Introduction

Wrangle Report

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

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, a Twitter account that rates people's dogs with a funny comment. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12etc.

Problem in that we solved in the dataset

Quality Issues

1 - Retweets: Some entries are retweets,We want to keep only the original.
2 - Name: replace with 'None' that rows where the value of 'name' is lowercase indicating that it's not an actual name.
3 - wrong format for tweet_id
4- Wrong data type: timestamp should be datetime instead of object.
5 - missing column for the fraction of rating_numerator and rating_denominator
6 - rename function to rename the column 'id' to 'tweet_id'
7- Rename column p1 and p2 and p3 to be clear
8- Missing value in table twitter_Archive in in_reply_to_status_id and in_reply_to_user_id

Tidiness Issues

1- Join image_predictions and df_tweet to twitter_archive.
2- merge one variable in four columns (doggo, floofer, pupper, puppo).