Crashes, Conditions, & Classification

By: Mia, Emily, and Aurora

The Data - US Accidents

Description

- Covers U.S. car accidents (2016-2023)
- Data collected from traffic sensors, cameras, and APIs
- 500k sampled accidents across 49 states

Variables

- Target: Severity level (1-4)
- Predictors: Humidity, temperature, distance, etc.

Source

- Dataset from Kaggle
- Data was collected using traffic feeds from departments of transportation, law enforcement, and road sensors

Incorporating Other Data: US Cities

Description

From:

1. U.S. Geological Survey
2. U.S. Census Bureau

Contains:

Information such as population and density for over 30,000 US Cities

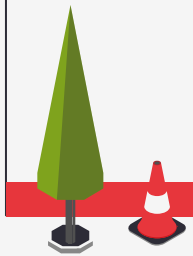
Important Variables

Joining On:

- City
- State

Adding from US Cities:

- Population
- City Density

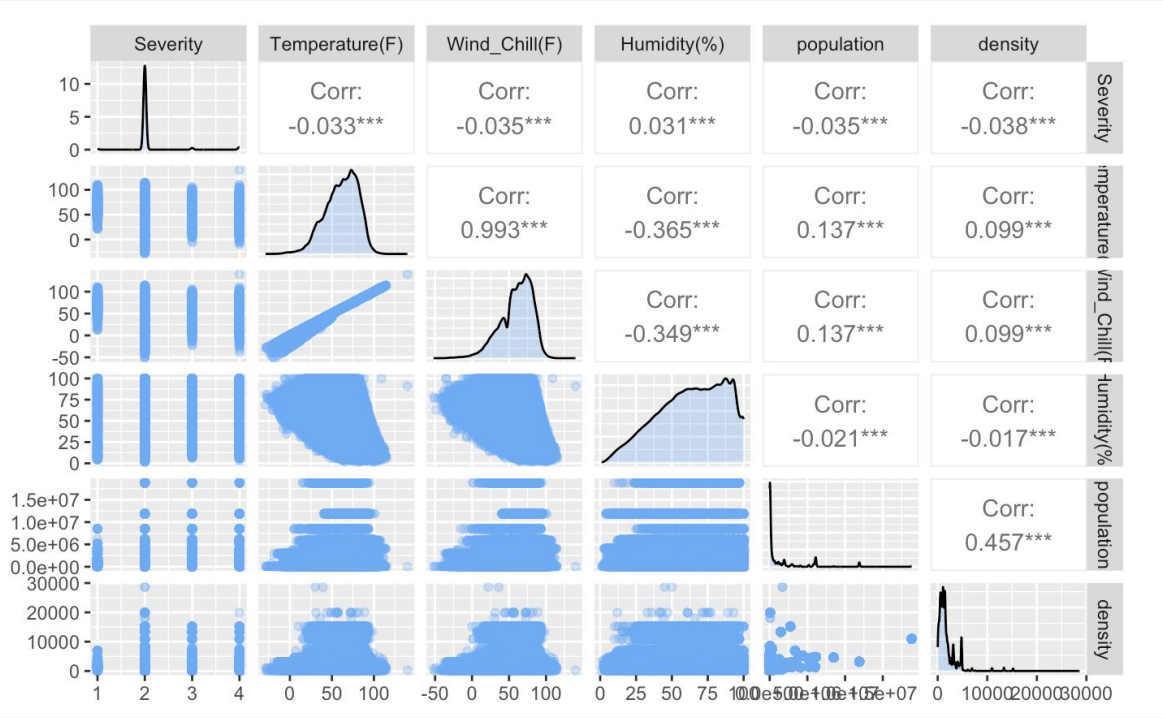


Data Cleaning

1. Remove NA values from data
2. Join the two datasets on ‘State’ and ‘City’
3. Select variables desired and standardize them

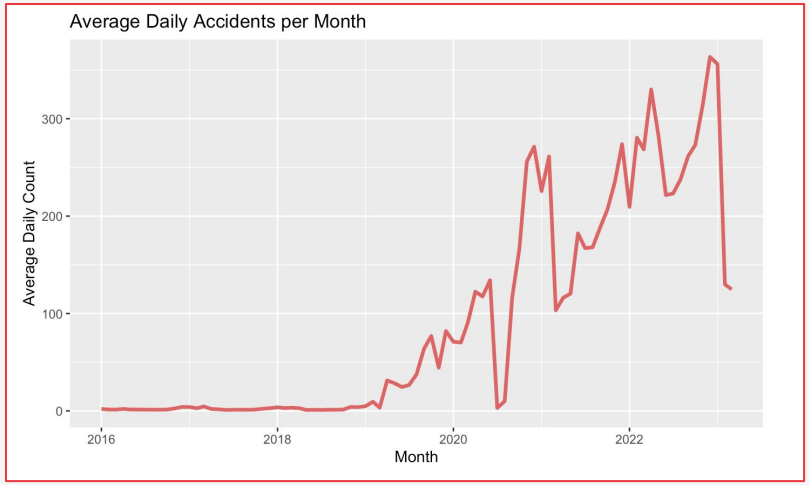
01 Exploratory Data Analysis

Pair Plot

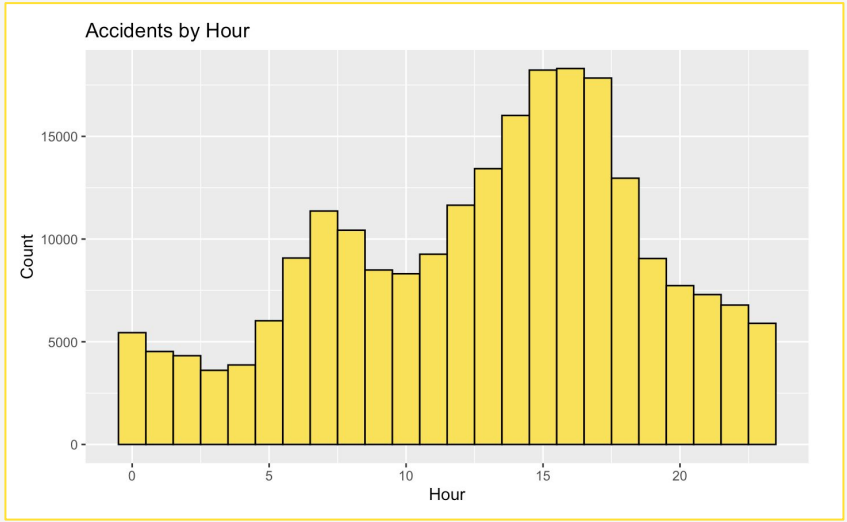


- Severity tends to have a weak correlation with the variables
- 2 Severity has a larger spread for all variables
- 2s and 4s occur in colder temperatures
- 1 Severity occurs in less dense areas

Accidents Over Time

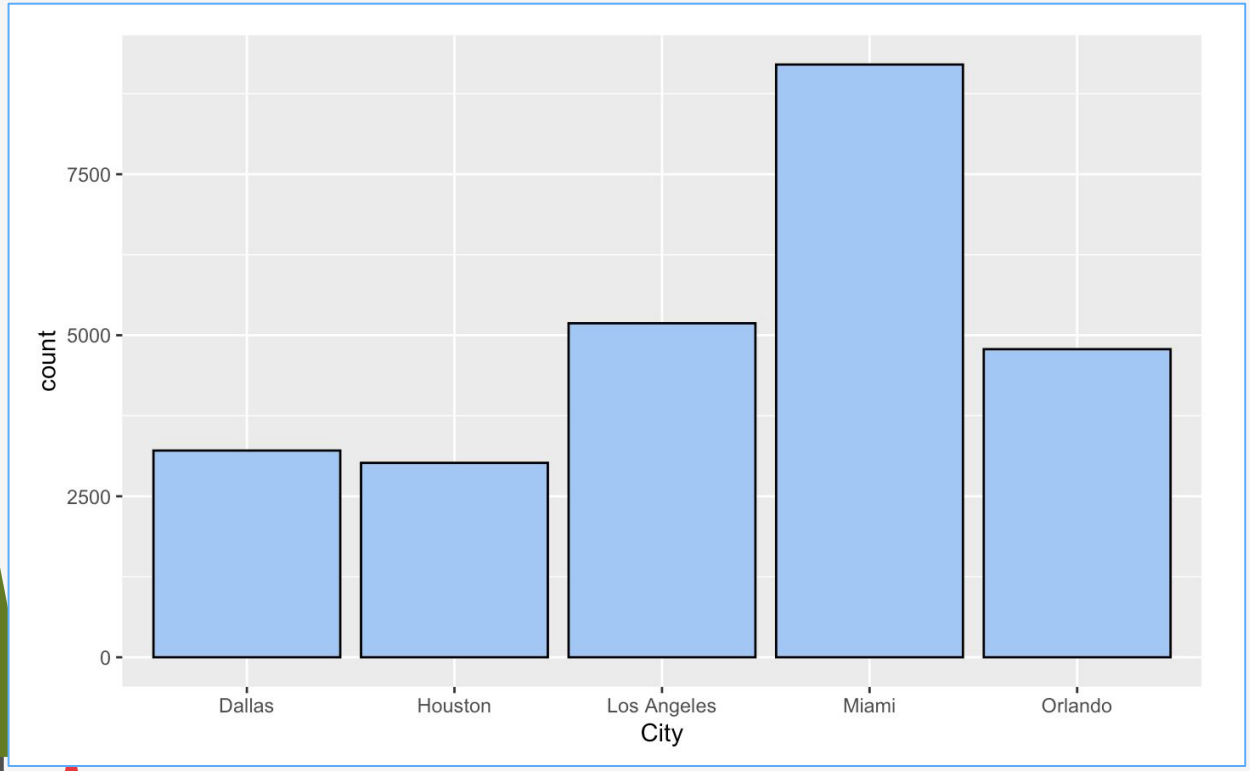


- Accidents have increased over the years
- Practically 250% increase



- Majority of accidents occur between 6-9 and 2-5

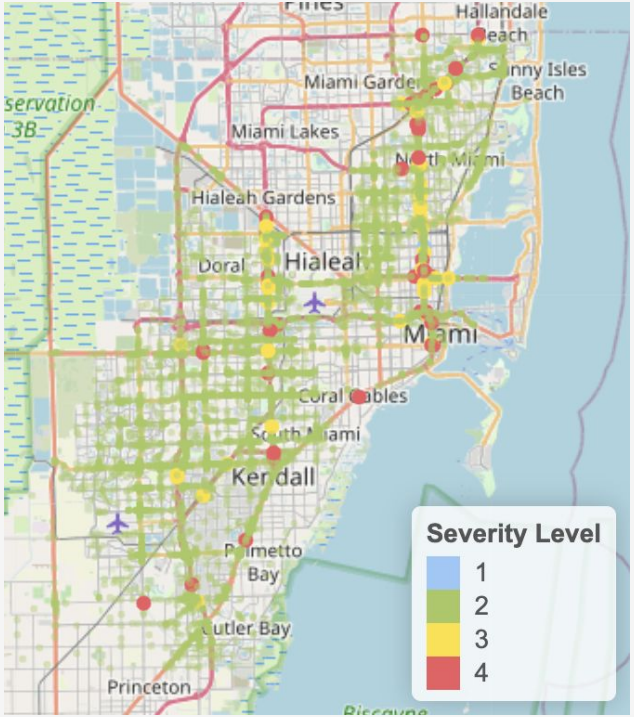
Top 5 Cities



- Observe that our top city is Miami, followed by LA
- All have over 2500 accidents recorded

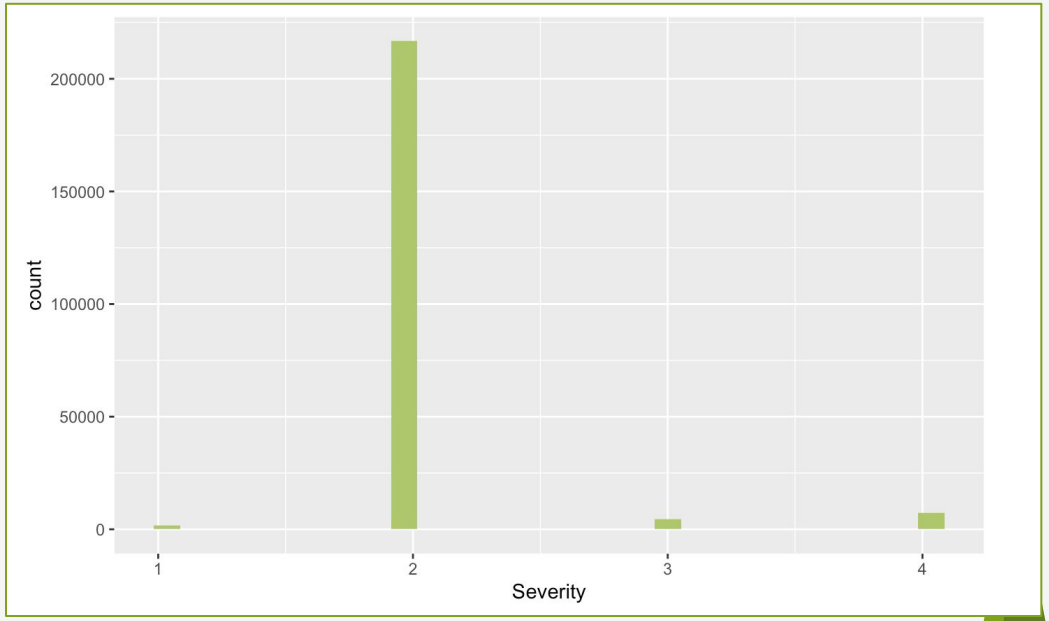
Accidents in Miami

- Distribution of accidents in Miami (top city in data)
- Can observe majority of accidents of 2
- Majority of all accidents occur at intersection
- Higher level accidents almost always at intersections



Distribution of Severity

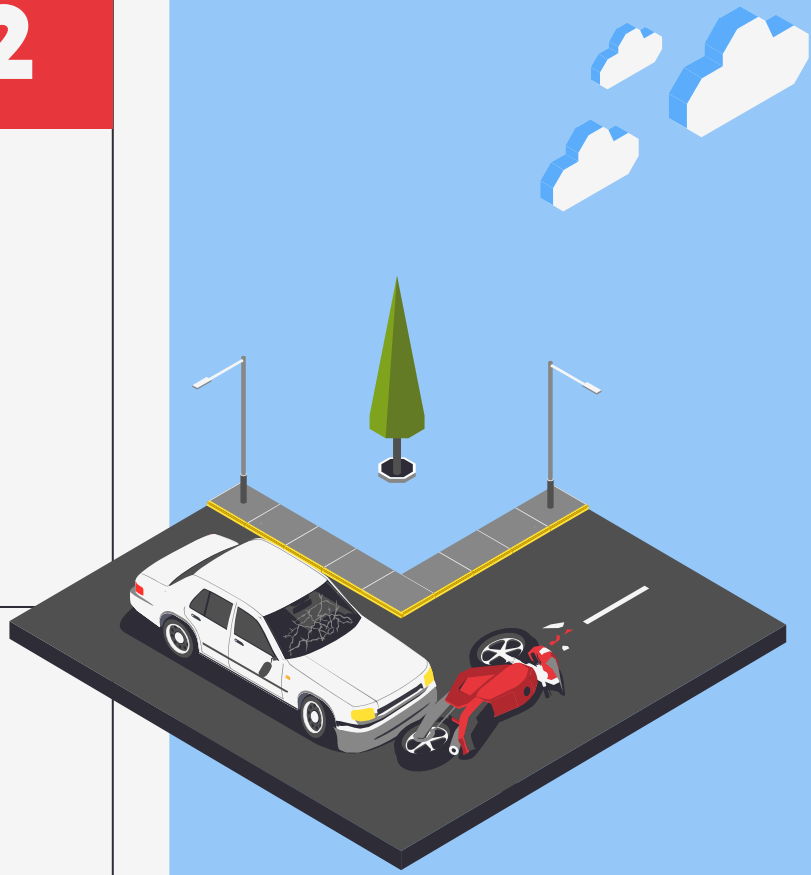
- 2 Severity have a strong domination
- We see 1 the least in the data
- Will skew our classification model



02

KNN Algorithm

- Mi
a



Process

Stratify by Severity

Limit #2 Severity to 7500 as too many occur

Separate

Separate data into 70% training and 30% testing

Training

Train our model with the training data

Improve

Look to how we can improve our model accuracy

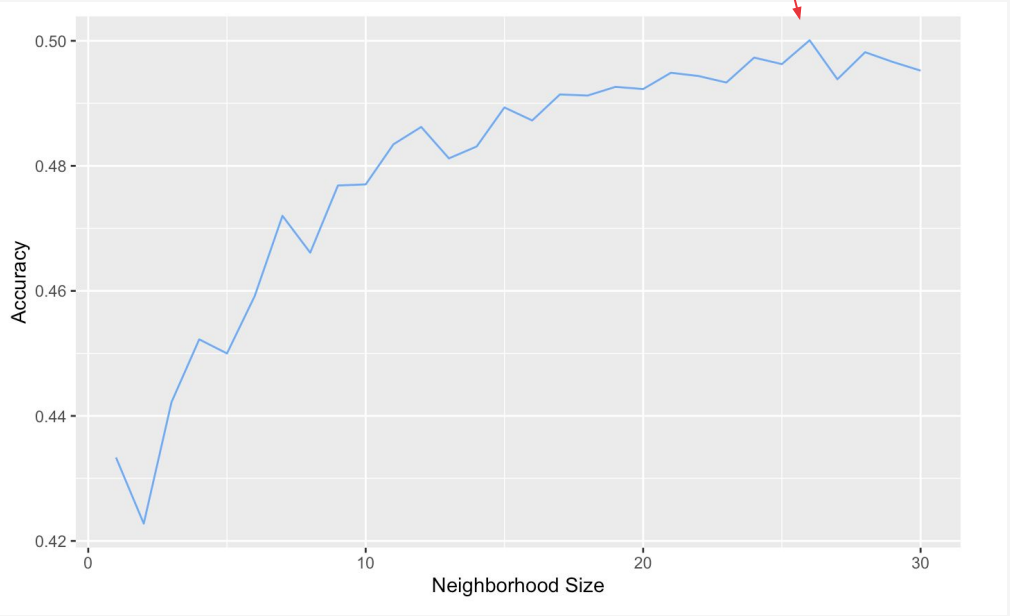
KNN Confusion Matrix

	1	2	3	4
1	170	151	78	77
2	139	1059	359	678
3	57	388	447	297
4	84	652	308	825

- Start with neighborhood of 1
- Accuracy of 0.433
- Severity 2 has highest identification rate of 0.47

Best Neighborhood

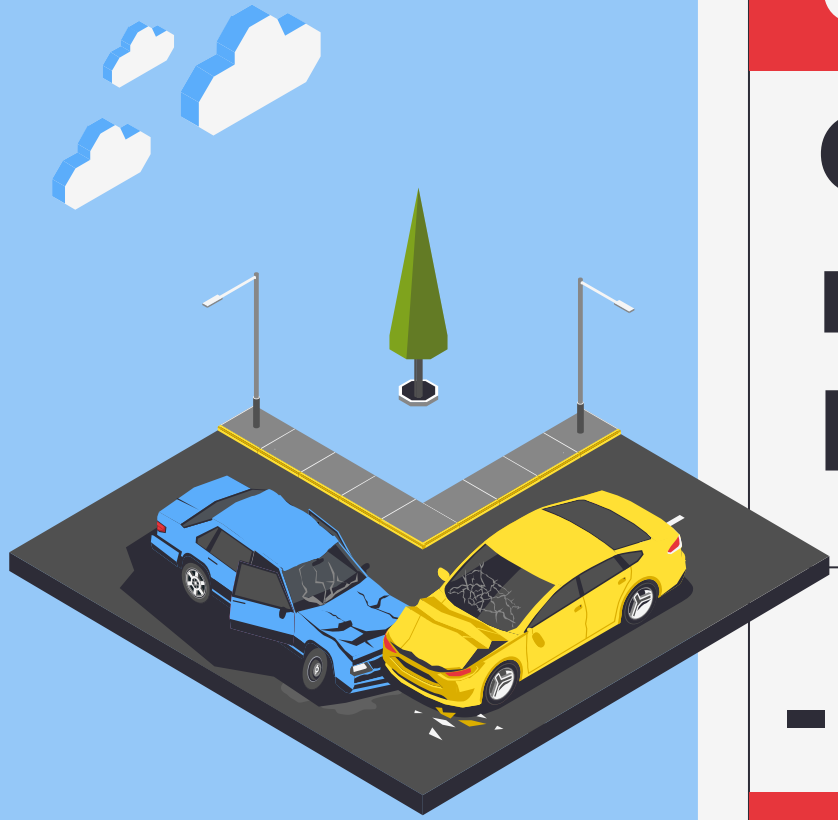
- Neighborhood size with the best accuracy rate is 26
- Increases significantly from 1-10, then slows down



Confusion Matrix with Optimal K Size

	1	2	3	4
1	105	45	13	23
2	206	1366	474	679
3	57	210	418	164
4	82	629	287	1011

- Neighborhood of 26
- Accuracy of 0.495
- Improvement of 0.06 from original matrix



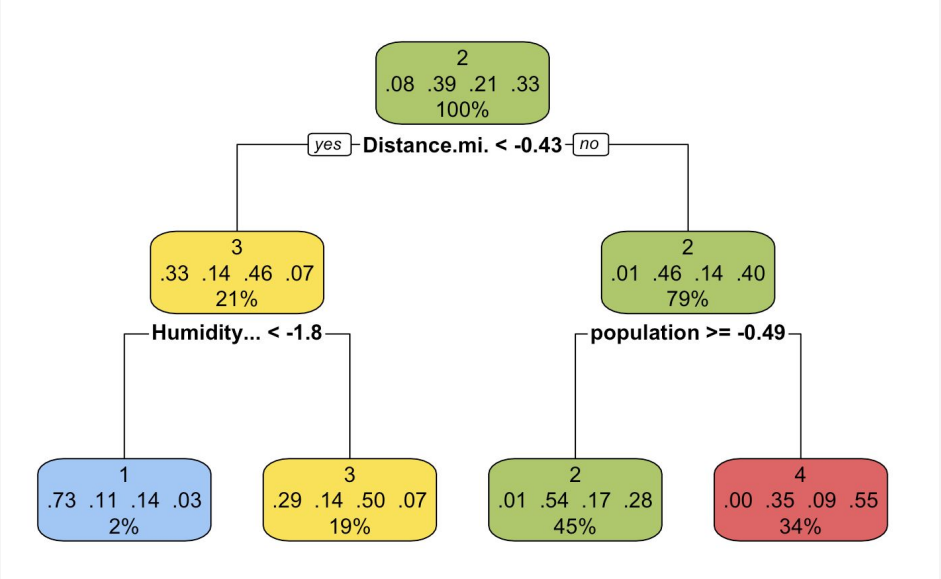
03

Classification Trees and Bagging

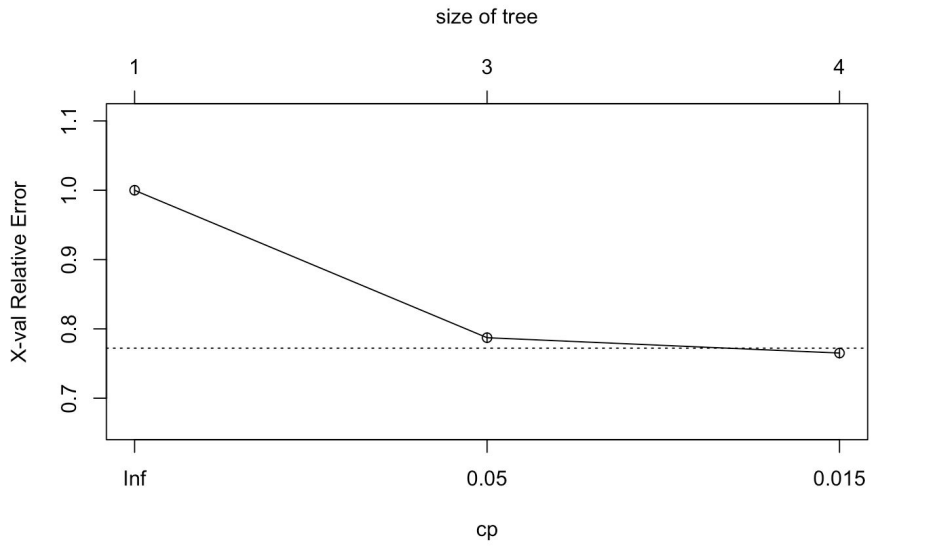
- Emily

Severity Patterns from Our Classification Tree

- Distance, humidity, and population were key split variables
- Low distance + low humidity → more severe crashes (Severity 3)
- Higher population → linked to more severe crashes (Severity 4)
- Low population → linked to moderate crashes (Severity 2)



Pruning the Classification Tree



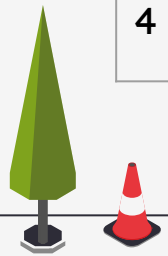
- Tried pruning to simplify the tree
- Minimum cross-validation error occurred at 4 splits
- Pruned version ended up identical to the original

Classification Tree Accuracy

Accuracy of 0.54

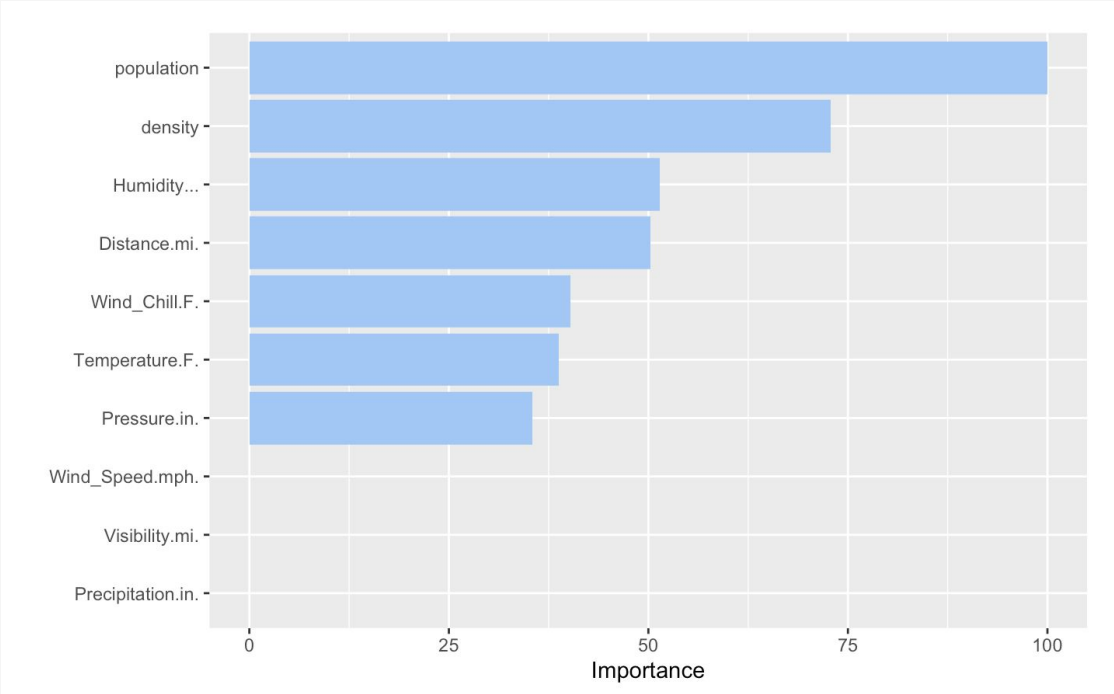
	1	2	3	4
1	76	33	329	11
2	11	1354	157	728
3	19	440	546	186
4	3	711	61	1104

- Strongest predictions for Severity 2 and 4
- Most confusion occurs between Severity 2 and 4
- Tree performs worse with Severity 1 and 3



Bagging: Importance of Variables

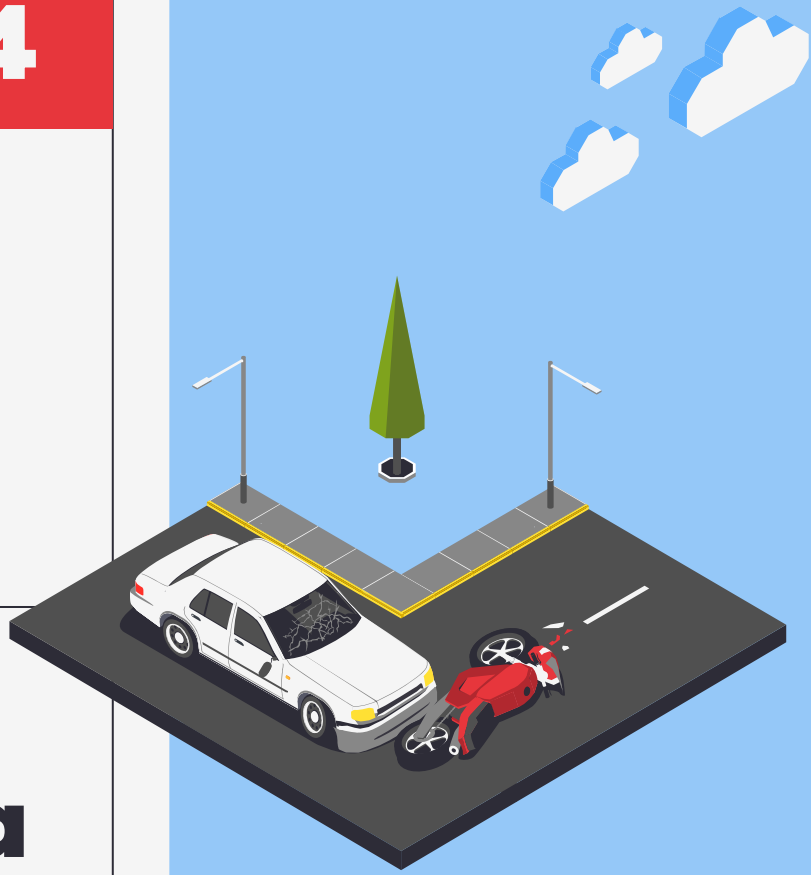
- Population is the strongest predictor
- Followed by density and humidity
- Weather features matter, but less than demographic ones
- Bagging model prioritizes urban and environmental context



04

Logistic Modeling

- Aurora



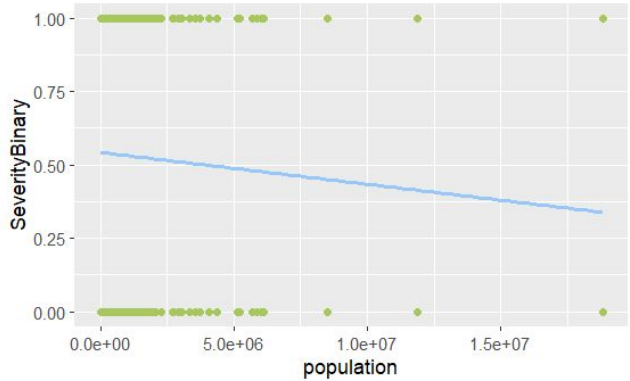
Binary Classification

Data Preparation

- Created binary variable “SeverityBinary”
  - Mild = Severity < 3
  - Severe = Severity >= 3
- Balanced Dataset
  - Downsampled Severity = 2
- Converted labels to numeric binary
  - Mild = 0
  - Severe = 1

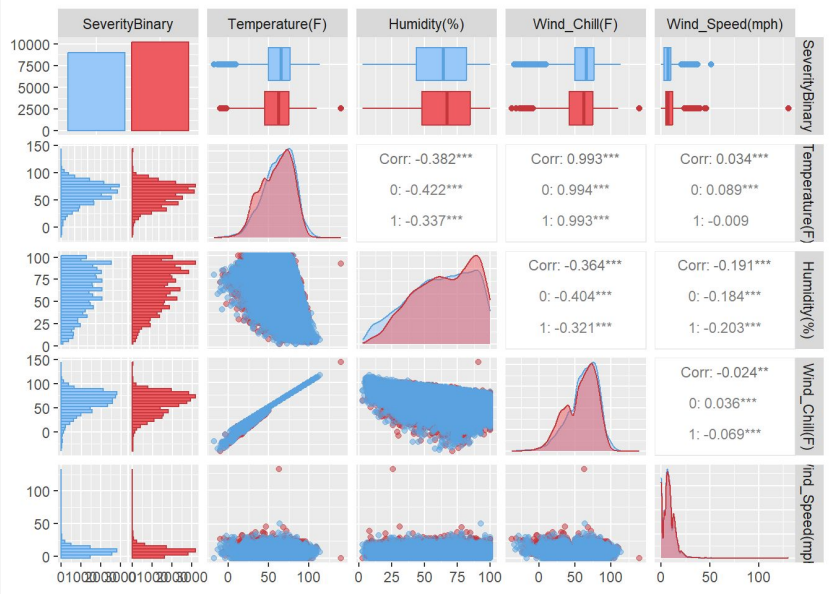
Data Splitting

- Split the data into training and test sets
- Graphed training data



Exploratory Analysis

- Used ggpairs (again) to look at interactions with the new binary variable



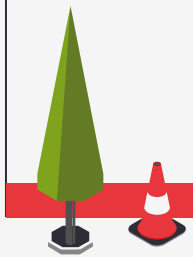
Logistic Regression Models

Simple Logistic

- Only shows impact of one feature
- Looked at how number of people might affect severity
- SeverityBinary~population
- Accuracy = 54.4%

Multiple Logistic

- Captures joint effects of multiple variables
- Looked at multiple weather conditions affect on severity
- SeverityBinary~ Temperature + Humidity + Wind Chill + Wind Speed
- Accuracy = 56.3%



What We Learned

- Multiple logistic regression provided better predictive performance than simple logistic regression
- Both models showed moderate performance, suggesting limited predictive power
- Weather data can hint at variance in severity, but isn't a strong enough predictor
- Population alone isn't enough of an indicator, though it would be worth diving deeper into things like population density and traffic density as other variables