Wrangle Report

Introduction

The dataset that I will discuss, analyze, and visualize is the WeRateDogs (Twitter user @dog_rate) tweets archive. WeRateDogs is a Twitter account that rates people's dogs along with a funny comment. Often the denominator for these ratings is 10, and the numerator is usually greater than 10. The higher the numerator rating, the "better" the dog. WeRateDogs has over 4 million followers and has received international media coverage. Data argument process: Gather, Assess, Clean.

1-Data Gathering:

I will be gathering data from three sources:

- The 'enhanced' Twitter archive WeRateDogs, a csv file provided by Udacity. This archive contains very basic tweet data for all 5000+ of their tweets, but not everything. The archive contains each tweet's text, which Udacity used to enhance by extracting rating, dog name, and dog 'stage' (doggo, floofer, pupper, and puppo). Of the 5000+ tweets, this archive is filtered for tweets with ratings only (there are 2356).

- An 'image prediction' file, or what breed of dog is in each tweet, according to a neural network. This shows the top three breed predictions alongside each tweet ID, the image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images). I will download the image predictions file programmatically from Udcity's servers using the Requests library.

- I use twitter_api.py and tweet_json.txt to read the code and comments, understand how the code works, then copy and paste it into my notebook.

2-Data Assessing:

The three data frames visually saved inside the jupyter notebook were then evaluated with the panda and because the data sets weren't very large, and a copy of each was exported into a single excel workbook. This allowed quickly scan the rows and use filters to select areas for more detailed investigation.

After this prof programmatic evaluation was performed inside jupyter with panda using the following functions, df.info (),df.head (), df.describe(),sample df (), df.value_counts ().Data sets were accessed according to two criteria, quality and accuracy. When a problem was detected it was notarized under one of these two criteria.

Quality refers to issues with the content of data, sometimes called dirty data. Standard standards for the completeness, correctness, accuracy and consistency of the data was used to identify quality issues. These issues miscellaneous and listed in the evaluation section of the jupyter notebook "wrangle_act.ipynb".

Tidiness refers to issues with data structure, sometimes called messy data. Basis for the evaluation is that each variable forms a column, and each note forms a row and each type of observation console make up the table. After evaluating the three data sets, it was decided to merge them into a single data frame.

3-Data Cleaning:

The final step in the disagreement process is to clean up the data for quality and arrangement issues. Cleaning the standard process of identification, coding, and testing was followed for each problem and handled in a logical manner the order, which is reflected in the notebook's numbering order “wrangle_act.ipynb” and is followed closely.

Standard practice is to clean up lost data first, then cleaning for ranking and finally quality.Most of the cleanups were done with software tools, like the built-in def or panda functionslike (drop,merge etc).

Conclusion

Data wrangling provides a clean data frame for future analysis and visualization, in our case we concluded with the “twitter_archive_master.csv’. This file can also be shared with others without having to wrangle the data.