Data Wrangling, Analysis and Visualization for WeRateDogs Dataset

By Ahmed M.Khalifa

Data wrangling is a core skill that everyone who works with data should be familiar with since so much of the world's data isn't clean. It's about gathering the data, accessing it and cleaning the unwanted data also fixing Issues

In this Project , we wrangling the WeRateDogs Dataset which is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

Step 1: Gathering the data

In this step, i have gathered data from Udacity "twitter_archive_enhanced.csv" through manual download and 'image_predictions.tsv' through programmatically. Also have gathered data from twitter api.

Step 2: Accessing data

In this step, Accessing data it's about discover the dataset with some functions like head(), info(),.... and detect the issues in Quality or Tidiness like i detected in this dataset which:

Quality

1.twitter_archive table had some columns have missing values, "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp".
2.image table had tweet_id should be an object but its shown as integer
3.tweet table had retweet_count and favorite_count should be an integer, its shown to be an object
4.In twitter_archive table , Some names are incorrect.
5.Column headers are not clear in image table.
6.The datatype of "timestamp" in twitter_archive table is object , its should have be datetime.
7.Some names in twitter_archive table start with lowercase letters.

Tideness

1.Columns 'doggo', 'floofer', 'pupper', 'puppo' in tweet table should be in 1 column.

2.The tweet table need to merge into the twitter_archive table.

Step 3: Cleaning

In this step, the Issues that detected are fixed one by one. The step included defining the problem, coding and testing it to know if the problem was fixed.

Step 4: Analysis and Visualisation

Once the Issues were fixed, i did some basic analysis like Popular type of dogs and ***Distribution*** of source devices , Also visualise those results