Git Repository

https://gitlab.bht-berlin.de/s90246/mental-health-and-social-media

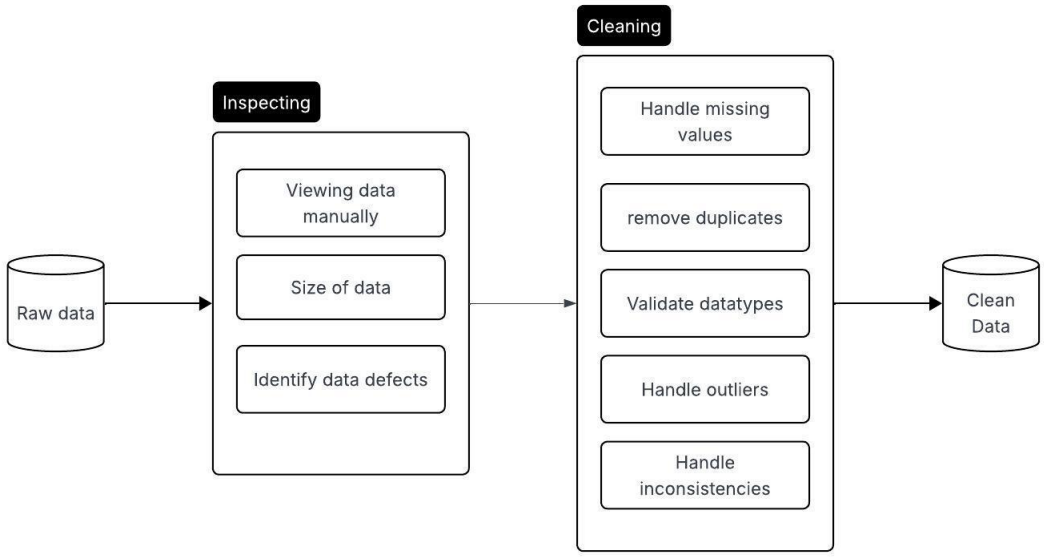
Daten Inspektion

Bei der ersten Inspektion der Daten habe ich festgestellt das die Daten zu sauber sind. Um trotzdem einen Lerneffekt zu erzielen habe ich die Daten mithilfe einer KI verunreinigen lassen.

Bei der Zweiten Daten Inspektion, des neuen verunreinigten Datensatzes, sind mir folgende Unreinheiten aufgefallen

- Fehlende Werte
- NaN Werte
- Werte mit Leerzeichen
- Dubletten
- Ausreißer
- Inkonsistenzen bei Strings

Data Cleaning Pipeline



Herausforderungen

Bei der Entfernung der Ausreißer hatte ich einen Bug eingebaut, bei welchem das DataFrame nach der Behandlung keine Kategorischen Spalten mehr hatte. Ich hatte die Boolesche Filterung direkt auf das Original DataFrame angewendet, anstatt die Maskierung in einer Variable zu speichern und dann auf das DataFrame anzuwenden.

Frage

Mein original Datenset hatte 500 Zeilen, nach der Reinigung hat es jetzt nur noch 445 Zeilen. Ist das Datenset jetzt noch zulässig?