The data has to be analyzed from 3 sources
1. Archive shared by Udacity
2. Extracting using twitter API
3. Extracting image prediction from a tsv file present at the URL provided

Data from archive file is the easiest of one to gather but must say this required most of effort in analyzing and cleaning. This data has some quality issues like formatting issues, missing values, incorrect values in some columns, multiple values in a column and tidiness issues like a column split to multiple columns.

Extraction of data using twitter API required some authentication approval from twitter. This process took 3 days for me as they circled back to me while seeking some more clarification on why I would be needing twitter developer access. After getting the approval, the code was pretty straightforward to extract the data, however this took significant time for running the code. Once done, analysis and cleaning part of this dataset was pretty simple. Data has no quality and tidiness issues.

Image prediction tsv file was hosted at an URL. The code to fetch that was pretty straight forward. However, the code has a few data quality issues, like incorrect datatype. The tsv provided to us was generating using neural network, that mapped the closest image it can identify to, for dog breed. Since it was machine learning working behind there are some invalid breed types of dogs are present like cup, pen, paper towel etc. This scenario was handled using a logical operation on the data and we arrived at best possible dog breed type for a particular column. However if for all the p1,p2,p3 values of a row where p{i}_dog is False we have omitted the row from the data.

There were many other quality issues still present in the clean dataset, but as this project required handing at least 10 just issues, this report is created.