Springboard—DSC Capstone Project 2

PREDICTING PUBLIC TRANSPORTATION SAFETY RISK

HTTPS://GITHUB.COM/TAMARAHORNE/SPRINGBOARD/TREE/MAIN/CAPSTONE%20PROJECT%202

TAMARA HORNE

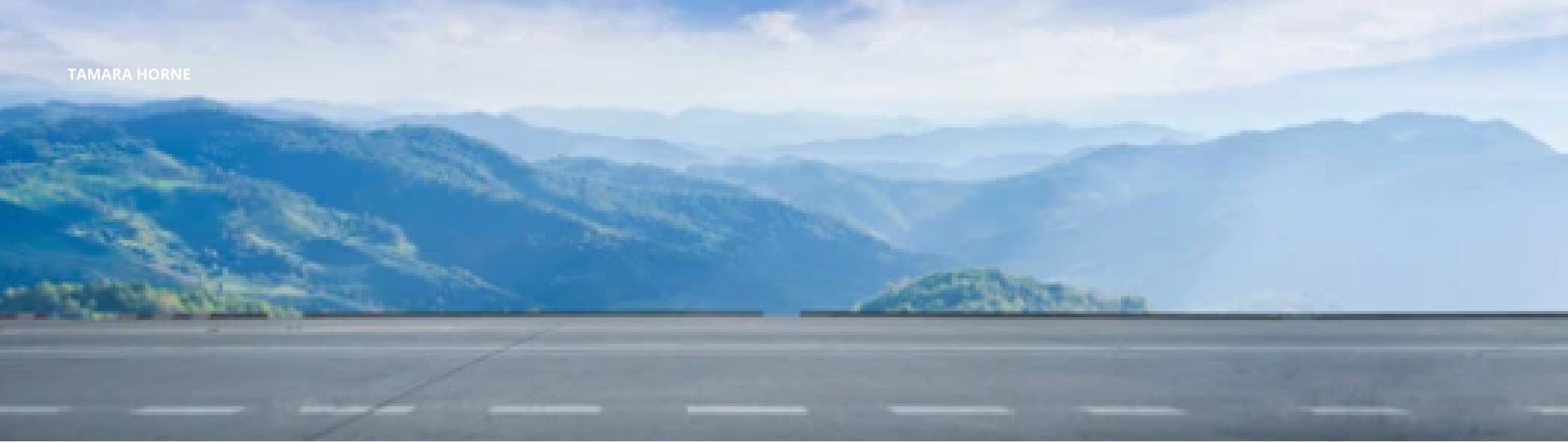
INTRODUCTION

WHAT'S THE SAFETY RISK?

Increasing Public Transportation Options Between Nashville's Convention Center and Airport

2023 | March

TAMARA HORNE



THE DATA

National Transit Database

- Monthly modal time series data
- 133,196 rows; 65 columns
- One row per month per agency per mode

Wrangling

- Filled NaNs in 4000+ with data found in the dataframe
- Filled NaNs in 27 rows with data from FTA

TAMARA HORNE

REDUCING THE COLUMNS

ID Columns

- ✓ 5 DIgit NTD ID
- ✗ 4 Digit NTD ID

Location Columns

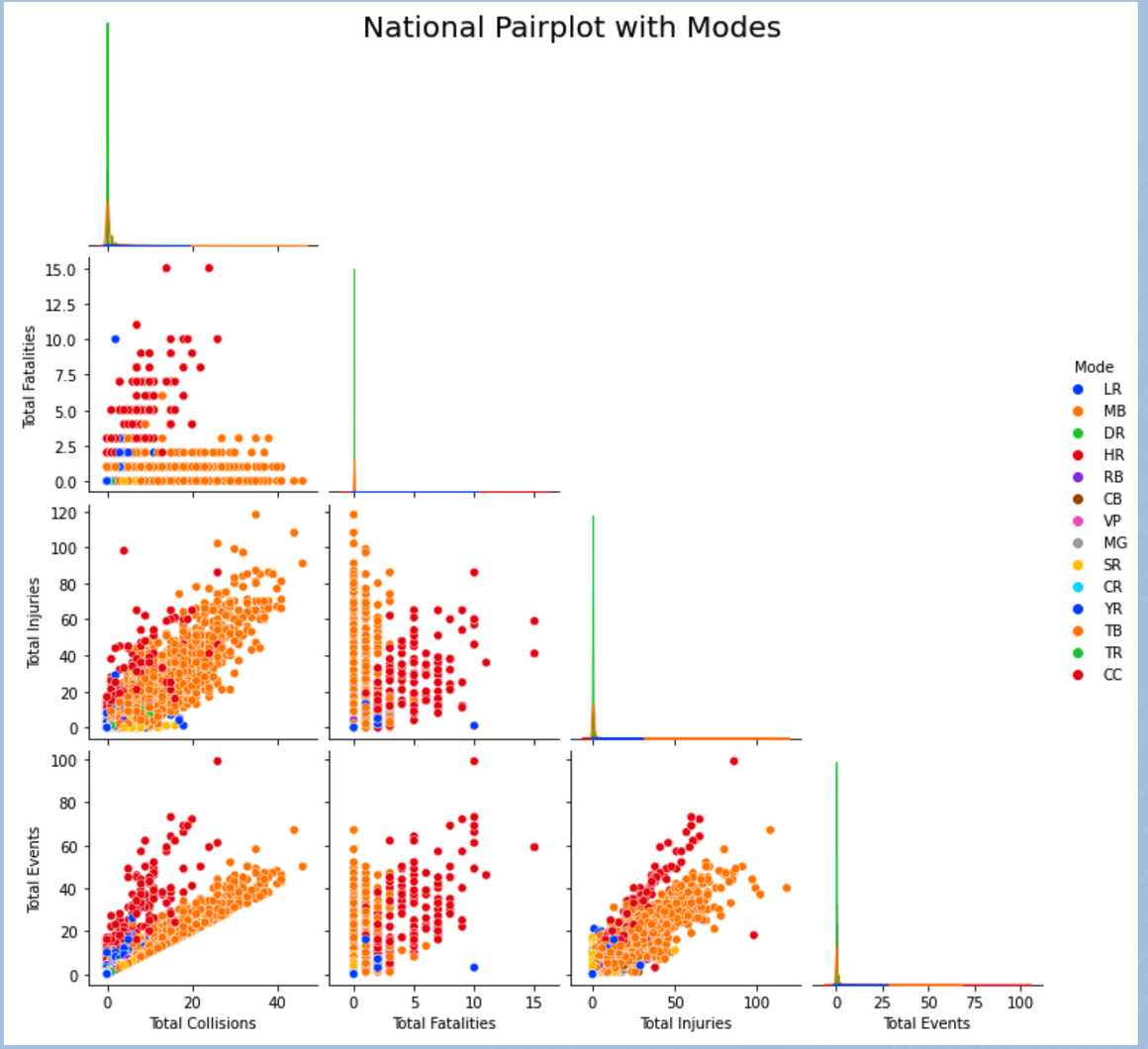
- ✓ Primary UZA Population
- ✗ Primary UZA Code

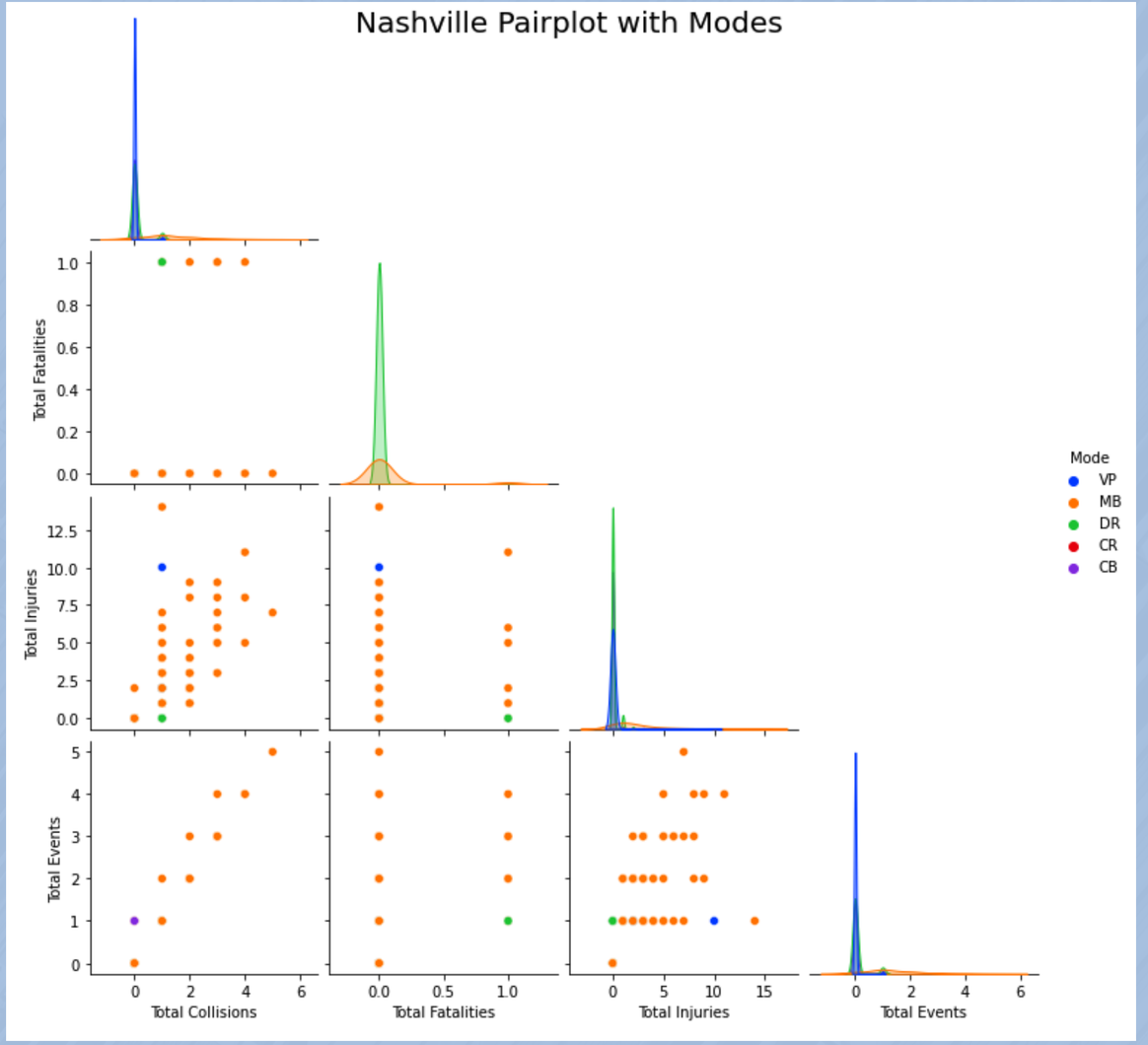
Totals Columns

- ✓ Total Fatalities
  Total Injuries
  Total Events
- ✗ All contributing columns

MANY SAFETY INCIDENTS FOR MB AND HR

TAMARA HORNE
2023 | March





TAMARA HORNE

THE TARGET

Safety Performance Targets as Reported to the National Transit Database (NTD)							
The targets listed below are based on reviews of the previous five years of MTA dba WeGo Public Transit's safety performance data.							
Mode of Transit Service	**Fatalities (total)**	**Fatalities (per 100 thousand VRM)**	**Injuries (total)**	**Injuries (per 100 thousand VRM)**	**Safety Events (total)**	**Safety Events (per 100 thousand VRM)**	**System Reliability (VRM / failures)**
Fixed Route Bus	0	0	35	.55	24	.45	5,500
Demand Response Bus	0	0	6	.27	6	.26	24,800
Demand Response Taxi	0	0	0	0	0	0	0

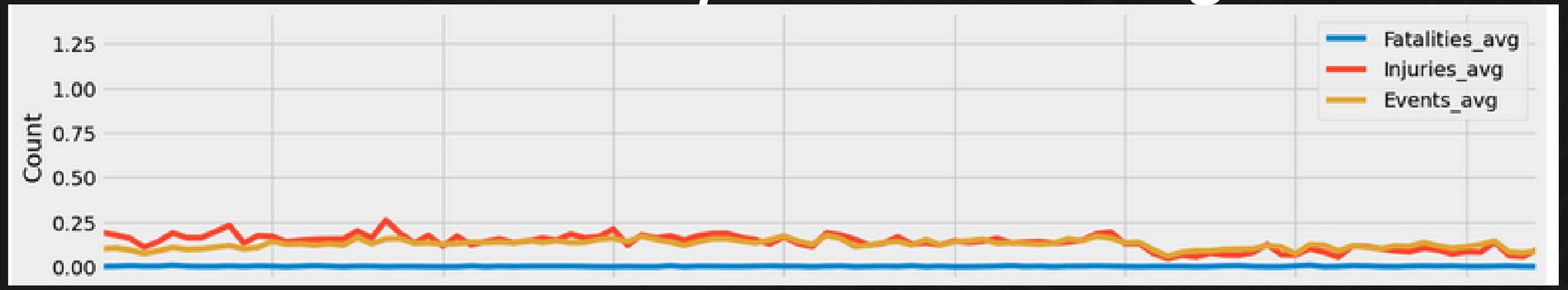
(TOTAL FATALITIES + TOTAL INJURIES + TOTAL EVENTS) / VEHICLE REVENUE MILES

2023 | march

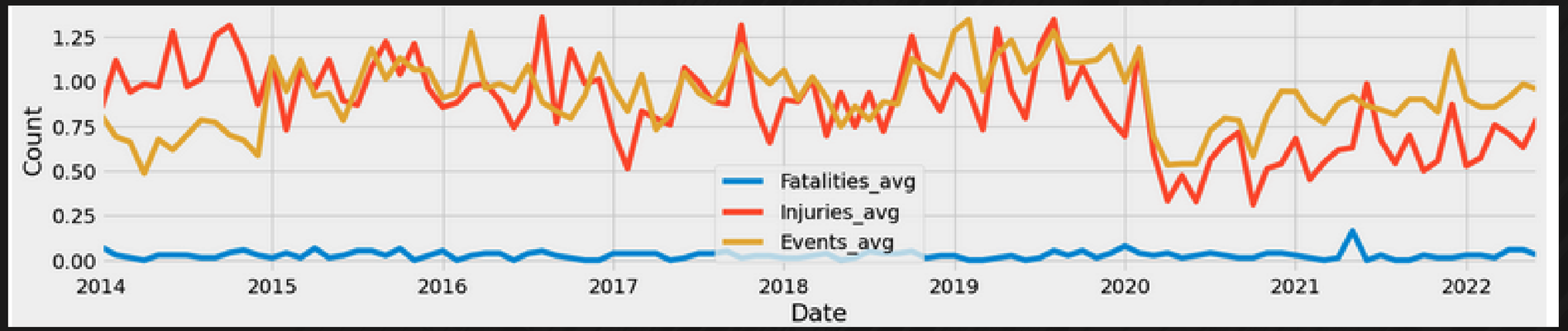
TAMARA HORNE

EXPLORATORY AGGREGATION

Locations with Modes up to and Including Nashville's



Locations with Nashville's Some or All of Modes Plus Light Rail



~modeling_df~	R Squared	MAPE	MAE	MSE
Linear Regression	0.645	0.127	0.058	0.007
Random Forest Regressor	0.902	0.071	0.033	0.002
K Neighbors Regressor	0.365	0.075	0.035	0.002
XGBoost Regressor	0.999	0.048	0.022	0.001

MODELING

First Aggregation

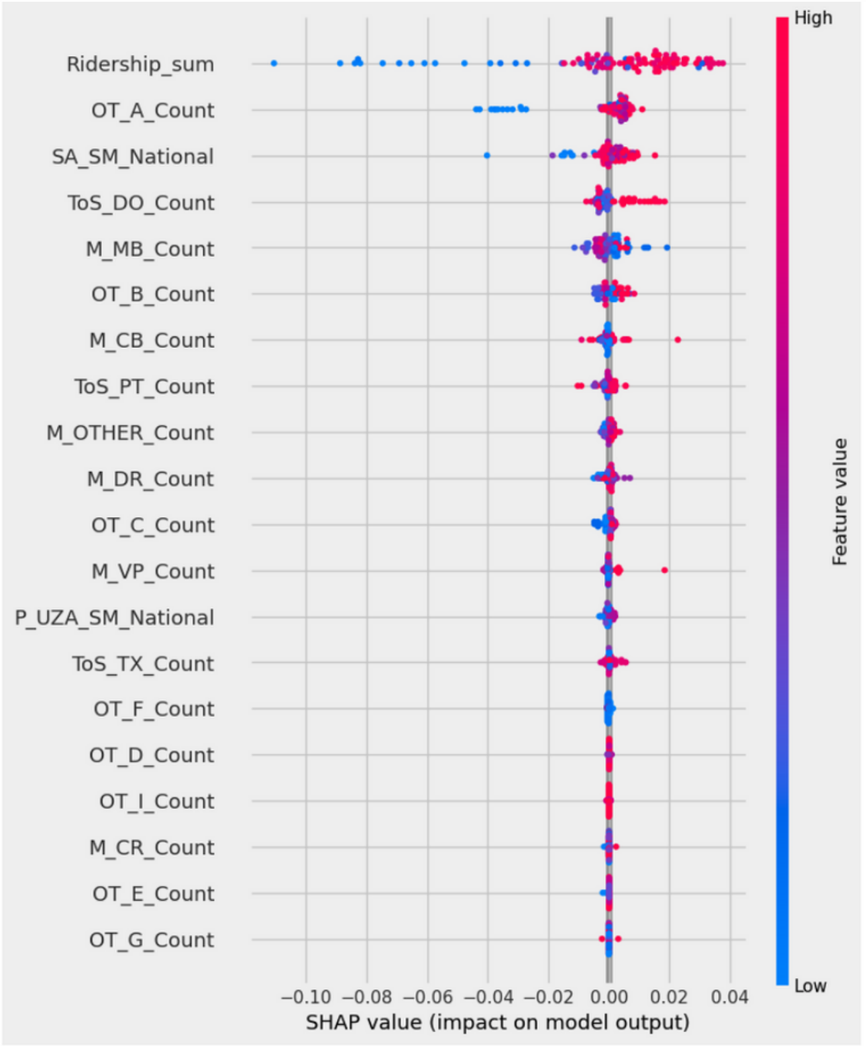
Date, nation

What features influence safety risk for the nation as a whole?

TAMARA HORNE

2023 | March

TAMARA HORNE



FEATURE IMPORTANCE

Ridership_sum

Lower values decrease safety risk

OT_A_Count

‘Independent Public Agency or Authority of Transit Service’ organization type.

Most frequently occurring organization type in the national data

Lower values decrease safety risk

SA_SM_National

Average service area square miles for the nation during each given month

Lower values decrease safety risk

~modeling_df2~	R Squared	MAPE	MAE	MSE
Linear Regression	0.251		0.425	0.656
Random Forest Regressor	0.833		0.522	0.871
K Neighbors Regressor	0.280		0.395	0.880
XGBoost Regressor	0.548		0.454	0.808

MODELING

Second Aggregation

Date, location

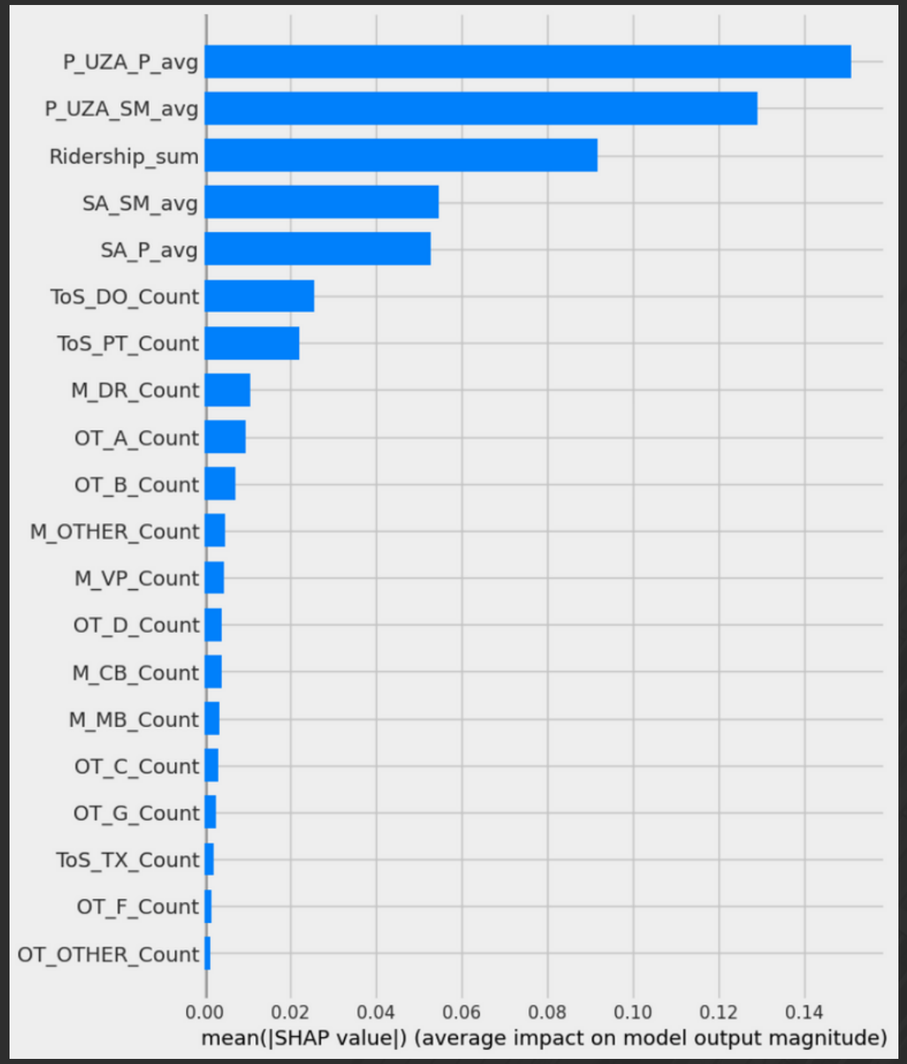
What features influence safety risk when data is aggregated to both date and location (Primary UZA Name)?

SHIFTED RESIDUALS MAKE THE RESULTS LESS TRUSTWORTHY

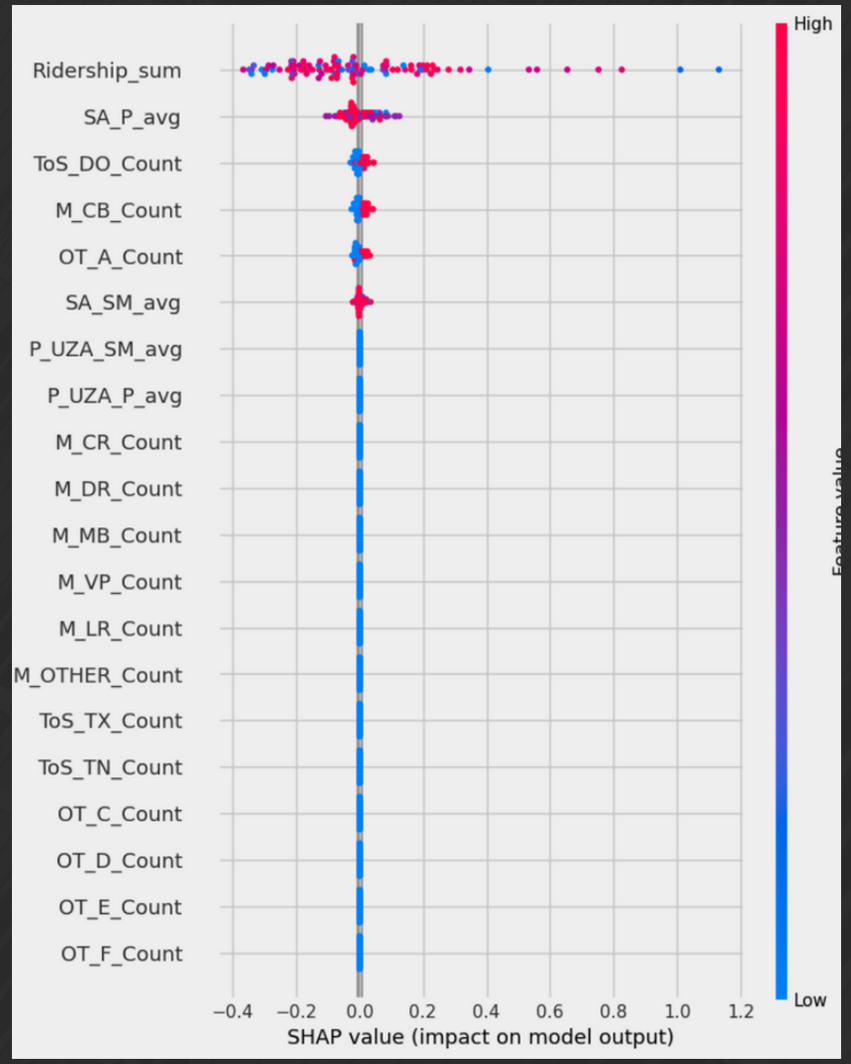
TAMARA HORNE
2023 | March

FEATURE IMPORTANCE COMPARISON

National



Nashville



TAMARA HORNE

2023 | March

~modeling_df3~	R Squared	MAPE	MAE	MSE
Linear Regression	0.210		1.403	3.587
Random Forest Regressor	0.883		0.680	2.335
K Neighbors Regressor	0.595		0.591	2.221
XGBoost Regressor	0.945		0.744	2.587

MODELING

Third Aggregation

Date, mode

What features influence safety risk for the individual modes?

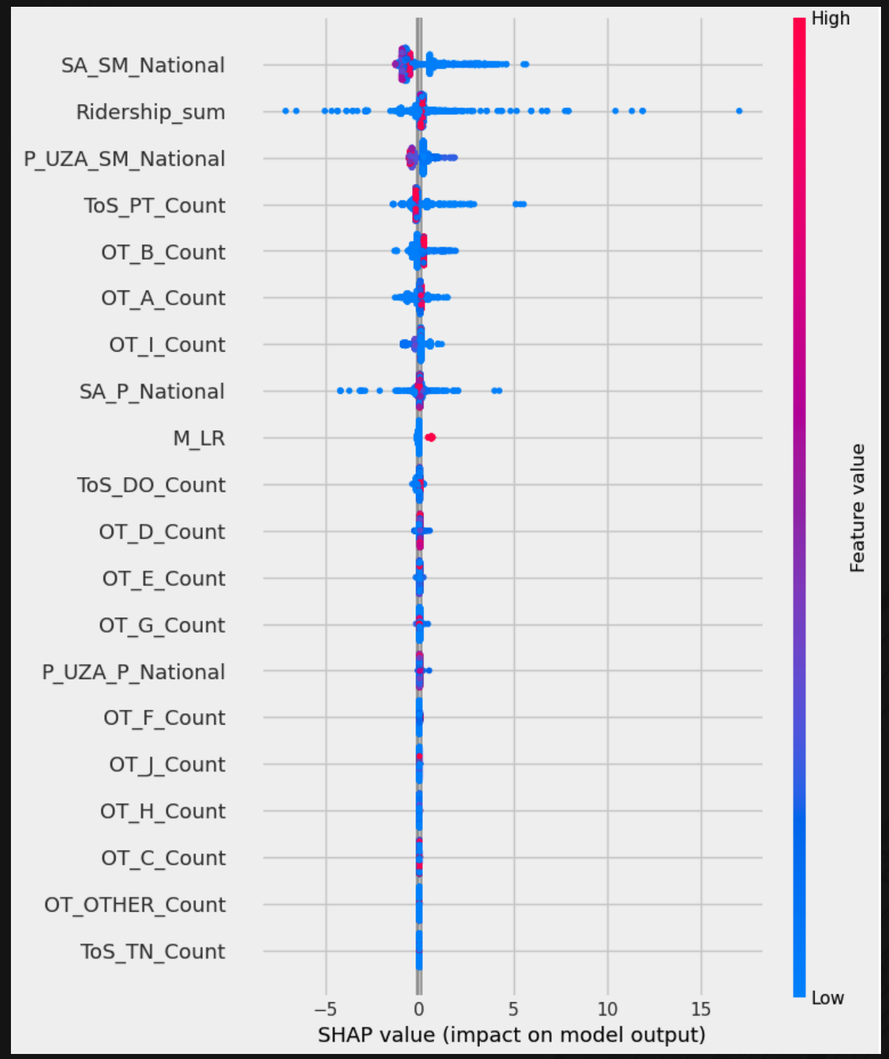
Which modes have the higest and lowest safety risk?

TAMARA HORNE

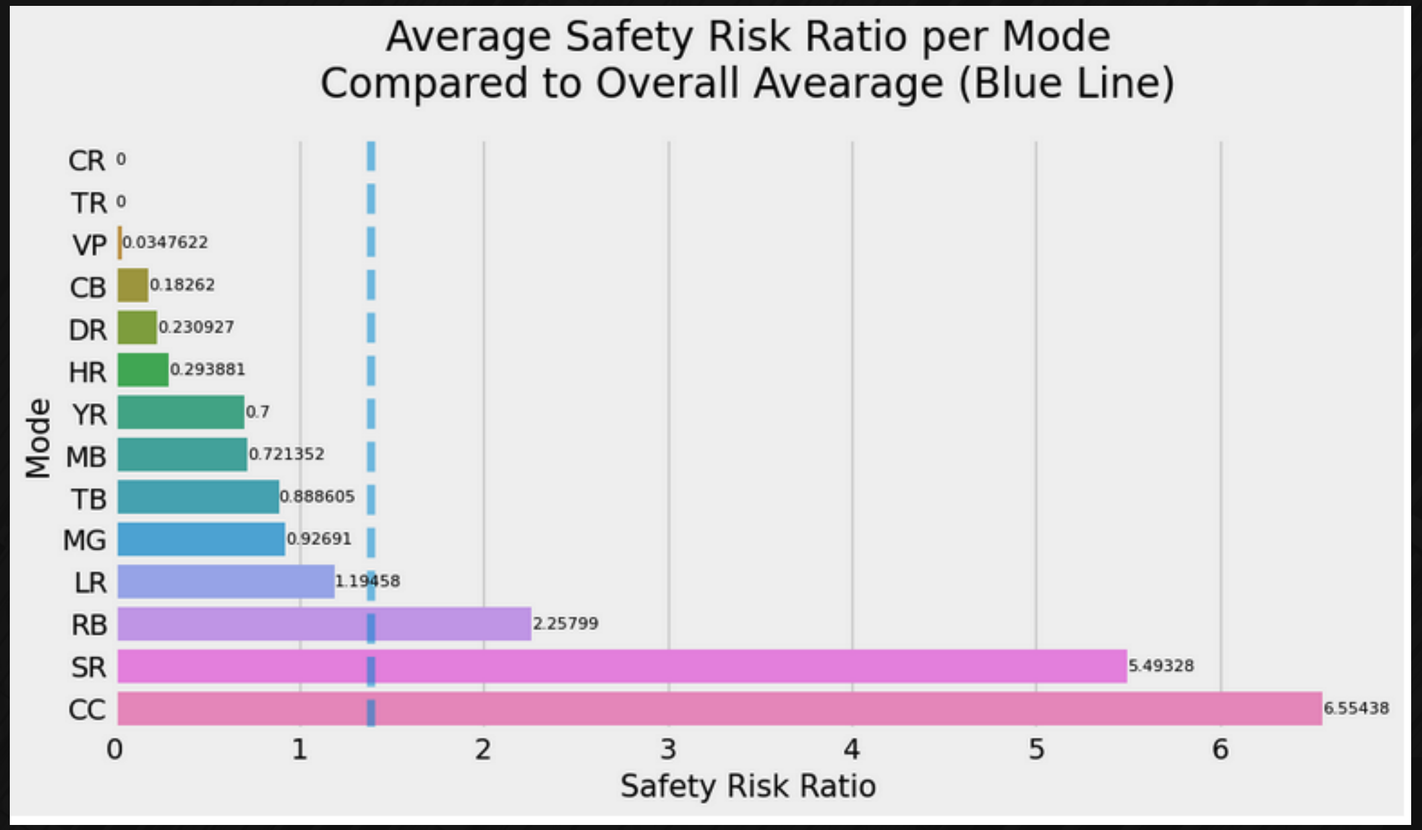
2023 | March

TAMARA HORNE

FEATURE IMPORTANCE



MODAL SAFETY RISK



TAMARA HORNE

CONCLUSIONS

Best Models

- XGBoost Regressor
- Random Forest Regressor

Aggregations

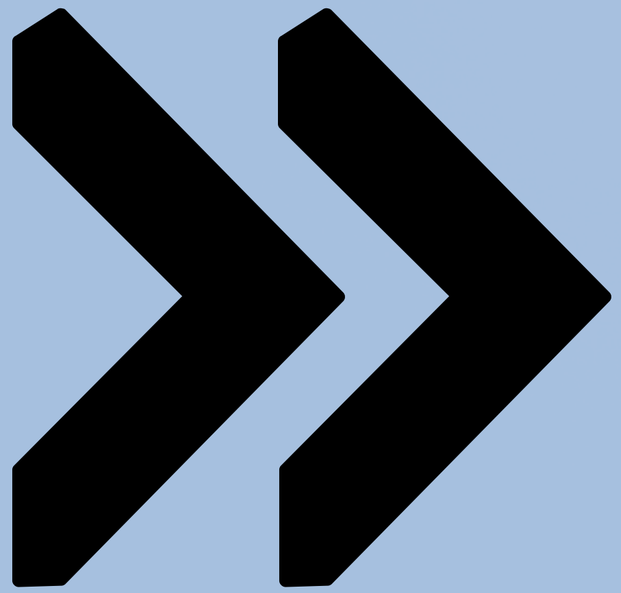
- Date, Nation
- Date, Location
- Date, Mode

Prediction Power

- Safety risk overall for the nation
- Safety risk for Nashville
- Safety risk for specific modes

TAMARA HORNE

FUTURE WORK



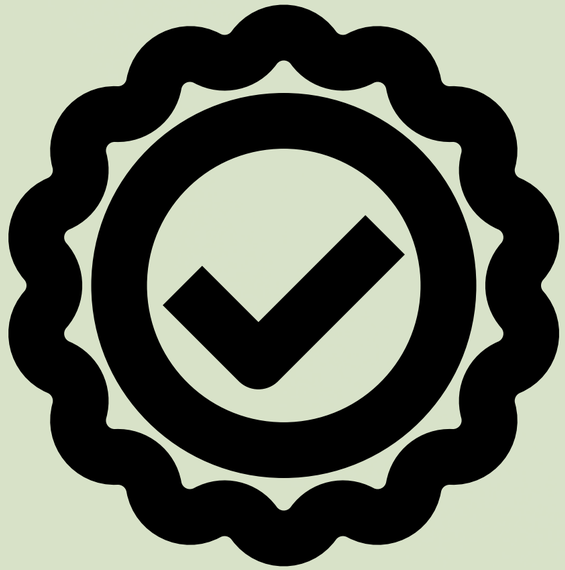
Small

- Diagnose shifted data in modeling_df2
- Explore two low ridership/high SHAP values for modeling_df2

Big

- Import data needed for System Reliability and incorporate into the definition of the target
- Time series modeling

TAMARA HORNE



RECOMMENDATIONS FOR THE CLIENT

- Adding light rail would increase safety risk
- Adding service to existing modes would be a good choice
  - Added vehicle revenue miles will differ according to mode; this will affect the target
  - Safety risk is lowest for CR and highest for MB

MORE INFORMATION

https://github.com/tamarahorne/Springboard/tree/main/Capstone%20Project%202

TAMARA HORNE
2023 | March