Wrangle report (Udacity Data Analysis Project Wrangle and Analyze Data)

Introduction

Wrangle and Analyze data project involves gathering data from various recourses. It is associated with tweets from the Twitter user @dog_rates, also known as WeRateDogs. Three main steps were processed in order to complete this project including, gathering data from different sources, assessing gathered data and clean data, since the data you need comes in different formats and ways, we had to assess this data and clean it in order to make it able to be analyzed and visualized. At the end, two visualizations were created from the dataset which will be discussed on the second report (act_report.pdf).

Step 1: Gathering Data

Data was gathered from 3 different sources:

1. The enhanced twitter archive file was provided and downloaded manually. This file includes various variables for each tweet including tweet id, timestamp, text, rating numerator and denominator, name, etc.
2. Additional data, including favorite count and retweet count, this file should be downloaded through Twitter API. However, due to the time we extracting data from the file tweet_json.txt.
3. The tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers. Using machine learning techniques, the breed of dog was predicted based on the picture.

Step 2: Assessing Data

After gathering data, assessing took place using methods such as: head(), sample(), info(), and value_counts() to identify 8 quality issues and 2 tidiness issues, these issues are:

Quality issues

- Incorrect datatypes for columns: tweet_id (should be string), rating_numerator and rating_denominator (both should be float)
- Incorrect names in 'name' column (such, a, quite, one...etc)
- Change 'timestamp' to be 'datetime' instead of 'object'
- Delete column that are unnecessary in our analysis
- Image_prediction contains duplicated 'jpg_url' values
- Name values contains string "None" instead of NaN
- Remove tweet that has been retweeted since its not original tweets
- Make rating standardized by dividing numerator by the denominator and save it in a new column

Tidiness issues

- Data should be grouped in one table
- 1 variable (dog stage) in 4 different columns (doggo, floofer, pupper, and puppo)

Step 3: Cleaning Data

After addressing both quality and tidiness issues we went to clean and fix them. astype() method were used to deal with incorrect types, in addition to that merge() method used to merge data in one table. Dropping unnecessary columns was essential using drop() method along with so many methods e.g.: replace(), extract(), islower() etc.

Conclusion

Usually data gathered from different sources and comes in different formats, in order to process and extract information from this data we have to assess it and clean it to put in a standard way before visualizing or analyzing it so that you can make the best use of it.