wrangle_report

June 27, 2022

0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

INTRODUCTION

During the process of Data Wrangling of tweet data from the Twitter user WeRateDogs, I have found several issues in the Quality and Tidiness of data collected by different means.

I have analyzed, cleaned and combined all the data into a new DataFrame and stored it in twitter_archive_master.csv file.

The project aims to gather data provided, to create analysis about the tweets and the predicted dog's breed.

The data wrangling procedure is as follows: 1. Gathering data 2. Assessing data 3. Cleaning data

1. Gathering Data I have gathered the files twitter_archive_enhanced.csv and image_predictions.tsv, which are provided by Udacity in the lesson: Data Wrangling

The twitter_archive_enhanced.csv file contains basic tweet data (tweet ID, timestamp, text, etc.) for 2356 of their tweets as they stood on August 1, 2017. As I needed further information from the WeRateDogs user, I gathered data the text_json.txt for the above mentioned period (querying by tweet_id present in twitter_archive_enhanced.csv)

The gathered data are loaded into three different DataFrame,namely: df1 = twitter_archive_enhanced.csv df2 = image_predictions.tsv df = tweet.json

2. Assessing Data The two types of Data Assessment performed,

Visual assessment: Each dataset is displayed in the Jupyter Notebook for visual assessment. I also used Excel worksheets. Programmatic assessment: Used the pandas function.

3. Data Cleaning I made acopy of each piece of data using .copy method. The reason is so that i could still view the original drty and messy datasets. They are:

df1_clean = df1.copy() df2_clean = df2.copy() df_clean = df.copy()

Further steps i managed to accomplish is as follows: Quality issues 1.Underscore ” present instead of space in dog breeds (p1,p2,p3) and Few names with '-' present retweetfavorite_count table

2.There are unintrested columns : retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_user_id

3.Erroneous datatypes : tweet_id and timestamp

4.Consolidation of dog style column into one

5.Droping not needed columns i.e. expanded_urls in df1 and img_num in df2 6.Rating denominator 7.Replacing doubtful words 8. Completing tag present instead of source name in source

In Conclusion, I have stored the wrangled data in twitter_archive_master.csv file ready for Data Analysis.