wrangle_report

April 21, 2020

0.1 Data Wrangling Report

0.1.1 Gathering Data

Three data sets are used in this project, and each of them are gathered with different methods. First, twitter-archive-enhanced.csv is downloaded with the given link in the project details. The tweet image predictions, image_predictions.tsv, is downloaded programmatically using the requests library. And the last data set is using Python's Tweepy library to query the Twitter API for each tweet's JSON data and store the JSON data in a file.

0.1.2 Assessing Data

First, I try to find problems from each data set visually. Then, I used programmatically functions to find more issues.

And I find out several issues in each data set: ##### df_predict table #### Quality 1. Prediction sometimes are lowercase, sometimes are uppercase 2. Some predictions are not dogs 3. Column tweet_id should be str type instead of int 4. There are retweets or duplicate tweets

Tidiness

1. Three predictions are too much while analyzing

df_archive table

Quality

1. Some dogs are not in any of the doggo, floofer, pupper or puppo category
2. Some names are not correct
3. Column tweet_id should be str type
4. Column timestamp and retweeted_status_timestamp should be datetime type
5. There are some retweets or duplicated tweets and replies

Tidiness

1. Columns for doggo, floofer, pupper, puppo are hard to analyze

df_api table

Quality

1. tweet_id should be str instead of int

0.1.3 Data Cleaning

1. Merge three data sets into one, and make a copy with it.
2. Drop the retweets and replies by filtering the NaN of each status id.
3. Drop the columns that are not going to use and tweets without image or not about dogs.
4. Melt doggo, floofer, pupper and puppo columns into one column
5. Using function to find the highest one from three predictions
6. Replace the wrong names with NaN values
7. Change all the column to correct data type