Predicting Motor Vehicle Crashes Severity

Amanda Curtis, Arthur Diep-Nguyen, Olti Myrtaj, Brandon Owens, Fabio Ricci

Erdős Institute Spring 2025

Overview

1. There are more than 5 million motor vehicles crashes (MVCs) every year in the US

2. Predicting the frequency and severity of MVCs as a function of built-environment features in the surrounding area can help

Built Environment

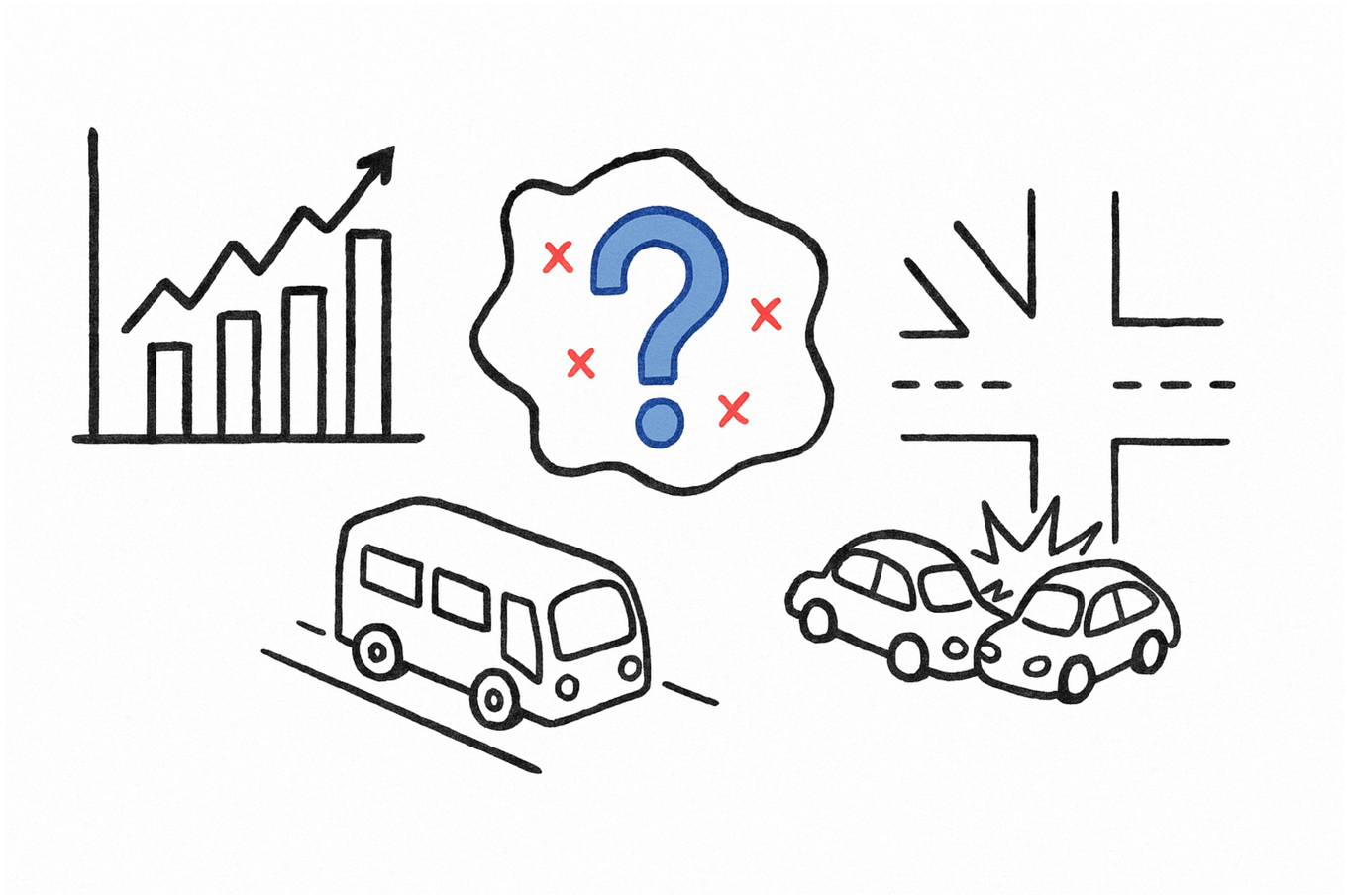
Features:

- Roads
- Crosswalks
- Transit Stops
- Walkways
- Stop Signs and Traffic Lights

https://publichealth.harriscountytx.gov/Divisions-Offices/Divisions/Environmental-Public-Health/Built-Environment-BE-Program/Built-Environment-101

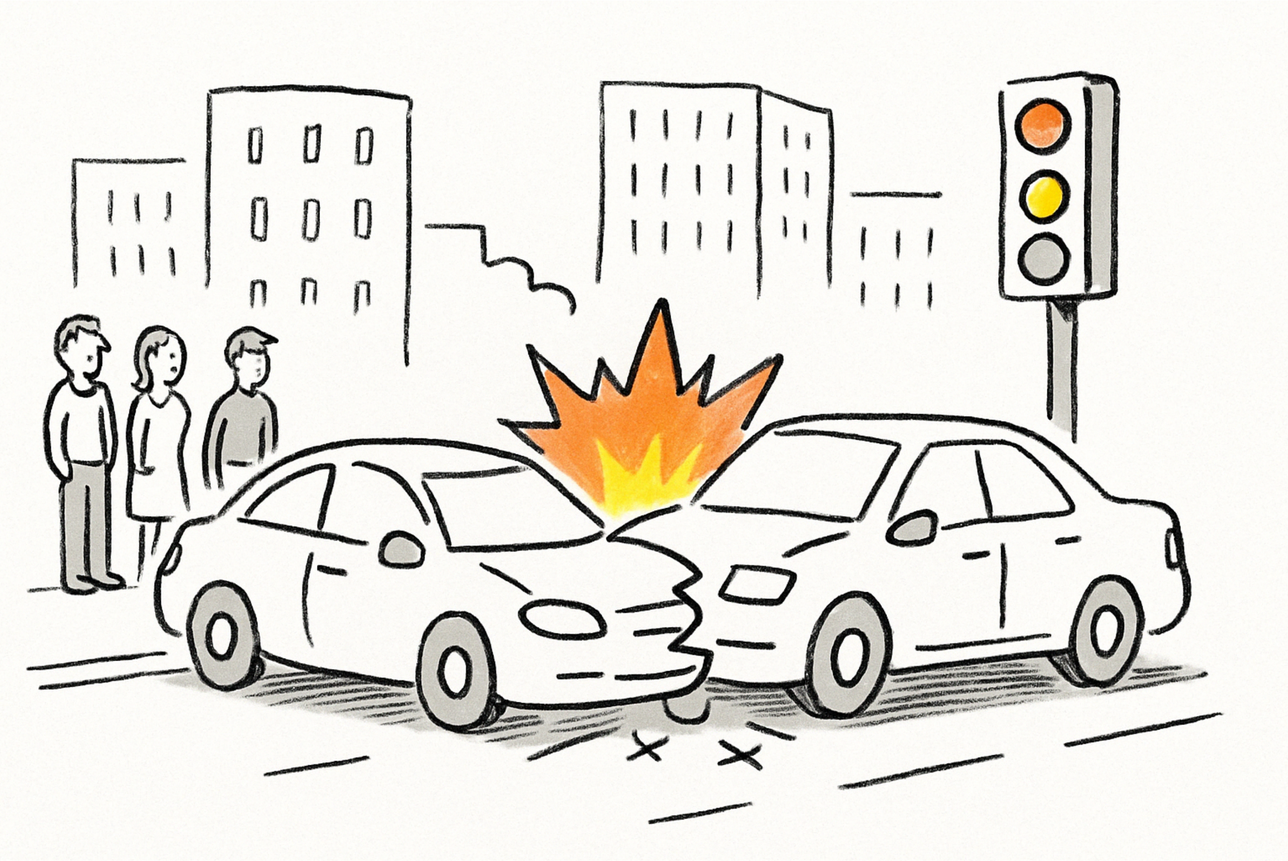
Research Question

How can we use features of the built environment in a given area to predict the “crash density”?

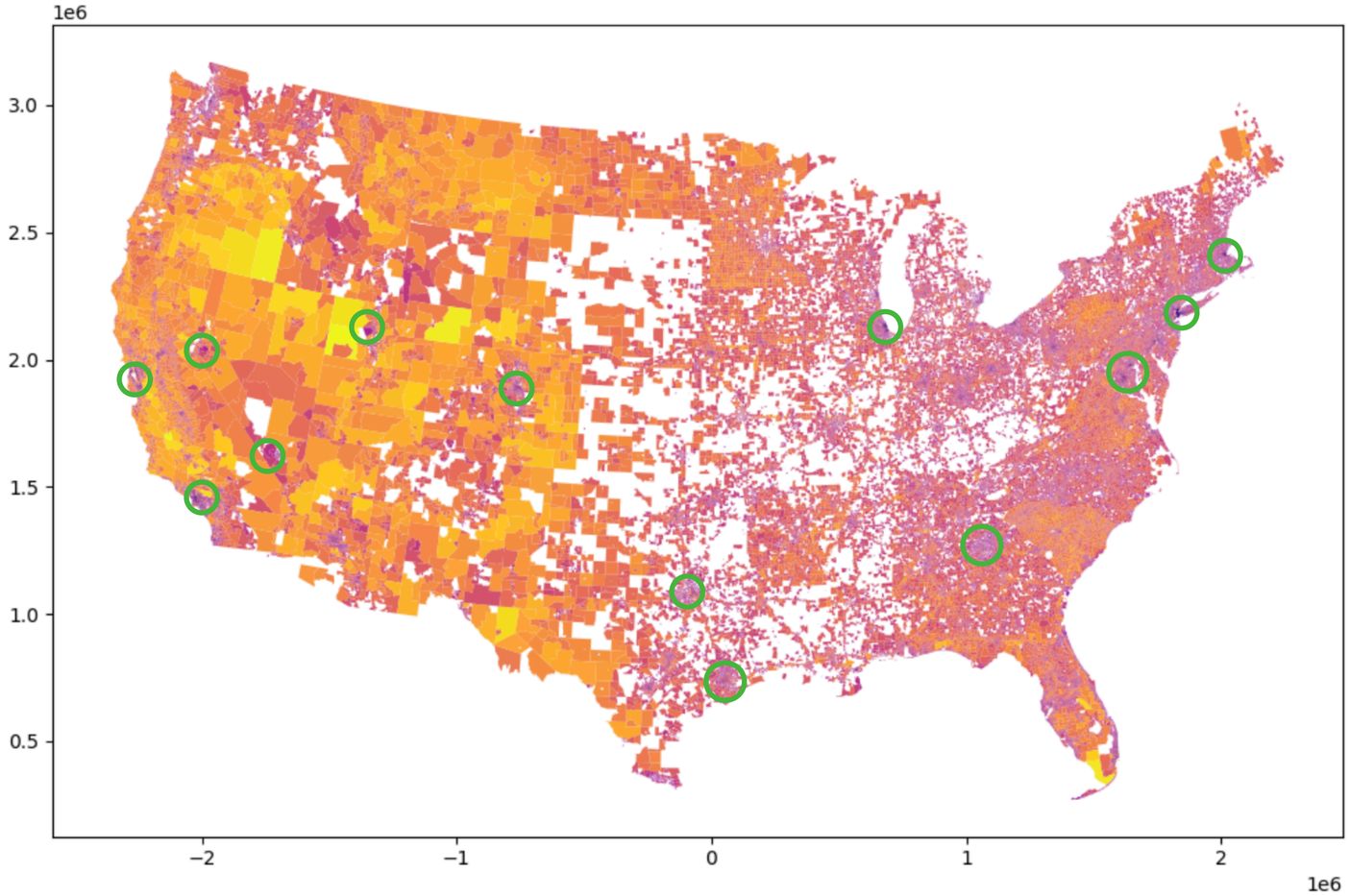


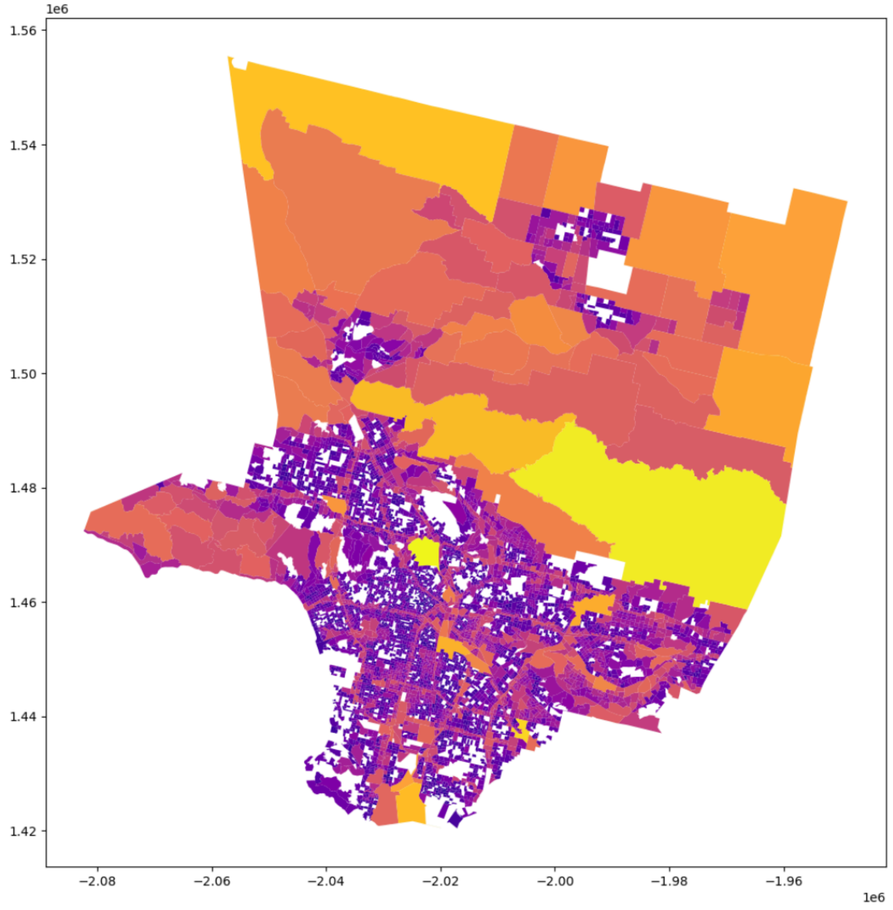
Defining "Crash Density"

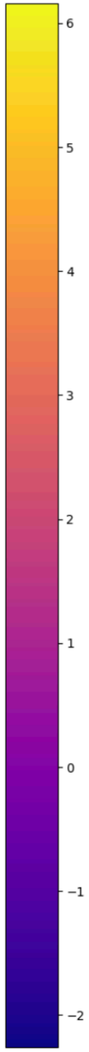
Engineered Target Variable = (Crashes * Severity)/(Population Density)



Visualizing Crash Density







Crash density is lowest in major metro areas

In LA County, crash density is greater along freeways and smaller in towns

Data Augmentation

EPA Data Set: Data based on Census Block Group

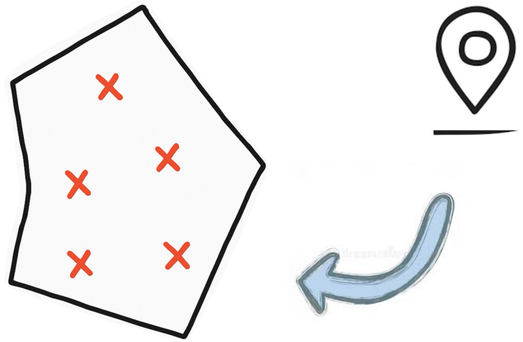
https://www.epa.gov/smartgrowth/smart-location-mapping

- > 200k census block groups
- Housing density, diversity of land use, walkability, neighborhood design, destination accessibility, transit service, etc.

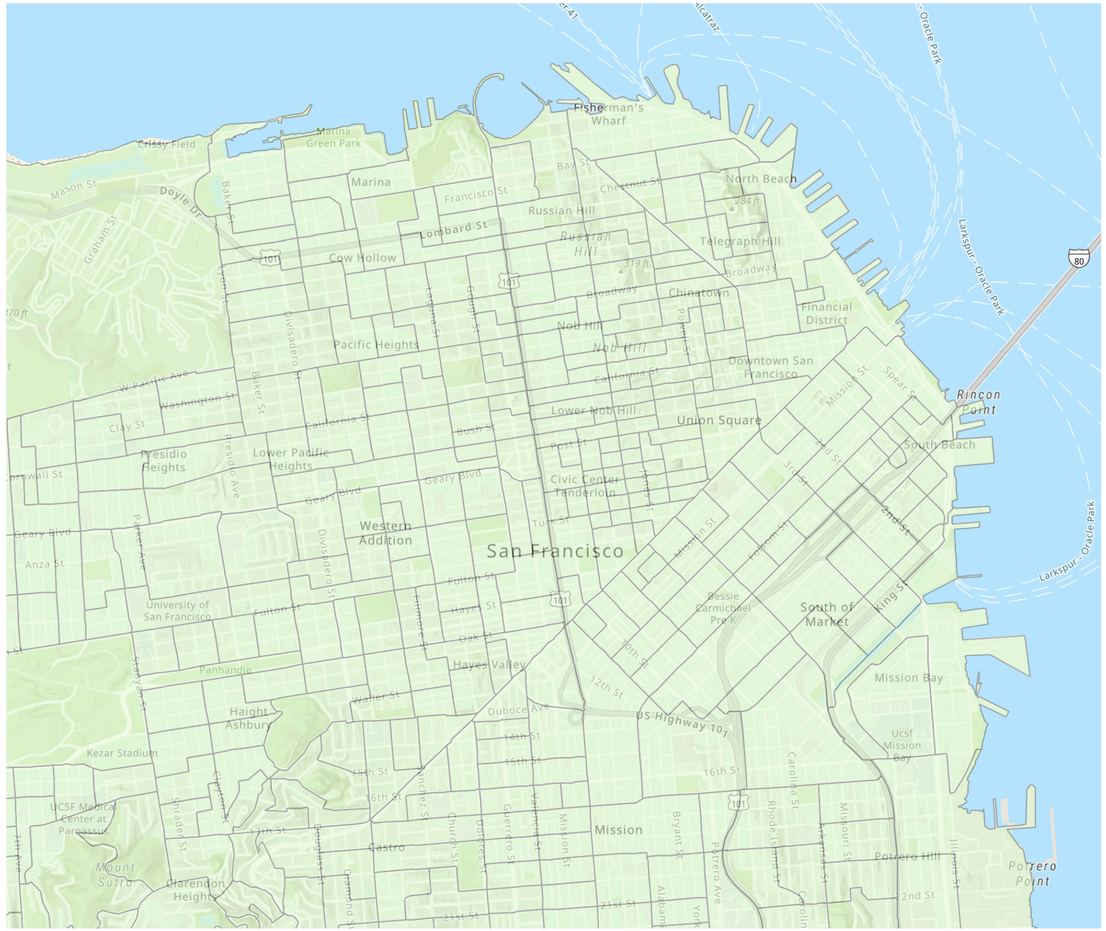
Kaggle Data Set: Single Crashes

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

- 7.7 million crashes from 2016-2023
- Location and severity of crashes

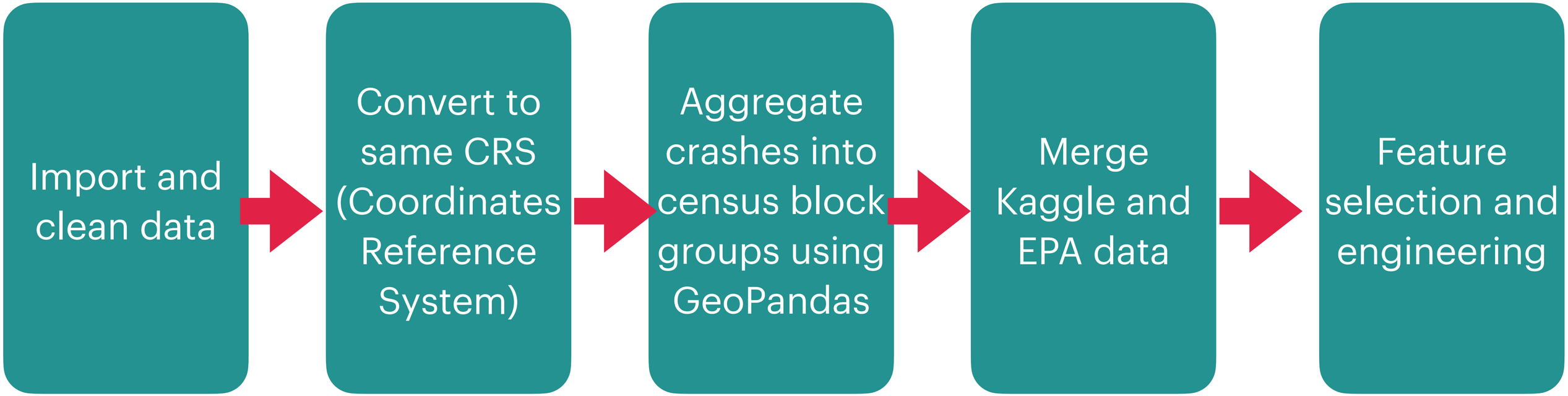


San Francisco, California - Census Block Groups

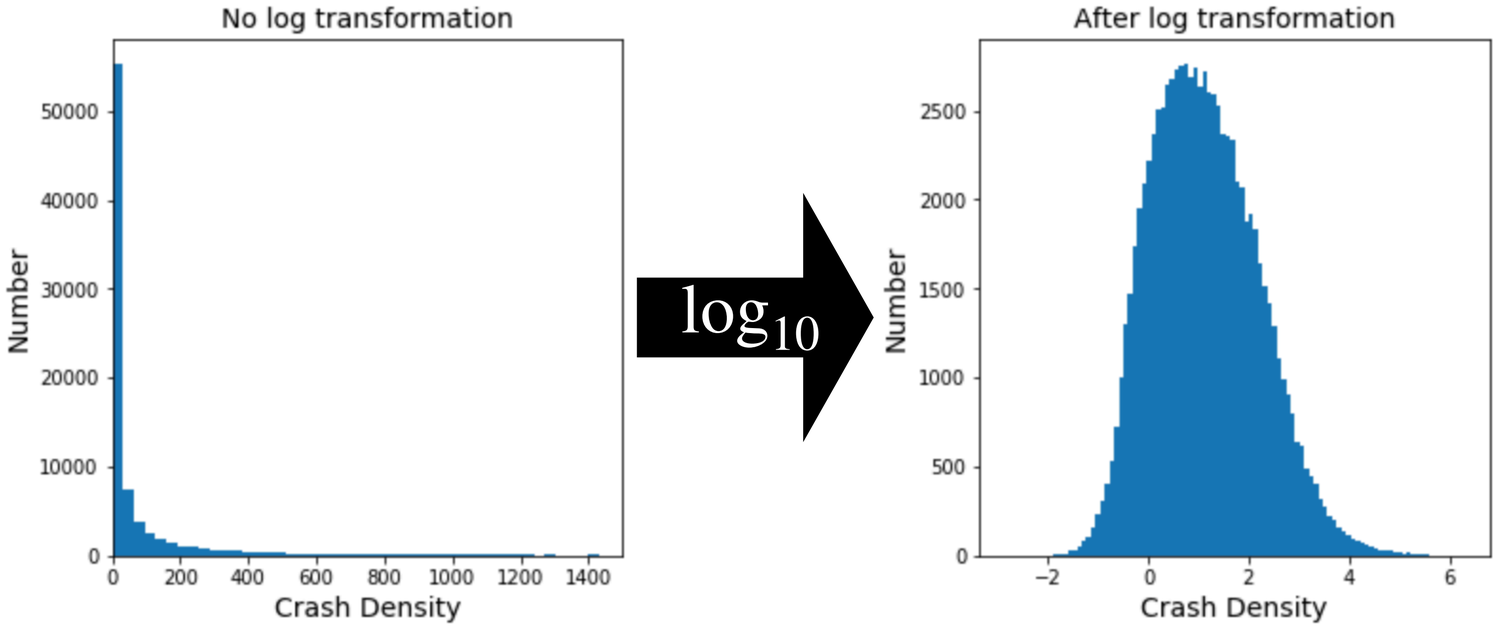


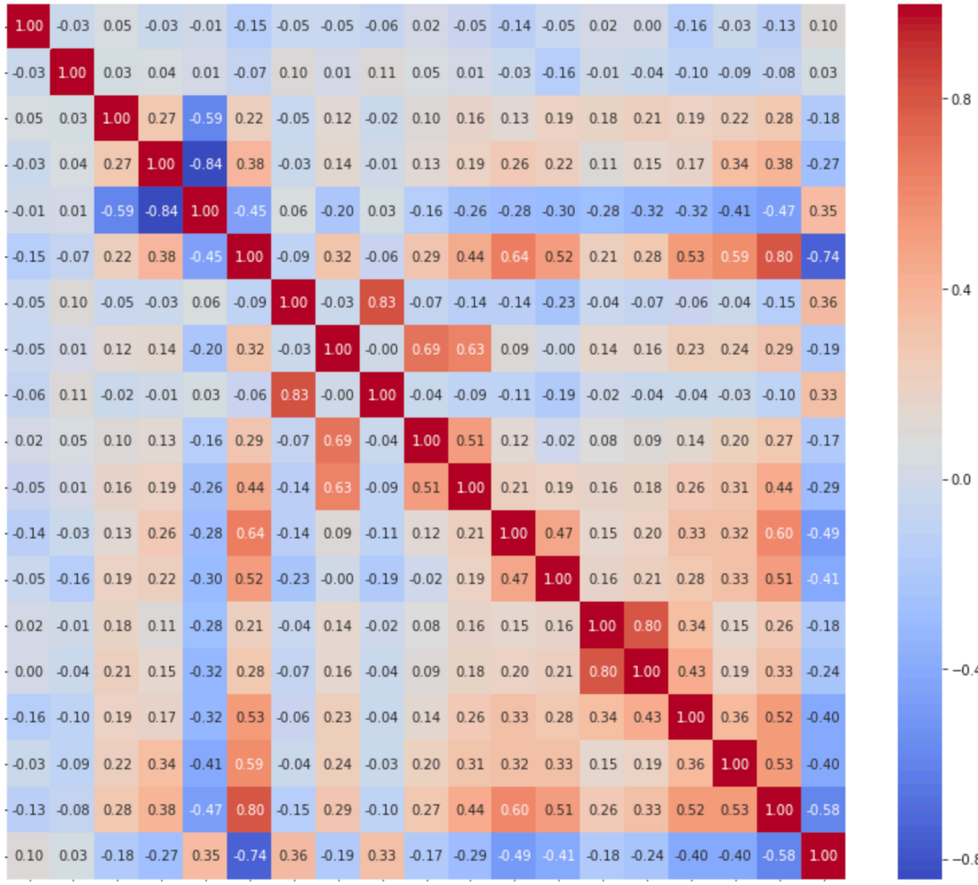
https://www.arcgis.com/apps/mapviewer/index.html?layers=2f5e592494d243b0aa5c253e75e792a4

Data Processing



Feature Selection

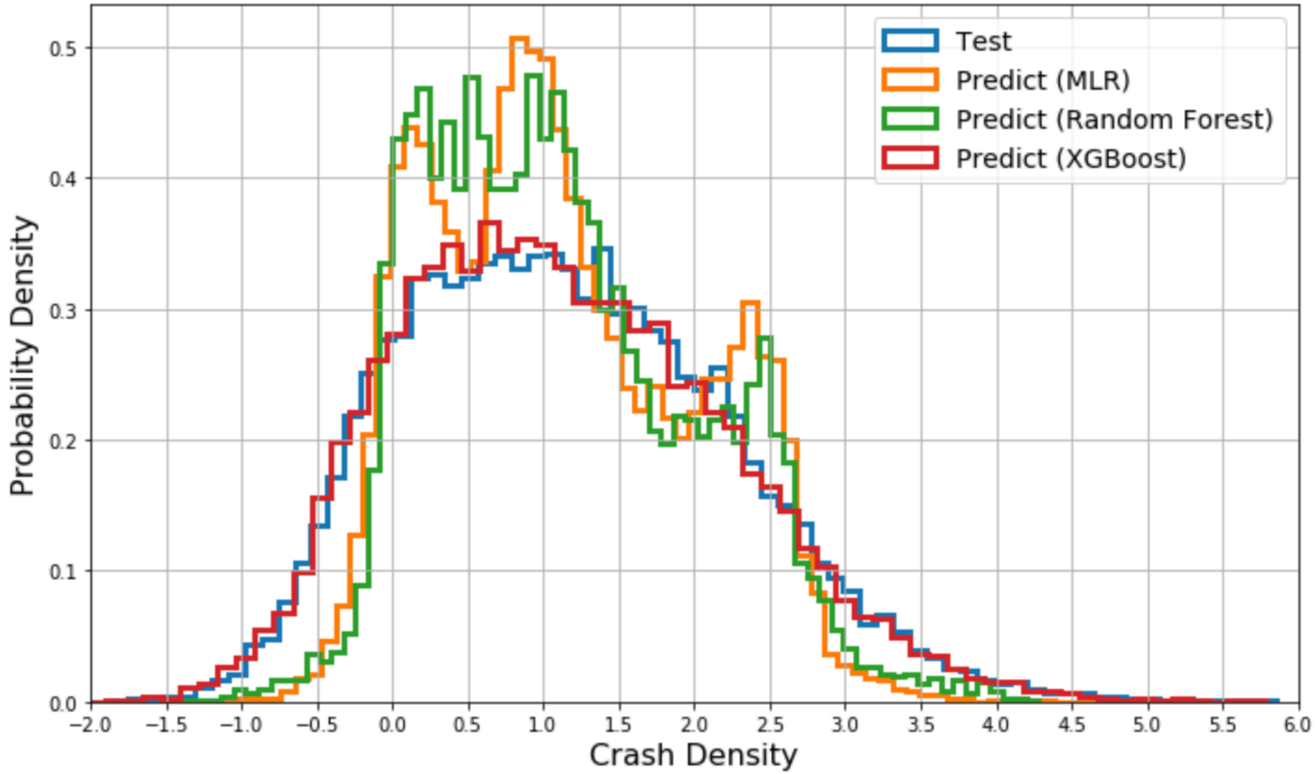




We apply a $\log_{10}$ scaling to highly skewed features to increase interpretability

Modeling Process

Merged and processed data → Tuned XGBoost → Crash Density Prediction

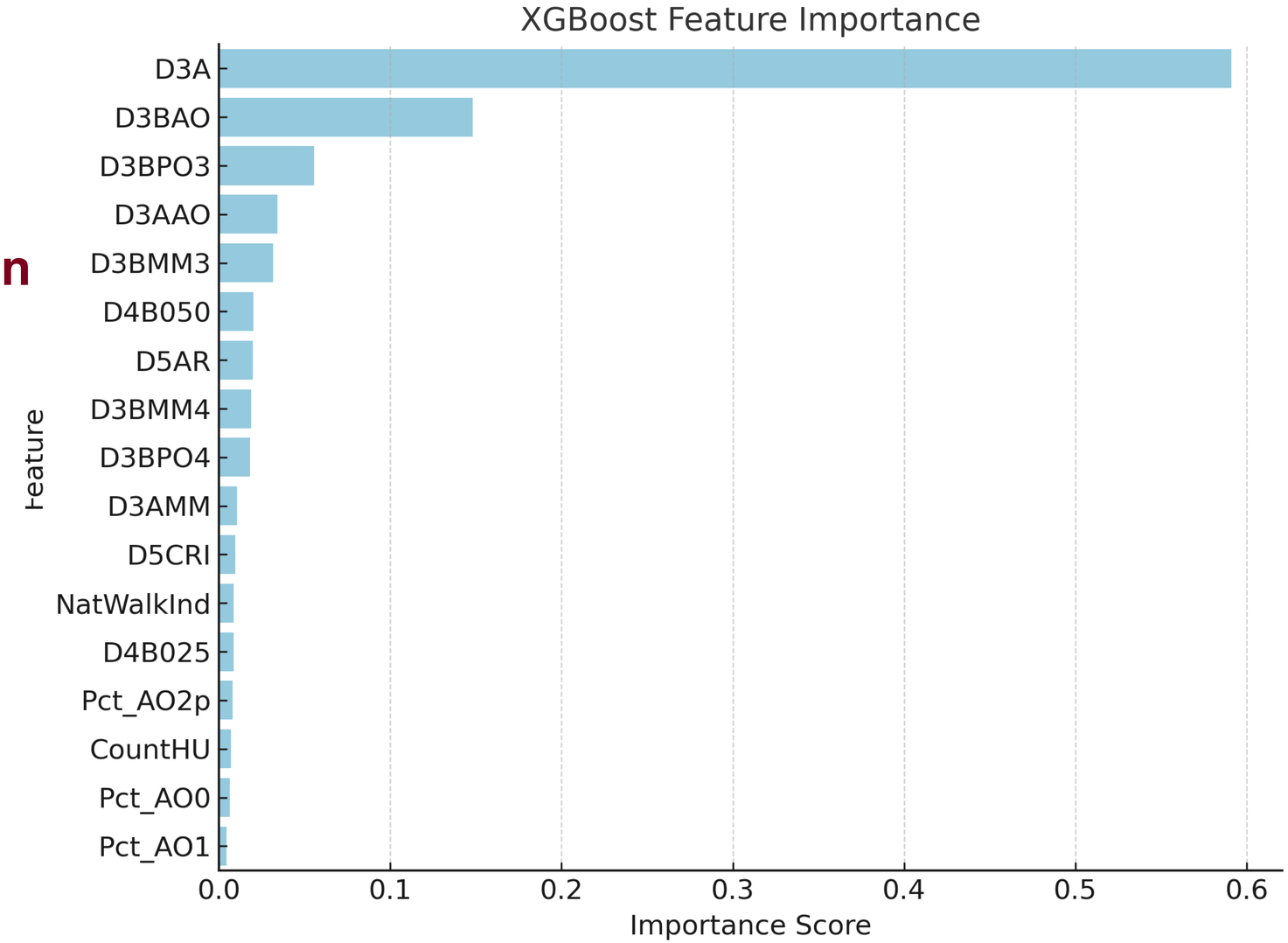


Model	RMSE
XGBoost	0.552
Random Forest	0.576
MLR	>0.6
Lasso	>0.6
Ridge	>0.6

Highlights of important features

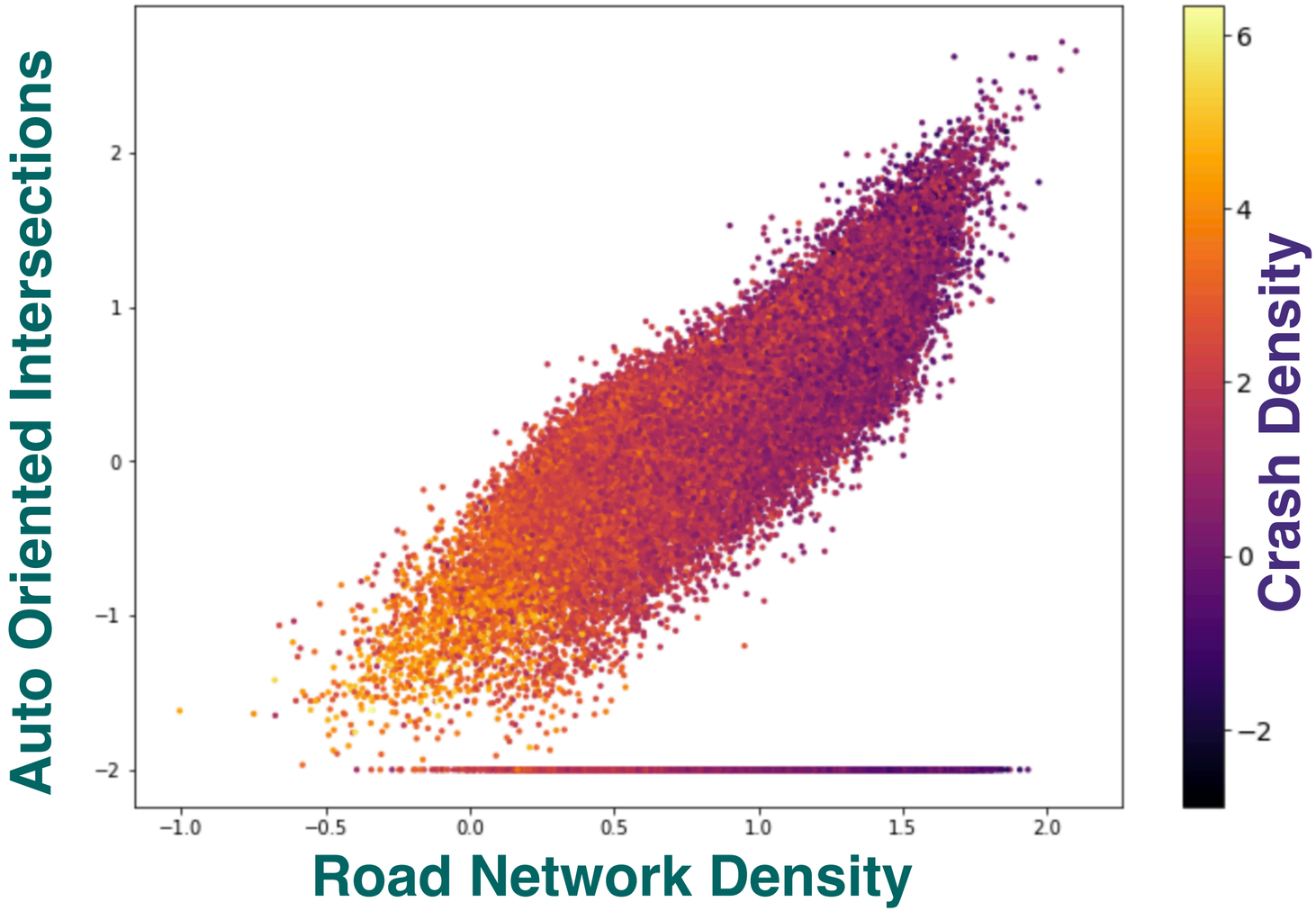
D3A = Total road network density

D3BAO = Auto-oriented Intersection density



Conclusions

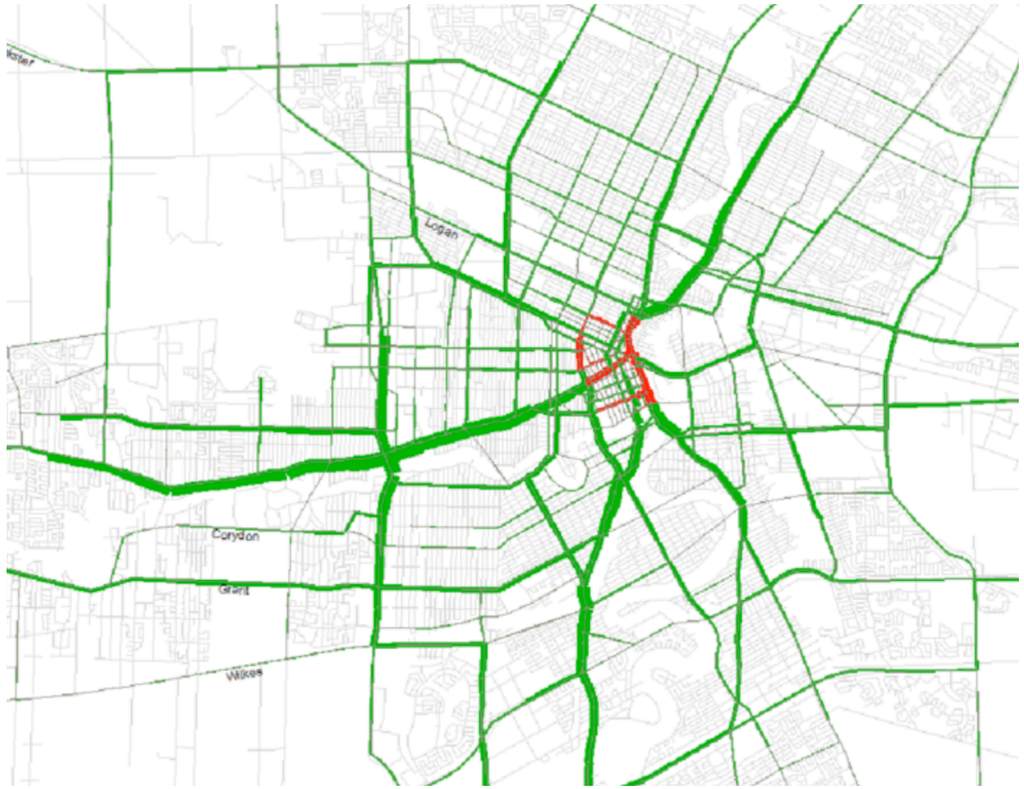
Densest road networks have lowest crash density



Future Work

Implement inference tests to study the impact built environment features have on Motor Vehicle Crashes

Design city plans that minimize number and severity of crashes using generative AI trained on CBG data



The Erdős Institute

Thank You

Thanks to Steven Gubkin, Alec Clott and Shravan Patankar