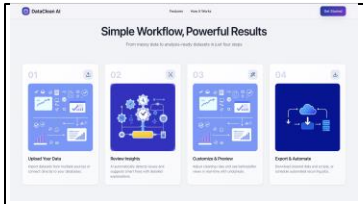
AI Dataset Cleaner

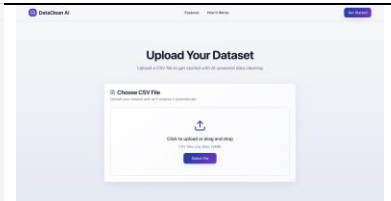
Name: Ethank Kuzmik and Elliott Corso
Project Track: Foundations of AI
Major: Data Analytics

Problem Statement : Data Scientists are constantly looking at different datasets that all need to be cleaned and analyzed. Cleaning and taking care of lower level tasks can become repetitive and time consuming. We wanted to create an application that can help speed this process up and give potential insights to jumpstart data analyzation.

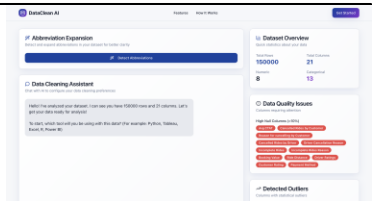
Interface



Our original prototype goals continue to guide our entire development process, helping us refine the system, improve usability, and deliver clearer, more seamless support for every user.



We created a simple easy and seamless way to interact with the user, once they press get started all they have to do is insert there CSV file of choice and give it around a minute to process. A nice right to the point process.

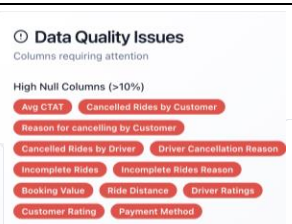


This is what the user sees right away, this gives them a chat bot where it can have help with code and ask it about the data and any questions that it might have. Giving the user and easy entry into the data science world.

User Outputs

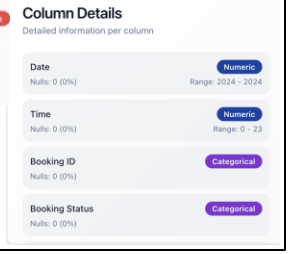
This is our simple and easy data overview, it gives just an idea of the quick stats of your Dataset and something to base you thoughts off of as you work with our AI.

Our outlier detector flags the first signs of irregular data, giving users an immediate, easy-to-spot starting point for review and helping them quickly understand which issues may matter most.



This gives users a clear view of which columns contain large amounts of null values. Instead of manually checking each one, they can immediately see where potential issues are most likely to occur.

This gives the under an general understanding of each column and they can see the data type and useful information from to help right away.



Methods:
We began with a simple prototype that produced our first cleaned samples, using multiple tools to understand the raw abnormalities in our dataset. To establish a strong baseline, we drew on abbreviation examples and open-source models from Hugging Face, which helped shape our early preprocessing rules. We then applied transformer-based methods to analyze and correct inconsistencies before integrating the model into a functional website. Using Cursor and a Vite/Vercel development workflow, we built a clean, user-friendly interface that makes our AI dataset cleaner accessible and practical for real use.

Results:
We now have a fully functional AI application that provides real-time feedback on any dataset you upload. Whether you need help cleaning your data or understanding unclear columns and abbreviations, the system can quickly interpret your input and offer clear, context-aware guidance. By pairing our transformer-based logic with an intuitive interface, the tool makes data preparation faster, easier, and accessible to users of all experience levels.

Resources:
https://huggingface.co/bert-base-uncased
https://huggingface.co/jaccob/clean-text
https://huggingface.co/dslim/bert-base-NER

Code

Ethan Kuzmik
LinkedIn: ethan-kuzmik
Email: ethan.kuzmik@gmail.com

Elliott Corso
LinkedIn: elliott-corso-3b1462346
Email: elliottcorso@gmail.com