Udacity
Data analyst Nanodegree program

Wrangle Report

Introduction

This project is part of Udacity data analyst nanodegree program – second term.
It is aiming on wrangling **WeRateDogs** Twitter data to create interesting and trustworthy analyses and visualizations.
And this report is concerned about summarizing all the wrangling of data coming of three different sources, all related to **WeRateDogs** twitter account.
Data wrangling consists of three main steps:

- Gathering data
- Assessing data
- Cleaning data

And this report briefly describes each of them.

Firstly, Gathering data

Our data comes from three different sources:
First source: the twitter archive data downloaded exclusively by **WeRateDogs** for Udacity.
Second source: the image prediction which is a prediction of the breed for each dog in the dataset, it's probably worth mentioning that this data came out of a neural network made by Udacity, and the data can be downloaded programmatically.
Third source: the twitter info, which is also downloaded programmatically using a twitter developer account and can contain as much information as need about the tweets, in this project we were only concerned about re-tweets and favorite counts.

Secondly, Assessing data

Assessing each of our three data paces was based on the same steps, below is a list of them:

- Checking a few number of entries of the data.
- Checking the number of entries, data types and nulls.
- Checking the unique values count for some interesting columns.
- Checking for duplicated content.

- **Finally, cleaning data**

- First Source: Twitter archive data

-Quality

1- The (retweeted_status_timestamp), (timestamp) shall be converted into datetime instead of object.
2- Splitting (timestamp) into three coloumns day - month - year might be useful for future analysis and then dropping it.
3- The (rating_numerator) has some invalid values.
Fixing it by eliminating values greater than 10, and less than twenty.
4- The (rating_denominator) has also invalid values.
Fixing it by setting all the columns values to 10.
5- The (name) column has some invalid data.
Trying to get the name from the original tweet.
6- Keeping only original ratings (no retweets) that have images, and dropping coloumns related to retweeting.
By removing those enteries with information about retweeting.
7- Too little dogs are calssified into (doggo), (floofer), (pupper), (puppo) columns, so we might want to check this.
Trying to get the dog stage from the original tweet.
8- Creating new column called (dog_stage) driven out of (doggo), (floofer), (pupper), (puppo) columns for future analysis and then dropping them.
This can programmatically be done, very easily.

- Second source: Image predictions data

-Quality

1- Keeping only one column for dog breed.
2- Keeping only one column for confidence.
Simply by dropping other rows.

- Third source: Twitter info data

-Quality

Nothing much to do here.

- ## Tidiness

1- Number of entries in the dog's predictions data (2075) is less than our main data frame (2356). This must be taken care of while merging (This step will be done after combining all data sets).
2- No need to take all the data in, since we won't be using all of them, so at the (analyses and visualize) phase, we will just drop them.
3- Just renaming (id) column into (tweet_id) for compatibility.

4- All the data parts must be joined together in one data frame.