Introduction

The important goal of this project is to wrangle weRateDogs Twitter data to create an interesting analysis and visualization. This project is part of the data wrangling section of the udacity Data Analysis Nanodegree programme   and is focused on data wrangling from weRateDogs Twitter account using Python library.

Project Details

Normally real-world data rarely comes out clean. Using python libraries, i had to gather, assess and clean the dataset given for this project, to carry out my analysis and visualization of the dataset. After fully assessing and cleaning the entire dataset, it would require exceptional effort so only a subset of its issues (eight quality issues and two tidiness issues at minimum) needed to be assessed and cleaned.

The tasks for the project include:

1. Data wrangling, which consist of:
   - Gathering
   - Assessing
   - Cleaning

2. Storing, analyzing and visualizing the wrangled data.
3. Reporting on my data analyses and visualizations (act_report.html).

Gathering the Data

The data for this project was in three different formats and they were obtained as mentioned below:

Twitter Archive File-weRateDogs: This was extracted programmatically by udacity and provided as twitter_archive_enhanced.csv to use.

image Predictions File: The tweet image predictions, breed of dog present in each tweet according to a neural network. This File(image_prediction.tsv) was hosted on the Udacity's servers and downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Twitter API & Tweet_JSON File: by using the supporting material provided by Udacity. I queried the Twitter API for each tweet's JSON data using python tweepy library and stored each tweet's of the entire dataset of JSON data in a file called tweet_json.txt file.

Data Cleaning

This part of the data wrangling was divided into three parts as per the dataset and was further divided into three steps Define, code and test blocks below it to make it easy to understand.

The first step is to create copies of all the three datam frame. so that i can do trial and error in the copied frame rather than the originals.

In the twitter archive data, I changed the datatype of the timesmap and made the dog names consistent i.e. First letter capital. AS mentioned earlier the standard for "rating_denominator" is 10, but on checking I found out that it includes some other numbers, which could be the mis parse. So, I check that the text corresponding to those rating and notice d that few of them were analyzed incorrectly due to the presence of another fraction in the text. I corrected the same for both the rating denominator and numerator.

In the image prediction i found out that there were 66 duplicated values. i drop the duplicated data. I also changed the column names to make it more descriptive and readable. Again, the dog breed of all the three-prediction column includes both upper and lowercases for the first letter. I made changes to make it consistent.

In the generated tweet JSON data, the most important column was retweet_count and favourite_count, others were actually redundant as the same were present in the twitter archive data. So, delete the unnecessary columns.

Storing Data

After cleaning the data i discovered that there was no need for three datasets. All the data could be easily made into a single file. So i joined 'tweet_json', 'image_predictions' to weratedogs, to twitter_archive_master.csv

Conclusion

A good data wrangler knows how to integrate information from multiple data sources, solving common information problems and resolve data cleaning and quality issues. A data wrangler also knows understands their data well and is always looking for ways to enrich the data. I have done this data analyses and visualization using python libraries.