Data Wrangling Report

by Pankaj

Jun,2020

In this report I specify the wrangling work done from gathering data to cleaning for data analysis of WeRateDogs tweets.

Gathering

I gathered data from 3 sources:

- The WeRateDogs Twitter archive from Udacity Servers.
- The tweet image predictions i.e., what breed of dog (or other object, animal, etc.) again from Udacity servers programmatically using the "request" python library.
- Each tweet's JSON data downloaded by Twitter API using "tweepy" python library.

Assessing and Cleaning the Data

The 3 saved data frames were analysed using pandas and excel. As all of them were not very large in size and can be easily visually assessed by opening in Excel. After this programmatic assessment was using pandas using various pandas functions df.info, df.head() etc.

There were many issues found out as the data was unclean. We can categorize the unclean data as:-

Unclean Data: There are two types of unclean data

- Dirty data also known as low quality data. Low quality data has content issues.
- Messy data also known as untidy data. Untidy data has structural issues.

These are some of the assessing and the cleaning done across the three datasets.

- All the rows containing non-null values like "retweeted_status_id" ,"retweeted_status_user_id" , "retweeted_status_timestamp", "in_reply_to_status_id" "in_reply_to_user_id" were removed.
- The 4 dog type was melted into one using pd.melt().
- Name of Dogs which started with small letters were converted to "None" as they seem to be incorrect.
- Rating numerator was changed to "rating" and outliers were removed while analysing the final data.
- Made a final one table by using pd.merge() on all the tables.
- The 3 twitter image predictions had 3 predictions so eliminated the 2 and took only the best possible.

- Many of the columns datatype were changed like "datetime" has object type which was changed to datetime using pd.to_datetime()

Conclusion

By Data Wrangling we got a nice and clean dataset which was stored as "WeRateDogDataset.csv" which was helpful in analysing the data graphically.