Data wrangling report

By: Eslam Ali Abou-shashaa
Dec. 2020
As an assignment for the Udacity Data Analysis Nanodegree; This report illustrates the main steps involved in the data wrangling of Twitter account "WeRateDog"

Data Gathering

in this step collecting data takes place. For this project, there were three main sources for the data ro feal with:
1-Twitter_aechive_enhanced.csv file this file was delivered by email and downloaded manually to our woking directory and then imported into our working environment using pandas function "pd.read_csv"
2-Image_prediction.tsv is the second file that has been hosted in a webpage and downloaded from its relevant URL using the requests library get function and pd.read_csv pandas function . this file encompassed image predictions for the dogs'breads obtained through a neural network on most of the tweets in the archive file
3- the final dataset was gathered from tweet-json.txt after then use json library to use file and use loop to read line by line in file to extract retweets count and favorite count

Assess

this step we use two type from Assess
1-visual this type i use google sheet to see all data and select issues
2-programmatic this type use functions like info,describe,value_counts
function info this function merged information like type of data in columns and number of None in columns
function value_counts this function appear number of character in columns
function unique this function appear all character in specially columns
from above step we find two type from issues
quality issues like some name don't correct like a
and some data is missing like data in in_reply_to_status_id , in_reply_to_user_id ,name , retweeted_status_id ,retweeted_status_user_id and retweeted_status_timestamp
timestamp is object should be data time
most rating is over rating

cleaning

we should cleaning this data to analysis it
let's cleaning the data
first remove all columns which don't have data like in_reply_to_s ,retweeted_status_timestamp to remove this columns we use drop function in pandas
convert timestamp to data time used is operation to_datetime function in pandas
make new columns to dog type and remove "puppo, doggo , floofer, poppur" columns
make the three table is one table