All of the data gathered for my project was taken from Udacity. The image predictions file was downloaded with the requests library and the twitter archive was downloaded manually. I was not able to get a developer account from Twitter so that data was also downloaded manually. Managing the JSON data was challenging because the twitter API data from Udacity must be read line by line by the json.loads function. Once I became aware of this by checking the Udacity knowledge base I took the needed data from the JSON txt files and put them into lists. These lists were then put into dictionaries that could be converted into dataframes. I suspected that since the names of the dogs were entered by software there could be errors and using sort_values this was the case. It appeared that the words starting in lowercase were not names and needed to be removed. The lowercase words were removed using a regular expression. All names 'None' were removed. I decided to keep the other values associated with the missing 'name' entries since it would still be useful for analysis. Timestamps were converted from object to datatime data so they could be more useful later. Datetime data was useful in the plotting of time vs other data points in the visualization section. Since IDs are normally not stored as integers they were changed to string in all dataframes. To remove any retweets or replies the series containing data for these types of tweets were null selected and the non-null were discarded. Tweet IDs with no URL were removed. To improve the dataframe structure each of the 4 columns for dog stage were merged into one column. The previous 4 coulumns were then dropped.Finally, all dataframes were merged using an inner merge by tweet_id. The final master dataframe contained values for each variable except for the names of the dogs.