Data Wrangling Report

Project : WeRateDogs

By : Nathapon Tansit

Data Gathering

There are three tables to be gathered from different sources, .csv file, web scraping and API.

First, ‘twitter-archive-enhanced.csv’ was downloaded from Udacity website then stored in a dataframe named ‘df_twitter_archive’ by using pandas.

Second, **requests** library was used to gather data from the assigned url. Requested file was in .tsv format and was later stored in a dataframe named ‘df_image_predictions’

Last, Twitter API. Since I failed to create Twitter developer account, I downloaded tweet_jason.txt then read text file as dataframe named ‘df_tweet_json’

Data Assessing

I used both visual assessment and programmatic assessment to assess the data. For programmatic assessment, I only used .info() and .tail()

I listed 8 quality and 2 tidiness issues and how I discovered these issues.

Issues	How to discover
Quality Issue	
Twitter Archive Enhanced Table	
1. 'tweet_id' must be string not int	df_twitter_archive.info()
2. Dog's name should not be 'a','an','the'	Visual Assessment
3. 'timestamp' must be in datetime format	df_twitter_archive.info()
4. delete retweets	From assigned project condition,
5. delete tweets that do not contain image	From assigned project condition

6. Drop unnecessary column 'in_reply_to_status_id', 'in_reply_to_user_id','source', 'text', 'retweeted_status_id', 'retweeted_status_user_id','retweeted_status_timestamp', 'expanded_urls','doggo', 'floofer', 'pupper', 'puppo' **Image Predictions Table**	df_twitter_archive.info() and Visual assessment
7. Drop unnecessary column 'img_num'	Visual assessment
8. 'tweet_id' must be string not int	df_image_predictions.info()
Tidiness Issue	
1. Dog types in 'df_twitter_archive' (doggo,floofer,pupper,puppo) should be in one column	Visual Assessment
2. 'df_image_predictions' and 'df_tweet_json' should be part of 'df_twitter_archive'	Visual Assessment and .info()

Data Cleaning

Copy Data

First, all dataframes must be copied to keep original data by using .copy() method.

Missing Data

Since missing data is not an issue for this project so I skipped this part.

Tidiness Issue

Tidiness Issue #1: Dog types in 'df_twitter_archive' (doggo,floofer,pupper,puppo) should be in one column. I created a categorical column named 'dog_type' containing the type of dogs. I began with find all possible dog types which are multiple,doggo,floofer,pupper,puppo and none. After that I defined a function to determine the type of dogs then applied to DataFrame.

Tidiness Issue#2: join all dataframe on ‘tweet_id’. I tried joining with .merge method but it turned out that ‘tweet_id’ in some dataframes are not string so I had to conduct this quality issue the back to joining table issue.

Quality Issue#1 and #7: ‘tweet_id’ should be string not int. I converted using .astype(str) to ‘tweet_id’ column in 'df_twitter_archive' and 'df_image_predictions'

Tidiness Issue#2: since ‘tweet_id’ in all dataframe are string. I used .merge to join table on ‘tweet_id’ column by left join. I joined 'df_image_predictions' to 'df_twitter_archive' first then joined ‘df_tweet_json’ later.

Quality Issue

Quality Issue#2: Dog's name should not be 'a','an','the'. I scanned all unique dog’s name and found that there are more strange name for dogs. Some strange names begin with lowercase letters while correct names begin with uppercase letters. So I filtered all strange names using .islower() method and some for loop. After that, I made some for loops to change strange name to ‘None’ if each rows match strange name.

Quality Issue#3: 'timestamp' must be in datetime format. I changed ‘timestamp’ to datetime format using pd.to_datetime()

Quality Issue#4: delete retweets. Retweet rows have data stored in ‘retweet_status_id’ column. I used .drop() method to drop any row that ‘retweet_status_id’ column is not null.

Quality Issue#5: delete tweets that do not contain image. Similar to issue#4. I drop all rows that ‘jpg_url’ column is null.

Quality Issue#6: drop unnecessary column. There are some columns that are not useful for further analysis. I dropped these column using .drop() method

Additional Issue : I found that 'img_num','favorite_count' and 'retweet_count' format should be int not float. I converted format using astype(int).

Data Storing

After cleaning data, I stored master dataset as "twitter_archive_master.csv" using .to_csv() method