This report tries to summarize the different steps we've followed to gather, assess and clean the data used in the Wrangle and Analyze project in the Udacity Data Analyst Nanodegree.

1. Gathering

We've gathered data in 3 ways.

1. We´ve used the pandas function read_csv to import the file provided by Udacity
2. We´ve downloaded a file from the Internet using the Python request library
3. Finally, we've connected with the Twitter api to create a json file

All three files have been imported into one different dataframe.

2. Assessing

The steps followed to assess the data in all three data frames is similar way:

1. Explore each dataframe visually
2. Explore each dataframe programmatically

Using both techniques we've identified this issues with the data:

Quality Issues

1. tweet_id is an integer but it should be a string as it is not used to compute anything
2. timestamp column should have datetime format
3. There are replies and retweets in our data
4. There are tweets without images
5. name column contains strings that aren´t names as 'a', 'an', 'the'...
6. source column contains html code
7. p1, p2, p3 columns have inconsistent capitalization
8. Not all dogs have been properly matched to a particular breed in columns

9. There are too many unneded columns

Tidyness Issues

1. One of the rules of Tiny Data is that each variable forms a column. In this dataframe we have one variable dog_stages split in 4 columns doggo, floofer, pupper, puppo.
2. Another rule about Tiny Data is that each type of observational unit forms a table. We are analyzing tweets, but we have the information split in three different dataframes. We should combine them.

3. Cleaning

Once identified each issue, we've proceeded to clean our data programmatically.

We've started with the Tidyness Issues:

1. Creating one dog_stage column from the 4 different ones
2. Merging the 3 dataframes

After solving these, we've proceeded to solve the Quality Issues previously mentioned with the follwowing actions:

1. Convert tweet_id from integer to string
2. Apply datetime format to timestamp column
3. Delete replies and retweets
4. Delete tweets without image
5. Fix wrong names
6. Strip html tags from source column
7. Fix inconsistent capitalization in p1, p2, p3 colums
8. Create one breed column
9. Delete unneded columns

4. Storing

Finally, after gathering, assessing and cleaning our data, we've stored it in a twitter_archive_master.csv file using the Pandas to_csv function.