wrangle_report

September 5, 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.

0.2 Introduction

Data Wrangling is the process of transforming raw data into more readily used formats. Methods differ from project to project depending on the data we're leveraging and the goal you're trying to achieve. Some examples of data wrangling include:

Merging multiple data sources into a single dataset for analysis

Identifying gaps in data (for example, empty cells in a spreadsheet) and either filling or deleting them

Deleting data that's either unnecessary or irrelevant to the project you're working on

Identifying extreme outliers in data and either explaining the discrepancies or removing them so that analysis can take place

0.3 Project Objective

The project objective was to perform data-wrangling analysis on the WeRateDog tweets from 2015 - 2017, which involves gathering, assessing and cleaning the datas

0.3.1 Gathering The Data

Data was gathered from three different sources

twitter_archive_enchanced.csv was directly downloaded from the udacity server.

tweet image prediction(image_predictions.tsv) was downloaded using the request libary

tweet_json was downloaded directly because twitter didnt approve my developer account.

0.3.2 Assessing The Data

The gather data was accessed visualy and programmatically in the notebook using various methods. Nine quality and two tidiness issues in the dataset was noted and processed to be cleaned.

Quality issues wrong datatype for the Timestamp column dtwitter archieve dataset
Keep only original retweet in twitter archieve dataset
Drop columns not needed in twitter archieve dataset
Names in the name column of twitter archieve data set are inconsistence (lower,sentence and upper case)
Extract tweet source from the html tag in the twitter archieve data set
Inconsistent dog name (lower case and title case) in p1 column of image prediction data set
Inconsistent dog name (lower case and title case) in p2 column of image prediction data set
Inconsistent dog name (lower case and title case) in p3 column of image prediction data set
Drop duplicate in jpg_url column in image prediction data set

Tidiness Issue twitter_archieve_df and tweet_df should be merged into one dataframe
tweet_df needs rearrangement (tweet_id should be the first column follow by the source)

0.3.3 Cleaning Data

I made copies of the original data before i cleaned it. i converted the timestamp column to the right datatype. i cleaned all the issues noted during the assessing stage

0.3.4 Storing Data

I saved the gathered, assessed, and cleaned master dataset to a CSV file named "twitter_archive_master.csv". Which i later used for my exploratory analysis