Student: David Singh Arjona

Project #4: Wrangle and Analyze Data

Wrangle Report

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. This report's purpose is to describe the wrangling process this data has been through.

Gathering Data

Data has been gathered from 3 sources:

1) 'twitter-archive-enhanced.csv': this file was provided by Udacity and downloaded from Udacity Classroom, it contains several useful variables, like 'id', 'timestamp', 'dog_type', 'dog_name' (when available), etc.
2) 'image_predictions.tsv': this file indicates which dog breed is most likely to correspond to the dog's picture, using machine learning techniques.
3) 'tweet-json': this file was generated using the Twitter API, it contains data like 'retweet_count' and 'favorite_count'. It was suggested to me that the JSON tweet file that was generated might be incomplete, so I used the JSON file available in Udacity Classroom.

Assessing Data

After the data gathering, the data showed tidiness and quality issues, which are the following:

Quality issues:

- Some datestamps indicates some tweets were posted after August 1st, 2017.
- Possible presence of duplicated data.
- Presence of retweets.
- A lot of columns do not give useful information.
- Several columns have naming issues.
- 'tweet_id' column has an incorrect format.
- Timestamp has an incorrect format.
- A rating system can be created, but some rating numerator need to be fixed before.

Tidiness issues:

- Separate information must be merged.
- There are 4 columns that indicate dog_types, they can be merged into one.
- A dog breed column can be created using the image prediction data.

Cleaning Data

The following techniques and methods were used to clean the quality and tidiness issues presented before:

- .shape
- .format()
- .value_counts()
- reduce()
- .merge()
- lambda functions
- .drop()
- .info()
- .head()
- .replace()
- .title()
- .astype()
- .extract()
- .slice()
- .to_datetime
- .loc[]
- .iteritems()
- for loops
- .append()
- .search()
- .group()
- Regular Expressions
- .describe()
- .iterrows()
- .columns
- .to_csv()

Conclusion

This project required various libraries and extensive gathering work, because several methods and techniques were new to me. Most of the cleaning work was made after the merge of the three data sets.