UDACITY

Wrangle Report

Ali Sammour

DAND - C.4

Project Overview

In the 5th project, I will wrangle WeRateDogs **Twitter** data to create interesting and trustworthy analyses and visualizations. this project took a lot of time and effort in gathering, assessing, and cleaning to analyses and visualizations.

Project Details

- Data wrangling, which consists of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing

Gathering Data

In this step, I treat 3 heterogeneous sources

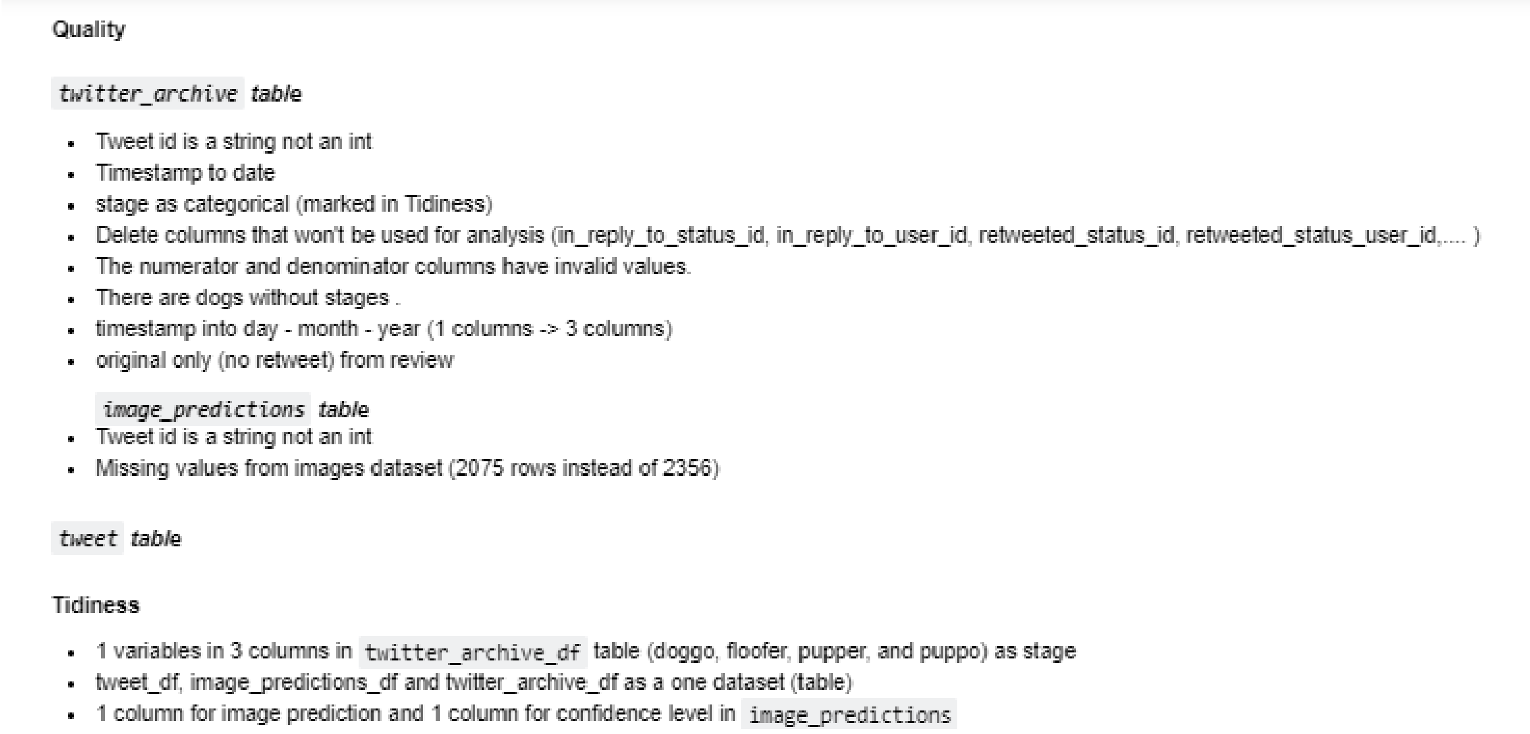
- twitter_archive_enhanced.csv
- image predictions (downloading programmatically)
- tweet-json.txt

In the last step here I didn't use Twitter API for 2 main reasons :

- I faced a problem when creating an email (no reply)
- This process consumes a lot of time

Assessing Data

After gathering each of the above pieces of data, I was assessing them visually and programmatically for quality and tidiness issues. I was Detected and documented it



Quality

`twitter_archive` *table*

- Tweet id is a string not an int
- Timestamp to date
- stage as categorical (marked in Tidiness)
- Delete columns that won't be used for analysis (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,....)
- The numerator and denominator columns have invalid values.
- There are dogs without stages .
- timestamp into day - month - year (1 columns -> 3 columns)
- original only (no retweet) from review

`image_predictions` *table*

- Tweet id is a string not an int
- Missing values from images dataset (2075 rows instead of 2356)

`tweet` *table*

Tidiness

- 1 variables in 3 columns in `twitter_archive_df` table (doggo, floofer, pupper, and puppo) as stage
- tweet_df, image_predictions_df and twitter_archive_df as a one dataset (table)
- 1 column for image prediction and 1 column for confidence level in `image_predictions`

Cleaning Data

I was cleaned each of the issues I documented while assessing. and I performed this cleaning in *wrangle_act.ipynb* as asked to me. The results are high quality and tidy master pandas DataFrame .

First issue I solved: Delete all retweet and keep the original

Second issue I solved: Convert tweet_Id in twitter_archive_df and image_predictions into string using astype , and timestamp into datetime using pd.datetime.

Third issue I solved: Merge tweet_df, image_predictions_df and twitter_archive_df because no need to 3 tables all info are related.

Fourth issue I solved: Create 1 variables in 3 columns in 'twitter_archive_df' table (doggo, floofer, pupper, and puppo) as stage (tidiness issue) then convert it to category datatype, after all this steps convert "none" stage to np.nan .

Fifth issue I solved: Convert time stamp into 3 coulmns Day , Month , Years to make easy visualization and analytics using ***dt*** functions.

Sixth issue I solved: Correct The numerator and denominator columns have invalid values because the numerator and denominator don't make sense.
This issue solution from [here]

Seventh issue I solved: Create 1 column for image prediction and 1 column for confidence level to visualization and analytics purposes.

note: analyzing, and visualizing details in act_report