# Ethics Discussion

Melody Goldanloo, Jenna Ramsey-Rutledge, Byron Selvage

Introduction:

Education data analysis is an important area for improving student outcomes and learning opportunities. However, it is also rife with ethical discourse due to concerns about privacy and bias. In our project, we explored graduation data related to school resources and other features.

Data Collection:

We decided to collect our data from Education Data Explorer. This API appeared the most ethical since it takes steps to anonymize student data and only provide users with aggregate statistics. We recognize that this data likely underrepresents smaller community schools in the state of Colorado since there is less data available in more rural areas. From the start of this project, we were aware of the lack of coverage, which is best showcased in the exploratory map below.
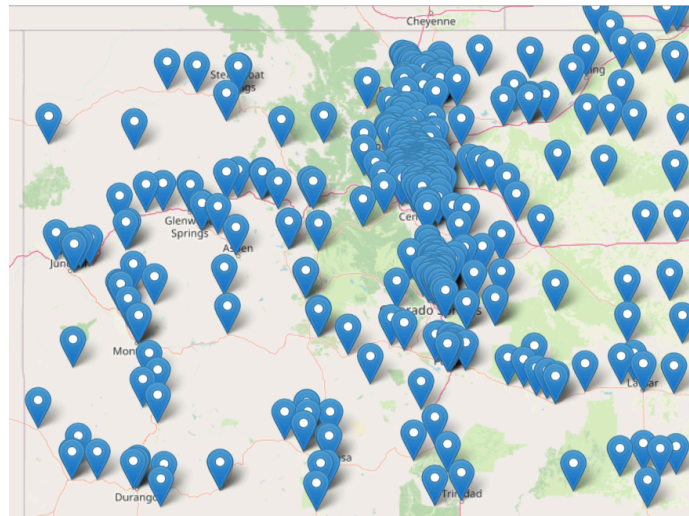
Another limitation of our data collection occurred from our time scale. Our graduation data was from 2021. These statistics were directly influenced by the COVID-19 Pandemic that disrupted the American School System. Therefore, these biases in location and lack of recency will affect the strength of our conclusions.

Analysis:

During our analysis, we deemed it important to avoid reproducing the bias inherent in the data. The locations of the schools and their demographics greatly impact graduation rates due to systemic injustices. In our pre-processing steps, we attempted to remove features that the model may falsely claim are the root causes of graduation rates. So, we removed the city and other distinguishing features from our data. These pre-processing steps allowed us to transform the data to align with privacy and fairness standards.

Model Development:

In the model development process, we wanted to use ethical practices to produce transparent and fair models. We created a decision tree model to aid in feature selection then used those features to train a linear regression model. We have stored the weights used in the decision tree. This allows us to see exactly what the model picked out as the important features and explain to stakeholders how our model works. Our team chose simpler models to be able to understand the decisions made by the machine. It is important for us to be able to explain how our algorithm works because determining graduation rates can affect future education funding and student outcomes. Making a transparent model is important for improving the model and producing fair results.

Conclusion:

When explaining and communicating our model, it is important to consider the biases and ethical considerations made along the way. We need to acknowledge where our data may be lacking in order to use the model to make informed decisions. In the future, it may be helpful to seek additional data collection methods or employ statistical techniques like bootstrapping to cover underrepresented communities.