Krishna Koirala

Introduction:

This file contains the documentation for data wrangling steps: gather, access & clean. The dataset I am using is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. Almost always greater than 10. 11/10, 12/10, 13/10, etc.

Gather:

Gathering data became one of the challenging part for me specially the api_tweet, took my few days. Getting only successful tweets corresponding to each tweet ids where some of the tweet ids were missing, was the challenging part. This step made me to learn except, continue, pass syntax in depth.

Data Sources: The twitter archive enhanced data was downloaded manually from following link.
twitter_archive_enhanced.csv

Image prediction data was downloaded programmatically from following link using request library.

URL: **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**

Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's **Tweepy** library and stored each tweet's entire set of JSON data in a file called tweet_json.txt . After writing Each tweet's JSON data to its own line, I read this .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

Access:

Accessing data was the easy part. Just see them closely and try to find the error as much as you can.

Clean:

Cleaning data is important step to make your data ready for further deep analysis. If cleaning is not done well, you might face problem in visualization and hence analysis. But it is very important to be careful of not to delete useful information from data in the process of cleaning. It is mandatory to make a copy of data set before you start cleaning it. Create copies of original DataFrames and perform cleaning on copy of data. Note that: simply assigning a Data Frame to a new variable name leaves the original Data Frame vulnerable to modifications.

Visualize:

The most interesting part in data analysis is the visualization for me. It gives the idea how target variables are related to other variables. That leads you tentative idea for model building part.