**Wrangle and Analyze Data Report**The report explains the wrangling efforts exercised in this project. In carrying out this project, I gathered three(3) datasets for analysis. The datasets are; WeRateDogs™ Twitter Archive, Tweet image predictions and   
Twitter data from Twitter API.

**We rate Dogs:**WeRateDogs is a [Twitter](https://en.wikipedia.org/wiki/Twitter) account that rates people's [dogs](https://en.wikipedia.org/wiki/Dog) with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson. WeRateDogs asks people to send photos of their dogs, then tweets selected photos rating and a humorous comment. Dogs are rated on a [scale of one to ten](https://en.wikipedia.org/wiki/Scale_of_one_to_ten), but are invariably given ratings in excess of the maximum, such as "13/10". Popular posts are re-posted on Instagram and Facebook. In 2017, Nelson started a spin-off Twitter account, Thoughts of Dog.

The account also has a branded game, a popular online store, and a book that was published in fall 2017. Nelson and his team of four receive 800 to 1,000 submissions daily and work to narrow them down to about one high-quality piece of dog content per day.

The entire project was completed on jupyter notebook; however, the project is being supported with two reports exported as PDFs.

The wrangling process is divided into 3 major steps;

1. Data Gathering
2. Data Accessing
3. Data Cleaning
4. **Data Gathering:**

● WeRateDogs™ Twitter Archive (twitter-archive-enhanced.csv):

Contains data extracted programmatically from tweet data sent by WeRateDogs to Udacity to be used exclusively in this project. The data provides the rating, dog name, dog stage and other information.  
▪ Source: [downloaded manually from Udacity](https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-%3Eenhanced/twitter-archive-enhanced.csv)

● Tweet image predictions (image\_predictions.tsv)

Produced by running every image in the WeRateDogs twitter archive through a neural network that classifies breed of dogs. This process resulted in a dataset full of image predictions (the top three only) along side each tweet ID, image URL, and the number that corresponded to the most confident prediction.

▪ Source: [Programmatically downloaded via Requests from Udacity](https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-%3Epredictions.tsv) server

● Additional Twitter data (tweet\_json.txt)

Obtained by querying Twitter’s API then stored in txt file. Gathering this data requires Twitter developer account, through a ready-made version of the file is provided by Udacity.   
▪ Source: Downloaded manually from Udacity

1. **Accessing Data**

After successfully gathering each of the dataset, they were accessed Visually and programmatically for quality and tidiness issues.

The following were concluded:

Tidiness:

i. Dog stage data were separated into 4 columns

ii. All data were related but divided into three (3) separate dataframes.

Quality:

▪ Enhanced Twitter Archive

i. some Dog names are invalid

ii. invalid tweet\_id datatype (should be integer instead of string)

iii. the datatype of the timestamp (column is object and should be datetime)

iv. some of the dogs are not classified as one of "doggo", "floofer", "pupper" or "puppo" and contain all "None" instead

v. Some dog names are invalid like 'a', 'an', 'the', 'such', etc.

▪ Tweet Image Prediction

i. the datatype of the id - columns is integer and should be str

ii. contains retweets (duplicated rows in column jpg\_url)

iii. the predictions are sometimes uppercase, sometimes lowercase

iv. also there is a "\_" instead of a whitespace in the predictions

▪ Tweet Data From Twitter API

i. the datatype of the id - columns is integer and should be str

**Cleaning Data**The previous issues were cleaned as appropriate resulting in a good quality and tidy master dataframe.