

wrangle_report

January 31, 2019

1 Wrangle Report

This report included the process of wrangling dataset in 3 steps:

1- Gathering Data. 2- Assessing Data. 3- Cleaning Data.

As well as, merging and storing final dataset and touched to some analysis and visualizations.

1.1 Gathering Data:

This project gathered 4 types of files: 1- CSV ('twitter-archive-enhanced.csv') 2- TSV ('image-predictions.tsv') 3- TXT ('tweet-json.txt') 4- API ('tweet_json.txt')

1.2 Assessing Data:

First, have a quick examine look to the dataframes to check quality and tidiness issues (Visual Assessment). Then started to check quality and tidiness programmatically using codes such as: `.head()`, `.info()`, `.shape`, `.value_counts()`, `.duplicated()`, `.sort_values()`, `.isnull()`. After that the extracted issues had been marked down for each dataframe.

1.3 Cleaning Data:

After checking each dataframe and holding the issues to fix, here it comes the role of cleaning dataframes from each issue assigned in the previous step, using the steps of cleaning data which are: define, code and test for each issue. This step contained some codes such as: `.replace()`, `.drop()`, `pd.melt()`, `.split()`, `pd.merge()`, `.astype()`, `pd.to_datetime()`, `.str.upper()`. In addition, after coding comes testing to ensure that the code is working correctly to get the desired results.

1.4 Storing Data:

The final step of wrangling data project is to combine and store all cleaned dataframes into 1 dataframe and store it in CSV file ('twitter_archive_master.csv')