wrangle_report

Gathering Data There are three main pieces of data involved in the wrangling process:

1. A twitter archive dataset from WeRateDogs in CSV format (provided and was added to workspace manually, called 'twitter-archive-enhanced-2.csv'). It contains tweet information like tweet id, timestamp, tweet text, source, dog rating numerator and denominator that were extracted from the tweet texts, dog stages, etc. It was read into dataframe 'twitter_arc'.

2. Dog breed or other object prediction information based on each tweet's images running through a neural network, in a TSV file format to download programmatically from a URL, It was read into dataframe 'image_predictions'.

3. Each tweet's retweet and favorite counts as well as tweet text display range indexes extracted from tweet JSON data by calling the Twitter python API Tweepy. (Tweet entire JSON data was saved as 'tweet_json.txt', subset dataframe created with the needed info is 'tweet_json'.)

Assessing

Visual Assessing: Opened the twitter-archive-enhanced.csv in Excel to do some initial visual

Programmatic Assessing: Used several panda's methods and functions like .info(), .describe(), .value_counts ….etc. to know more about data and extract quality and tidiness issues

Cleaning Data

This part of the data wrangling was divided in three parts: Define, code and test the code. These three steps were on each of the issues described in the assess section.

First and very helpful step was to create a copy of the three original dataframes. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original.

-quality issues

1.twitter_archive_enh contains retweets and reply, not all are original tweets.

Define : keep original ratings (no retweets,no reply) that have images

2.twitter_archive_enh has unused columns, i.e. 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'.

Define : Delete columns that won't be used for analysis

3.timestamp column is str instead of datetime

define : replace column datatype from str to datetime

4.source formate

define : source column is html tag <a> we can extract the source of the tweet

5.name has values that are the string "None" instead of NaN

6.Looking programmatically, some names are inaccurate such as "a", "an", "the", "very", "by", etc

define : filter out all non-capitalized names and assign to NaN, update "None" names to NaN as well

7.The rating_numerator column should of type float and also it should be correctly extracted.

Define : Extract rating scores correctly from tweet text and convert it to float

8. Remove values other than 10 for rating_denominator

define: Remove values other than 10 for rating_denominator

tidiness issues

1.Variable dog stage are in 4 columns instead of one in twitter_archive_enh.

define : extract dog_stage from text and remove une used stages columns

2.All tables should be part of one dataset

define : All three tables/dataframes are describing each tweet's info thus should be combined to one.

Output

one dataset’tweet_master.’.