# Wrangle Report

Udacity | @WeRateDogs archive on Twitter

**Introduction**

For my Udacity course on data wrangling, I learned how to gather, clean, and assess tweet data based on the Twitter archive for the user WeRateDogs, which posts dog pictures along with a caption and rating out of 10. The account is known for giving the majority of its ratings a score above 10, which makes the analysis interesting and funny. I thought gathering data was the biggest challenge in this project because it involved exploring the Twitter API and applying newly-learned packages like “requests” and “os” in order to download files programmatically.

**Gathering Process**

I had to collect data from three different sources. The easiest one (“twitter-archive-enhanced.csv”) was using Pandas to read a flat .csv file. The intermediate one (“image-preds.tsv”) was using the requests library to make a GET request with a URL in order to write contents of its to a file. The hardest task was creating a Twitter developer account, setting up authentication, exploring its API functionalities, and parsing the contents into JSON with a specified format. After writing “tweet_json.txt” with one tweet per line, I parsed out the wanted columns.

**Cleaning Process**

Since I spent a few hours figuring out how to collect the data in the correct format, I saw obvious areas that needed cleaning in terms of quality and tidiness/structure. I identified incorrect formats, such as timestamp casted as string, and messy data, such as dog names as “a” and “the” instead of the actual name. There were also columns indicating the presence of retweets, which should not be included in the analysis. After defining what needs to be fixed, I made changes and validated that the datasets were cleaned and ready for analysis.

**Analyzing Process**

Gathering and cleaning data feels tedious yet rewarding, but creating insights is always the best part for me. I enjoy knowing which dog breeds get the most love in terms of ratings and number of tweets and imagining if I fit in with the majority. If I were an actual twitter user, I could see myself displaying my preferences for cuter dogs over others.

**Conclusion**

This Twitter data wrangling and analyzing project was a challenging way to apply my skills from practicing problems to doing a project from end-to-end. I plan on exploring more APIs, probably related to social media or data on an interesting industry, to try other analyses. Now that I know how to programmatically collect data from online, I have access to find and solve more interesting data and problems.