wrangle_report

November 1, 2020

1 Wrangle report

1.1 1. Gathering Data

First step was downloading all the necessary data. In the notebook all the data was downloaded by request library from Udayitie's server. The *twitter-archive-enhanced.csv* file was downloaded from [https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv] and is about 915692 bytes. The *image-predictions.tsv* file was downloaded from [https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv] and is about 335079 bytes. The extended twitter archive file was also downloaded from udacity, since after almost two weeks I still didn't get access from Twitter to it's API. The *tweets.txt* file was downloaded from [https://video.udacity-data.com/topher/2018/November/5be5fb7d_tweet-json/tweet-json.txt] and is 10609234 bytes.

Each file was read into a dataframe with the pandas library.

1.2 2. Assessing Data

Data was assessed by using the pandas dataframe methods like head(), tail(), sample(), info(), isnull(), value_counts(), apply() and slicing the datframe. Here are some but by far not all issues: ### Quality: **twitter archive** - columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp contain NaN values or missing values - columns doggo, floofer, pupper, puppo contains sometimes a valid classification and sometimes all None - column name has sometimes None value or a name for the dog - column img_num contains sometimes a value other than 1, column doesn't have any information - column timestamp seems to have a consistent format, but datatype might be better in datetime

predictions - columns p1, p2, p3 contains sometimes uppercase beginnings - some entries are not predictions of of dogs

twitter API - columns in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, in_reply_to_screen_name, retweeted_status, quoted_status_id, quoted_status_id_str, quoted_status, geo, coordinates, place, contributors, possibly_sensitive, possibly_sensitive_appealable contain NaN or missing values - display_text_range contains the start and end_value - column retweeted is never True, favorited only 8 times - column lang contains undefined language abbreviations (und, eu), likely retweets …

all - tweet_id or id is in integer, for joining good, string is better after that

1.2.1 Tidiness:

twitter archive - column source has HTML tags with an URL and the device description

predictions - there are three prediction, but only one (the most likely one) is necessary - all dataframes should be merged, hence it's all about tweets about dogs

general issue - some features are twice available and also implicitly given by other features

1.3 3. Cleaning Data

The first thing was joining all three dataframe on the twitter_id. After that it was an iterative approach of all listed issues. Sometimes other issues were found and fixed adhoc (some columns contained only ONE unique value). Also by fixing one issue, two were sometimes solved. Also manual change was sometimes neccessary (a dog name had to changed manually). Final step included removing all columns that didn't have any information or were collections of already available data in the dataframe.

1.4 4. Storing and Insight

After finishing the cleaning step, the dataframe was saved by the dataframe method to_csv() as *twitter_archive_master.csv* Followed by reading this file in a dataframe and answering and also visualizing the questions: - What devices do users have? - What are the top 10 most favorite tweets? - What are the top 10 most retweeted tweets? - What day of the week and time of the day to people tweet the most? - What are the top 10 dog names (despite noname)?

[ ]: