Wrangle Report , By Rami Salman

1- data gathering:

1.1 - archive_df data

twitter-archive-enhanced.csv file is given in the classroom and downloaded manually. This file is the archive of WeRateDogs contains data about the account tweets with 2356 row (tweet).

1.2 - image predictions data

image-predictions.tsv given file in a link to download programmatically using python . This file contains predictions of images(dogs) using neural network that can classify breeds of dogs.

1.3 - json data from twitter api

This is the most challenging part of data gathering , this is the first time for me to get data from ready api. It takes about two hours of contacting with twitter to create my developer account then I got the credentials and start collecting data , it takes about 35 minutes of collecting data of tweet id's . Then I saved the collected data in txt file before convert them to csv file to use it , the selected columns/features are the following : 'id','retweet_count','favorite_count','created_at'.

2- data assessing

2.1- archive_df data

2.1.1- structural problems - *Tidiness* issues:

- rating_numerator and rating_denominator can be stored in one column , name it by rating_percent , which generated by the following formula : rating_numerator/rating_denominator.
- doggo, floofer, pupper, and puppo columns can combined in one column ! all of these are categorical data , we can combine them in one categorical column named it by stage.

2.1.2- Quality problems - dirty issues:

- the following columns have a lot of nulls : in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp . At the same time , there is no need for theis columns .
- id's (tweet_id , in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id) type is numeric(int or float) , id should be string better than numeric.
- retweeted_status_timestamp and timestamp are dates , but they stored as objects (string).
- source column storing a very long text with tag containing the source , we can store the source in few words instead (for example : iphone , chrome , android ... etc.)
- doggo, floofer, pupper, and puppo columns have None value rather than null .
- expanded_urls has some missing values.

2.2- image predictions data

2.2.1- structural problems - *Tidiness* issues:

- columns names are not readable (shortcuts are not known!).

2.2.2- Quality problems - dirty issues:

- for image_url column , the first part (https://pbs.twimg.com/) is common for all rows , so we can save the link after this part to reduce the space of saved data.
- tweet_id type is int not string.

2.3- json data from twitter api

2.3.1- structural problems - *Tidiness* issues:

- created_at is already saved in archive_df table.

2.3.1- Quality problems - dirty issues:

- id type is int not string.

3- data cleaning:

I cleaned many of supposed problems in assessing part, as the following.

3.1 structural problems - Tidiness issues:

4 problems are cleaned .

- archive_df data

3.1.1-

- rating_numerator and rating_denominator can be stored in one column , name it by rating_percent , which generated by the following formula : rating_numerator/rating_denominator.
- mathmatically , the numerator should be less than or equal to denominator to calculate the ratio , but in the given data set this rule is'nt applied , see the following article : https://knowyourmeme.com/memes/theyre-good-dogs-brent
- now , I will drop rating_numerator and rating_denominator columns.

3.1.2- doggo, floofer, pupper, and puppo columns can combined in one column ! all of theis are categorical data , we can combine them in one categorical column named it by stage.

- image predictions data

3.1.3- columns names are not readable (shortcuts are not known!).

- json data from twitter api

3.1.4 -created_at is already saved in archive_df table.

3.2 - Quality problems - dirty issues:

8 problems are cleaned as the following

- archive_df data

3.2.1-the following columns have a lot of nulls : in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp . At the same time , there is no need for theis columns .

3.2.2- tweet_id type is int not string

3.2.3- timestamp is date , but stored as objects (string)

3.2.4- source column storing a very long text with tag containing the source , we can store the source in less words instead : Iphone , Vine , Twitter Web Client and TweetDeck .

- image predictions data

3.2.5- for jpg_url column , the first part (https://pbs.twimg.com/) is common for all rows , so we can save the link after this part to reduce the space of saved data.

3.2.6- tweet_id type is int not string.

3.2.7- predictions contains under score (_) , I will remove them

- json data from twitter api

3.2.8- id type is int not string.