UDACITY

Data Wrangling Project

ALX-T Data Analyst Nanodegree Program

Elaborated by : Ferdawes Haouala

Title : Wrangle Report (project 2)

Date : 06/09/2022

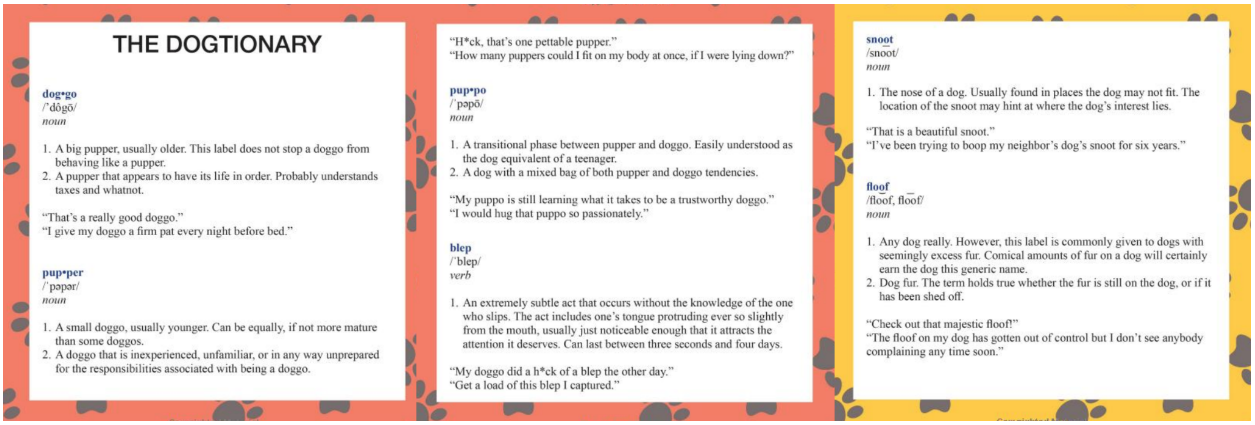
Table of content

1. Introduction : What is the project

- Aim : Gather, assess and clean data to create trustworthy analysis
- How : Using the data of “WeRateDog” (Twitter user) and collecting additional data using various method :

Name of the data set	Content
`image_predictions.TSV`	3 predictions along with their confidence interval and a boolean test
`twitter_archive_enhanced.csv`	Twitter data of the account “WeRateDog”
`tweet_json.txt`	Additional twitter data (retweet and favorite count)

- Additional resources :



THE DOGTIONARY

dog•go
/ˈdôgō/
noun

1. A big pupper, usually older. This label does not stop a doggo from behaving like a pupper.
2. A pupper that appears to have its life in order. Probably understands taxes and whatnot.

“That’s a really good doggo.”
“I give my doggo a firm pat every night before bed.”

pup•per
/ˈpəpər/
noun

1. A small doggo, usually younger. Can be equally, if not more mature than some doggos.
2. A doggo that is inexperienced, unfamiliar, or in any way unprepared for the responsibilities associated with being a doggo.

“H*ck, that’s one pettable pupper.”
“How many puppers could I fit on my body at once, if I were lying down?”

pup•po
/ˈpəpō/
noun

1. A transitional phase between pupper and doggo. Easily understood as the dog equivalent of a teenager.
2. A dog with a mixed bag of both pupper and doggo tendencies.

“My puppo is still learning what it takes to be a trustworthy doggo.”
“I would hug that puppo so passionately.”

blep
/ˈblep/
verb

1. An extremely subtle act that occurs without the knowledge of the one who slips. The act includes one’s tongue protruding ever so slightly from the mouth, usually just noticeable enough that it attracts the attention it deserves. Can last between three seconds and four days.

“My doggo did a h*ck of a blep the other day.”
“Get a load of this blep I captured.”

snoot
/snoot/
noun

1. The nose of a dog. Usually found in places the dog may not fit. The location of the snoot may hint at where the dog’s interest lies.

“That is a beautiful snoot.”
“I’ve been trying to boop my neighbor’s dog’s snoot for six years.”

floof
/floof, floof/
noun

1. Any dog really. However, this label is commonly given to dogs with seemingly excess fur. Comical amounts of fur on a dog will certainly earn the dog this generic name.
2. Dog fur. The term holds true whether the fur is still on the dog, or if it has been shed off.

“Check out that majestic floof!”
“The floof on my dog has gotten out of control but I don’t see anybody complaining any time soon.”

- Names of tables:

Original name of the file	Name of the table in jupyter notebook	Copy of the data set for cleaning
`image_predictions.TSV`	prediction	clean_prediction
`twitter_archive_enhanced.csv`	ratings	clean_ratings

`tweet_json.txt`	twitter_data	clean_twitter_data

New names of tables To solve tidiness issues :
Clean_twitter_data+clean_rating = twitter_data_dog table
clean_prediction+clean_rating = dog_table

2. Methodology & steps

2.1. Gathering data:

Steps :

Name of the data set	Method used to gather data
`image_predictions.TSV`	downloaded programmatically using the Requests library through a link provided
`twitter_archive_enhanced.csv`	Read it using pandas read method
`tweet_json.txt`	Using Twitter API I saved the data in a TXT file then I open it using JSON method. Then, I dropped columns that are not demanded.

For API there are hidden steps like applying for Twitter API to have elevated access.

2.2. Assessing data:

From **visual and programmatic assessment** these issues where found :

➢ **Tidiness issues:**

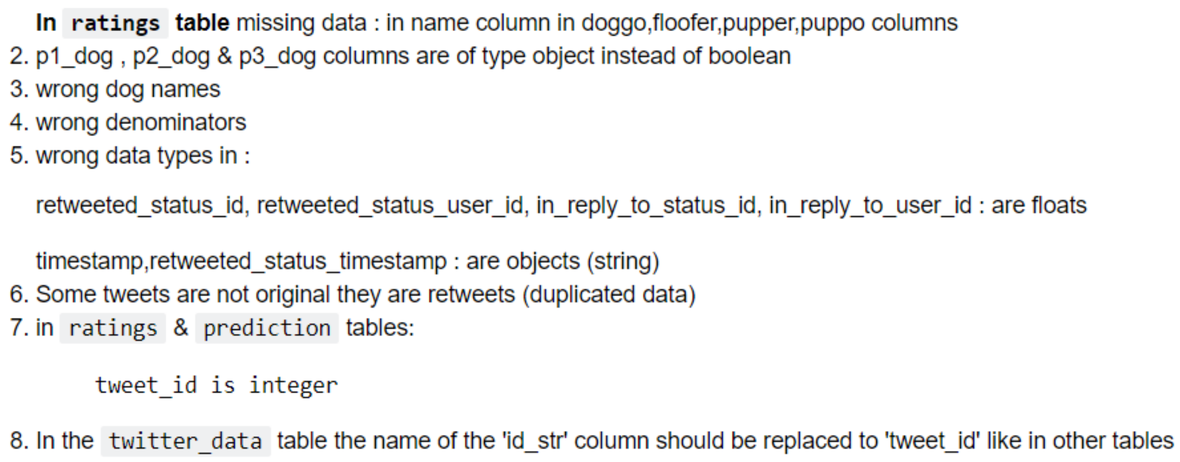
1. doggo floffer pupper poppo columns should be all in one column called race_dog

2. In time stamps we have 2 variable date and time : should be seperated

3. Add (race_dog, name) columns from ratings table to pediction table which will be dog table
4. Add ('retweet_count', 'favorite_count'and 'retweeted') columns from clean_twitter_data table to ratings table which will be twitter_data_dog table

➢ **Quality issues:**

1. Missing data in twitter_data table (but I can not solve this issue)



In `ratings` table missing data : in name column in doggo,floofer,pupper,puppo columns

2. p1_dog , p2_dog & p3_dog columns are of type object instead of boolean
3. wrong dog names
4. wrong denominators
5. wrong data types in :

   retweeted_status_id, retweeted_status_user_id, in_reply_to_status_id, in_reply_to_user_id : are floats

   timestamp,retweeted_status_timestamp : are objects (string)
6. Some tweets are not original they are retweets (duplicated data)
7. in `ratings` & `prediction` tables:

   ```
   tweet_id is integer
   ```
8. In the `twitter_data` table the name of the 'id_str' column should be replaced to 'tweet_id' like in other tables

2.3. Cleaning data:

I used define, code and test methodology to structure the cleaning steps.

Issues	Method used in cleaning	Explanation
Deleting retweets	.isna()	- Leaving the rows that have null values in “retweeted_status_id” - Deleting columns relevant to retweets in ratings table
Create a column for the race of the dog	Concatenate (+) .replace()	- Concatenating the values of doggo, floofer, pupper and puppo column in a new list. - Replacing the values with the correct names.. - Adding that list as a column in the ratings table.
Creating date and time columns from timestamp column	.str.extract()	- Extracting the date and time using a regular expression - Creating the date and time columns with the values

		extracted.

Changing data types	.astype()
	.fillna()
		In case there are nan values

Renaming id_str column to tweet_id	.rename()	

Create dog_table and twitter_data_dog table	.merge()
	.drop()
		Merginfg dg
		Dropping irrelevent columns

Deleting tweet not having images	.notna()	Leaving only rows having values in img_url column
Replacing wrong name/denominator values	.str.replace() .loc[]	

2.4. Storing data :

Using .to_csv() method

Name of data frame	Name of the CSV file
twitter_data_dog	twitter_archive_master
dog_table	dog_infos_from_twitter

2.5. Analyzing and visualizing:

The results will be presented in “act_report”

2.6. Deriving insights :

The results will be presented in “act_report”

3. Challenges and limitations

I couldn’t find the missing values for:

- expanded_urls column
- name column

- doggo, floofer, pupper, puppo columns
- twitter_data and prediction tables in comparaison to ratings table