Wrangle Report

Introduction

In this project, I must implement what I learned in part data wrangling from the Udacity Data Analysis program. The project is about to gather data from WeRateDogs which is a Twitter account that rates dogs, including dog's name, stage, and a number of retweets, etc. in this project I had to go through gather steps, assessing, cleaning, and analysis.

Data Gathering

In this project I had to gather data from 3 sources:

1. Manually, from Udacity servers which were WeRateDogs twitter archive csv.
2. Programmatically, which was The Image prediction files.
3. Twitter API but unfortunately, my application was not approved. So, I collected the data from Udacity's project resources where was including two files
    I. Twitter_api.py: the code for Twitter API, its included in my project code as commented to not interrupt my kernel when running.
    II. Tweet-json.zip: it is a zip folder that contains a text file that should be storing the result of API Twitter data when running the code.

Assessing Data

In this section, I had to assessing the three dataframes and identifying several quality and tidiness issues. And to assess them there is two way programmatically or visually. The notes were:

Quality:

- tweet_id column is int datatype it should be object datatype in twitter_archive
- The timestamp column has dates in string form in twitter_archive
- Delete columns that will not be used for analysis in twitter_archive
- Convert 'None' values with 'NaN' for all dog stages intwitter_archive
- Rename column name to dog_name in twitter_archive
- Wrong names or it could be missing names in name column in twitter_archive
- Some dogs have multiple stages in twitter_archive
- Convert capital letters in p1, p2, p3 to lowercase in image_predictions
- retweets, favorites columns in df_tweet are object datatypes they should be integers

Tidiness:

- The values doggo, floofer, pupper, puppo in twitter_archive should be in one category datatype.
- There are 3 different dataframes better to store it in one dataframe

Cleaning Data

The final step in the wrangling process is cleaning the data, as I learned from Udacity is to divide the cleaning into three parts: Define, Code, and test. And implement it to each issue.