We rate dogs: Wrangling report

By: Abdullah Walied Allama

Data gathering:

The data was gathered from three resources:

1) The WeRateDogs Twitter archive: WeRateDogs downloaded their Twitter archive and sent it to Udacity via email. It contains basic tweet data for all 5000+ of their tweets (tweet ID, timestamp, text, etc.)

2) Retweet count and favorite count have been provided using the twitter API.

3) Every image of the dogs was ran through a neural network that can classify breeds of dogs.

Data assessment and cleaning:

After assessing the data programmatically and visually, the quality issues that were found were:

- Some dogs prediction names are inaccurate such as (box-turtle, hen, water bottle, toilet seat, etc...)
  Those rows were removed
- An unnecessary "0000+" in the end of the timestamp column in df_twitter_archive
  The “0000+” were removed
- Missing data on the stage of a dog (doggo, floofler, pupper, or puppo)
  Nothing can be made about these rows since we do not have another data set to replace those rows
- Rating denominator is equal to 0 in 1 row
  That row was removed
- Rating numerator is equal to 0 in 2 row
  They were removed
- Some values in the rating numerator are too large
  Rows with the numerator larger than 30 were removed
- Some values in the rating denominator are too large
  Rows with the denominator larger than 30 were removed
- "None" present in twitter archive dog stages
  "None" were replaced with nulls
- Retweets are present in the twitter archive data set
  Those rows were dropped
- Join the dog stage columns into one column
- The number of rows in each dataset is different
  Will be solved by merging the three data sets
- Merged the three data sets into one master data frame

- Some columns are not needed in the analysis
  Drop those columns
- For the sake of the analysis, the rows where the rating denominator is not equal to 10 will be dropped and drop the rating denominator column
  - The reason for not removing them in the other data set is because it may be needed there but it is not needed in the master data frame for the sake of the analysis
- Some columns names need to be changed for the sake of accuracy
  The columns names that were changed were:
  (rating_numerator':'rating', 'p1':'breed_prediction
  ’,'p1_conf':'prediction_confident')