wrangle_report

February 19, 2018

1 Introduction

This project is a part data wrangling section of Udacity's Data Analyst Nanodegree program. In this project I am required to wrangle data using tweet archive of Twitter user [@dog_rates](https://twitter.com/dog_rates), also known as WeRateDogs. It is twitter account that rates people's dog with a humorous comment about the dog. The numerator is almost always greater than the denominator. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for 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.

2 Project Details

- In this project I will perform all three steps of data wrangling i.e. :
  - Gather
  - Assess
  - Clean
- I would also store, analyse, and visualize the wrangled data.
- Finally I would report the wrangling efforts, analyses, and visualization

2.1 Gather

To gather data I performed the following steps :

1. The WeRateDogs Twitter archive. I downloaded this file manually using the 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 URL
3. I used Python's Tweepy library to query the twitter API using tweet_id from twitter_archive_enhanced.csv. I then stored each tweet's entire set of JSON data in a file called tweet_json.txt file.

2.2 Assess

After gathering each of the above pieces of data, I must assess them visually and programmatically for quality and tidiness issues. Detect and document at least eight (8) quality issues and two (2) tidiness issues in your wrangle_act.ipynb Jupyter Notebook.

I identified eleven (11) quality issues and five (5) tidiness issues

2.3 Clean

I cleaned each of the issues you documented while assessing. The result should be a high quality and tidy master pandas DataFrame (or DataFrames, if appropriate). The Cleaning should take care of the following points :

- Only wanted original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
- 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.

2.4 Storing, Analyzing, and Visualizing Data for this Project

I stored the final cleaned dataframe in a file 'twitter_archive_master.csv'. Additional files, which are mainly pngs of visualization were also saved. I was required to produce atleast three (3) insights and one (1) visualization from the final data saved. I took five (5) insights and produced two (2) visualization.