Wrangle Report

Mohamed Abojabl

December 2020

1 Introduction: Data Wrangling Steps: Gathering, Assessing, and Cleaning

The data-set that i will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user 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..

2 Gathering Data:

The WeRateDogs Twitter archive. The "twitter archive enhanced". file was provided to Udacity students ("like me"). This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students ("Like me"). Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data I find interesting.

3 Assessing Data:

1- Quality Issues: Data-set twitter enhanced:
- There are unnecessary columns.
- There are retweets.
- 'expanded URLs' column: tweets/ retweets without images
- 'timestamp' is not in date time format
- 'name' has missing data but not Nan, and some names are wrong.
- 'tweet id' should be in type object.

- Some tweets include more than one rating with decimal numbers.
- ‘pupper‘, ‘puppo‘, ‘floofer‘ and ‘doggo‘ columns: For 1976 IDs there is no dog ”stage” information.
- ‘pupper‘, ‘puppo‘, ‘floofer‘ and ‘doggo‘ columns: There are some IDs with more than one dog ”stage” information (two dogs are rated).

2- Data-set twitter: - ‘id‘ should be type object.

3- Dataset image pred:
- Dog breeds are not lower or uppercase always. - ‘tweet id‘ should be type object. - ‘img num‘ column isn’t needed

Tidiness Issues:

- twitter enhanced: Transform the 4 columns ‘dogger‘, ‘floofer‘, ‘pupper‘ and ‘puppo‘ into one variable ‘stage‘.
- image pred: the prediction shall be collected in only one column ‘breed pred‘
- image pred: the prediction confidence should be collected int one column ‘pred confidence‘
- image pred: ‘jpg url‘, ‘breed pred‘ and ‘pred confidence‘ should be in one table ‘twitter enhanced‘.
- twitter: ‘favorite count‘ and ‘retweet count‘ column should be in one table, ‘twitter enhanced‘ Dataset.

4 Cleaning Data:

1- After the assessment, I cleaned the data, and stored it in a csv file ’twitter archive master.’