Info challenge

Washington Fatal Crash File

Xirui Han, Johannah Ryan, Tam Thu Doan, Kevin Weiner

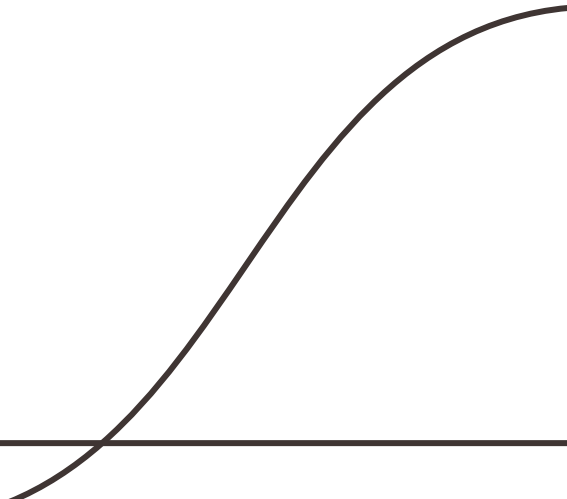
Table of contents

Project Objective

This project aims to analyze the relationship between fatal crashes and communities in Washington state. The project plan comprises two stages, which include dataset preparation and exploratory data analysis (EDA).

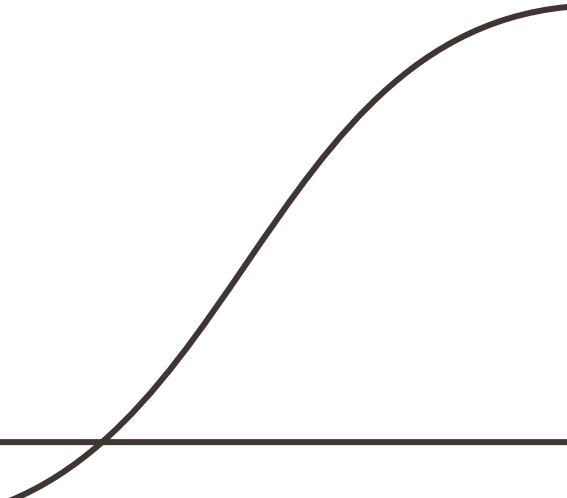
Phase 1

In Stage 1, the project will gather Washington Fatal Crash Files, parse and load them into DataFile, check the consistency of the files with the data dictionary, add missing ZIP codes to the dataset, and remove duplicates and invalid data. The cleaned data will be saved as an output.

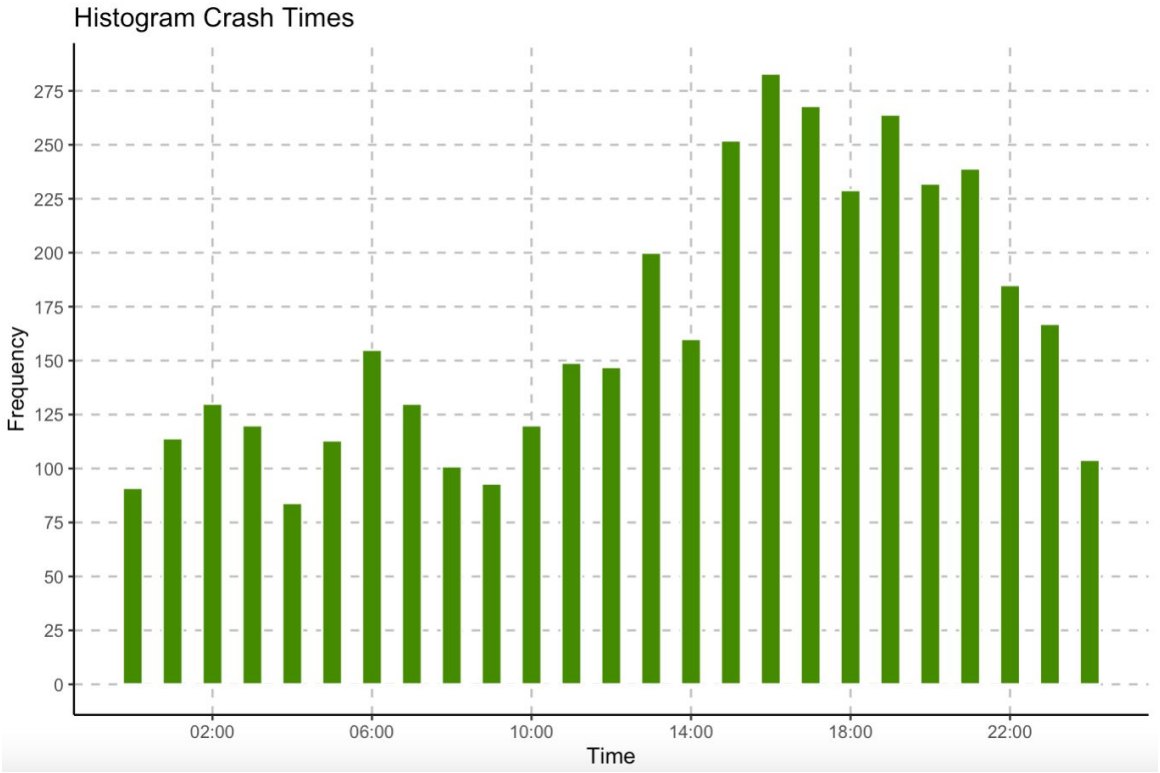


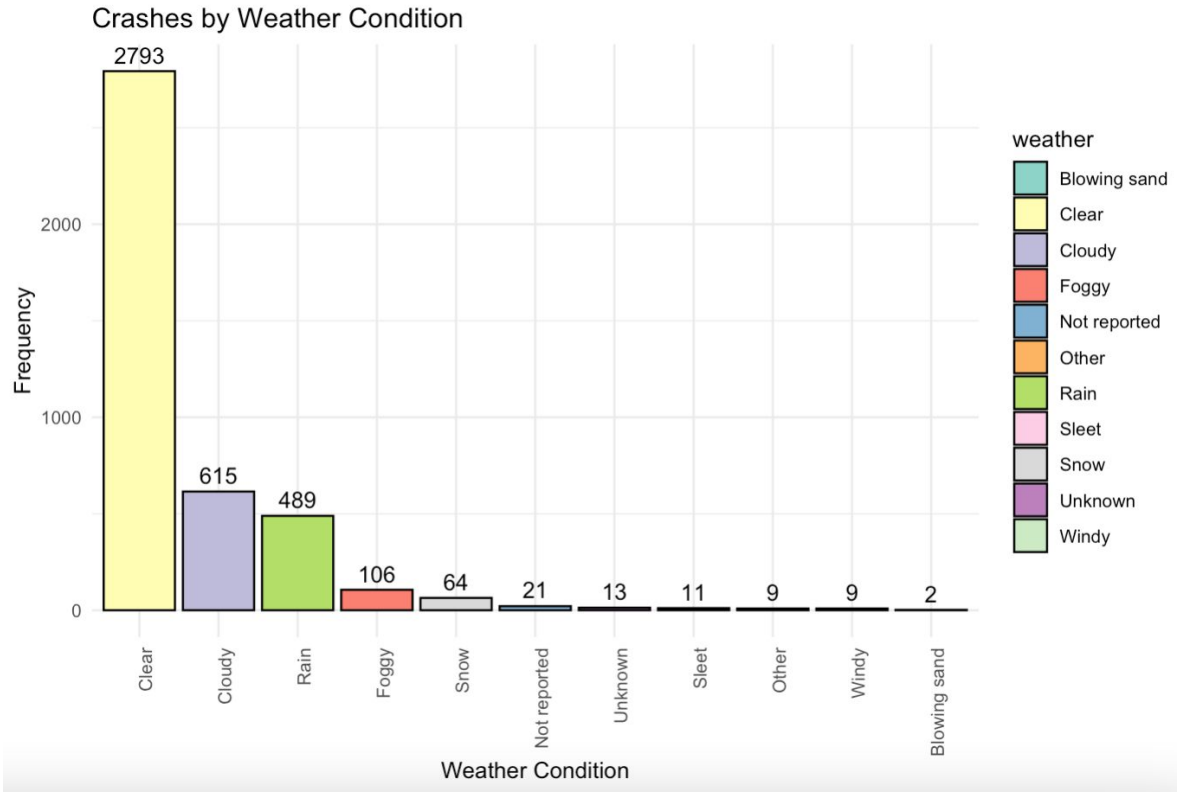
Phase 2

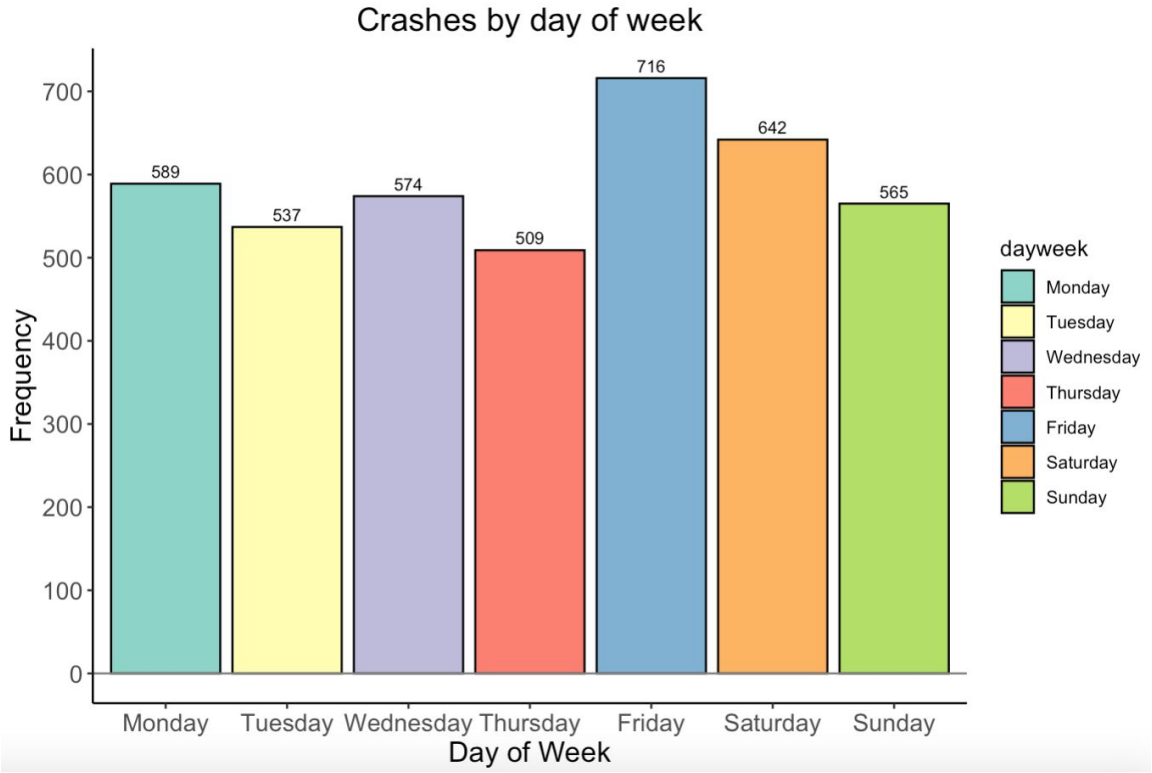
In Stage 2, the project will conduct EDA to identify patterns and relationships. The data will be summarized using descriptive statistics, cross-tabs stats, and correlation analysis. Highly correlated variables will be identified, and a data dictionary will be created to include new or planned calculated variables.

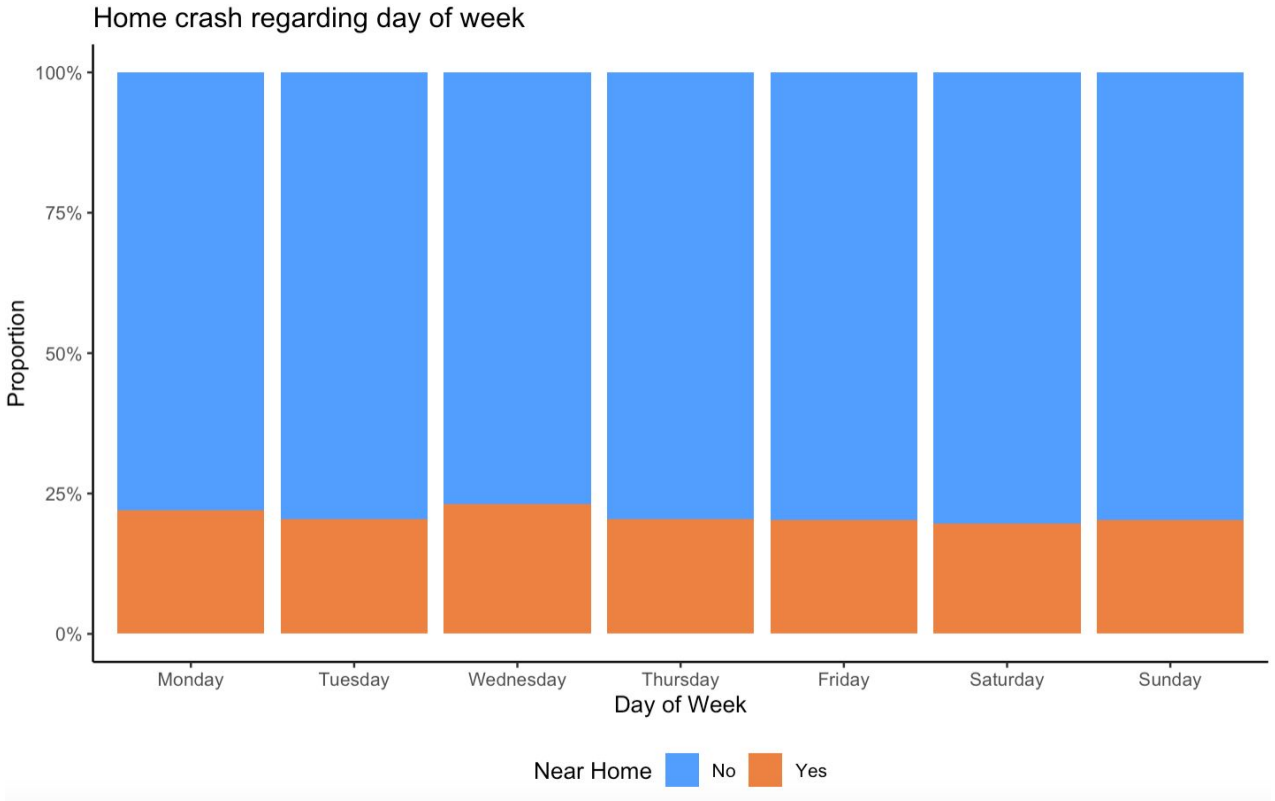


General Factors


Histogram Crash Times
Frequency
275
250
225
200
175
150
125
100
75
50
25
0
02:00
06:00
10:00
14:00
18:00
22:00
Time

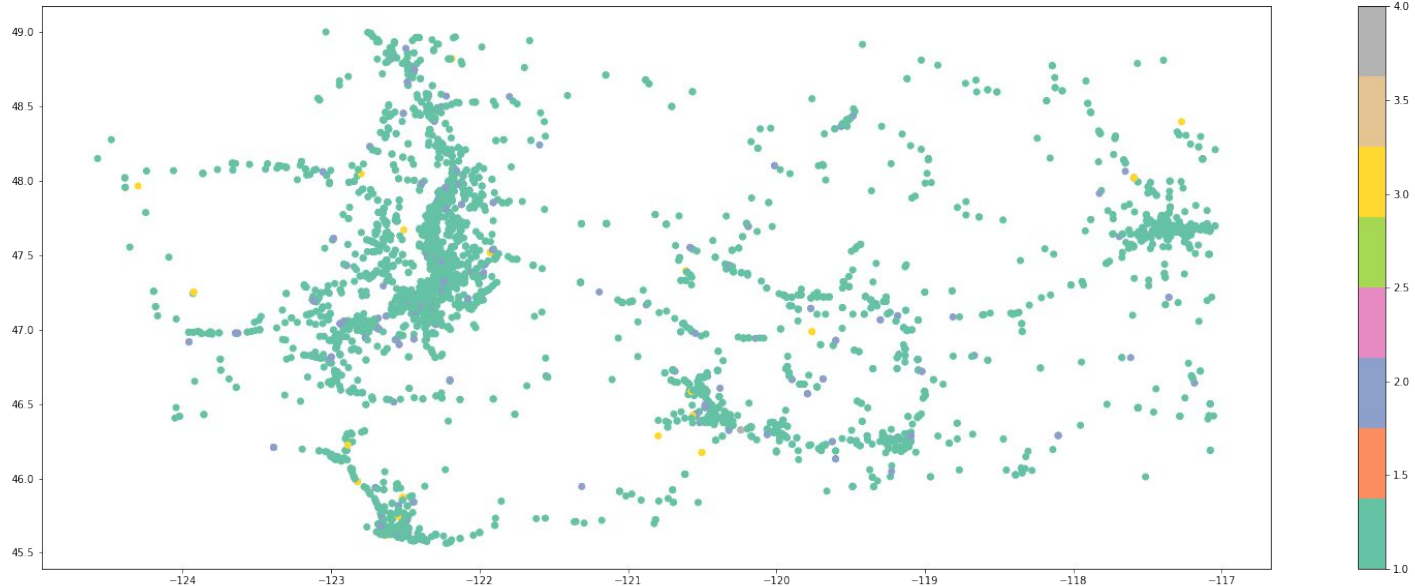

Crashes by Weather Condition
2793
615
489
106
64
21
13
11
9
9
2
2000
1000
0
Frequency
Clear
Cloudy
Rain
Foggy
Snow
Not reported
Unknown
Sleet
Other
Windy
Blowing sand
Weather Condition
weather
Blowing sand
Clear
Cloudy
Foggy
Not reported
Other
Rain
Sleet
Snow
Unknown
Windy


Crashes by day of week
Frequency
0
100
200
300
400
500
600
700
589
537
574
509
716
642
565
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week
dayweek
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

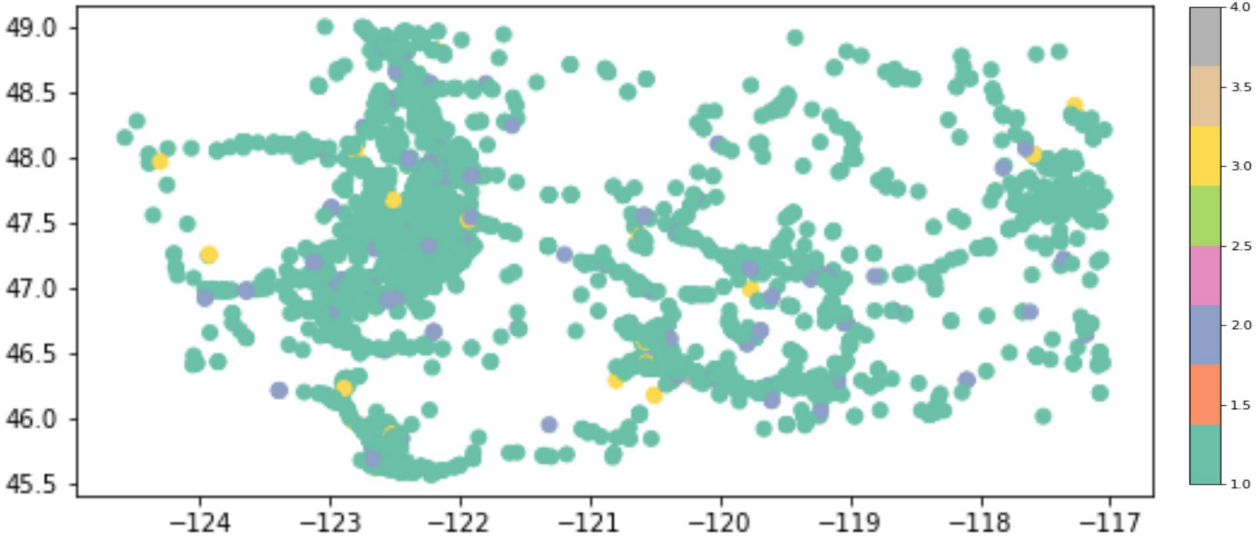

Home crash regarding day of week
100%
75%
50%
25%
0%
Proportion
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week
Near Home
No
Yes

Defining Community

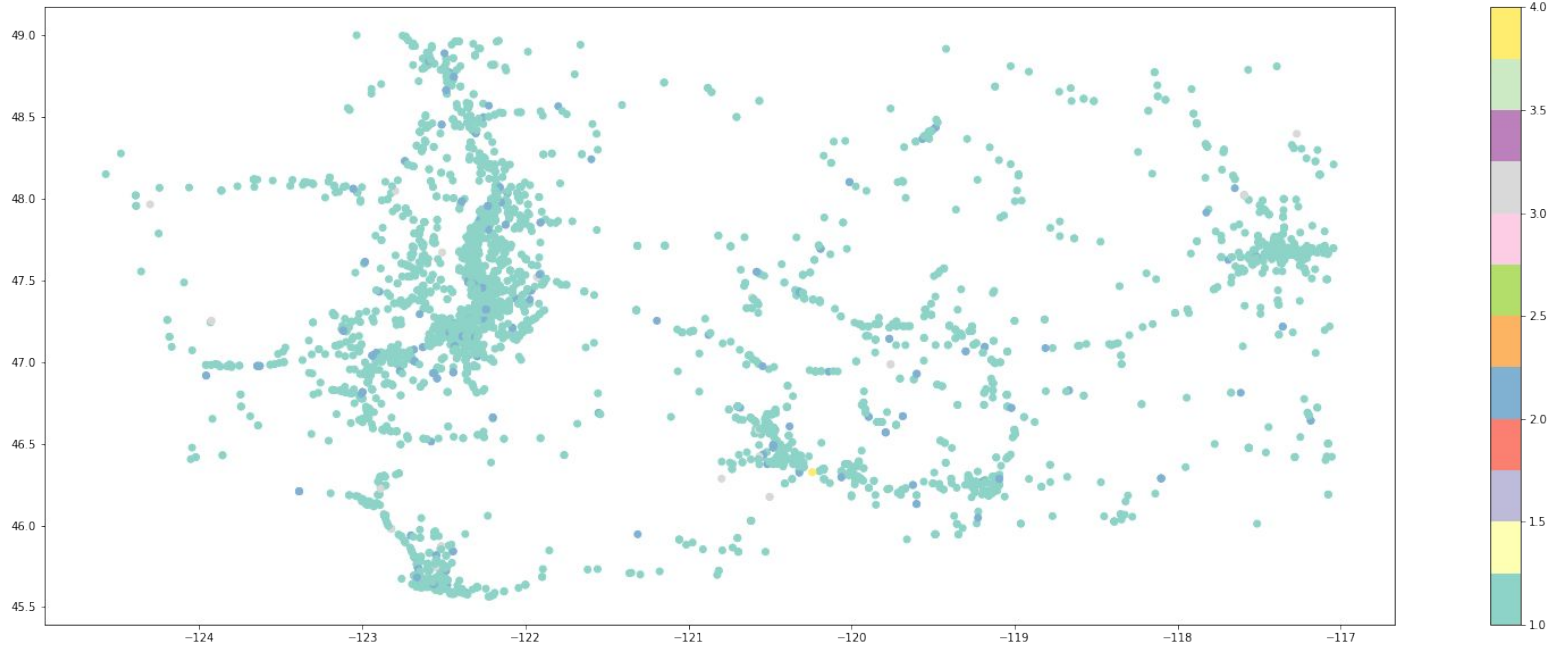
Fatal Crashes at a Glance

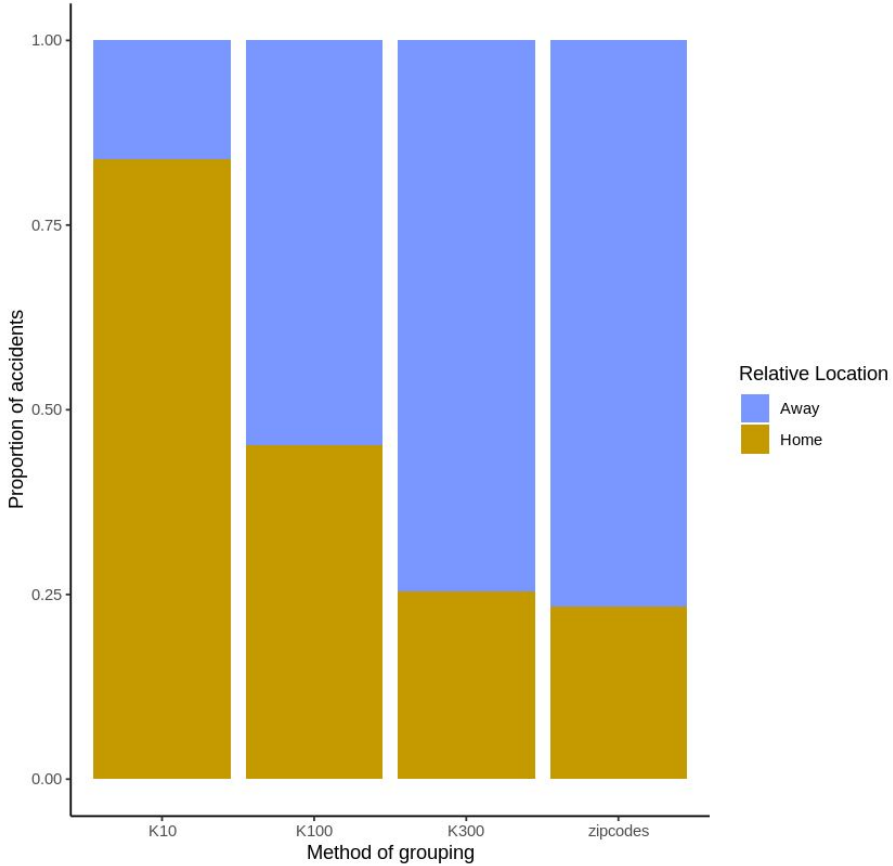


Fatal Crashes at a Glance



Fatal Crashes at a Glance

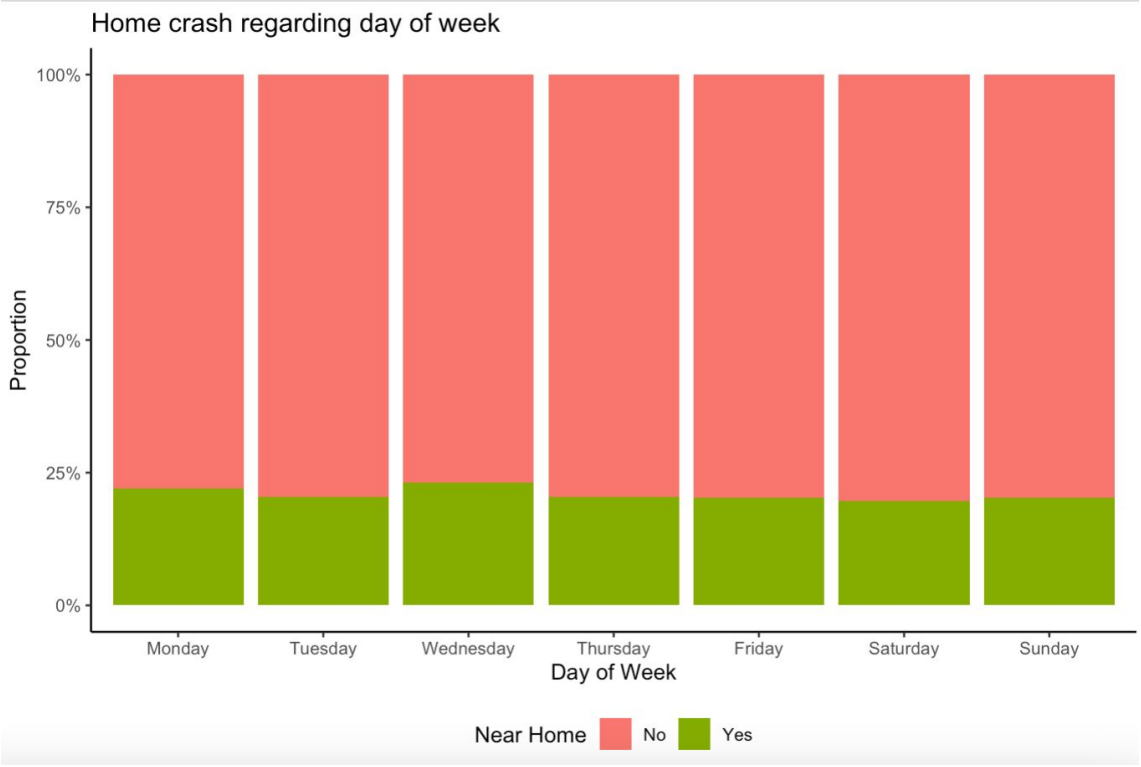




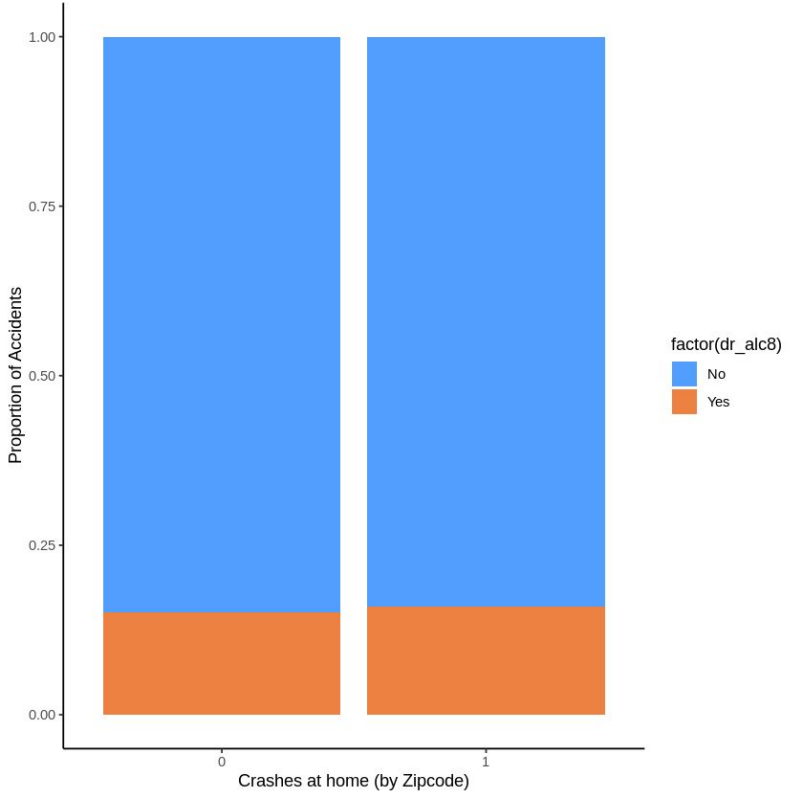
K	Silhouette Score
10	0.507
100	0.421
300	0.453

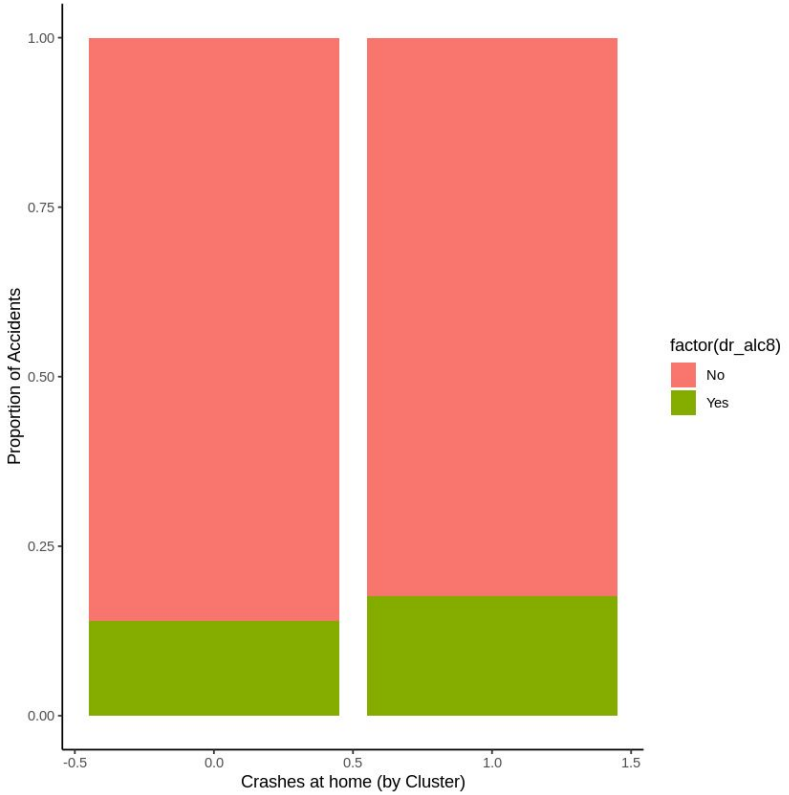
Crashes at Home

Clustering VS zipcodes

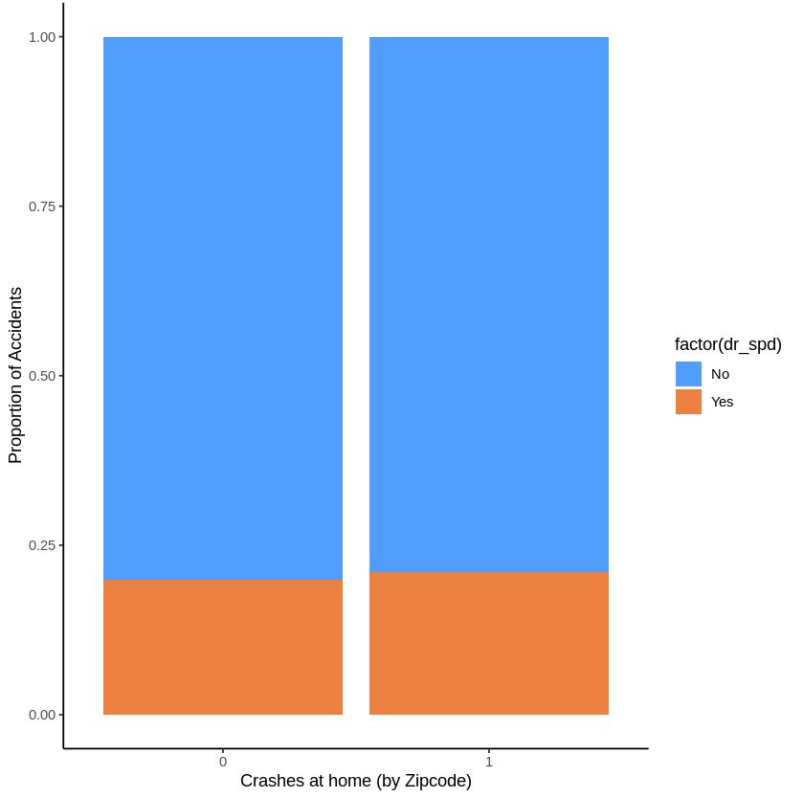


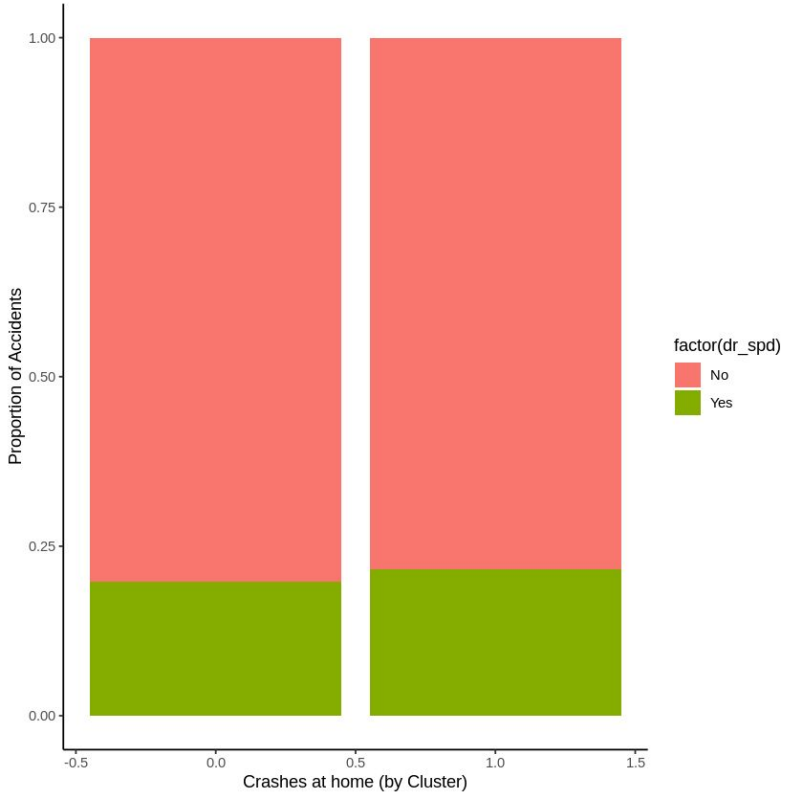
Alcohol Involvement



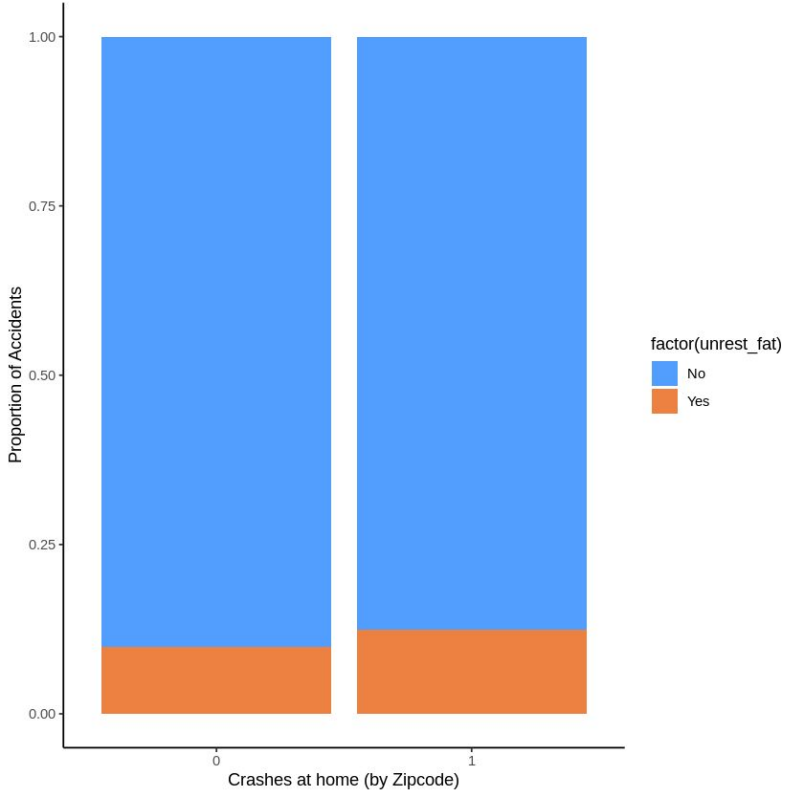


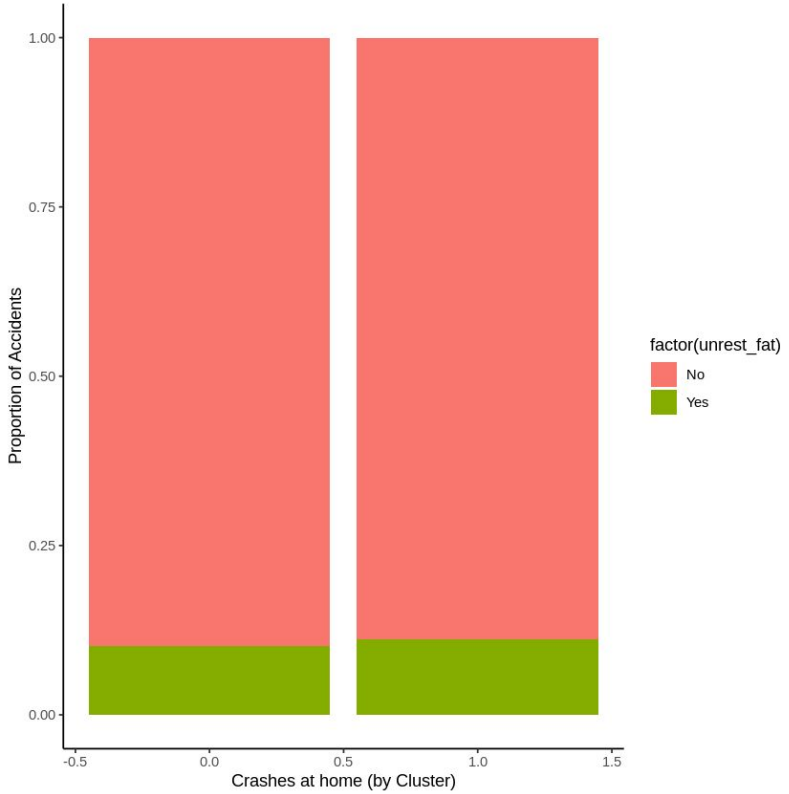
Speeding


1.00
0.75
0.50
0.25
0.00
Proportion of Accidents
0
1
Crashes at home (by Zipcode)
factor(dr_spd)
No
Yes

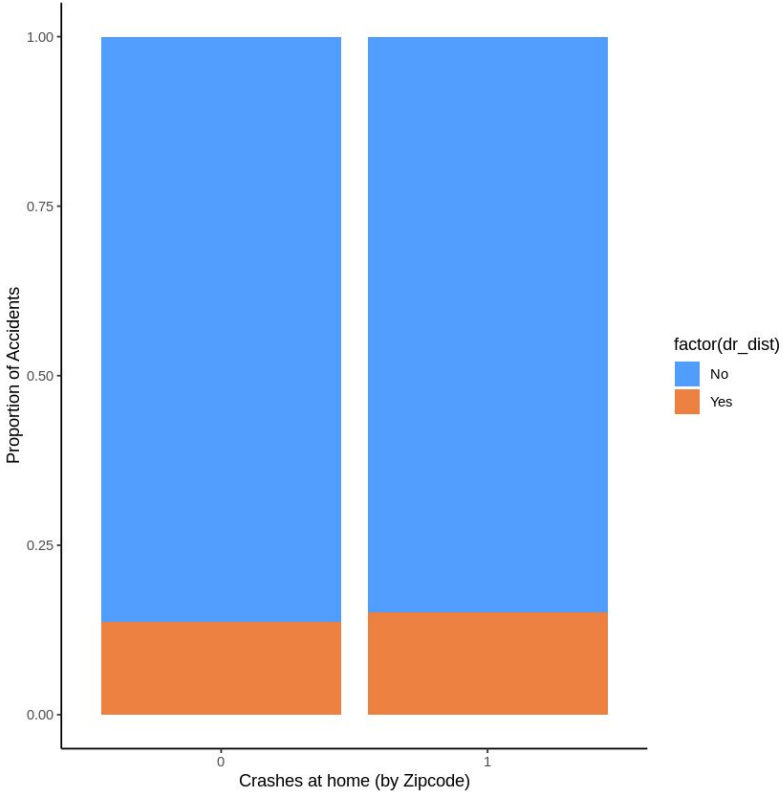

1.00
0.75
0.50
0.25
0.00
Proportion of Accidents
-0.5
0.0
0.5
1.0
1.5
Crashes at home (by Cluster)
factor(dr_spd)
No
Yes

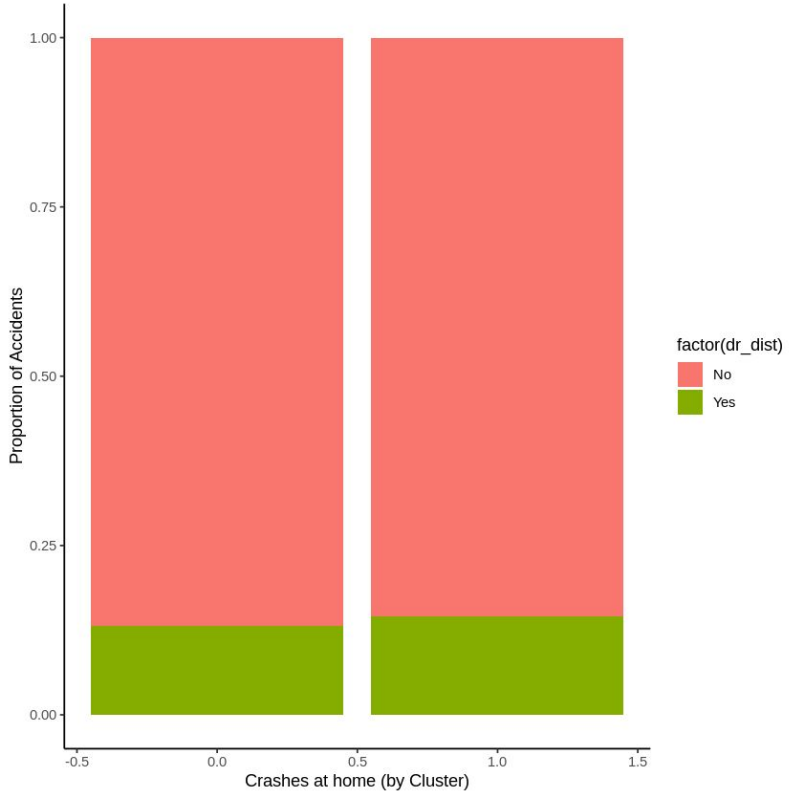
Lack of seatbelt



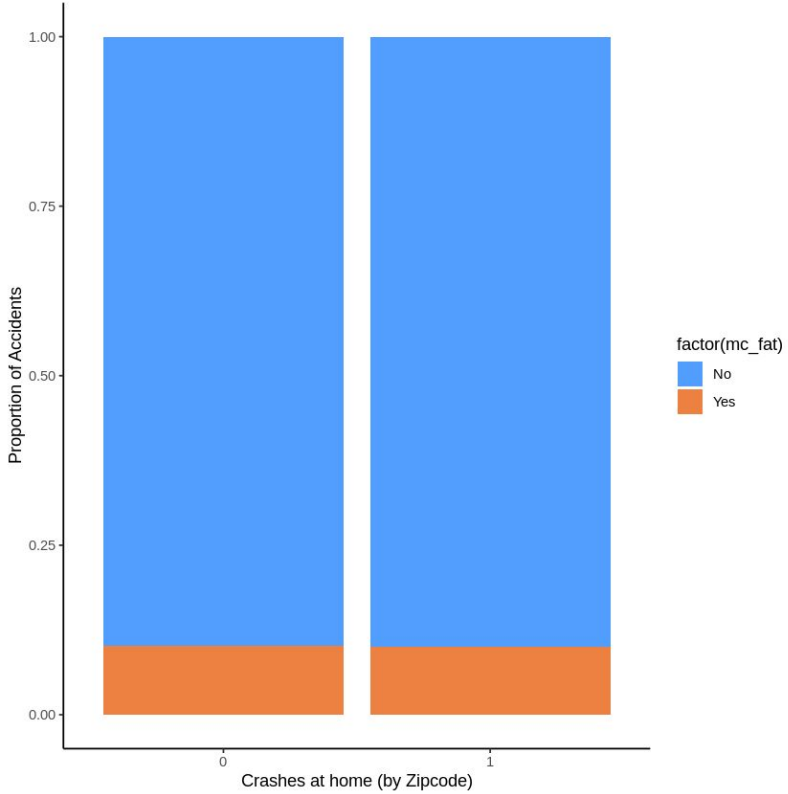


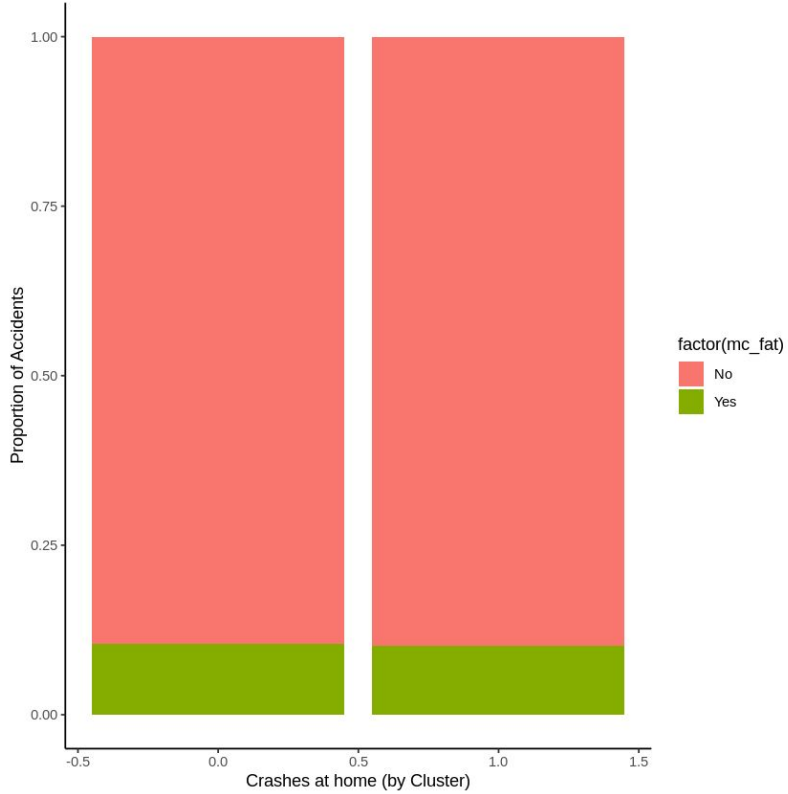
Distracted Driver





Motorcycle Fatalities



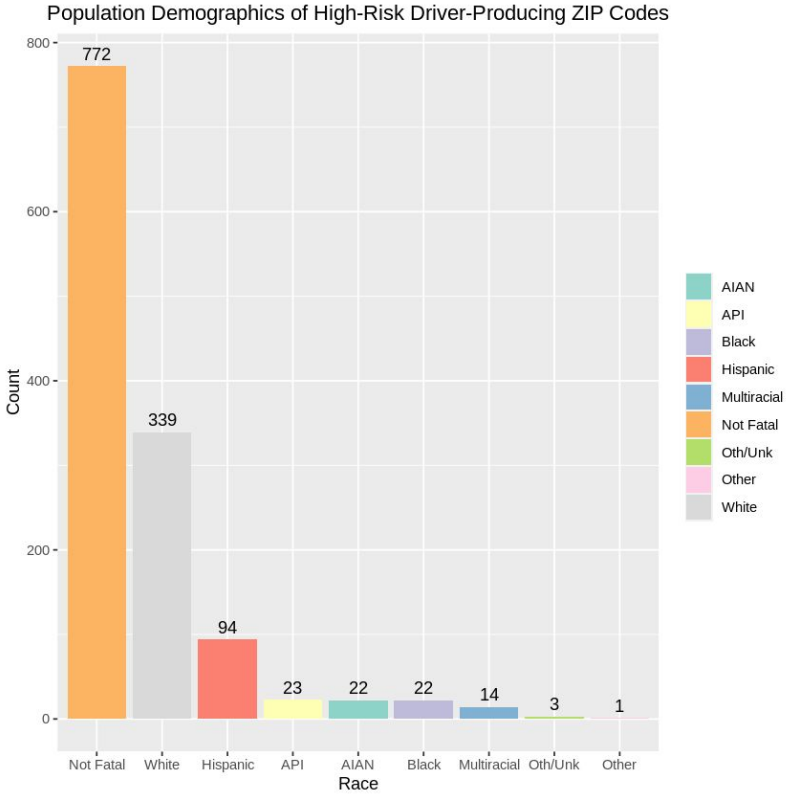


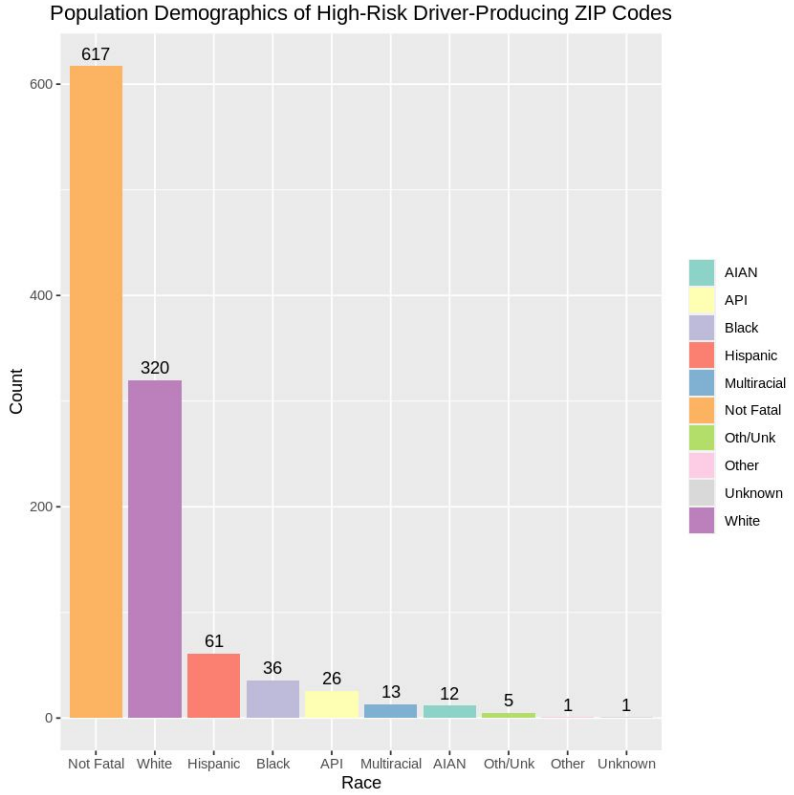
High Risk Communities

Method of define high risk drivers/communities:

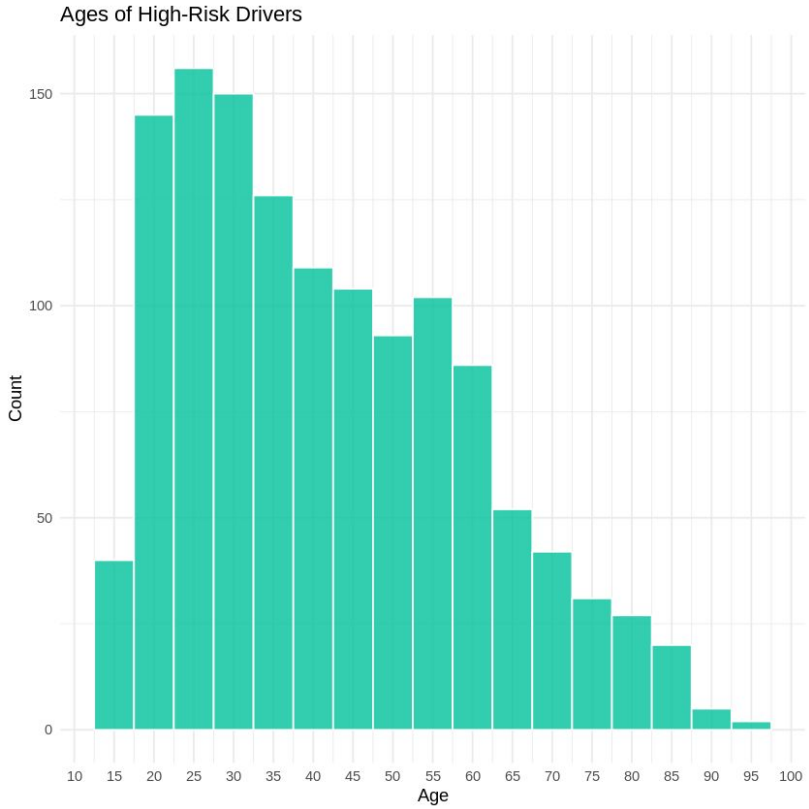
- Frequency : Number crashes in dzip/ total number of crashes
- Fatality rate:Number of Fatalities in Crash/Number of Persons in Crash

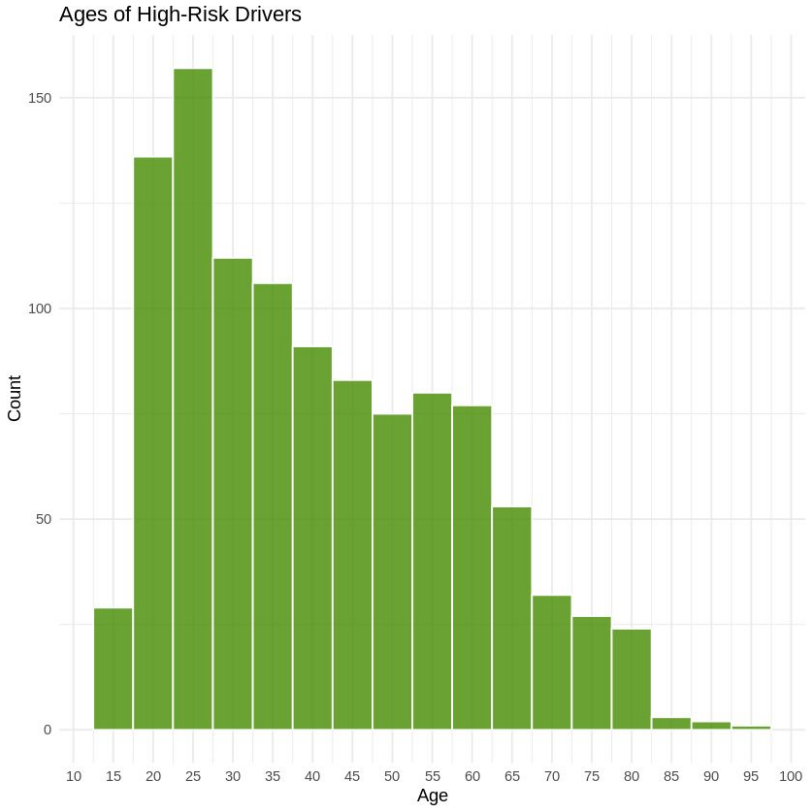
Race


Population Demographics of High-Risk Driver-Producing ZIP Codes
Count
800
600
400
200
0
772
339
94
23
22
22
14
3
1
Not Fatal
White
Hispanic
API
AIAN
Black
Multiracial
Oth/Unk
Other
Race
AIAN
API
Black
Hispanic
Multiracial
Not Fatal
Oth/Unk
Other
White

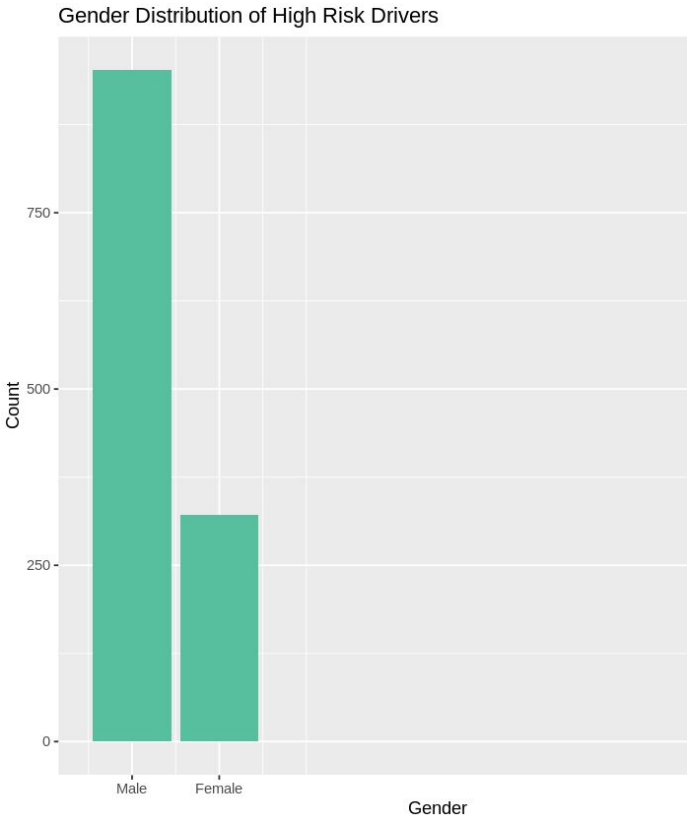

Population Demographics of High-Risk Driver-Producing ZIP Codes
Count
600
400
200
0
617
320
61
36
26
13
12
5
1
1
Not Fatal
White
Hispanic
Black
API
Multiracial
AIAN
Oth/Unk
Other
Unknown
Race
AIAN
API
Black
Hispanic
Multiracial
Not Fatal
Oth/Unk
Other
Unknown
White

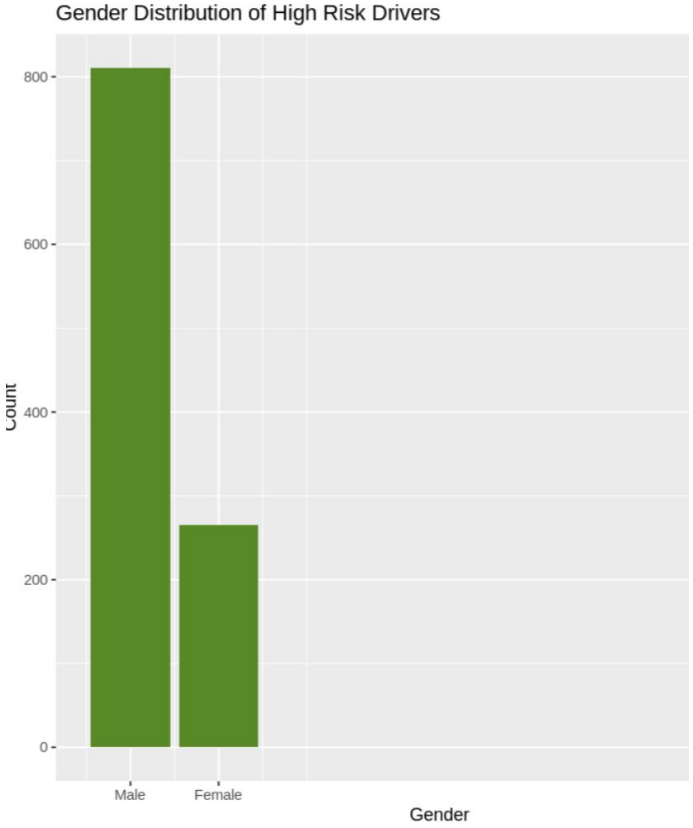
Age


Ages of High-Risk Drivers
Count
150
100
50
0
10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Age

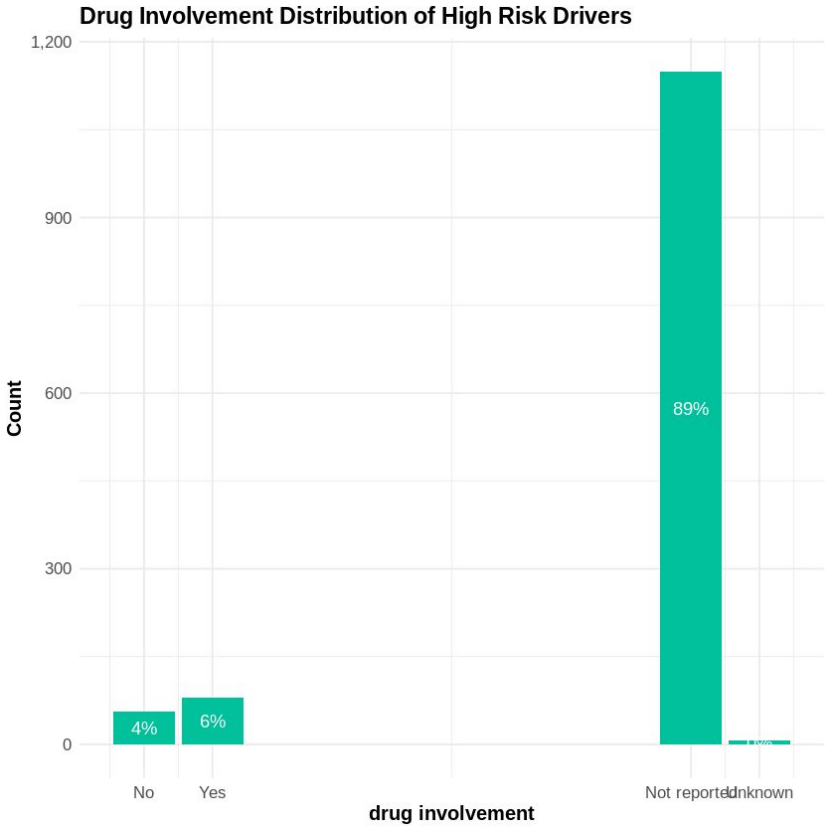

Ages of High-Risk Drivers
Count
150
100
50
0
10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Age

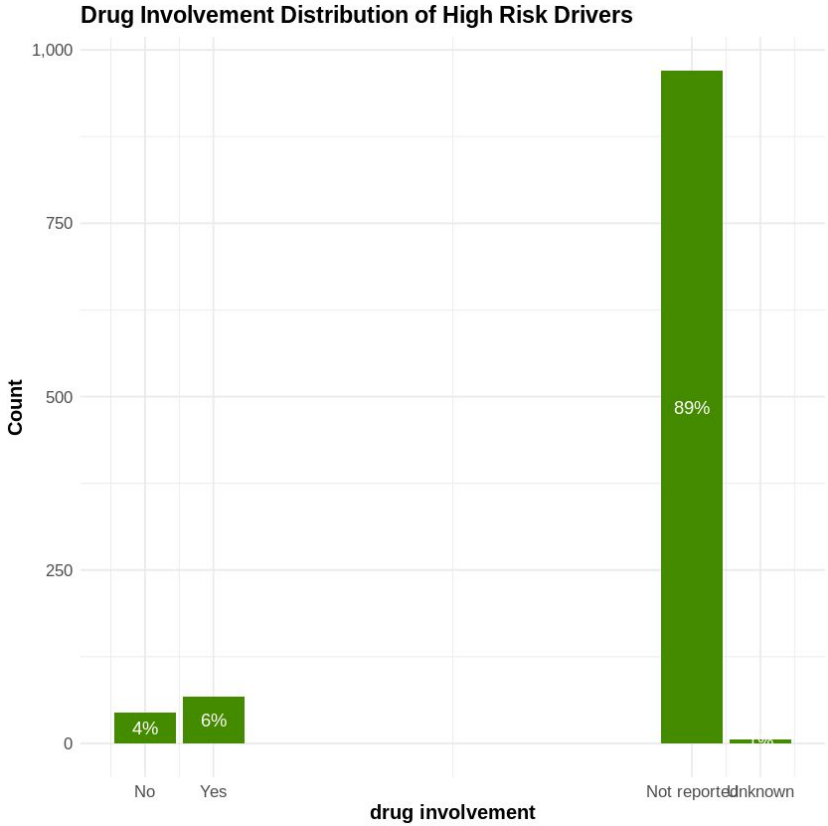
Gender


Gender Distribution of High Risk Drivers
750
500
250
0
Count
Male
Female
Gender

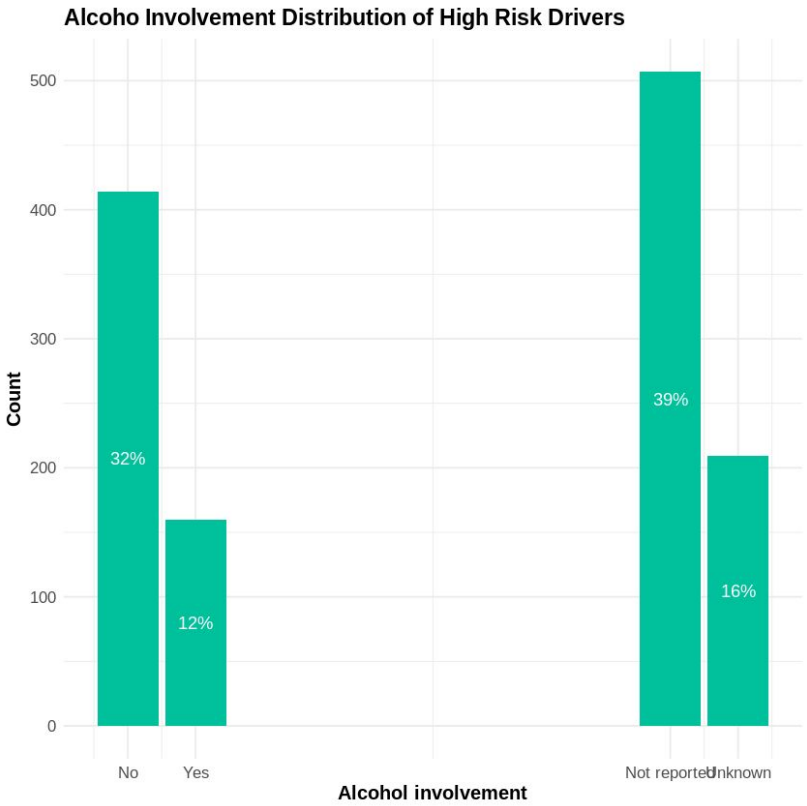

Gender Distribution of High Risk Drivers
800
600
400
200
0
Count
Male
Female
Gender

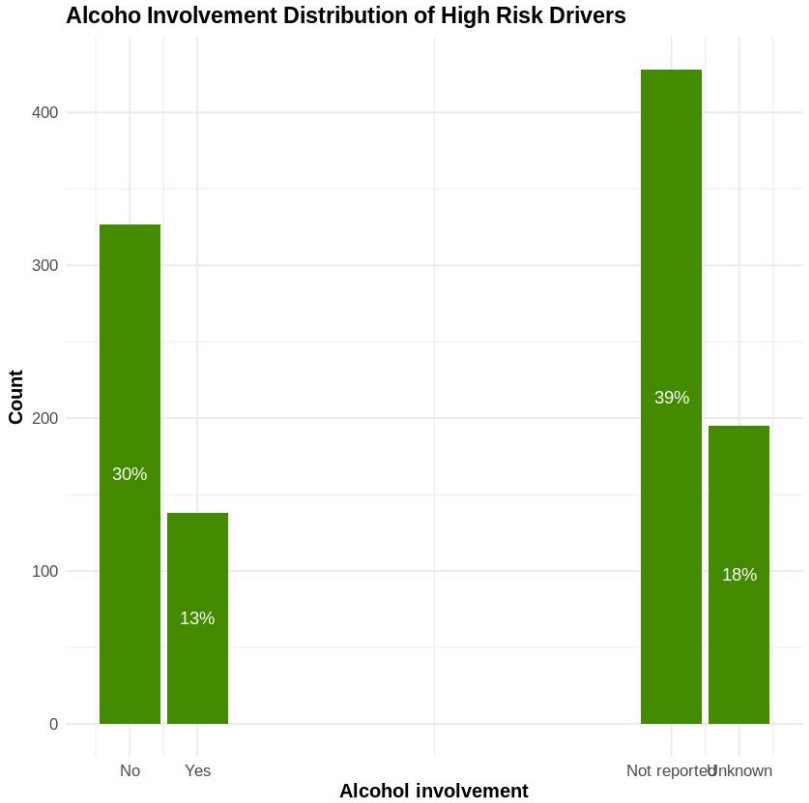
Drug


Drug Involvement Distribution of High Risk Drivers
1,200
900
600
300
0
Count
4%
6%
89%
No
Yes
Not reported
Unknown
drug involvement


Drug Involvement Distribution of High Risk Drivers
1,000
750
500
250
0
Count
4%
6%
89%
No
Yes
Not reported
Unknown
drug involvement

Alcohol


Alcoho Involvement Distribution of High Risk Drivers
Count
500
400
300
200
100
0
32%
12%
39%
16%
No
Yes
Not reported
Unknown
Alcohol involvement


Alcoho Involvement Distribution of High Risk Drivers
Count
400
300
200
100
0
30%
13%
39%
18%
No
Yes
Not reported
Unknown
Alcohol involvement

Conclusion

- Similar trends across community definitions
- Crashes close to home tend to involve more:
  - Alcohol use
  - Driver Distraction
  - Speeding
  - Lack of seatbelts
- Drivers involved in more fatal crashes tend to be:
  - Male
  - 20s
  - Involving drugs

Further Directions

- Geographic distance clustering algorithm
- Domain knowledge for determining K
- Alternate groupings: counties
- Non-fatal crash data to redefine risk