Data Wrangling report

Raw data collected for this project were from different formats and not suitable for analysis. The three datasets gathered weren't well-structured which made it impossible to proceed to the next step right away. Visual assessment was performed using both, pandas and Microsoft excel. Visually assessing it pointed out the obvious issues such as unwanted columns, variables that need to be separated, and missing data. Programmatic assessment made it easier for me to check for incorrect data types, duplicated rows, and inconsistent variable within the columns.

There were four major issues during the data wrangling process:

1. Gathering the additional data via Twitter API didn't work out well. Although the script given worked, it kept giving error messages and wouldn't finish running. I decided to proceed with the given JSON file.
2. The ***df_archive*** table had quality issue in the *text* column. The string contains URLs, ratings. I was able to remove both using the regex and replace method. In this column also had the keywords for dog stages which made it possible for me to extract and later drops the other four columns representing each of the dog stages (*doggo,floofer,pupper, and puppo*).
3. The df_archive table also has a lot of invalid names. Although there are ways of cleaning it programmatically, I decided to visually assess it using both pandas and Microsoft excel. I put all the names in an array and replace them with the Numpy np.nan method.
4. In ***df_prediction*** table, I felt like the p1-p3 dog names and confidence were all unnecessary for analysis. It should all be combined in one column *breed* by counting the MAX value out of the three confidence columns.

Outside of these four issues, the rest are easy to observe and clean. Some of the main issues were the incorrect data types found in every data frame, handling missing values, and inconsistent units. Copies of each data frame were also made in the beginning before the cleaning, in case I need to refer back to the original gathered data.