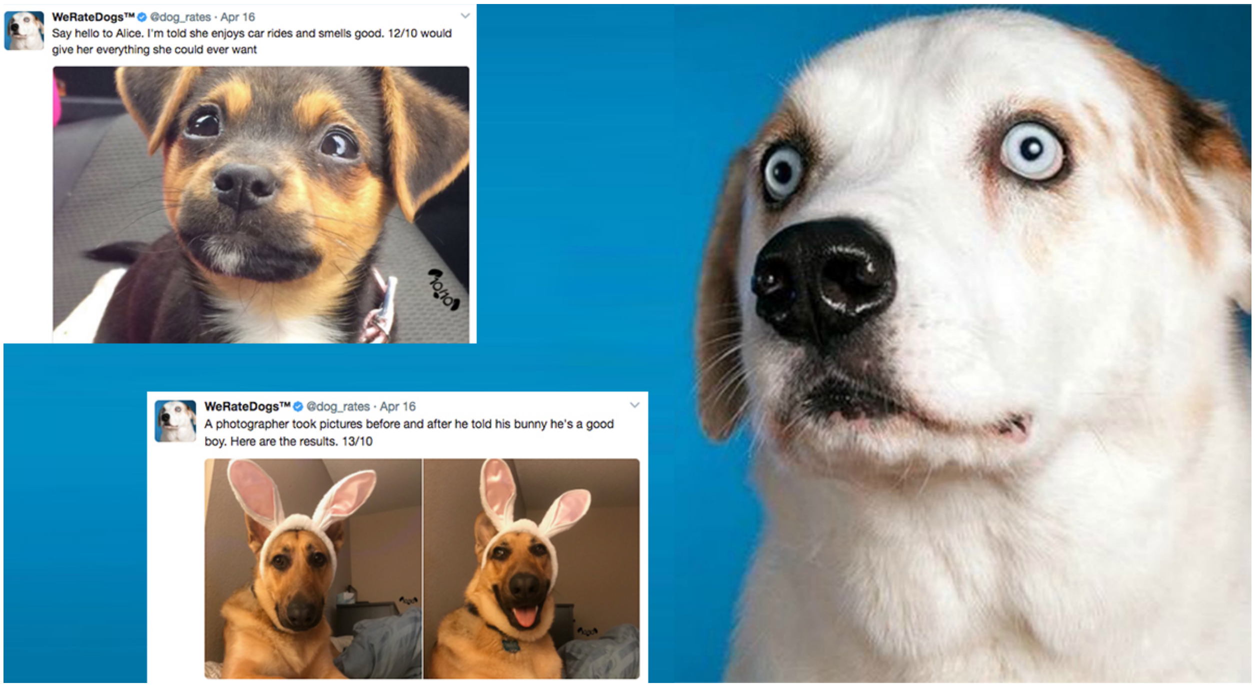
Wrangling Report



WeRateDogs™ @dog_rates · Apr 16
Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Introduction

The report briefly describes the wrangling effort that is done for the ”WeRateDogs” dataset. The dataset is 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. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why?

The work is done in the following steps:

1. Gathering Data
2. Assessing Data
3. Cleaning Data
4. Store Data
5. Analyzing and Visualizing the Data

Gathering Data

I collected the data from three different sources. The “Enhanced Twitter Archive” basically contains the tweet text and some information extracted from this text. The second one is the tweets information itself from Tweeter. I only took two pieces of information from it which are the “retweet count” and “favorite counte”. The last one is the image prediction which contains the output of a neural network that specifies what is in the picture of the tweet; is it dog or not?

I convert each source into an individual dataframe so that I can work with them in a jupyter notebook.

Assessing Data

I assess the data visually and programmatically. The visual assessing is through looking at the complete information in excel. I assess the data programmatically using the panda library methods like: head(), tail(), info(), describe(), value_counts(), duplicated() …

During the work, I documented some issues:

1. The datatype of the column are not correct like for the “timestamp”
2. There is records related to retweet and reply and we should not take them into our analysis
3. The column "source" contains html code
4. The column “text” contains short url links
5. The column “text” contains “&amp;”
6. The following columns contains 'None' value (doggo, floofer, pupper, puppo, name) and have the same information about the “dog_stage”
7. The column name contains “None” value and incorrect names
8. Some column names are not descriptive
9. The “rating_denominator” contains value other than 10
10. The datatype of the column “tweeet_id” is not “object”

Cleaning Data

I was cleaning the data using the strategy of “Define”, “Code” and “Test”. The following are the steps of cleaning:

1. Change the “timestamp” to datetime format.
2. Remove the records related to retweet and reply.
3. Extract the source from the html code in “source” column
4. Extract the short url from the text column
5. Replace the “&amp;” in “text” column with “&”
6. Combine the 4 columns of dogs into only one
7. Replace the “None” value from “name” column with empty data
8. Replace the wrong names with the correct one from the “text” column
9. Adding retweet count and favorite count in the master dataframe
10. Update the column name with a descriptive ones
11. Adding image prediction in the master dataframe

12. Update the “rating_denominator” with the correct ones from the “text” column and remove the records that do not have a denominator equal to ten.
13. Remove the unneeded columns.
14. Change the type of “tweet_id” to “Object”

Store Data

I stored the data in 'twitter_archive_master.csv' file taking into consideration removing the index from it.

Analysing the Data

I created multiple plots and got insights from them.