@WeRateDogs

Wrangle report

Twitter Data Project

Data wrangling steps

1. Gathering
2. Assessing
3. Cleaning

—

Gathering

For this project 3 files were needed :

1- WerateDogs twitter archive CSV file was given from Udacity, manually downloadable, then uploaded on the working space.

2- Tweet Image prediction TSV file, which is hosted on Udacity servers, was downloaded using the **Requests** library with a pre-given Url from Udacity.

3- downloaded a JSON file from Twitter, using twitter API query then looping and extracting project related data from the file, storing it as new Dataframe.

—

Assessing

Each file was read and opened, I queried the file in different ways using: .head() - .info() - .describe() - .value_counts()
Came across many issues which were listed as follows :

Quality issues

- tweet_ids should be changes into a String
- timestamp is not defined as date
- Wrong and missing Dog names, (None, a,an) change these to no name
- Drop tweets with no images there are 2075 images and 2356 Tweets
- img_num should be changed into a string
- delete retweets and semi empty columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
- remove column (doggo,floofer,pupper, puppo)
- Clean the content of the source column, make it more readable.

Tidiness issues

- change multi-column (doggo,floofer,pupper, puppo) into 1 column
- merge all tables into a 1 dataframe

—

Cleaning

Ech assessed issue was addressed within a 3 step process defining, cleaning and testing with docstring explaining the codes.

Quality issues :

- **Tweet ids should be changes into strings**

  Define:

  - changeing type using astype.(str)

  Test

  - calling dtype() to check if type has changed

- **df timestamp should be changed into datetime**

  Define

  - changing using pd.to_datetime() in format of %Y=year, %m= month, %d = day

  Test

  - we can check the datatype using type method

- **Wrong and missing Dog names, (None, a,an...etc) changing so they will be treated as one group.**

  Define:

  - changing a list on wrong names into 'no name' using replace()

  Test

  - check if names replaced - still exist

- **changing img_num from an int into a string**

  ### Define

  - changing type using .astype(str)

  ### Test

  - we can check the datatype using type method

- **deleting retweets and semi empty columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp**

  ### Define:

  - we can delete any column using .DROP() method

  ### Test

  - call info() to check change

- **remove column (doggo,floofer,pupper, puppo)**

  ### Define:

  - remove using drop()

  ### Test

  - call info to check all columns

- **Clean the content of the source column, make it more readable.**

  ### Define:

  - extracting the source using regex.findall in a lambda function

  ### Test

  - check using value_sounts()

- **changing img_num from an int into a string**

  ### Define

  - changing type using .astype(str)

  ### Test

  - we can check the datatype using type method

—

Tideness issues :

- **create 1 new columns to have all dog stages : doggo,floofer, pupper, puppo**

  Define:

  - create 1 new column with the 4 variables, and combine using Numpy. make None values as empty

  Test

  - check change using value_counts()

- **merge all 3 data frames into 1 dataframe "df2" : df_copy , image_df_copy , tweets_api_copy**

  Define:

  - merge all using 'merge()' on tweet_id in an inner join

  Test

  - showing the new dataframe

—

Storing data

The final merged data frame was stored into as CSV file named twitter_archive_master.csv

Analyzing and Visualizing Data

3 visualizations were made, used in insights using matplotlib library and different plots.

- check see the correlation, if any, between the retweets and favorite count using a scatter plot
- Visualizing the count of life stages for each dog stage, using a pie chart with percentile
- showing the used source with the most impact "retweets"