Wrangle Report

Introduction

The purpose of this project is use what I learned in data wrangling lesson from Udacity Data Analysis Nanodegree program. The dataset which will be wrangled is the tweets archive of Twitter user @dog_rates, also known as WeRateDogs.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for us to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

The goal of this project is to wrangle the WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The challenge lies in the fact that the Twitter archive is amazing, but it only contains very basic tweet information that comes in JSON format. So, I need to gather, asses and clean the Twitter data for a worthy analysis and visualization.

Key Points

- 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**.

- **We do not need to gather the tweets beyond August 1st, 2017**. We can but note that we won't be able to gather the image predictions for these tweets since we don't have access to the algorithm used.

- Fully assessing and cleaning the entire dataset requires exceptional effort so only a subset of its issues **(eight (8) quality issues and two (2) tidiness issues at minimum) need to be assessed and cleaned.**

- 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.

Project details

The tasks of this project are as follows:

- Gathering data.
- Assessing data.
- Cleaning data.

Gathering data

The data for this project consists of three different dataset that were obtained as following:

1. **Twitter archive file:** the **twitter_archive_enhanced.csv** was provided by Udacity and downloaded manually.
2. **The tweet image predictions**, i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information
3. **Twitter API & JSON:** by 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 file. I read this .txt file line by line into a pandas data frame with tweet ID, favourite count, retweet count, followers count, friends count, source, retweeted status and URL.

Assessing data

Once the three tables were obtained, I assessed the data as following:

- Visually, I used two tools. One was by printing the three entire data frames separate in Jupyter Notebook and two by checking the csv files in Excel.

- Programmatically, by using different methods (e.g. info, value counts, sample, duplicated, group by, etc). Then I separated the issues encountered in quality issues and tidiness issues.

Cleaning data

This part of the data wrangling was divided in three parts: Define, code and test the code. These three steps were on each of the issues described in the above section.

First and very helpful step was to create a copy of the three original data frames.

There were a couple of cleaning steps that were very challenging, I will mention here some of them:

1- Other interesting cleaning code was to melt the dog bloods in one column instead of four columns as original presented in twitter archive.

2- In the Twitter Archive table, I had to correct some numerators that were **actual decimals**. This issue was brought to my attention after I checked **the Slack channel for the program DAND**, and found some colleagues asking about it. So, I decided to include it in my cleaning steps.

Conclusion

1. For gathering data there are several packages that help scraping data off the web, that help using APIs to collect data (Tweepy for Twitter) or to communicate with SQL databases.

2. Python's libraries are more powerful with dealing with big data (more than Excel).

3. Python's libraries can deal with a large variety of data (unstructured data like JSON (Tweets).

4. It's very important for the analysis to always consider using the programmatic approach in his/her analysis, to make his/her analysis more generic and can be used either for frequent times without a lot of efforts to modify it or can be used to integrate in other projects.