wrangle_report

June 5, 2022

0.1 Wrangle Report : "WeRateDog"

Data wrangling is the process of gathering, assessing and cleaning data. Real world data rarely comes clean and in this progect, I wrangled the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog

I imported the following libraries which i used in this project - import pandas as pd - import numpy as np - import requests - import os - import json - import re - import matplotlib.pyplot as plt - import seaborn as sns - %matplotlib inline

The first step in the wrangling process is gathering. In this project, three datasets were made available. Enhanced Twitter Archive, image predictions file and an additional data via the Twitter API. These three datasets were gathered differently. I downloaded The Enhanced Twitter Archive (a CSV file) manually and loaded it into a Pandas DataFrame, the image prediction file (a TSV file) was downloaded programmatically from Udacity server using the Requests library and url and finally i downloaded the additional data from Twitter API (a JSON file) and this was read manually (using the file.readline() method) an alternative method provided by Udacity to assess data without a twitter developer account

After saving all three datasets, i assessed them visually and then programmatically, using the .head(), .tail(), .info(), .sample(), .describe() methods among others. I noted two tidiness issues and eight quality issues which i documented. The tidiness issue had to do with the structure of the data, some columns needed to be melted into one column and the three datasets ought to be one for the data to be considered tidy, and the the quality issues had to do with the content, there were cases of wrong data types, missing values, invalid names and even some unneccessary data not needed for analysis.

The next stage was the cleaning of the data. I made a copy of the original dataset and followed the programmatic Data Cleaning process i.e. Define, Code and Test. I converted my observations from the assess step into defined problems, translated these definitions to code to fix these problems, then tested the three datasets to make sure the operations worked. I concatenated the three DataFrames on a common attribute twitter_id, to create the master_df.

finally, I stored the cleaned master_df in a CSV file named twitter_archive_master.csv

In [ ]: