Data Wrangling

Introduction

This is a data wrangling report of the We-Rate-Dog project and I will show all stages involved in cleaning the data of this project. It is expected that the data needs to be cleaned properly in order to gain useful insights from it.

Gathering data

We are gathering data from 3 sources:

1- twitter_archive
2- image_predictions
3- tweet_info

Assessing data

At this stage we are looking at our data and assessing it programmatically to find:

- Quality issues
- Tidiness issues

Upon finding our issues we document it by listing in the jupyter environment to help us cleaning our data as required and in order not to be confused or waste time trying to remember our issues.

Tidiness issues:

1- Columns 'doggo', 'floofer', 'pupper', 'puppo' in twitter_archive should belong to one colomn – stage
2- tweet_info and image predictions to twitter_archive table.

Quality issues:

1- Some columns have huge amount of missing values, for example, "in_reply_to_status_id"

2- The datatype of "timestamp" is not correct.

3- The varaible "expanded_urls" also has few missing values, which means some records had no images

4- The standard for "rating_denominator" is 10, but it includes some other numbers, which could be the misparse.

5- The "rating_numerator" also has some incorrect values.

6- Some dog name are incorrect.

7- Extra characters after '&' in twitter_archive_clean['text'].

8- The columns' names are not clear and straightforward such as p1,p2.

Data cleaning

Define issue #1 (Tidiness)

Combining dog types in one column as one variable and filling the nan values.

Define issue 2 (Tidiness)

Add tweet_info and image predictions to twitter_archive table.

Define issue 1 (Quality)

Remove all the unnecessary columns directly

Define issue 2 (Quality)

Change the datatype of 'timestamp' to datetime

Define issue 3 (Quality)

Remove the records with no images information

Define issue 4 (Quality)

10 is the default value of ‘rating_denominator’, then correct the wrong values based on the corresponding text information.

Define issue 5 (Quality)

Correct the 'rating_numerator' values from the text information

Define issue 6 (Quality)

Change the frequent incorrect dog name to None

Define issue 7 (Quality)

Removing extra characters after '&' in text column

Define issue 8 (Quality)

Change the column names for better readability in image_predictions.

After cleaning our data we store it in csv file twitter_archive_clean.

Conclusion: Data cleaning and analyzing is an interesting process and without doing it properly we wouldn’t find interesting from our models. The process requires thinking about our data and viewing it from different perspectives to be able to clean it properly and reach interesting facts.