Data wrangling WeRateDogs project

Introduction

In this project, I completed the gathering and cleaning process of three data frames. I had to gather the data fro m different sources. There were various quality and tidiness issues that I tried to fix.

Gather

- Gathering Data for this Project composed from three different sources. The WeRateDogs Twitter archive . I download this dataframe from the Udacity server.
- The tweet image predictions. This file (image_predictions.tsv) hosted on Udacity's servers and should b e downloaded programmatically using the Requests library.
- Finally, by using the Twitters API, I made a new data frame, with many details for every tweet ID from t he WeRateDogs Twitter archive.

The most challenging part was the Twitter API because I had never used it before. I had some problem, not to d ownload the data, but to store them in a data frame. Finally, I solve this problem.

Accessing Data

In this step I accessed data both visually and pragmatically. There are three data frames and it is not impossible to take a look at all the data visually. At the end, of course, the programmatically prospect give us a very deep lo ok at our data.

Cleaning

Issues list

Quality.

1.The tweets with a number >0 in_reply_to_status_id, in_reply_to_user_id columns should be deleted because we need only original tweets. Not retweets and replies

2.retweeted_status_timestamp, timestamp should be datetime instead of object (string)

3.We only want tweets that have images so I delete all the tweets without.

4.Missing values from images dataset (2075 rows instead of 2356)

5.Some tweet_ids have the same jpg_url

6.The 'name' column has many invalid values like, a, an, the.

7.We should make the 'source' column easier to read.

8.We need to rename some columns to be more user-friendly

Tidiness

1.We should melt the columns doggo,floofer,pupper,puppo in one columns named Dog_styles

2.We dont need all those prediction columns. I will keep the first True prediction.

3.All tables should be part of one dataset

In this step I had to fix on problem at a time, by following the define-code-test progress. The most challenging iss ue was to merge the dog stage columns into one because there were several tweets with two doge stages. I had to find them and change their stage to ‘multiple’

Analysis

After I completed the cleaning process, I continue with some analysis. I made three plots, showing that:

1. There is a strong correlation between favorites and retweets. More retweets mean more favorites.
2. The doggo and puppo stages collect more ‘favorites’ from the audience.
3. The most popular breed is the Chow and the most unpopular is Chihuahu

Conclusion

It was a very interesting project. It was very helpful and challenging for me. Through all the wrangling process a became more familiar with many important commands, that I will use for sure at the future. I understand there th ere are several ways to solve a problem and that I had to be very careful and patient in order to find as much as possible issues the data have.