Wrangle Report

Introduction

Real-world data rarely comes clean. Using Python and its libraries,we will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. we will document our wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that we will be wrangling (and analyzing and visualizing) 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? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Problem in that we solved in the dataset

Tidiness problem

1- In Twitter_Archive dataset, {doggo,floofer,pupper,puppo} are all dog stages so to reduce redundancy I will merge them into one column called dog_stage.

2- There are three separate tables, I will merge them into one table.

Quality problems

1- Change source to have only the type of device, because the source is not clear.

2- Missing value in table Twitter_Archive in in_reply_to_status_id and in_reply_to_user_id, so I drop these columns.

3- in name, there are dogs with "None" instead of null, so python can recognize it as an empty cell.

4- TimeStamp should be in type "DateTime".

5- tweet_id should be string not integer.

6- Rename column p1 and p2 and p3 to prediction so it become more clear.

7- Let numerator and denominator' in one column I called it rating, we don't need it to be separated.

8- We don't need the retweet, so it should be removed