wrangle_report

September 5, 2022

0.1 WRANGLE REPORT

In this report i've gathered the information from twitter and assest it in order to find new valuable information. The whole process have been divided in three parts:

gathered
assess
clean

0.2 Gathered

3 archives had been dowloaded from different sources and melted.

twitter-archive-enhanced.csv image-predictions.tsv json_tweets.to_csv

First of all i've downloaded the database file from the udacity website. I' ve downloaded the image predictions in a folder using the appropiate code. I've downloaded the tweets from the twitter api with the developer profile. it contains the tweet ID, text, retweets, and other relevant data. Gathered the info and put it in a folder.It was a though transition beacuase it was my first time using an api.

0.3 Asses

Quality issues from Twitter Archive: After all files are gathered, the contents are evaluated. The first method is to display a random sample of the file. Other methods include looking at data types and maximum/minimum values.

1-change variables name of the column to make them more readable.

2-retweeted_status_id retweeted_status_user_id retweeted_status_timestamp shows whos have made a Rt (drop those clums).

3-The dog types should have NaN instead of None.

4- be carefull about the description in the Text column, the message to be valid should start like:'This is...

5-change timestamp format into datetime.

Quality issues from Image Predictions 6-Drop irrelevant columns.

7-Some names are strange find out the correct names etc.

8-Duplicated pics between jpg_url and img_num.

Quality issues in JSON tweets 9-Some tweets are repeated.

0.3.1 Tidiness

Tidiness is related to the presentation of the data

0.4 Clean

The quality and tidiness issues listed in the previous section are addressed by programatic and sometimes manual methods. The first step was to merge all folders into one data with the tweet ID as the primary key. For this step is is always useful to work outisde the main dataframe by creating copies or arrays before modifying the contents.

0.5 Conclusion

The data needed for analyzing is often located somewhere else and in an ufamiliar format. Our objective is to extract, assess, and clean it for our use. In some instances it is made easy by the host's API. After gathering the data, exploration is useful to identify quality and tidiness issues. Finally, the data is cleaned based on the latter step.

```
In [ ]:
```