Wrangle Report

Introduction

The dataset wrangled for this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter page with about 8.7million followers that regularly shares pictures of dogs along with a catchy description and often a rating typically out of 10 for the dog in the picture, sometimes exceeds 10.

The WeRateDogs Twitter project goals included:

- Wrangling the twitter data through the following processes:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on the data wrangling efforts and data analyses and visualizations

Gathering Data

I would like to say first that my requests to twitter for a developer account has not been approved. Hence I had to work with the Udacity-provided tweet-json.txt.

My wrangling efforts for the WeRateDogs Twitter project included gathering data from the following sources:

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students. 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 tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers.
- The Udacity-provided tweet-json.txt

Assessing Data

Once the data was gathered, I began to assess the data on both quality and tidiness issues. The methods I used include .head() , .info() , .value_counts() among others

Quality and Tidiness issues that were cleaned include the following

- Removal of 181 Retweets
- Fixing of missing data in expanded_urls (Tweets without images)
- Fixing non-dog images
- Fixing incorrect ratings and as well incorrect dog names
- Fixing missing values in dog names (represented as None)
- Fixing erroneous datatype (tweet_id, timestamp)
- Merging three data frames
- Dropping unneeded columns.

Cleaning Data

The process of fixing and resolving issues identified in the Cleaning process. The (define, code, and test) steps were used in the cleaning process. First, copies of the DataFrames were created before cleaning. Then, the steps of cleaning were applied on each issue. Each issue was defined and a code to fix the identified issue and a test followed

I used basic python function like duplicates , drop, value_count , describe , info and others in this process. I accessed a number of websites for syntax and possible solutions.

Storing

The clean DataFrame called archive_clean which 1910 rows and 25 columns and was stored in a csv file called 'twitter_archive_master.csv'. At this point, the data was successfully wrangled and therefore ready for analysis and visualization.

Analysis & Visualization

The analysis and visualization part of this project is presented in 'act_report.pdf'

Sources or References

- Data Analysis Nanodegree/Data Wrangling Lessons
- https://docs.python.org/3/library/datetime.html
- https://www.tutorialspoint.com/python_data_structure/python_2darray.htm
- https://ipython.org/ipython-doc/3/api/generated/IPython.display.html
- https://matplotlib.org/3.1.1/api/axes_api.html
- https://stackabuse.com/reading-and-writing-json-to-a-file-in-python
- https://knowledge.udacity.com/questions/36432