Wrangle Data Report

Introduction:

By using Anaconda with Python and its libraries, I gathered the provided data from a variety of sources and in a variety of formats. I assessed its quality and tidiness, then performed data cleaning. The dataset that I wrangled, analyzed and visualize 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 Steps:

The steps I took are:

- Gathering data
- Assessing data
- Cleaning data
- Storing, analyzing, and visualizing the cleaned data

Data Gathering Step:

Here we have three different data sets which consist of:

1. **twitter_archive_enhanced.csv:** which is a csv file that contains different information like, tweet_id, number of retweets, dog type, and a lot more.
2. **image_predictions.tsv: t**he tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network.
3. **tweet_json.txt:** 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

Assessing Data Step

After performing each of the above stesp of data, for quality and tidiness issues, I needed to assess them visually and programmatically by using different methods like, value counts, , duplicated, group by and etc.

Archive dataset:

1. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integers instead of float.
2. retweeted_status_timestamp, timestamp should be datetime instead of object (string)
3. In several columns null objects are non-null
4. We only want original ratings not the retweets that contains images
5. Drop unused columns

Images dataset

1. Some tweet_ids have repeated jpg_url and other tweets have two different tweet_id one linked to the other
2. Delete columns that will not be used for analysis

json_tweeets dataset:

1. Duplicated tweet_id exist
2. Change tweet_id data type to int64

Cleaning Data Step:

After perfroming data assessment, I performed data pre-processing to clean each of the issues that I documented. The final data should be, a high quality DataFrame.

Conclusion:

This project touched on very important aspects of the daily life of data analyst performs 70% of the time. The project stresses that if the data is not cleaned well, you might end up with wrong conclusion/recommendations. Data wrangling is a key to conduct successful analysis.