

SEPTEMBER 07, 2022

Data Wrangling Report

Gathering, Assessing, Cleaning and
Visualizations

PRESENTED TO

Udacity

PRESENTED BY

Dania Alaa
Adimi






Table of Contents

1 About the project

2 Data Gathering

3 Data Assessing

4 Data Cleaning

5 Conclusion

About the project

This project was all about Data Wrangling, which consists of data gathering, assessing and cleaning. In this project, I wrangled and analyzed the tweet archive of Twitter user @dog_rates, also known as @WeRateDogs. [@WeRateDogs](#) is a Twitter account that rates people's dogs with a humorous comment about the dog.

In the next pages, I will be documenting each step. For further details, you can check the notebook `wrangle_act.ipynb`

Data gathering

My wrangling efforts for the [@WeRateDogs](#) Twitter project included gathering data from the following sources:

The WeRateDogs Twitter archive: The `twitter_archive_enhanced.csv` file was provided by Udacity. This archive contains basic tweet information such as tweet ID, timestamp, text, etc, for over 5000+ tweets. These tweets were filtered and we only kept those who have ratings. So this dataset contains 2356 entry.

The tweet image predictions: i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided by Udacity and I had to download it programmatically.

Additional data from the Twitter API: Using Twitter API and Python's Tweepy library, I gathered more data about the tweets such as each tweet's retweet count and favorite ("like") count.

Data Assessing

Once our data is gathered, we need to assess it right? going through some visual and programmatic assessts here's what I could come up with:

Quality Issues

`df_twitter_enhanced`

- We need to remove retweets and columns that come with it.
- timestamp should be datetime not an object.
- Extra HTML text in the source column needs to be removed.
- In the name column, we have null objects are not declared as non-null objects. So we should turn None to NaN.
- Incorrect names or missing names in name column such as, a, an, the, very (entries that are not names), most of them in lowercase.
- puppo, floofer, pupper and doggo columns should be Boolean.
- The ratings is not standard since there are different denominators. We better add a new column that contains the resulted rating.

Data Assessing

df_image_predictions

- Remove duplicate jpg_url entries.
- Refine p1, p2 and p3 columns and confidence associated by combining.

df_image_predictions

- user_favourites, and user_followers values are identical for all rows, which is specious, we should deal with this.
- Drop the datetime column because we already have timestamp in df_twitter_enhanced.

Tidiness

- Four columns are present for variable dog stage – pupper, puppo and doggo. So we should merge them.
- Join archive, predictions and tweet_json tables into one dataframe df_master_archive.

Now It's time for cleaning!

Data Cleaning

After the assessment, I went through cleaning the data through the following means "Define", "Code" and "Test". I made a copy of dataset in order to keep the original versions in case we want to trace back to them, then I cleaned each dataset separately according to the previous step, merged them all together and stored them in a dataframe called `df_archive_master`.

Conclusion

I had so much fun working on this project. Thanks again Udacity for this opportunity!