Introduction

This project was part of the Udacity Data Analysis Nanodegree program, the purpose of it is put in practice what I learned in data wrangling data section. The data from the WeRateDogs twitter account rates dogs with humorous commentary.

Project Details

The tasks in this project are as follows:

- Data wrangling, which consists of:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing your wrangled data.
- Reporting on 1) data wrangling efforts and 2) data analyses and visualization.

Gathering Data for this Project

Data was gathered from three different pieces:

1- The twitter archive enhanced csv file was downloaded manually that provided by Udacity.
2- The image predictions tsv file is hosted on Udacity servers which I downloaded programmatically using the Requests library.
3- The Twitter API & JSON by using the tweet IDs to query the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt. I read this txt file line by line into a pandas DataFrame only including the desired variables; retweet count and favorite count.

Assessing Data for this Project

After the data was gathered, I assessed the data as the methods (e.g. head, info, sample, value counts, duplicated).

Then I identify the issues encountered like 8 quality issues and 2 tidiness issues.

Cleaning Data for this Project

In this task, firest was create a copy of a three dataframes as a df clean. I divided the performed in three stage:

1- Define: definition the issues that identify in assess task to be clean.
2- Code: do the cleaning code.
3- Test: to testing the code.

Storing Data for this Project

I was stored the clean DataFrame(s) archive_df_clean in a CSV named twitter_ archive_master.csv.