶	
the same url repeated multiple times and separated by comma in expanded_urls	
image_pred_df: some rows do not have dog breeds but other animals/ objects instead (invalid data)	The rows in the image prediction dataframe that had non-dogs predictions were removed. Then only the highest prediction of a dog breed was kept and the rest of the data was removed.

After that, the tidiness issue were faced:

Tidiness point	How it was cleaned
Twitter_archive_df :dog types should be merged into one column	Any none found was replaced by null then the four columns were merged together into one column
tweet_df and twitter_archive_df should be merged	The dataframes were merged together on tweet id
image_pred_df and twitter_archive_df should be merged	The new image prediction dataframe was merged with the new twitter archive dataframe

Quality issues that were faced:

Quality	Cleaning process
twitter_archive_df: timestamp has wrong datatype	Timestamp datatype was changed to datetime instead of a string
incorrect numerator and denominator for Sam (tweet id:810984652412424192)	All of these issues were addressed manually by knowing the numerator and denominator from the tweet text then changing them.
correct numerator and denominator for tweet: 666287406224695296	
incorrect numerator and denominator for tweet: 682962037429899265	
incorrect numerator and denominator for tweet: 740373189193256964	

After cleaning the data and merging the dataframes, the data was then saved externally to be used later on for extracting information.