wrangle_report

Lanmixue Mao April 1, 2018

1 Gather

Gathered data from three different sources:

1) twitter-archive-enhanced.csv

This dataset was downloaded from a link in Udacity Data Analyst Nanodegree data wrangling project portal.
This dataset has been read by pandas into a dataframe called 'tweet'.

2) image_predictions.tsv

This dataset was collected by using Requests library. This was downloaded programmatically from the URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
When read this TSV file via pandas read_csv function, it has been separated by '\t'.
This dataset has been read by pandas into a dataframe called 'image'.

3) tweet_json.txt

This dataset was downloaded by using Python's Tweepy library. This tweet's JSON data matched the tweet id which get from image_predictions.tsv to get the rest retweet_count and favorite_count.
Used the timer from timeit to calculate the downloading time.
Read this TXT file by using pandas read_csv function into a dataframe called 'info', and seprated by " ".

2 Assess

Assessed the tweet, image and info datasets by using head(), info(), describe(), value_counts(), duplicated(), sort_values(), and isnull() functions.

The datasets have eight main quality issues:

1) Missing data in tweet dataset: expanded_urls, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
2) Erroneous datatype in tweet dataset: retweeted_status_timestamp, timestamp, text, and source columns
3) Incorrect dog name, such as 'a, an, actually, by, his, my, one, the, very'
4) Multiple records in jpg_url in image dataset
5) Inaccurate rating_numerator and rating denominator value
6) Missing data in image dataset (2075 instead of 2356)
7) Missing data in info dataset (2069 instead of 2075)
8) Lowercase image dataset p1, p2, p3 columns and tweet dataset name column

The datasets have two mian tidiness issues:

1) doggo, pupper, puppo, and floofer columns can be merged into one column
2) info table and image table should be part of the tweet table

3 Clean

According to quality and tidiness issues assessed above, cleaned the dataset and exported to a CSV file named “twitter_archive_master.csv”.

Quality:

1) Remove replied tweets and retweets
2) Change tweet_id to int data type, Change timestamp to date data type, change text to string data type, change source to category data type
3) Filter the text column to check if 'a', 'an', etc have corresponding name. If not, change the name to ‘None’
4) Find the duplicates in jpg_url, remove the duplicate rows
5) Calculate the rating based on rating_numerator divide rating_denominator, remove the rating greater than 2.0
6) Remove tweets that don't have image url
7) Remove tweets that don’t have corresponding tweet id
8) Lower case name column in 'tweet' table and p1, p2, p3 columns in ‘image’ table

Tidiness:

1) Concatenate the doggo, floofer, pupper, and puppo columns to a breed column, drop the doggo, floofer, pupper, and puppo columns
2) Merge the info table and image table to the treatments table, joining on tweet_id