wrangle_report

August 1, 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.

The first step of the wrangling process is gathering data. I have obtained the data of the project through three different resources: 1. The WeRateDogs twitter archive: I downloaded this data through a file that called “twitter-archive-enhanced.csv”, from the resources section to our environment using pandas library. And then initialize it to data frame named “twitter_arc”. 2. The tweet image prediction file: It’s a table of information that include a data about the images of the tweets. Also running these images of the dogs in an image prediction model (neural network) to predict the breed of the dog. I downloaded this data from a web link using request library. And then I wrote the content of this file in file called “image_predictions.tsv”. Then upload it. 3. Twitter API: Using the tweet ids in the WeRateDogs Twitter archive, We queried the twitter API for each tweet’s json data using python’s tweepy library. And we take retweet count and the favorite count from the tweet’s data. Then we store each tweet’s json data in a file called “tweet_json.txt” file. After that, we extracted the json data from the text file and upload it in a data frame called “tweet_data”. The second step of the wrangling process is assessing the data. I assessed the data through two ways: 1. Visually: We have assessed our data by only going through the data visually and scrolling through it. Trying to explore some issues. 2. Programmatically: We tried to explore our data by using some code to view specific portions and get some summaries of the data. And we figured out some issues we need to solve and clean: Quality Issues: “twitter_archive” table: Quality issues 1.Unnecessary columns on each data

2.Incorrect data types

3.Repeated columns

4.Unknown columns

5.many retweets

6.not related information

7.one columns with many values

8.werate dogs twitter retweets

Tidiness issues 1.Null values

2.Repeated values of the wrangling process is: Cleaning I started the cleaning step by merging all the table (Tidiness issues and quality issues). Then, I focused on the insights I got especially on the most popular dog name and source.

At the end, we store our clean data in a csv file “twitter_archive_master.csv”.