UDACITY

WRANGLE AND ANALYZE DATA

WRANGLE REPORT

PRESENTED BY: Sunny Paul

WRANGLE AND ANALYZE DATA

INTRODUCTION

The dataset that was wrangled, analyzed and visualized 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

PROJECT DETAILS:

An Enhanced Twitter Archive has been provided by Udacity to facilitate Wrangling process, the archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which was used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, the archive was filtered for tweets with ratings only (there are 2356).

Additionally, Udacity provided an image predictions file that can classify breeds of dogs. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

It is part of the requirements for this project to gather additional data via the Twitter API **Tweepy** and store the Retweet Count and Favorite Count of the tweets provided in the Enhanced Twitter Archive. The gathering of this data has been performed in this project.

KEY POINTS:

Key points to keep in mind when data wrangling for this project:

- We only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
- Assessing and cleaning the entire dataset completely would require a lot of time, and is not necessary to practice and demonstrate your skills in data wrangling. Therefore, the requirements of this project are only to assess and clean at least 8 quality issues and at least 2 quality issues in this dataset.
- Cleaning includes merging individual pieces of data according to the rules of tidy data.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.
- We do not need to gather the tweets beyond August 1st, 2017. You can, but note that you won't be able to gather the image predictions for these tweets since you don't have access to the algorithm used.

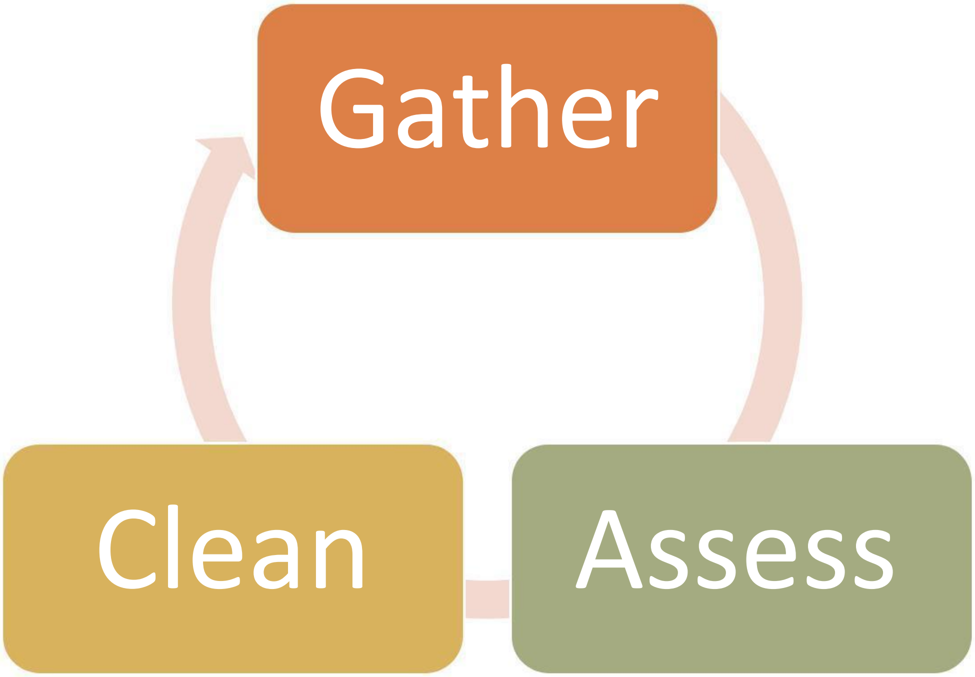
PROJECT TASKS

- ### DATA WRANGLING, WHICH CONSISTS OF:
  - Gathering data
  - Assessing data
  - Cleaning data
- ### STORING, ANALYZING, AND VISUALIZING WRANGLED DATA
- ### REPORTING:
  - Data wrangling efforts
  - Data analyses and visualizations

This document will demonstrate the wrangling efforts exerted to analyze and visualize the tweet archive of @dog_rates

DATA WRANGLING PROCESS

The process of data wrangling consists of 3 main steps:



GATHERING DATA FOR THIS PROJECT

There are (3) sources of data gathering that were performed for this project:

1. The WeRateDogs Twitter archive provided by Udacity. Download this file manually by clicking the following link: `twitter_archive_enhanced.csv`
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (`image_predictions.tsv`) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Each tweet's retweet count and favorite ("like") count at minimum, and any additional data. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called `tweet_json.txt` file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.
   Note: *Twitter API keys, secrets, and tokens are not to be included with the project submission.*

ASSESSING DATA FOR THIS PROJECT

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues. Detect and document at least **eight (8) quality issues** and **two (2) tidiness issues**

The assessment performed for this project is not definitive, there are other quality issues that have not been addressed.

ASSESSMENT SUMMARY

Quality Assessment

No.	Issue	Column(s)	Dataset
1	None is given for missing dog name	name	twdf
2	Incorrect dog names exist in the name column like none, a, the, an, my	name	twdf
3	timestamp datatype is incorrect	timestamp	twdf
4	Unnecessary characters (+0000)	timestamp	twdf
5	Missing values in the dog breed columns which was filled with 'None'	doggo, floofer, pupper, puppo	twdf
6	There are missing values (nulls)	in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp	twdf
7	Rating numerator values of 0 or 1 might have been miss-typed (Remove 10 tweets)	rating_numerator	twdf

No.	Issue	Column(s)	Dataset
8	Rating denominator value should always be 10, there are many incorrect denominators	rating_denominator	twdf
9	(9) missing tweets that could not be read	tweet_id	df_json
10	Tweet_id datatype needs to be converted to string	tweet_id	twdf
11	Rating numerator is not reflecting actual rating, need to extract rating from tweet text	rating_numerator	twdf

Tidiness Assessment

No.	Issue	Dataset
1	Dogs breed is splitted into 4 columns	twdf
2	'Unnamed: 0' column is not needed	df_json
3	Data can be structured into 1 dataset of tweets archive	twdf, image_predictions, df_json

CLEANING DATA FOR THIS PROJECT

Cleaning actions that were performed against the quality and tidiness issues are listed below

Quality Cleaning

No.	Issue	Action
1	None is given for missing dog name	Replace invalid dog names (none, a, the, an) with null
2	Incorrect dog names exist in the name column like none, a, the, an, my	
3	timestamp datatype is incorrect	Change the timestamp datatype from object to timestamp
4	Unnecessary characters (+0000)	
5	Missing values in the dog breed columns which was filled with 'None'	Replace the none values for dog breeds with null
6	There are missing values (nulls)	It makes sense to have null values in the below columns since they are related to retweets. in_reply_to_status_id in_reply_to_user_id

No.	Issue	Action
		As we only want original ratings (no retweets) that have images, I will remove the rows that have tweet IDs in the in_reply_to_status_id column, and then drop the retweet related columns. According to Twitter Dev, **in_reply_to_status_id**: If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's ID.
7	Rating numerator values of 0 or 1 might have been miss-typed (Remove 10 tweets)	Delete the (10) Tweets that are not related to dogs rating: (Note: removing replys resulted in shrinking this list to 7 tweets only). The tweets to be deleted are: 835152434251116546 798576900688019456 675153376133427200 670783437142401025 667549055577362432 666287406224695296 666104133288665088
8	Rating denominator value should always be 10, there are many incorrect denominators	Set the value of the rating denominator to 10 across the dataset
9	(9) missing tweets that could not be read	Remove the (9) missing tweets that could not be read (stored in df_errors)
10	Tweet_id datatype needs to be converted to string	Tweet_id column has been converted to string across all the (3) datasets twdf_clean, image_predictions_clean, df_json_clean
11	Rating numerator is not reflecting actual rating, need to extract rating from tweet text	Rating numerator has been corrected by extracting the actual ratings from the tweets text.

Tidiness Cleaning

No.	Issue	Actions
1	Dogs breed is splitted into 4 columns	1. Join the 4 dog stages column into (stage) column

No.	Issue	Actions
		2. Convert the stage column type to categorial
2	'Unnamed: 0' column is not needed	Drop the 'Unnamed: 0' and 'Unnamed: 0.1' columns from twdf_clean, ,image_predictions_clean and df_json_clean
3	Data can be structured into 1 dataset of tweets archive	Create a comprehensive dataset by merging the (image_predictions_clean and df_json_clean into twdf_clean) datasets and drop the unneeded columns.

STORING DATA FOR THIS PROJECT

The clean DataFrame(s) are stored in a CSV file with the main one named **twitter_archive_master.csv**

Additional files were also created to store dataframes prepared after downloading the tweets data to avoid re-downloading process which takes long time (about 40 minutes):

- **twdf.clean.csv**
- **image_predictions.clean.csv**
- **df_json.clean.csv**
- **df_errors.csv**

REFERENCES:

https://www.digitalocean.com/community/tutorials/how-to-authenticate-a-python-application-with-twitter-using-tweepy-on-ubuntu-14-04

https://stackoverflow.com/questions/42384118/how-to-get-distinct-value-while-using-apply-join-in-pandas-dataframe

https://seaborn.pydata.org/examples/horizontal_barplot.html

https://guides.github.com/pdfs/markdown-cheatsheet-online.pdf