Quality issues:

1. tweet_masterdata:

- Erroneous datatypes (timestamp, retweeted_status_timestamp).
- The names of dogs are not always available, None, a, also consistency of names should be either upper case or lower case.
- Data type for tweet_id is int64, but should be str.
- Unreasonable rating_ numerator is noted, although it's probably twitter humor, but there are extreme rating, e.g., 1776, 960, 666, 420. Moreover, it's noted that some ratings are not aligned with twitter texts.
- Several columns should be deleted, for example, retweet related columns are probably unnecessary as we only want original ratings with images.

2. image_prediction:

- Names of columns (variables) are not very clear on the content. For example, it's not clear what stands for p1, p2. So "rename" of these columns are needed.
- The data formats at p1, p2, and p3 are not very consistent in terms of lower case or capital letter. The consistency of format is needed to identify potential duplicates.
- 66 duplicated image urls are identified and the reason is possibly due to retweet. Thus these duplication should be excluded from the analysis.
- Data type for tweet_id is int64, but should be str.
- We will only use columns of p1 and p1_dog, as information with very low confidence at p2 (13%) and p3 (6%) are unable to provide convincing conclusions.

3. tweet_overview:

- Noted 836 id and id_str are not the same, for example: id (892420643555336193) is id_str (892420643555336192). So we should focus on variable **id**, and exclude **id_str** from analysis.
- Erroneous data type for id is int64, but should be str.
- Several columns should be deleted or extracted to make this table more clear with useful info. For example, geo, contributors, coordinates, place has no or only 1 data on it and should be removed. Moreover, retweet related columns are probably unnecessary as we only want original ratings with images.

Tidiness issues:

1. tweet_masterdata:

- For the types of dogs should be transferred into 1 column, instead of 4 columns (doggo, floofer, pupper and puppo).

2. all:

- To help us analysis of all the useful data, we merge all the three files into a complete file based on common column called tweet_id. However, the merging of three files should be based on the tweet_id at tweet_masterdata file as it has the most complete tweet_id.