wrangle_report

September 7, 2022

0.1 Reporting: wrangle_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.

Data wrangling is the process of removing errors and combining complex data sets to make them more accessible and easier to analyze. Due to the rapid expansion of the amount of data and data sources available today, storing and organizing large quantities of data for analysis is becoming increasingly necessary. As we all know that data wrangling involves gathering, accessing and cleaning, I'm going to take you through a few of the things i did working on this project.

1.Data gathering: I gathered the files provided for this project which are the twitter_archive_enhanced.csv and image_predictions.tsv using the requests package. I noticed that retweet counts and favourite counts(likes) were not present in the default data provided for this project which led me to applying for a twitter developer account because it was necessary to query twitter's Api using tweepy, and store the data as text_json.txt in other to get the missing data. The gathered data are loaded into different DataFrames:

TwitterArchive : Loaded data from twitter_archive_enhanced.csv ImagePrediction : Loaded data from image_predictions.tsv.

2.I have performed two types of data assessment i. visual assessment: where i loaded the data in excel and i scanned through ii programmatic assessment: where i loaded the data in a jupyter notebook and i used pandas to access the data.

After i was done assessing, i singled out 8 quality issues and 2 tidyness issues from both dataframes combined, some of which are; wrong datatypes, duplicates, erroneous dog names, and some records having more than just one dogstage name just to mention a few, and how to solve them. Then i moved to

Data cleaning:At the data cleaning stage, first thing i did was make a copy of my dataframes, then i started writing the codes to clean the quality and tidyness issues that i singled out earlier.

in conclusion, i would say the data is now clean to the best of my knowledge as i have fixed the quality and tidyness issues that i observed. i have written the steps in markdown on how i attempted the cleaning. this was not easy, this wouldn't have been possible without help from google, stackoverflow, w3schools and geekcodes.