# Data Wrangling Report

Noura Shebl

This project consisted of three main steps:

1- Gathering
2- Assessing
3- Cleaning

**Gathering of Data:**

I gathered data from three different sources.

1- A csv file that was already provided in the project.
2- A tsv file that I downloaded programmatically
3- The data from twitter's API was extracted, and written to a
4- A text file (tweet_json.txt) in json format. This txt file contained each tweet's tweet id, favourite count, and retweet count.

**Assessing the data:**

In this phase, I explored the different aspects of every dataset gathered using panda library features, such as df.info(), df.describe(), df.head(), df.sample() and many more other function.

I assessed the datasets both visually, by scrolling through the data, and programmatically, using these functions.

As a result I noticed several items regarding both quality and tidiness, that required attention to achieve a clean dataset for better visualisations and analysis.

The following are the issues I figured out:

**Tidiness**:
1- twitter_archive table: The 4 columns for dog types doggo, poppy, pupper and floofer should be in a single column called dog type.
2- The three tables should be all in one dataframe.

**Quality**
1-Data type of tweet_id to string and Data type of timestamp to datetime.
2-Retweet enteries to be removed
3-Enteries without images (jpg_url) are to be removed
4-The text in p1, p2 and p3 to be all in lower case
5-Redundant columns to be removed
6-Columns containing all null values to be removed.
7-A new ratio column in the final table
8-"a" in name of dogs to None, because "a" is not a name

**Cleaning the data:**

By far this was the part that required more effort than the first two steps. There were some functions that I couldn't remember, I searched for some references, such as stackoverflow, and previous Udacity lessons that discussed these functions during the course. I find searching and figuring out things for myself both educative and satisfying when I actually figure it out and understand how to do it properly.