ACT Report of *WeRateDogs* Twitter Account

1. Introduction

Data wrangling is the process of transforming “raw” data for making it more suitable for analysis, and it will improve the quality of your data.
Data wrangling consists of 3 phases:

- Gathering
- Assessment
- Cleaning

And we will go through each one of them in a bit more details.

2. Gathering

This the first step in the data wrangling, where we collect our data from different sources such as predownloaded files, APIs, scraping data from the internet or downloading online files. The collected data usually comes in different forms and encodings, so we need to handle each format with appropriate methods to not mess the data.

The data used in this project was collected from 3 different sources with 3 different formats.

1) **The archived data from twitter which was a given file by Udacity with CSV format**, that I managed to read using **Pandas** library that read the “twitter-archive-enhanced.csv” file, using **csv_read** function. Then I stored this data into a dataframe called **archive_df**.
2) **The image predictions file which I downloaded programmatically form the given URL in the project details, the file was in TSV format.** I managed to download this file using Python **Requests** library which has a **get** method that acquired a response from the URL, then I saved its content in a filed called “image-predictions.tsv”, then it was read using the same **csv_read** function, but with the tab as a separator, these data was stored in the **image_predictions_df**.
3) **Data retrieved from querying the Twitter API,** for this step, I used **tweet_json.txt** given file, which had JSON objects that I read into a list of dictionaries using the python library **json**. Then I read it into a dataframe called **api_df.**

3. Assessment

This assessment process is about investigating all our data, to find which issues that exist, so we can document them, to later deal with them in the cleaning process. So, we can make our data clean and suitable for further analysis.

The assessment process includes investigating the quality and tidiness of the data. Where quality issues are related to the dirty data, but the tidiness issue are related to what is messy in the data.

This step was done visually by investigating all the dataframes using pandas functions such as **head()** and **sample()**, then further investigation was done programmatically to find more about the null values, data types, statistics information about each dataframe.

The assessment process results in the following issues:

- **Quality related issues:**

1) There are 78 replied tweets which is invalid data for our criteria, and 181 retweets which are also invalid data.
2) The in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp columns should be removed as they are related to the replies and retweets which are invalid for our data.
3) tweet_id should be str not int, as we won't perform any numerical operations on it.
4) The timestamp should be datetime type not string.
5) Inaccurate extraction of the name of the dogs in some tweets (for example: a, not, an)
6) The source column contains lots of information that is needed to be cleaned (HTML tags), also the source should be considered as category (finite sources).
7) Inconsistent representation of the missing data (NaN and None str)
8) Missing values in the name and dog stage (Insignificant information)
9) Some of the columns don't have descriptive names like p1, p1_conf, p1_dog ... etc
10) 66 images were duplicated
11) The dog breeds have inconsistent representation (Some of them begin with lower letters while the other with capital ones, also some of them are separated with _ and the others with space).

12) The API has only 2354 while the archive has 2356, which means 2 missing IDs. (They can't be retrieved)

- **Tidiness related issues:**

1) The 4 columns that represent the stages of the dogs (doggo, floofer, pupper, puppo) should be considered as only 1 column as they represent values of one variable.
2) The 3 columns that represent the breed prediction of the dogs should bs merged into only 1 column, and these existing columns should be values in the new column (variable).
3) The 3 columns that represent the confidence of the predictions should bs merged into only 1 column, and these existing columns should be values in the new column (variable).
4) A single observational unit is stored in multiple tables (archived_data, api_data, image_predictions_data) should all be gathered in 1 table.

4. Cleaning

The cleaning process includes resolving each issue that has been identified in the assessment phase. Each of the issues was handled separately through 3 stages (Define, Code, Test). But before starting cleaning, a copy was made of each dataframe to not mess the original data.

After the end of cleaning, the data was stored to a csv file “twitter_archive_master.csv”, Then it was used for analysis and visualization to discover insights, which is discussed in the act report.