The third project of the Udacity Data Analyst Nanodegree was about showcasing data wrangling skills using Python3 and Jupyter Notebook (JN). The project was divided into three parts of the wrangling workflow: Gathering, Assessing, and Cleaning.

I gathered data in three different ways. The first dataset, we_rate_dogs, was downloaded manually and then loaded to the JN using the Pandas method: read_csv. The second dataset, dog_breed, was loaded programmatically from a Udacity URL. The "dog_breed" dataset was a dataset created in another course using Neural Networks to identify breeds of dogs from tweets in the first dataset. The third dataset, tweets_and_likes, was the most challenging and fun way to gather data. I connected to the Twitter API using a module called "tweepy", then I downloaded programmatically the JSON files of all present tweets in the first dataset. After that, I programmatically created the dataset, tweets_and_likes, by accessing the JSON files one by one and filling in the dataset with observations of every tweet: tweet_id, retweet_count, and favorite_count.

After gathering the three datasets, I started the assessment part. I divided the different type of fixes by quality issues or tidiness issues of the three datasets. The main tidiness issues were to make sure every dataset represent one observation, every row represents one subject (a tweet in this case), and every column represents one variable (tweet ID, retweet count, rating, etc.). The assessments were first described informally after every visual or programmatic piece of code. Then, all the issues were formally stated in a bullet points structure in order for an easy walk through during the cleaning part.

The cleaning part was structured defining the issue that needed "cleaning", then a code was written to fix the issue either manually or programmatically. Most issues were fixed programmatically except for the quality issues in the rating columns (rating_numerator, and rating_denominator). After coding the fixes, a test was created in order to make sure that the issue was fixed. These steps were helpful in order to create a meaningful exploration of the data.