Introduction

The objective of this project is to put in practice what I learned in data wrangling section from Udacity Data Analysis Nanodegree program. I use WeRateDogs dataset from user @dog_rates archive in tweeter.

WeRateDogs is a popular Twitter account, and it has a 9 million flowers, tweeters rate dogs to show how lovely the dog is.

In this report I will describe my wrangling efforts

Project Steps Overview

Gathering data

Assessing data

Cleaning data

Gathering data

In this project the data was gathered from three sources: Enhanced Twitter archive: This data with file name twitter_archive_enhanced.csv is from udacity and I downloaded it manually.

- Image prediction: I get the image prediction data from udacity server, and I downloaded it programmatically.
- Twitter API: I tried to contact with twitter to get a twitter developer account to get access to @dog_rates account but they did not give a permission, so I used the given data from Udacity.
- I queried a twitter API and save it in a tweeter json.text file.

Assessing Data

This part has a tow issues Quality issues and Tidiness issues.

Quality issues.

Tweeter_Archive_df

- Invalid timestamp data type (must be datetime).
- There are invalid dog names like None, a, an and the.
- There are unnecessary columns must be deleted.
- Wrong use of doggo, floofer, pupper and puppo columns (must be in one column).
- There are 181 retweeted by retweeted_status_id and retweeted_status_user_id .

Image_Predictions

- Some names in P start with an uppercase letter and some of them start with lowercase.
- There are 66 row jpg_url duplicated
- There are unnecessary columns must be deleted.

Tweet_Jason

- Invalid timestamp data type (must be datetime).

Tidiness issues.

- The 3 tables are related but they appear individually
- The doggo, floofer, pupper and puppo columns are appear individually

Cleaning data

I used many pandas methods to clean the quality and tidiness issues from the three dataset then I merge them in one dataset called twitter_archive_master.csv.

Conclusion

By the end of this project, I got a good practice of what I learned during this Data Analysis Nano Degree course. I putted all my effort to gather, assess and cleaned data to get best visualization of this data. Now I can handle the unstructured data to make a well-structured wrangling project.