Gathering:

1. Imported the `'twitter-archive-enhanced.csv'` file into a dataframe using pandas.
2. Used requests library to get the content of `'image_predictions.tsv'` then wrote that content to a tsv file and imported it into a dataframe using pandas.
3. Downloaded the `'tweet_json.txt'` file and appended it line by line to a list and then made a dataframe of that list using pandas.
4. Merged all dataframes into one dataframe called `'master_df'`.

Assessing:

- Checked for missing values and wrong data types.
- Checked for suspicions numerical values.
- Visually assessed the first 5 and last 5 rows of the dataframe.
- Checked for identical rows.
- Checked for rows with problematic data types.
- Checked for retweets.
- Checked for duplicated rows.
- Checked for posts that didn't contain dog images and determining which prediction is the most accurate.

Issues Found:

Quality:

1. values have a or none instad of NaN
1. there is alot of null values in a lot of columns
2. `timestamp` and `created_at` columns are object type
3. `text` and `full_text` columns are the same
4. `source_x` and `source_y` columns are identical
5. `possibly_sensitive` and `possibly_sensitive_appealable` are object type
6. `display_text_range` contains lists
7. `created_at` and timestamp are the same(nearly)
8. `entities` column has alot of empty lists
9. `doggo,floofer,pupper and puppo` columns are useless
10. `name` column has a lot of empty values
11. Some image predictions are not dogs

Tidiness:

1. created_at has more than one value
2. extended_entities and user have more than one variable

Cleaning:

1. Made copy of the original dataframe.

2. Changed values that are equal to a or none to NaN. (Issue #1(quality))
3. Dropped columns with too much NaN values and entities column and `doggo,floofer,pupper` and `puppo` columns. (Issue #2(quality) and issue #9(quality) and issue #10(quality))

4. Made the `created_`at column only contain weekday and be categorical and rename it to `weekday`.and chnage type of `timestamp` column to object. (Issue #3(quality), Issue #8(quality) and issue #1(tidiness))
5. Dropped `text` column and leave `full_text` and drop `source_x` and and `leave source_y`. (Issue #4(quality) and issue #5(quality))
6. Changed `possibly_sensitive` and `possibly_sensitive_appealable` type from object to bool. (Issue #6(quality))
7. Changed type of values in `display_text_range` column to tuple to change all lists in it to tuples. (Issue #7(quality))
8. Splitted `extended_entities` and `user` columns into different columns and drop any columns that can't be used. (Issue #2(tidiness))
9. Imputed the `name` column. (Issue #11(quality))
10. dropped the rows where `p1_dog` column is false.(Issue #12(quality))
11. Tested after fixing each issue.