We Rate Doge – Project 4

Introduction

In this forth project, we will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. The dataset that you will be wrangling (and analyzing and visualizing). We will get data from three sources.

Gather Data

Gathering Data for this Project Gather each of the three pieces of data as described below in a Jupyter Notebook titled wrangle_act.ipynb:

1. The WeRateDogs Twitter archive. Which provide bu udacity
2. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically.
3. Twitter API.

Assess Data

We will assess the Quality issues with content and tidiness issues with structure. The quality and tidiness issues in 3 tables in figure below.

A. The first table "twitter – archived – table"

Quality issues:

1. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integers/strings instead of float
2. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id have a null value
3. The numenoter has an incorrect value such as zero
4. The denmenoter has an incorrect value
5. The name column has incorrect name such as a,an,the
6. In some columns the null value represented as non-null
7. retweeted_status_timestamp and timestamp should be a time format instead of object

Tidiness issues:

1. Melt four dog stages as one column instead of four.

B. The Second table "image – prediction – table"

Quality issues:

1. p1, p2, and p3 have incorrect writing (capital and small,underscors and dashes)

Tidiness issues:

1. There are a 324 non dog, so we need to remove it

C. The Third table "API -– twitter – table"

**No issues found in this table.

Clean Data

We clean the data by using the Python libraries such as pandas, NumPy. Cleaning by removing unnecessary rows and columns, change the data types, rename the incorrect data.

**We can look to a wrangle_iypn to see the used codes.

Merge all data frames

We merge all clean data frames to make the analysis easy. Ans save the master data frame in 'df_merge.csv'

References

https://stackabuse.com/reading-and-writing-json-to-a-file-in-python
https://stackoverflow.com/questions/46429088/how-to-view-an-image-with-matplotlib-when-using-requests-getimage-url
https://matplotlib.org/stable/gallery/subplots_axes_and_figures/axes_margins.html
https://pandas.pydata.org/docs/reference/api/pandas.melt.html
https://pbpython.com/pandas_dtypes.html
https://stackoverflow.com/questions/31511997/pandas-dataframe-replace-all-values-in-a-column-based-on-condition/31512025
https://github.com/maysazqarqaz/WeRateDogs-Data-Wrangling/blob/main/wrangle_act.ipynb