Wrangle & Analyze WeRateDogs Data

The goal of this project is apply skills that I have learned in udacity course of wrangling data which is a part from udacity nanodegree of data analysis.

I will use data from twitter account called weRateDogs , this account is for rating people dogs from.

Project outline

• Gathering data

Collecting data from 3 sources Twitter archive file, tweet image predictions and Twitter API File using python libraries for extracting data

• Assessing data

Discovering data issues by displaying it visually or programmatically, tidiness and quality .

• Cleaning data

Applying cleaning methodologies of python by dropping columns, merging files, filling.

twitter_archive:

this file(twitter-archive-enhanced.csv)was downloaded from Udacity , and it contains the following columns tweet id timestamp and any other information.

Assessing data

Tidincess and quality Issues in twitter_archive:

1-retweeted_status_timestamp is type 'object'

2-doggo, floofer, pupper, puppo have string none value it should be NaN

3- there are a lot of names that don't make sense such as 'a', 'all', 'old', 'infuriating', 'the' also 'None'

I noticed that programmatically also I have noticed that by MS Excel i found 'not' , 'an'

4-tweet id need to be string

Cleaning data

1. **Converting it to datatype object**
2. **I replaced all none values by NaN using replace method**
3. **Converting tweet id to type string because later I have to merge the files so it has to be the same type not integer.**

Image predictions

Tidincess and quality Issues:

1-there are 66 duplicated urls

2-Image predictions dataset should be emerged with twitter_archive

3- tweet id need change it to string

4- create new columns named dog_types

cleaning

1- drop duplicated urls
2- I have created new columns called dog_types
3- Change tweet id to string

twitter_api:

I didn't find issues in this file so just converting tweet id to string.