WRANGLING REPORT

Introduction

In this project, we wrangled the tweet archive data of Twitter user @dog_rates, also known as WeRateDogs. The main purpose of the data wrangling project is gathering data from different resources, assessing the data to find some problems and cleaning what we found already in the assessment part.

1. Gathering Data

The gathering process includes three parts for this project.

1) The WeRateDogs Twitter archive file: twitter_archive_enhanced.csv

2) The tweeted image prediction file: This file will be downloaded using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3) Twitter API and JSON file: Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

2. Assessing Data

After gathering data from different sources, I assessed them visually and programmatically for quality and tidiness issues.

- Visual Assessment

- I used Jupyter Notebook and Google Sheets to see all data.

- Programmatically Assessment

- I used some functions from the Pandas library.
  - info,
  - value_counts,
  - describe,
  - sample,
  - duplicated, etc.

Using visual and programmatic assessment, I found out some quality and tidiness issues in my dataset and took some notes to keep them for the cleaning part. These notes are in the wrangle_act.ipynb.

3. Cleaning Data

This part includes three different parts for each quality and tidiness issues: **Define**, **Code** and **Test**. Before the cleaning process, the copies of all three data frames were created. Afterward, the steps of cleaning parts were applied iteratively for all quality and tidiness issues.

Storing Data

After finish the cleaning process, the clean data was stored in the twitter_archive_master.csv file.