# wrangle_report

February 15, 2022

## 0.1 Reporting: wrangle_report

The goal was to collect data about a particular Twitter tweet archive. Three datasets from three different sources were gathered. One source being from a file on hand, the other had to be downloaded programmatically from a server using an HTTP library, and the final third was data accessed by and downloaded through an API. After gathering the needed data and uploaded into a dataframe structure, each dataset had some issues of its own, after assessing each data source both visually and programmatically. The issues were spread across different dimensions whether that's validity, consistency, or accuracy issues. And categories such as content- or structural-related issues. Fixing of these different types of issues can facilitate the process of analyzing the entirety of the gathered data later on. Some of the issues that were fixed were associated with columns not having the appropriate data type, some because of the lack of consistency for representing essentially the same values but without a defined standard. There was also an issue of having some foreign data, either relevant but didn't come from the original source such as from a retweet rather than the tweet itself or the data was completely irrelevant and somehow found its way into the dataset; this type of issue was fixed by identifying and removing records that didn't represent the intended goal for the gathered data. These few issues fall in the quality/content category. Furthermore, some issues were structural, which was having the data structured in a way that could hinder seeing insights when analyzing the data. Another issue that converges with both categories of issues is when columns—especially after cleaning and restructuring them—are not relevant to the dataset we are wrangling and potentially analyzing. These irrelevant columns being present made it inconvenient to skim through the data for visual assessment, and they also would add up to the decrease in efficiency when writing or reading datasets in which such columns exist(especially for vast datasets). Such columns were dropped after being deemed unnecessary anymore. By the end, the whole act of wrangling enabled me to create clean and insightful visualizations and the possibility of grouping by features of the dataset for aggregation, given the new and well-structured data.