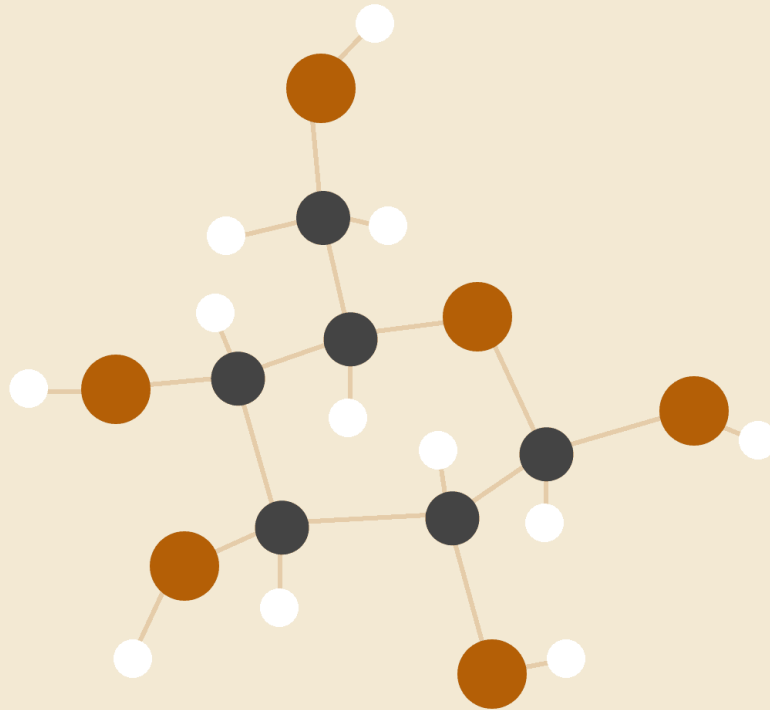


Wrangling report

In God we trust, Others must bring Data



Joshua Regan

16th July. 2022

Data Wrangling: WeRateDogs twitter archive

INTRODUCTION

In a world of data driven decisions, the quest to source this data to the ends of

HYPOTHESIS

Our datasets are;

df_1 -> twitter_enhanced_archive.csv

df_2 -> image_predictions.tsv

Df_3 -> tweet-jason.txt

Visual and programmatics assessment on all datasets/tables listed above yields some the issues which include but are no limited to;

Quality Issues

1. Invalid names in ``name`` column for the **df_1**.
2. **tweet_id** had **dtype = int64** instead of **string**.
3. Columns such as; **in_reply_to_status_id**, **in_reply_to_user_id**, **retweeted_status_id**, **retweeted_status_id**, **retweeted_status_timestamp**. Have non-null entries less than 25% of the whole dataset.
4. Timestamp had **dtype = object** instead of **datetime**.
5. Dog ratings were not standardized, instead they were left at numerator and denominator ratings.
6. **expanded_urls** also had Nan entries.
7. **jpg_url** also had duplicate entries (this is not possible since every tweet is unique).
8. By visual assessment, we observe that dog names with **p1_dog == False** are not actually dog names so **pn_dog == False**, tells us the name in question is not a dog.

Tidiness Issues

9. The four dog_stages each having a column makes the table messy.
10. Merge the right tables (df_1 and df_2)

This

MATERIALS

1. As provided for in Python, the pandas library has built-in functions that come in handy when issues like the ones stated above and others in our dataset
2. Df.info -> this gives a summary description of our DataFrame's data type, non-null entries, etc..
3. Other methods like value_counts() are applied to the columns as pd.Series.

The next logical step is the cleaning step

CLEANING PROCEDURE

The Data Wrangling syntax is as follows

1. Define
2. Code
3. Test.

These steps have a logical dependence. And follow the format below

Issue #	Define	Code	Test
Issue 1			
Issue 2			
...			
Issue n			

RESULTS

At the end of a well conducted wrangling the datasets should

1. Be neater (of better quality)
2. Analysis worthy
3. Anyone should be able to recreate the analysis using your Approach.

CONCLUSION

I cleaned the datasets and saved them in the files **df_1_clean**, **df_2_clean** and **df_3_clean** and the final master dataset was created by concatenating all three datasets into one DataFrame.

REFERENCES

1. <https://stackoverflow.com/questions/32444138/concatenate-a-list-of-pandas-dataframes-together>
2. <https://pandas.pydata.org/docs/>
3. https://github.com/ahmed-gharib89/wrangle-and-analyze_data