WeRateDogs Twitter Archive - Wrangle Report

Introduction:

The aim of the project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. This report contains the steps performed and with a brief summary of the data wrangling efforts in this project.

The key steps performed in this project were:

1. Data wrangling
    - Gathering data
    - Assessing data
    - Cleaning data
2. Storing the data
3. Analyzing and visualizing the data (not covered in this report)
4. Reporting
    - Wrangling efforts (current report)
    - Analyses and visualisations (not covered in this report)

Data Wrangling:

Real-world data rarely comes clean. Using Python and its libraries, I have gathered data from a variety of sources and in a variety of formats, assessed its quality and tidiness, then cleaned it, so it can be used for analyses and visualisation.

Gathering data:

The data for this project has been gathered from three different sources.

1) Twitter archive data (file: twitter_archive_enhanced.csv)

   This file has been downloaded manually from Udacity website and read the data into pandas dataframe. The file contains the information such as twee id, the dog's rating and its stage along with some additional information like timestamp, urls, source etc.

2) Tweet image predictions (file: image_predictions.tsv)

   The file has been downloaded programmatically using Requests library, the download link is provided by Udacity. The file contains the tweet id and dog breed predictions data from 3 different prediction sources.

3) Twitter API additional data (File: tweet_json.txt)

   This file has been downloaded manually from Udacity website and read the data into pandas dataframe. The file contains the information such as twee id, retweet count and favourite count etc.

Assessing data:

After gathering all the required data from three different resources, the data has been assessed visually and programmatically for quality and tidiness issues. Identified at least eight quality issues and two tidiness issues in the data gathered.

The visual and programmatic assessment showed several quality and tidiness issues such as redundant rows, columns, incorrect data type, very suspicious dogs rating, multiple columns with similar information where it is possible to present the information in one single column and data in multiple files than in a one single file etc.

Cleaning data:

The data has been cleaned for the quality and tidiness identified during the assessment phase. The data cleaning has been done in three steps Define, Code and Test.

Before cleaning the data, copies of the original dataframes have been created, so the original data will always remain available and if there are any issues in cleaning process the original data can be still accessed.

Twitter archive data (file: twitter_archive_enhanced.csv):

- The dog stages 4 columns 'doggo', 'floofer', 'pupper', 'puppo' has been converted to one single column 'stage' and dropped these 4 columns.
- Removed the redundant data: Dropped all rows containing retweets where these columns will be non-null and then dropped the retweet columns. Similarly dropped all rows containing ‘in_reply_to’ entries where these columns will be non-null and then dropped the ‘in_reply_to’ columns.
- Dropped all rows with missing data in 'expanded_urls' column
- Dropped all rows with 'rating_denominator' coulmn is not equal to 10 and also the rows with ‘rating_numerator' value greater than 30.
- Changed the 'timestamp' to datetime data type

Twitter API additional data (File: tweet_json.txt):

- Merged the 'retweet_count' and 'favorite_count' columns from the Twitter API additional data to Twitter archive data and removed any rows with missing information and dropped the redundant columns.

Tweet image predictions (file: image_predictions.tsv)

- Merged the Tweet image predictions dataframe to Twitter archive data and removed any rows with missing information and dropped the redundant columns.

- Created a single column 'Breed' covering the breed of the dog (from p1, p2, p3) and another column 'Confidence' covering the highest confidence from p1_conf, p2_conf, p3_conf and dropped the redundant columns

Storing the data:

A single dataframe was created by cleaning the quality and tidiness issues and was saved to a new ‘twitter_archive_master.csv’ file.