Data Wrangling Project

In the Project, I have analysed the tweets about dog rating. For that I had 3 different data from 3 different sources. These data sources consist of tweet details and dog pictures with text and rating. For gathering I used traditional csv file upload, web source upload and twitter API.

After gathering process, I tried to access to the data. I used both programmatic and visual assessment. However, visual assessment needs detailed attention, it can be useful sometimes. But programmatic assessment is easy to work with big data. I diagnosed 8 quality and 2 tidiness issues. These issues were cleaned in clean section. Before to pass cleaning section I want to note my findings that should have been cleaned.

Quality	
Issues	**Tables**
Choose where retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp do not have value	twitter_archive
The type of timestamp is object, it should be datetime.	twitter_archive
Retweeted_status_timestamp column is object	twitter_archive
Source column is html	twitter_archive
The mean of denominator is 10.45 which is supposed to be 10 /max()	twitter_archive
Text column contains text and URL	twitter_archive
Some dog names are not correct.	twitter_archive
Twitter_data id column object should be integer	image_predictions
Some tweets do not contain dog pictures	twitter_data

Tidiness	
Issues	**Tables**
Doggo, floofer, pupper and puppo columns in twitter_archive table should be merged into one column named "dog_stage".	twitter_archive
twitter api table columns (retweet_count, favorite_count, followers_count) should be added to twitter archive table.	twitter_data

Cleaning process contains 3 steps for each issue. First step is defining the issue, this means we should clearly understand what problem is, how to handle it and what do I want to achieve. After defining the problem, I wrote relevant code to fix the issue and after the coding process, I tested them to be sure if the result is correct.

Cleaning step is the last step of data wrangling process. We can say that after wrangling the data we are ready to analyse and search for meaningful findings. I used cleared data to build visualisations that gave wide opinion about dog rating. Before starting the analysing, I saved the data to csv file called ‘twitter_archive_master.csv’.

In short about Visualization, I want to note that generally I used bar chart. Because I had categorical data as dog names or source etc. But I also preferred to create a scatter plot and a line chart.