Wrangling and Analyze the tweets of We Rate Dogs twitter account

This report summarize the data wrangling process of WeRateDogs twitter account data.

Introduction:

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. Data splitted in three tables: tweets archive, tweets image predictions and tweets counts.

Tweets Archive table:

It contains basic tweet characteristics, some of them are: tweet id, time stamp, text, dog rating, dog name and dog stage.

Image Predictions table:

Every dog image ran through a neural network that can classify the dog breed, for each image there is three predictions with confidence for each prediction and a boolean expression indicates weather this prediction is a breed of a dog or not.

These data were gathered in image predictions table along with tweet id and image url.

Tweet Counts table:

It contains main counts of each tweet: retweet count and favorite count.

1- Data Gathering:

The datasets for this project exists in three different source, each source have its own way to obtain the dataset from it.

Tweets Archive dataset:

This dataset is given as a csv file called twitter_archive_enhanced.csv which can be downloaded manually.

Image Predictions dataset:

This dataset is hosted on Udacity's servers, it requested using Python Requests library and its response had to be converted to a human readable pandas data frame.

Tweets Counts dataset:

This dataset obtained using Twitter API and Tweepy library, it returns the data in JSON format and it converted to pandas dataframe using json library, iteration and append concepts.

2- Data Assessing:

The assessment of data were done both visually and programmatically, some quality issues were found such as missing urls, erroneous data types, wrong dog names and incorrect ratings.

Tidiness issues were also found such as dog stage splitted in multiple columns because it’s inconsistent with the tidy data principle: Each variable must have its own column.

Tweets counts columns also treated as a tidiness issue because there is no need to be in a different table rather than tweets archive table.

3- Data Cleaning:

The cleaning of data were divided into three section for each issue: define, code and test.

Define is a brief description of how we will solve issue, Code is the actual code for solving the issue and test is done to ensure that the issue were solved correctly.