Data was obtained from three sources – a provided csv, a tsv populated by a data produced from a neural network, and a json file created from querying Twitter's API. This data was in three different formats (csv, tsv, and json), which were then read into the Jupyter Notebook. Quick code checks were performed to verify that the data was read in as intended.

Both visual and programmatic assessment was then performed on the data. Visual assessment was initially done in Jupyter Notebook. The csv file was then review in Excel and the tsv file in Notepad++. Visual quickly assessment indicated that the source column contained additional text that needed to be removed. All three dataframes also had a tweet identification number, which indicated that they should be combined to help create a tidy dataset. Furthermore, the dog name variants (doggo, floofer, pupper, and puppo) are also portraying the same categorical variable, dog name, but are in different columns. To tidy these up, they should all be combined. Since this data is not used in any subsequent analysis, we opted to drop these columns for simplicity. Additionally, the prediction columns (p2, p2_conf, p2_dog, p3, p3_conf, and p3_dog) were dropped since they were cluttering the dataframe and were not used in any further analysis.

Programmatic assessment allowed us to quickly analyze the scope of some concerns. There were 181 retweets and 78 replies. There were numerous tweets that did not have an associated photo. There were also some unreliable numerators and large denominators. The unreliable numerators were dropped from the dataset, and the denominators were all set to 10 since that is the ranking system used on WeRateDogs. Additionally, the timestamp was a string instead of a datetime object, so that was converted to allow for easier analyze of time data. Lastly, several "cat" predictions were seen in the p1 prediction column. These were replaced with their corresponding p2 data.