WeRateDogs

Wrangling data report

Gathering data

Data for Analysis was gathered from different sources and with different methods:

1. getting data programmatically by given url link and with help of Python library request
2. getting data via API (Twitter API) and by storing them in DataFrame.
3. getting by reading prepared file from file.csv source

Assessing data

While data obtained via Twitter API and via url were clean and ready for use, data archive in file.csv format were raw and required cleaning.

Assessing data was done by both, manual check up and programatically while using Python methods.

Assessing data

Were identified 8 quality issues and 2 tidiness issues:

quality issues of particular columns::

- name: some names are not defined and some consist of 1-2 letters articles (a,an,the ..) or random word(such,..)
- name: has None values which are not recognized as missing values
- name: some Names are missing
- doggo, floofer,pupper,puppo: includes None when it is not True. Better to represent in boolean type
- timestamp: object type instead of date type
- source,text: object type instead of string type
- text: object type instead of string type
- retweeted_status_timestamp: object type instead of date type
- number of columns has missing values

tidiness issues:

- source: consists of unnecessary info
- timestamp is presented in format which hard to work with, better separate day from time

Cleaning data

Cleaning was done with help of different Python methods and functions to prepare data for a further Analysis

Storing data

Cleaned data were stored in separate DataFrame and written in separated file.csv for a further Analysis.

After that explorative Analysis has been done, list of questions to answer formulated and Analysis performed.