WRANGLE & ANALYZE DATA

Iintroduction:

Real-world data rarely come clean. The data is an account in twitter known as WeRateDogs its user @dogs_rates, this account rating dogs using people comments by uploading photos of the dogs and let the people rating it by retweet or favorites, We will use Python and its libraries, we will gather data from a variety of sources, assess it by the lock for a quality and tidiness issues, then clean it. This is called data wrangling. we will document our wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The Data wrangling as it below:

- Gathering data
- Assessing data
- Cleaning data

And then Storing, analysing, and visualizing

Gathering data:

We gathered the data from these three sources:

1. The WeRateDogs Twitter archive. we Download this file (twitter_archive_enhanced.csv) was provided by Udacity to students

2. 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 (image_predictions.tsv) is hosted on Udacity's servers, we downloaded programmatically using the Requests library

3. And this file tweet-json.txt provided by Udacity, we should use Twitter API and Tweepy library, but we could not get the consumer and secret keys from twitter developer.

Assessing data:

In these files, we find quality issues and Tidiness issues it declared below :

Quality

In twitter archive data frame, we find that many columns useless and the data type of the columns not logical or reasonable (e.g. datatype of timestamp object) and unreadable in column names and source

image

we found here many missing data and the same problem we faced in twitter archive data frame the Datatype issue

Tidiness:

Three related taples should be in one table as one data Frame also in twitter archive data frame has 4 columns should be in one column to achieve tidy data requirements which is:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table

Cleaning data

In this step we solve and fixing the issues that we find it in the assessing step programmatically using library's, also we following this iterate (Define, code and test) in the first we copy the data frames and then using this iterate to clean the data

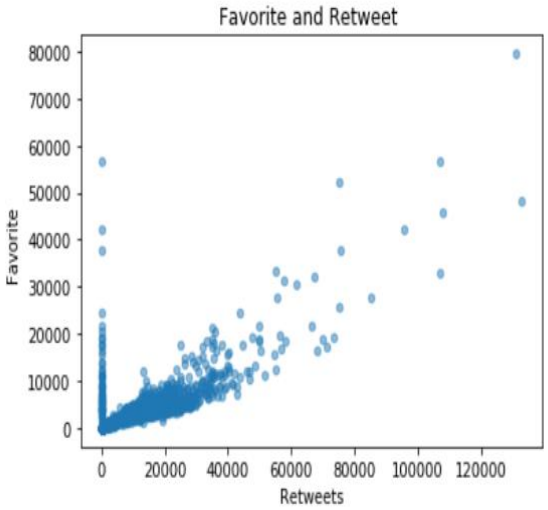
Storing, analysing, and visualizing

After finishing the clean step and completely no issue is there then we stored the data in a file named twitter_archive_master to commit our work there.

Then we ask sample questions for analysis and visualize.

Q1: the Favorite and retweet relation

positive correlation and that what expected as long as the people love the pic as long, they will talk about it And express their feelings



Q2: what is The popular 10 dogs by favorites

This is the cute dog that have the most of like Unfortunately we don't know his name

Q3: the dog has the most People Talk

and there is the dog that has the most people talk maybe because of his talent for swimming