Introduction

Data wrangling is a core skill that everyone who works with data should be familiar with since so much of the world's data isn't clean. Data wrangling has three main steps Gathering the data then Assessing the data and finally Cleaning the data
In this project we have a dataset that we have wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. And briefly i will try to explain how we wrangled this dataset

Software used

Jupyter Notebook with The following packages (libraries)(pandas, NumPy, requests, tweepy, json) and A text editor (Atom)

1-Gathering the data

First, we had the twitter_archive_enhanced.csv file which we downloaded manually, then we downloaded image_predictions.tsv file which we downloaded programmatically using Requests library in python and for our third data to be gathered is each tweet's, retweet count and favorite ("like") count at minimum. 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.

2-Assessing Data

Assess data for: Quality: issues with content. Low quality data is also known as dirty data. Tidiness: issues with structure that prevent easy analysis. Untidy data is also known as messy data. Tidy data requirements: Each variable forms a column, each observation forms a row, each type of observational unit forms a table.... using two types of assessment: Visual assessment: scrolling through the data in your preferred software application (Google Sheets, Excel, a text editor, etc.). Programmatic assessment: using code to view specific portions and summaries of the data (pandas' head, tail, and info methods, for example).

so after gathering each of the above pieces of data, we started to assess them visually and programmatically for quality and tidiness issues.so we detected and documented eight (8) quality issues and two (2) tidiness issues

3-Cleaning Data

There are two types of cleaning:

Manual (not recommended unless the issues are one-off occurrences) Programmatic The programmatic data cleaning process:

Define: convert our assessments into defined cleaning tasks. These definitions also serve as an instruction list so others (or yourself in the future) can look at your work and reproduce it, Code: convert those definitions to code and run that code, Test: test your dataset, visually or with code, to make sure your cleaning operations worked.

So here in our project we first made a copy of our datasets then we started to define our completeness issues first which is part of quality issues and solve it then we went to our tidiness issues then we continued in the rest of the quality issues to keep our data tidy and clean so we can analyze it after this.
Then we stored the clean DataFrames in CSV files
After this we made some visualization and analysis to the data