Data Gathering

I gathered data from 3 different sources:

- The twitter-archive-enhnaced.csv
- The image-predictions.tsv
- The Tweet-json.txt

I imported Pandas, Numpy, Matplotlib, Seaborn, Requests Tweepy, OAuthHandler and json. I read in the twitter-archive-enhanced.csv file provided by Udacity which I had previously downloaded and renamed it df_1.
Next I used the request library to download the tweet image predictions before reading it into the dataframe as well. And renamed it df_2.
I used Tweepy to query additional data through Twitter's API and since I got the Elevated access from twitter, I used my Consumer key and secret, and Consumer access key and secret to gain access to the twitter data. Next I needed the Txt document to be in a list format so I converted it using df.list.append (json.loads(line)). I set them line by line into a pandas dataframe with twet_id, favorite count and retweet count and renamed it df_3.

Assessing Data

To assess the data, I made use of both programmatic and visual assessment. I used some programmatic assessments like:
df.head(),
df.info(),
df.nunique(),

df.sample(),
df.duplicated().sum(),
df.describe()
and for the df_1 I also used df_1.name because I wanted to further check the few errors that I found in that column.
Then finally I used :
all_columns = pd.Series(list(df_1)+ list(df_2)+ list(df_3))
all_columns[all_columns.duplicated()]
To check for the recurring column in the three datasets and I found that to be ‘tweet_id’.

Quality Issues

These three datasets were not just messy, they were also untidy. I found 8 quality issues and 2 tidiness issues from the three datasets, though the data is not completely cleaned, I tried my best to make it look a bit presentable.

The 8 quality issues:

1. Dropping all retweeted columns.
2. Tweet_id is a string not an int for the three datasets.
3. Timestamp is a datetime not an object.
4. Replace None in the name column with Nan.
5. Drop columns that are irrelevant to the analysis - source and expanded urls.
6. Drop incomplete columns - in_reply_to_status_id and in_reply_to_user_id.
7. Changing p1, p2, and p3 name columns to lowercase
8. Changing ‘id’ to tweet_id.

Tidiness Issues

1. Merge floofer, pupper, doggo and puppo into one single column(dog_stage).
2. The three datasets should be merged on tweet_id to form one single table.

Cleaning the Datasets -

I made a copy of the three datasets and made use of Define, Code, Test during cleaning.

Analyzing and Visualizing Data

I used a bar chart to check the value counts of the dog stage and it helped me find the one with the highest and lowest number in the dataset.
I also made use of the scatter plot to see the relationship between the favorite count and the retweet count.

Insights:

- Pupper has the largest number in the dataset.
- Floofer has the least number in the dataset.
- There is a positive relationship between favorite count and retweet count.