Data Wrangling Project – WeRateDogs

By: Khaled Salem

1. Introduction

This is a report that describes briefly and concisely efforts done in the Jupyter Notebook file regarding Udacity's data wrangling project. The data wrangling project's main aim is to gather, assess and clean WeRateDogs archived data from Twitter.

2. Data Gathering

Data was gathered from multiple sources and in uploaded using a variety of tools. The three files used were:

- Enhanced Twitter archive: this is given filtered data containing basic information such as timestamp and dog names which are extracted from tweets; however, these records were dirty and messy and hence required improvement and refinement. This was uploaded by using pandas module function read_csv.
- Image predictions file: images found attached as urls in tweets of which the majority shows dogs was run in a developed algorithm that uses neural networks to classify what breed of dog is shown in tweet was programmatically downloaded using the source url using requests library.
- API file: tweet's json data was downloaded by querying the Twitter API using tweepy library.

All three files were loaded into notebook as archive_df, image_predictions and api_df respectively.

3. Data Assessment

Firstly and before starting with any cleaning effort, data was assessed both visually and programmatically. The former was done by looking at the entire dataset while the latter was done through coded commands in the notebook. Expectedly, there were noticeable quality and tidiness issues found in all three datasets.

Quality (content) issues highlighted:

- id column data type is in int format
- Timestamp as object (string format) in archive_df
- Dog names are inconsistent with some having uppercase while others are in lowercase as starting letter
- Missing data misrepresented as None in name column in archive_df
- Zero values in numerator and denominator ratings
- Inconsistent denominator rating values

- There are four sources found in source column in archive_df dataframe
- 181 retweets present in archive_df
- Missing values in expanded_urls in archive_df indicate missing images

Tidiness (structural) issues highlighted:

- Several unnecassary columns in api_df dataframe
- Retweet columns in archive_df are no longer required
- Four columns which doggo, floofer, pupper and puppo can be transformed with an identifier column
- All dataframes should be merged into one master dataset

4. Data Cleaning

Issues highlighted in the preceding step were addressed in the data cleaning step in the same order.

Addressing quality issues:

- ID data type was converted from int to string in all dataframes
- Timestamp data type was converted from string/object to datetime
- Lowercase entries in dog names denoted regular words and was replaced by None
- Missing data misrepresented as None in name column needs is addressed and converted to NaN
- Any numerator rating or denominator rating equal to zero was removed
- Denominator rating standardized to a value of 10 by dropping all values that are not equal to 10
- Only four type of sources in source column extracted from html
- Any retweet rows were removed from archive_df
- Missing values in expanded_urls column signifies missing images in tweets therefore was removed

Addressing tidiness issues:

- Unnecessary columns dropped in api_df and id column renamed
- Retweet columns in archive_df dropped
- Dog stages were classified and melted into single dog_classification column
- All dataframes merged creating one master dataset with all relevant columns in it