Data Wrangling Report

Project Objectives

- Perform data wrangling (gathering, accessing, and cleaning) on provided datasets
- Analyze the cleaned up data
- Produce insights based on the analyzation

Gather

There are three files to gather for this project:

- WeRateDogs Twitter archive, ‘twitter-archive-enhanced.csv’, manually downloaded from the course website
- Tweet image predictions, ‘image-prediction.tsv’, downloaded programmatically using Requests library
- Each tweet’s entire JSON data, ‘tweet_json.txt’, downloaded with Tweepy

Access and Clean

After closely examining the data, we discovered a few issues that need to be fixed before the data can be used meaningfully. Please refer to the table below for observations and fixes.

Quality

Dataframe	Observation	Fix
tweets_clean	Erroneous datatypes (tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweet_status_timestamp columns)	Change tweet_id to string, change timestamp to datetime, and drop the other columns
	Missing information on in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweet_status_timestamp	Resolved due to fixes of other issues
	rating_numerator and rating_denominators are not matching the rating given in text	Extract numbers from text and rewrite the rating_numerator and rating_denominator

	Rating_numerators having a very large range, some are not actual ratings (e.g. 1776 is not a rating for a dog, but the year of the Declaration of Independence of the U.S.)	Inspect rating_numerators and decide if deletion is needed
	745 dogs are named “None”, some are named “a”	Change dog names from “None”/”a” to empty string
	rating_numerator and rating_denominators are wrong for tweet_id 666287406224695296, should be 9/10 instead of 1/2	Change rating to 9/10
	rating_numerator is wrong for tweet_id 883482846933004288, should be 13.5 instead of 5	Fixed along with other issues
	rating_numerator and rating_denominators are wrong for tweet_id 810984652412424192, 24/7 is not a rating. There is no rating for this dog	Remove row with tweet_id 810984652412424192
	Some tweets are not original tweets	Drop all retweets
	Some tweets contain videos instead of images	Remove all tweets without an expanded_urls or expanded_urls contains the word “twitter”
images_clean	Erroneous datatypes (tweet_id)	Change tweet_id to string
	Mix use of upper and lower cases first letters of prediction (p1, p2, p3)	Change every string in p1, p2 and p3 to lower case
likes_clean	Number of entries does not match number of entries of df_tweets - there are some deleted tweets	Resolved due to fixes of other issues
	Erroneous datatypes (retweet_count, favorite_count)	Change retweet_count and favorite_count to int

Tidiness

Dataframe	Observation	Fix

tweets_clean	Four columns of dog stages	Merge doggo, floofer, pupper and puppo columns to a dog_stages column
	Retweet counts and favorite counts should be part of the df_tweets table	Merge likes_clean and tweets_clean
	Ratings are not standardized due to different denominators	Add a rating column with standardized rating
likes_clean	retweet_count and favorite_count are similar ways to calculate how popular a dog is	Make an column named likes as the sum of the retweet_count and favorite_count
all tables	df_tweets, df_images, and df_likes should be part of the same table	Merge tweets_clean, image_clean and likes_clean

Result

After accessing and cleaning the three datasets, we have one cleaned and ready to be analyzed dataframe.

```
twitter_archive_master.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1966 entries, 0 to 1994
Data columns (total 24 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   tweet_id            1966 non-null   object
 1   timestamp           1966 non-null   datetime64[ns, UTC]
 2   source              1966 non-null   object
 3   text                1966 non-null   object
 4   expanded_urls       1966 non-null   object
 5   rating_numerator    1966 non-null   float64
 6   rating_denominator  1966 non-null   float64
 7   name                1966 non-null   object
 8   dog_stages          1966 non-null   category
 9   retweet_count       1966 non-null   int32
 10  favorite_count      1966 non-null   int32
 11  likes               1966 non-null   int32
 12  rating              1966 non-null   float64
 13  jpg_url             1966 non-null   object
 14  img_num             1966 non-null   int32
 15  p1                  1966 non-null   object
 16  p1_conf             1966 non-null   float64
 17  p1_dog              1966 non-null   bool
 18  p2                  1966 non-null   object
 19  p2_conf             1966 non-null   float64
 20  p2_dog              1966 non-null   bool
 21  p3                  1966 non-null   object
 22  p3_conf             1966 non-null   float64
 23  p3_dog              1966 non-null   bool
dtypes: bool(3), category(1), datetime64[ns, UTC](1), float64(6), int3
2(4), object(9)
memory usage: 299.9+ KB
```