1. WRANGLING REPORT

The data wrangling project was on wrangling data from @WeRateDogs Twitter Account which was obtained from three sources. I basically wrangle the WeRateDogs data as instructed in the project details, the image predictions data from udacity servers and the data within tweet-json.txt which contains additional tweets from WeRateDogs twitter data such as retweet count and favorite count and other useful information.

DATA GATHERING

The first step I started off with is importing packages I would need in the project such as numpy, pandas, requests, os, json, re, BeautifulSoup and some others. The first source is the Twitter Archive data. The Twitter Archive data was downloaded manually from the udacity project page and imported to the workspace. The data is in csv format so I load it into a pandas dataframe straight up using the read_csv() function.

The second source is an Image Predictions data provided in a URL which I downloaded programmatically using the Python Requests library. I started off by using requests to request for the file using the URL provided. I then write the contents to a file I opened in 'wb' (write binary) mode and then save the file to my project root folder. I then read the file I just saved into a pandas dataframe.

Due to some missing variables in the Twitter Archive data, I was instructed to fetch additional data from WeRateDogs twitter account using tweepy API as the last source of the data to be used in the wrangling project. I applied for a Twitter Developer account, but due to one reason or the other I can't get the Developer Account so I had no choice but to use the alternative method provided on the Project Details page in the form of a text file containing the data I am to fetch.

I opened the file from within my code and then used a for loop to loop through all the lines and started extracting data I needed then putting in dataframe.

DATA ASSESSMENT

I assessed the data looking out for data quality & tidiness issues. I started with visual assessment using Excel Viewer(VS Code Extension). The use of Excel Viewer is good as I scrolled through the pandas Dataframe without having to leave VS Code while checking visually for any issues which I documented. I programmatically assess the data frames using different pandas methods. I did this to inspect for invalid data such as the name of certain dogs being words and not actual names. I also noticed type issues with some columns, handling missing data points and tidiness issues in dog stages are represented in

4 columns violating the rule 'Each variable is a column'. Also, the retweets count, and favorite count were on a different table than the Twitter Archive hence violating the third rule of tidy data that says 'each Type of observational unit forms a table'.

CLEANING DATA

I started addressing each of the issues I documented in the assessment stage taking my three data from an unstructured, untidy, and dirty format to a structured and clean data. The first step I did was address invalid data types issues. Based on the information given on the project details, I dropped certain records where tweets were retweets. I removed, invalid data, combined dog stages into one column. Then I merged the three data frames into one after dropping all columns with enormous nulls and unnecessary ones that contained data that I did not need.

STORING DATA

I stored the merged dataset with the name twitter-archive-master.csv for future references.