Data Wrangling Report for WeRateDogs dataset

Introduction

In the project, data wrangling is performed on WeRateDogs dataset, which is the tweet archive of the Twitter user @dog_rates. Data wrangling consists of three steps - gathering, assessing and cleaning. Data is gathered from three different sources, which include files on a local device and Udacity's server, as well as Udacity's server. Subsequently, data is assessed with visual and programmatic assessments to look for data quality and tidiness issues. Finally, the data is cleaned based on the issues found in assessment.

Gather

Data is gather from the following three sources:

CSV file on a local device	Contains basic tweet data for Twitter user @dog_rates until August 1, 2017
TSV file on Udacity's server	Contains tweet image predictions is present in each tweet according to a neural network, which is downloaded programmatically from Udacity's server using URL.
Twitter's database	Contain each existing tweet's retweet count and favorite count, which is retrieved using Twitter API and stored in a .txt file in JSON format.

Assess

Data is assessed with visual and programmatic assessment to look for data quality and tidiness issues. Visual assessment is carried out by inspecting the data frames in Jupyter notebook, whereas programmatic assessment is carried out by using Pandas functions such as df.info(), df.describe(), series.unique() and etc. The following quality and tidiness issues are detected.

Quality

tweets table

- Some of the tweets are retweets
- Erroneous datatypes (tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id & retweeted_status_user_id columns) - should be string
- Erroneous datatypes (timestamp & retweeted_status_timestamp columns) - should be datetime
- None instead of NaN in doggo, floofer, pupper and puppo columns

- Invalid numerator values such as 960, 75, 27, 50, 26
- Numerator rating should be float64 as some ratings have decimal values
- Invalid denominator values such as 0, 11, 20, 50, 2
- Tweets with tweet_id of 832088576586297345, 682808988178739200 & 810984652412424192 do not have dog rating
- Invalid dog names such as a, the, infuriating, officially, unacceptable and O
- None instead of NaN in name column

images_predictions table

- Inconsistent naming convention for values in p1, p2 and p3 columns (for e.g. Maltese_dog, golden_retriever, soft-coated_wheaten_terrier & Chihuahua
- Erroneous datatypes (tweet_id & img_num columns) - should be string

tweets_additional_info table

- Erroneous datatypes (tweet_id column) - should be string

Tidiness

- 'doggo', 'floofer', 'pupper', 'puppo' should be combined into one column and None instead of NaN in these columns
- All three data frames should be combined into one data frame

Clean

Data is clean based on data quality and tidiness issues found in the assessment stage. All the data cleaning procedures are carried out programmatically using relevant Pandas functions, and follow a systematic approach, which involves three stages - define, code and test. Each issue is tackled one at a time, and tidiness issues are fixed first before fixing quality issues.

Conclusion

Data wrangling is made up of three steps, namely gathering, assessing and cleaning. The aim of data wrangling is to remove dirty and untidy data in the dataset. Dirty or untidy data will result in data analysis or visualization producing wrong insights or results, which will lead to wrong decisions. As we can safely say that most of the dataset contains dirty and untidy data, it is crucial to carry out data wrangling before any data analysis or visualization can be carried out.