wrangle_report

February 13, 2019

1 Wrangle Report

By Abdullah Saud AlFawwaz

1.1 Gathering Data

Data is gathered from 3 resources:

- Gathering data from file on hand:

  ```
  Read data from existing file twitter-archive-enhanced.csv by using pd.read_csv and store it
  ```

- Download file by URL:

  ```
  Download file image_prediction.tsv programmatically from the Internet and store data as "im
  ```

- Gathering data from twitter API using Python's Tweepy library:

  ```
  Read data from twitter API using Pythons Tweepy and store data as "json_tweets".
  ```

1.2 Assess Data

1.2.1 Quality

1- Missing values from images dataset which are 2075 rows and the archive are 2356.
2- Retweets need to be removed
3- Timestamp should be datetime instead of object (string)
4- There are same jpg_url for more than one tweet_ids
5- Incorrect dog names, The most popular name is 'a' which is not corrected name.
6- Drop some columns which is not usefull for analysis.
7- Some columns need to rename it to understandable name.
8- Nulls represented as 'None' in columns 'name', 'doggo', 'floofer', 'pupper','puppo'.
9- Some tweets have additional characters in the text which is not meaningful.
10- The numerator and denominator columns have invalid values

1.2.2 Tidiness

1- Dog stage variable in different columns: doggo, floofer, pupper, puppo
2- Marge 'json_tweets' and 'image' to 'archive' dataset

1.3 Cleaning

Before Cleaning the data I make copy for the datasets to save the orginal data and clean the copies.

1.3.1 Difines:

- Delete the duplicated jpg_url images>
- Delete the tweets which is haven't image from the dataset
- Delete retweets from the dataset
- Convert timestamp to datetime data type.
- Delete these columns which is not usefull for analysis.
- Rename the columns to understandable name.
- Marge the cleaned datasets together
- Melt columnes ['doggo', 'floofer', 'pupper', 'puppo'] under stage column.
- Replace incorrect name to Nan.
- Replace None to Nan in the datasets.
- Remove the Additional characters in tweet text.
- Reorder the columns and drop the additional columns to make the dataset easy to read.

1.4 Storing

```
Storing the cleaned master data as "twitter_archive_master.csv
```

```
In [ ]:
```