WeRateDogs project

Introduction

The project is about data in a twitter profile that its main duty to rate dogs of the fans, this page give rating for the dogs more than 10 from 10 witch is so amazing for the dogs lovers>

Project stages

Gathering data:

There was three sources of the data used in the project

1- The tweets archive witch was downloaded from the classroom
2- The image prediction file that was made by neural networks, and this file was downloaded programmatically
3- The twitter json file wish was planed to be downloaded programmatically by the twitter API, but unfortunately the developer account wasn't verified yet so I was have to make the project with the file in the classroom

Assessing data:

The assessing stage was on two types to assess visually & by pandas functions

The first one was by using Microsoft excel to see the data and scroll through it to see its composition and notice as the eye can detect errors in it

Then it was the pandas functions rule witch would detect issues as follows

Tidiness issues:

1- I noticed that there was four columns that by logic have to be merged in one column, these columns are (doggo ,floofer, pupper, puppo) as they are originally values not variables
2- The three datasets have to be combined in one dataset

Quality issues

1- Some columns have missing values
2- The timestamp wasn't made as datatime but as object
3- Name contain a lot of invalid names
4- The dogs names aren't consistent. A lot of differs in formatting lower/ upper case
5- Some numerator incorrect values
6- It was terrible to understand what is p1, p2 in the beginning
7- Filter the sources to more easier to read by extracting it from the links text
8- Retweets exists witch make the data inaccurate

Cleaning data:

- Making a copy of the data
- Merged the three datasets in one dataset
- Removing the unwanted columns
- Merging 4 dog stages in one column with a funny name (doggy meter)

Saving the dataset