



Info challenge

# Washington Fatal Crash File

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



---

# Table of contents

**01**

**Introduction**

**02**

**Process**

**03**

**Analysis**

**04**

**Conclusion**

---

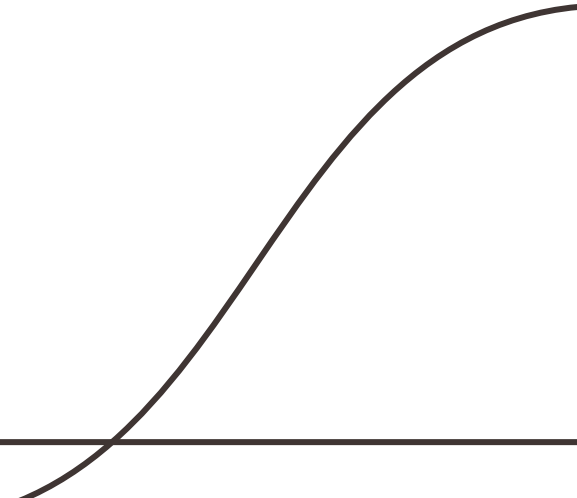
---

# 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).

---

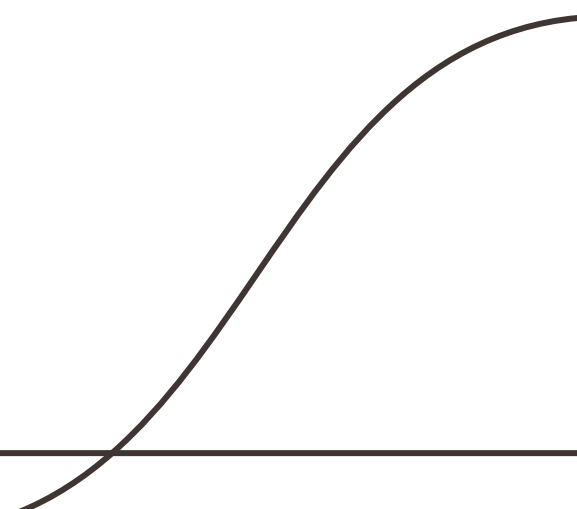
# Approach

- Analyse General Trends for Fatal Crashes
  - Connect Data to Grouped Location
    - Reverse geocoding
    - Latitude and longitude distance clustering
  - Analyse trends by groups
    - Crash Location
    - Driver Location
- 

---

# Phase 1

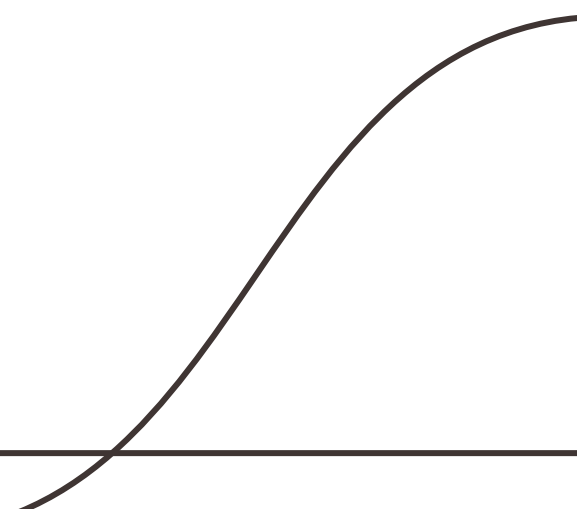
In Stage 1, the project will gather Washington Fatal Crash Files, parse and load them into DataFile, check the consistency of the files with the data dictionary, add missing ZIP codes to the dataset, and remove duplicates and invalid data. The cleaned data will be saved as an output.



---

# 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.

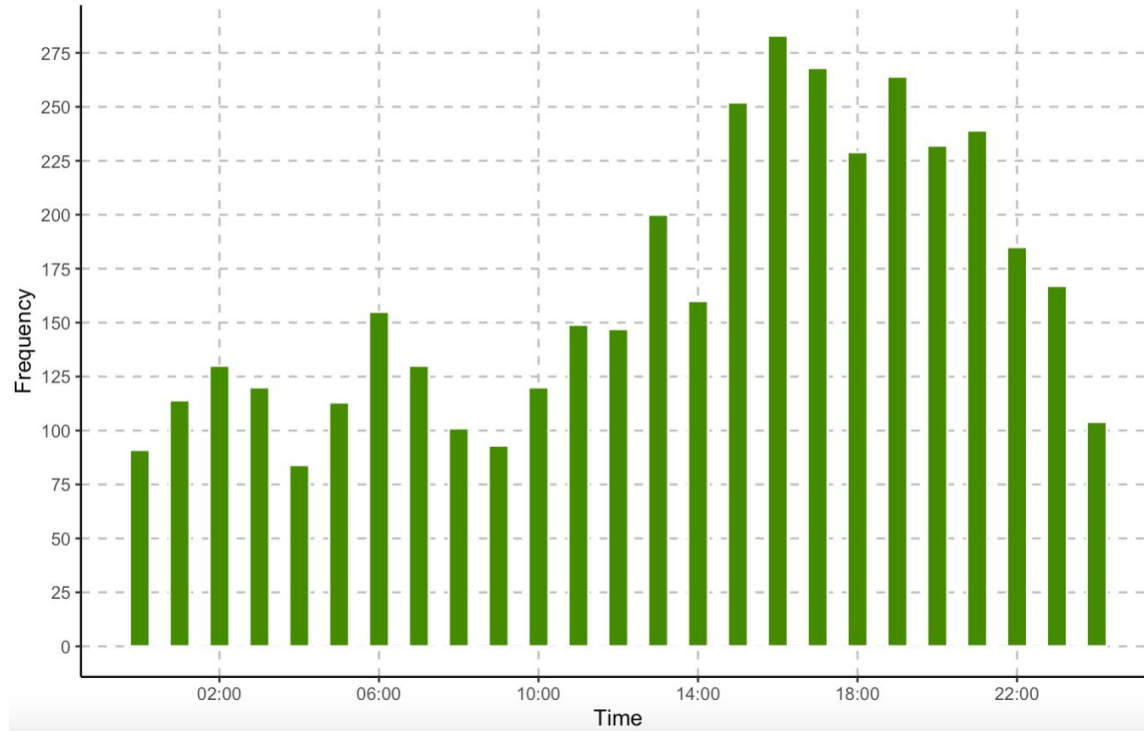




# General Factors

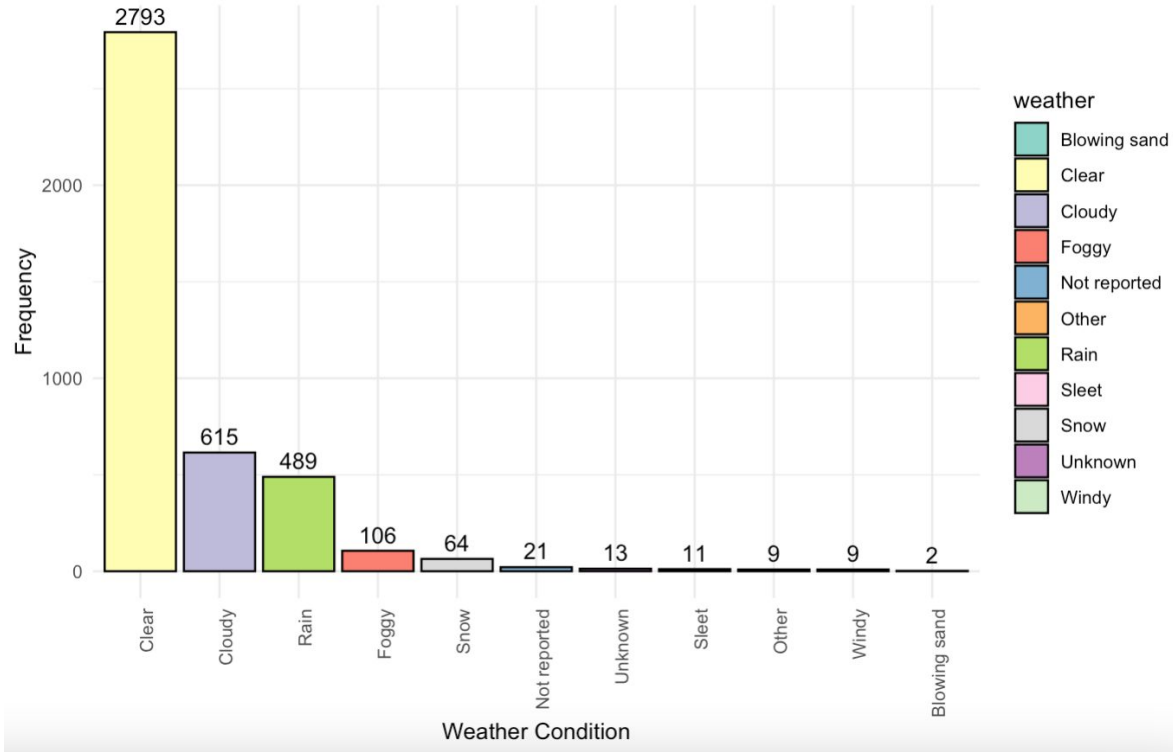


Histogram Crash Times

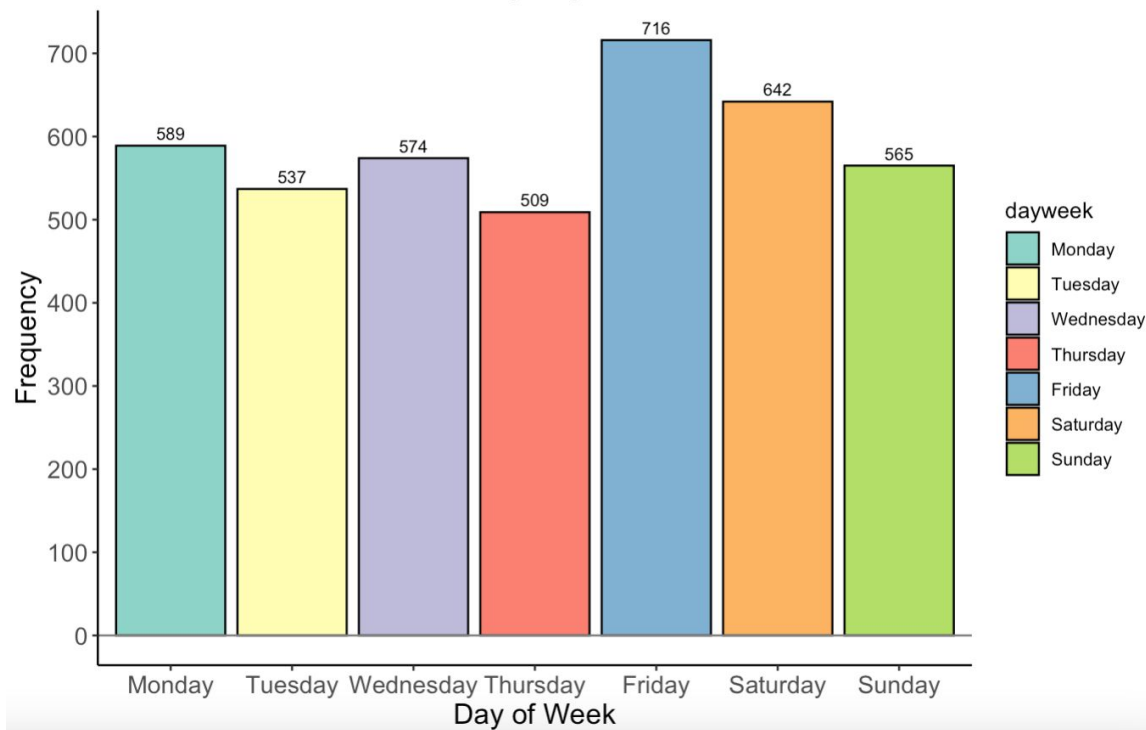


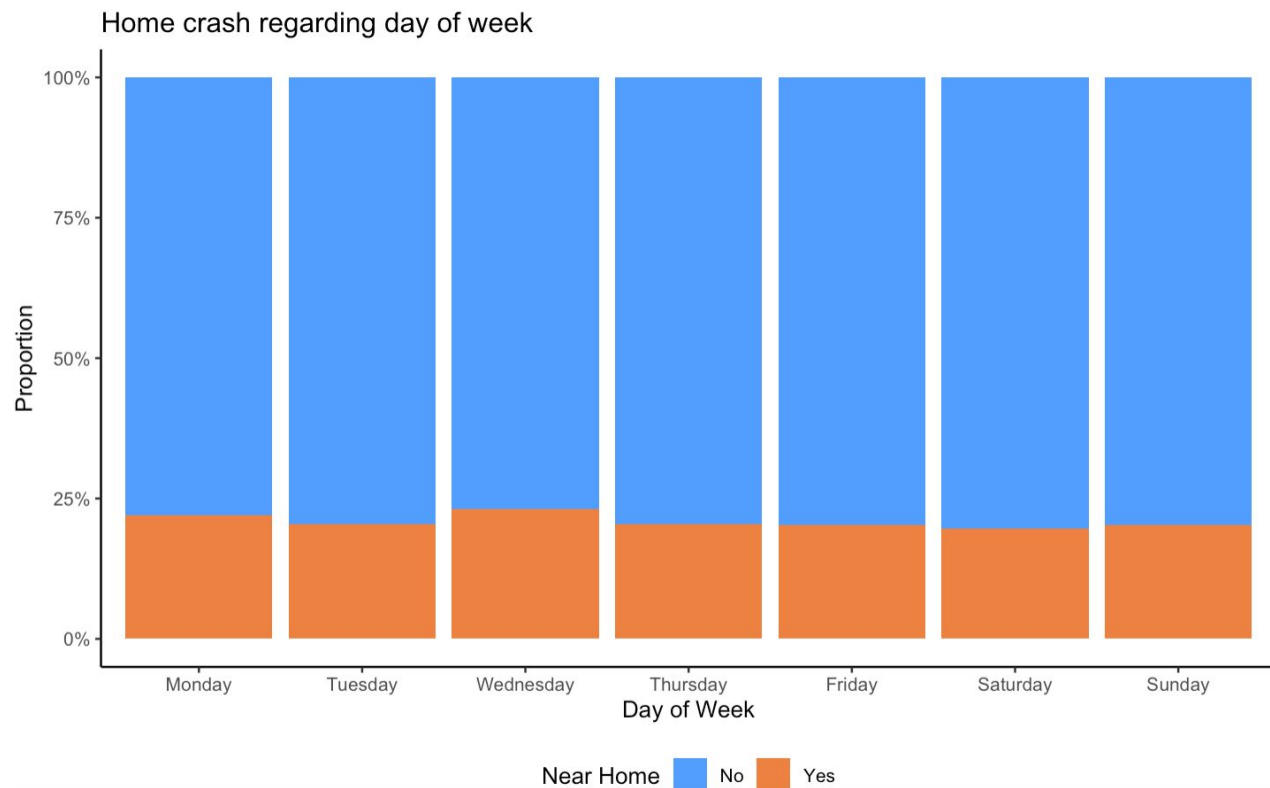


Crashes by Weather Condition



Crashes by day of week



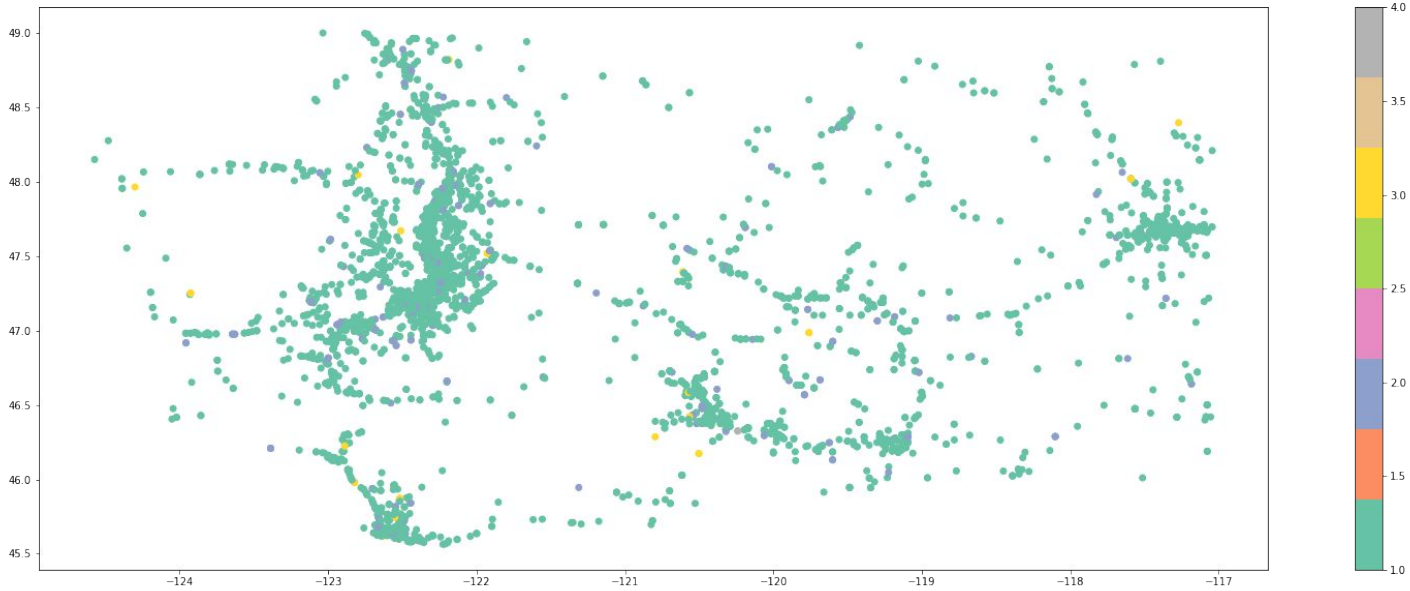


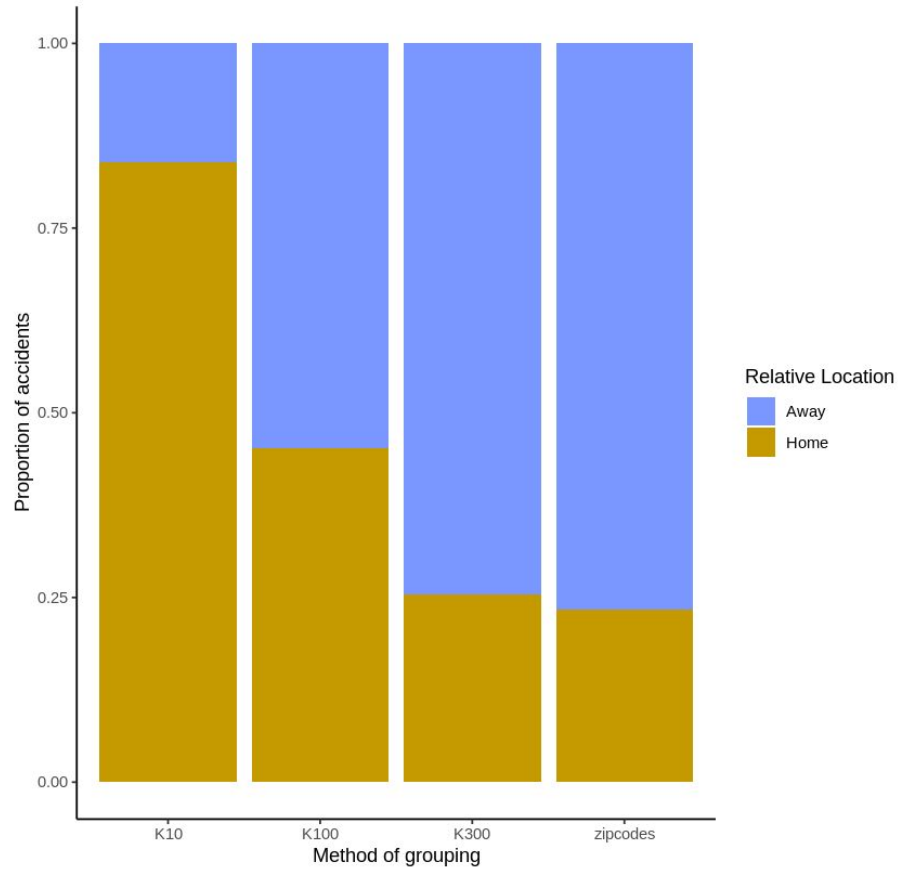


# Defining Community



# Fatal Crashes at a Glance





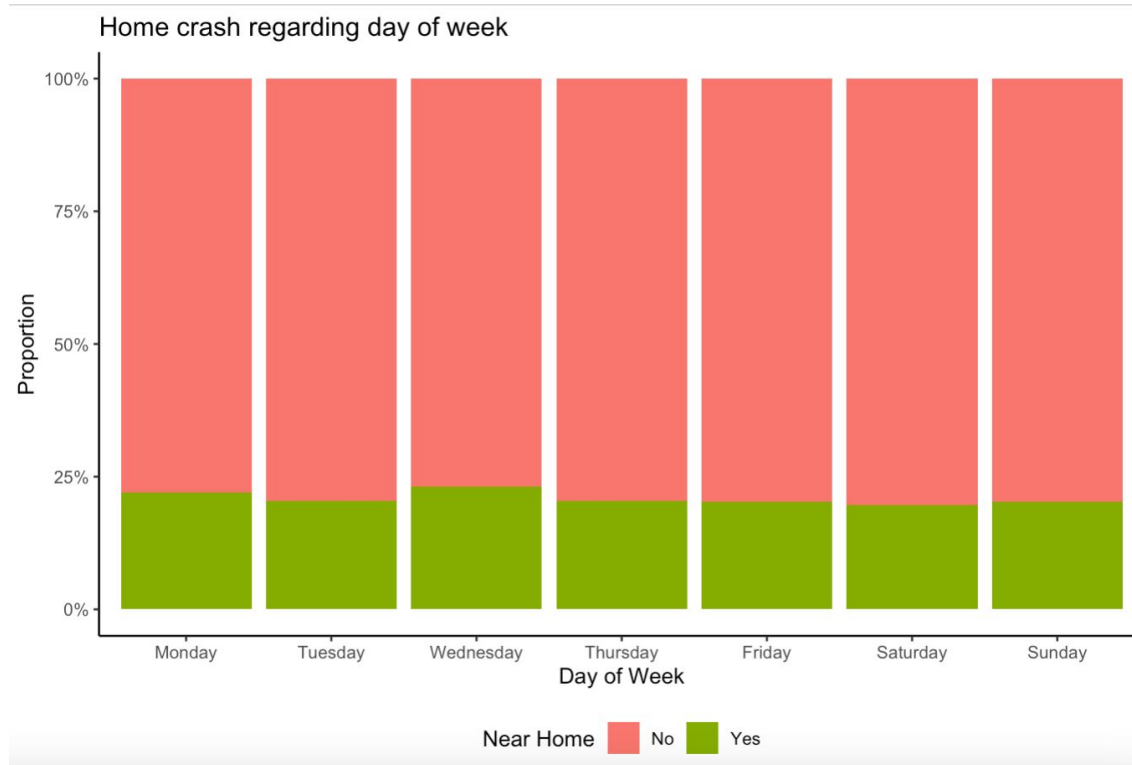
| K   | Silhouette Score |
|-----|------------------|
| 10  | 0.507            |
| 100 | 0.421            |
| 300 | 0.453            |



# Crashes at Home

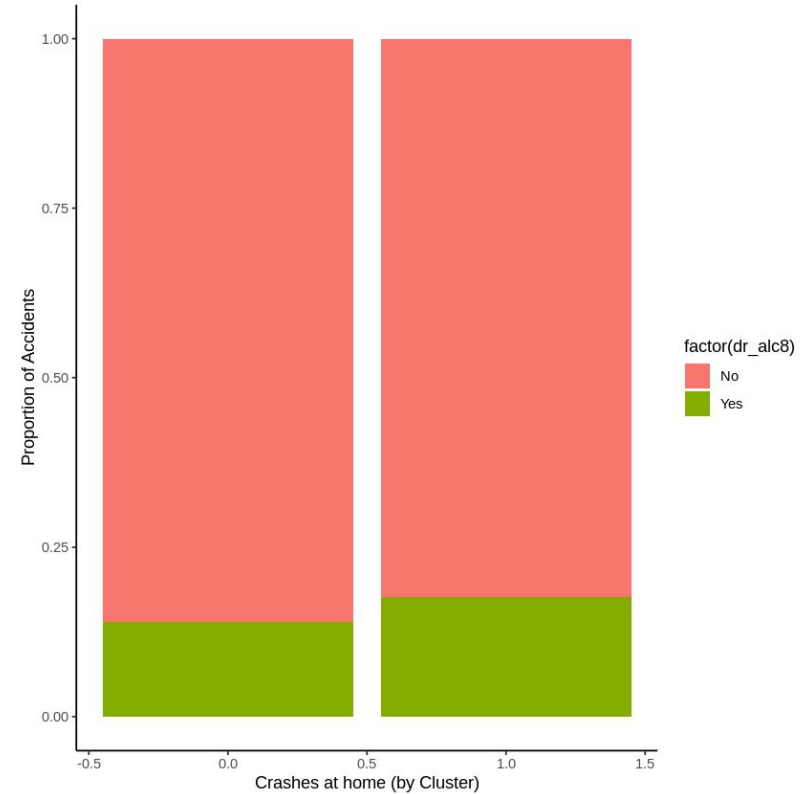
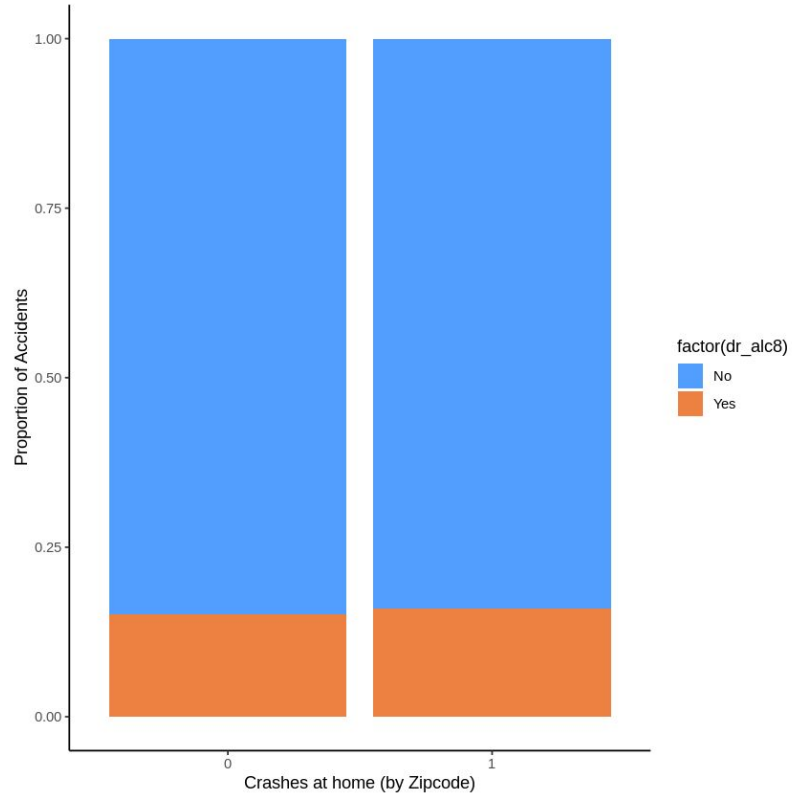


# Clustering VS zipcodes

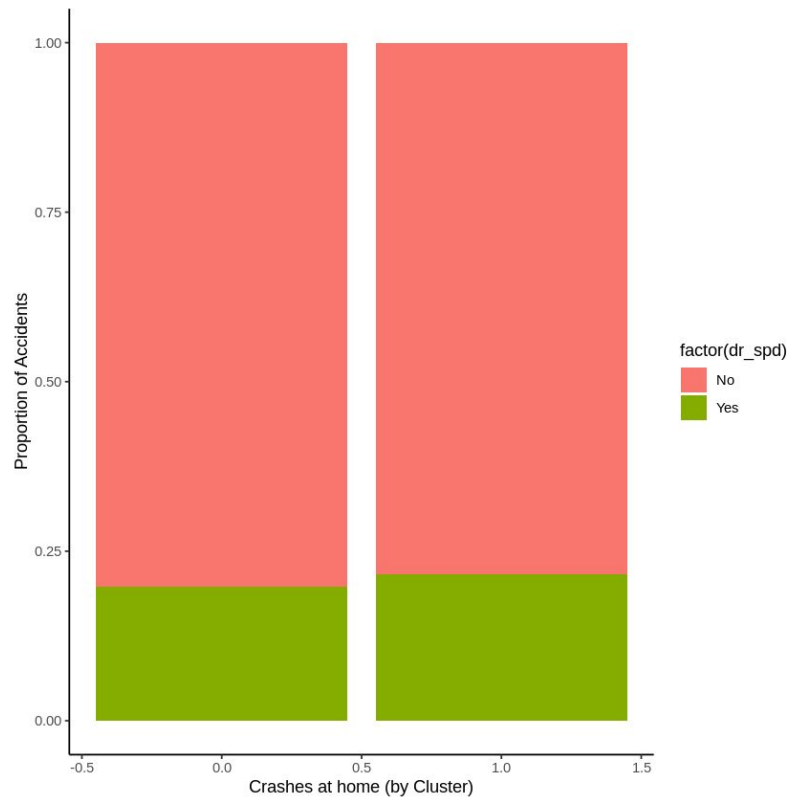
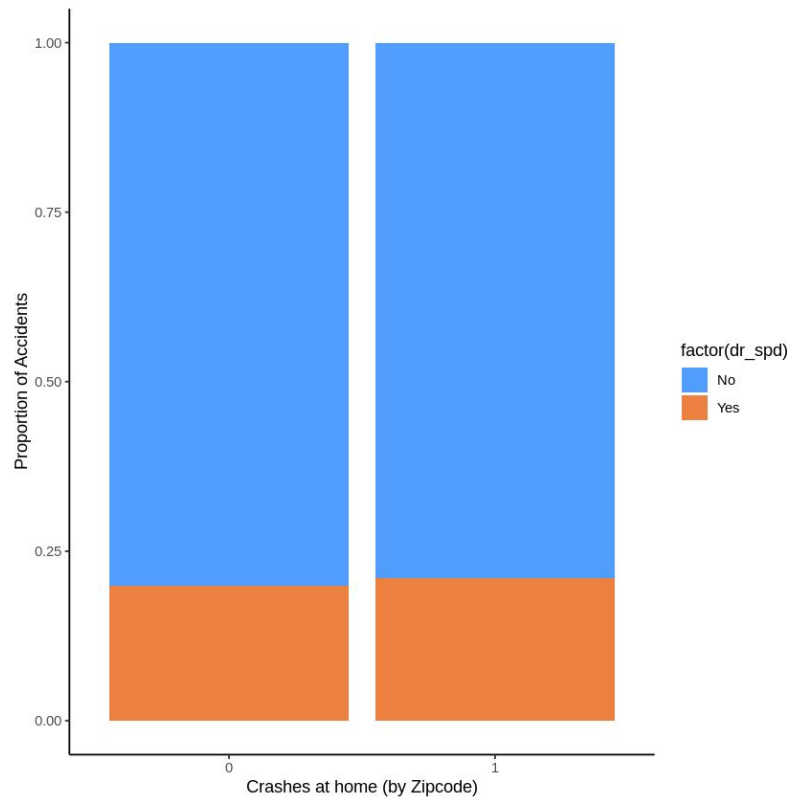




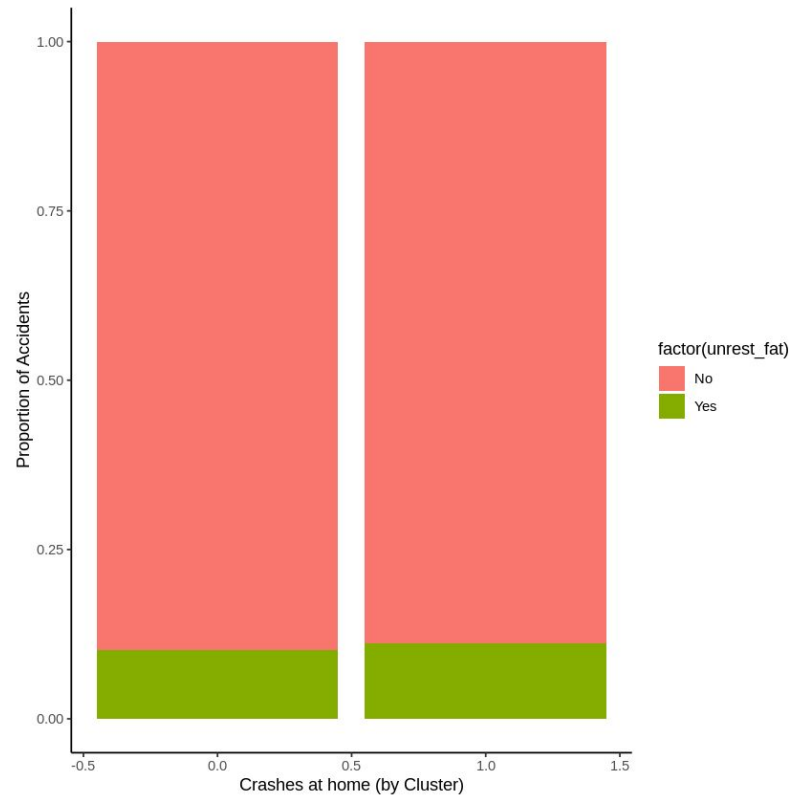
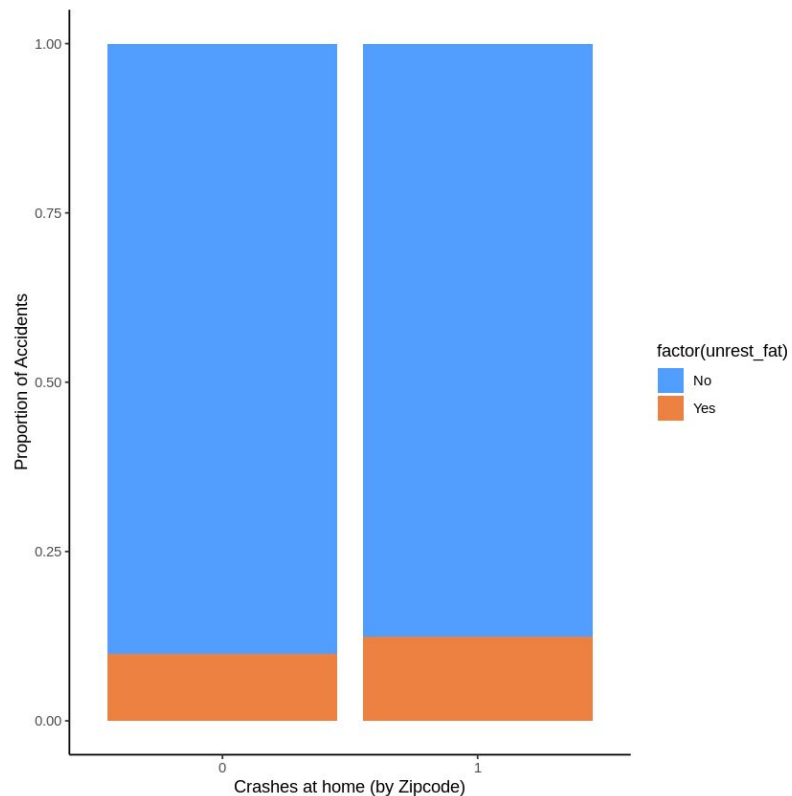
# Alcohol Involvement



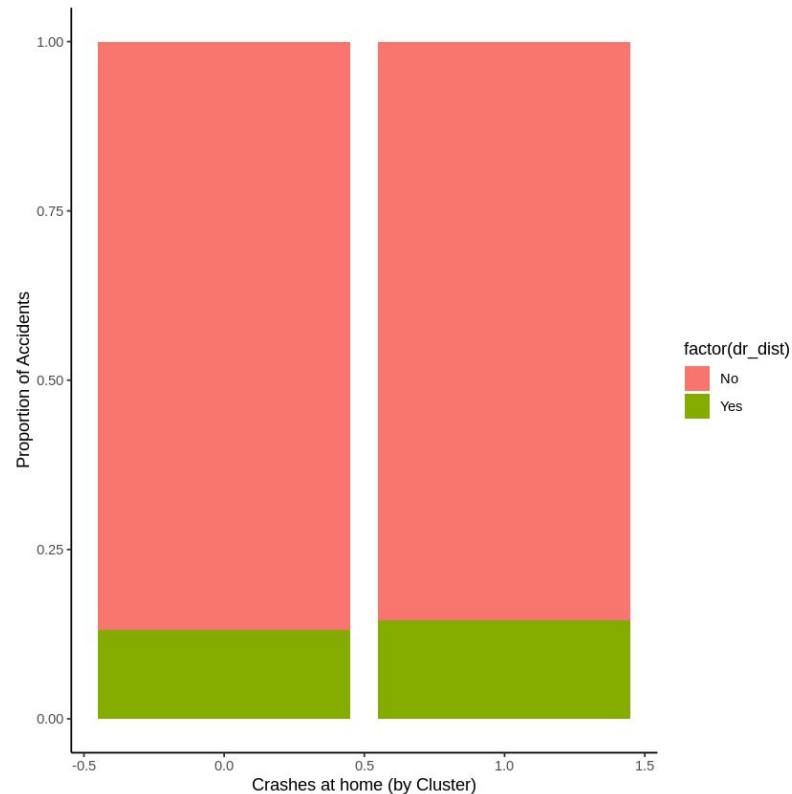
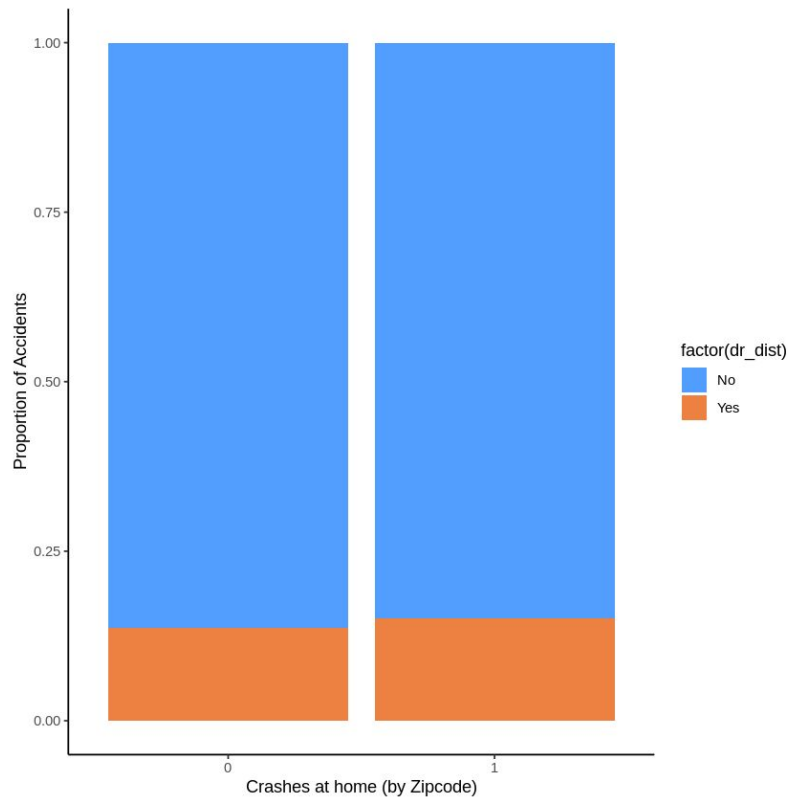
# Speeding



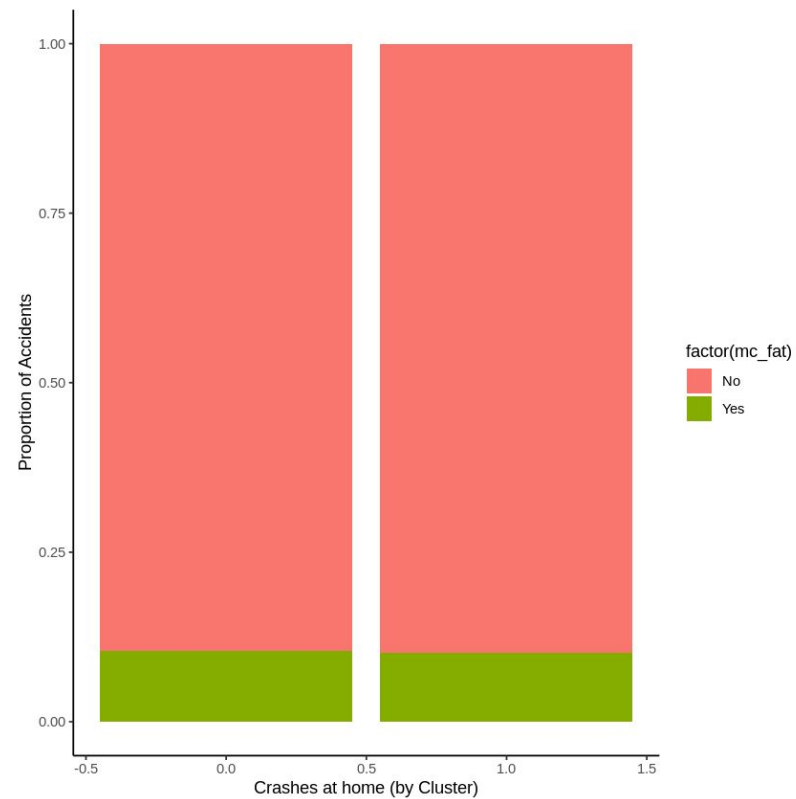
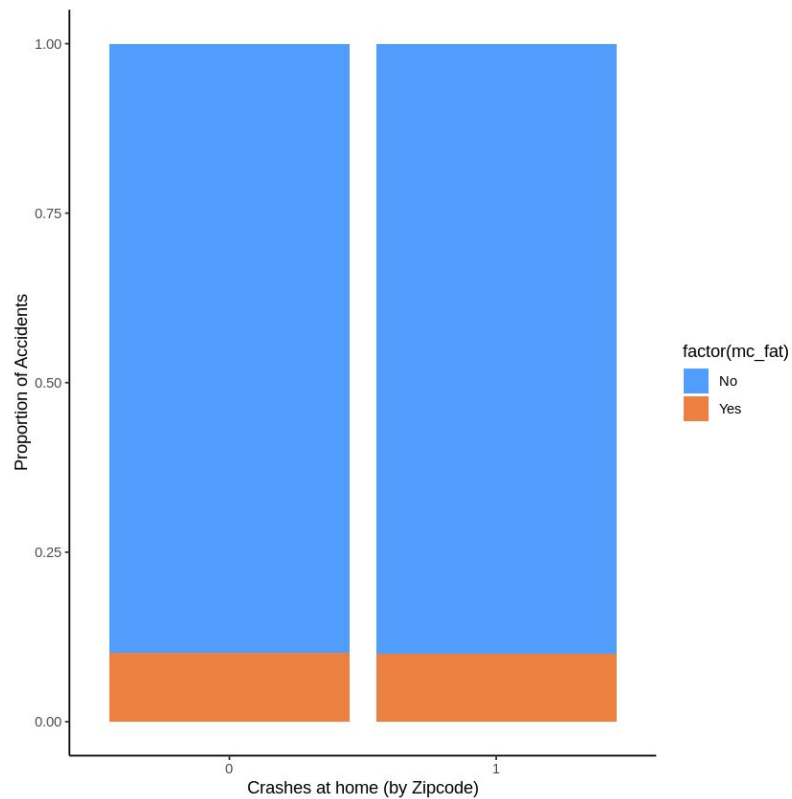
# Lack of seatbelt



# Distracted Driver



# Motorcycle Fatalities





# High Risk Communities

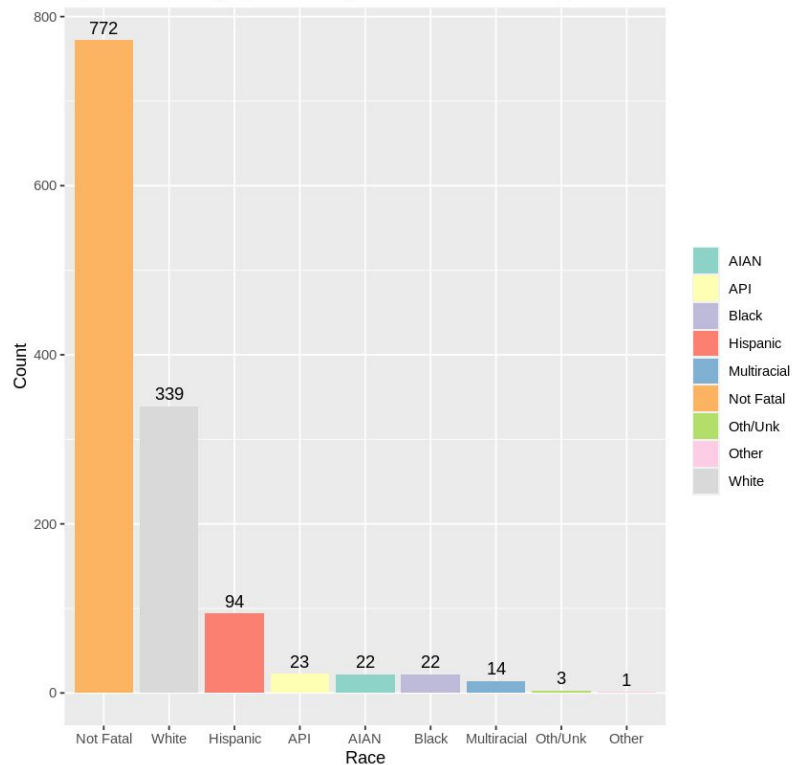


## Method of define high risk drivers/communities:

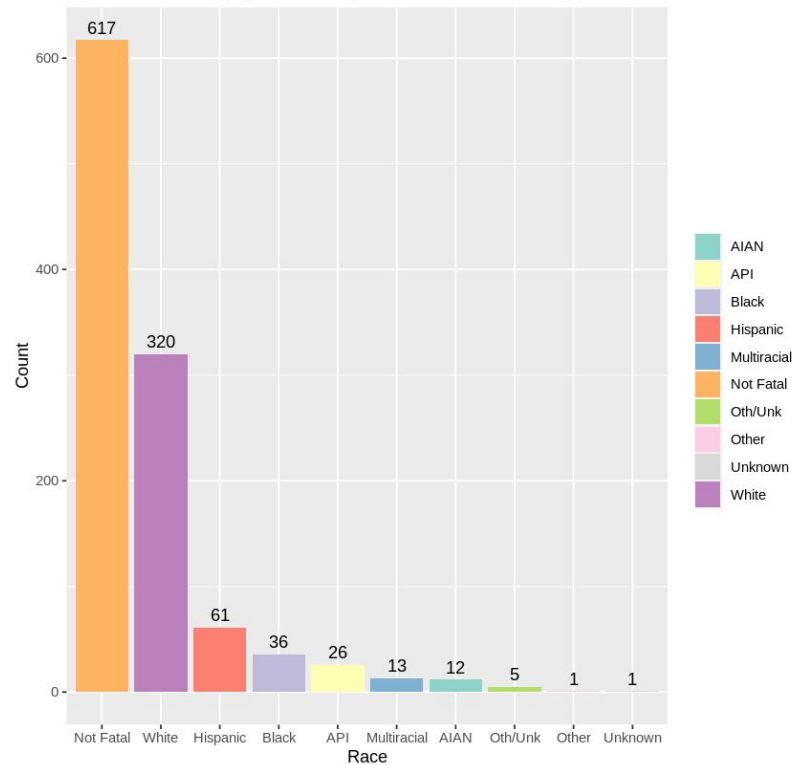
- **Frequency : Number crashes in dzip(or cluster)/ total number of crashes**
- **Fatality rate: Number of Fatalities in Crash/Number of Persons in Crash**

# Race

Population Demographics of High-Risk Driver-Producing ZIP Codes



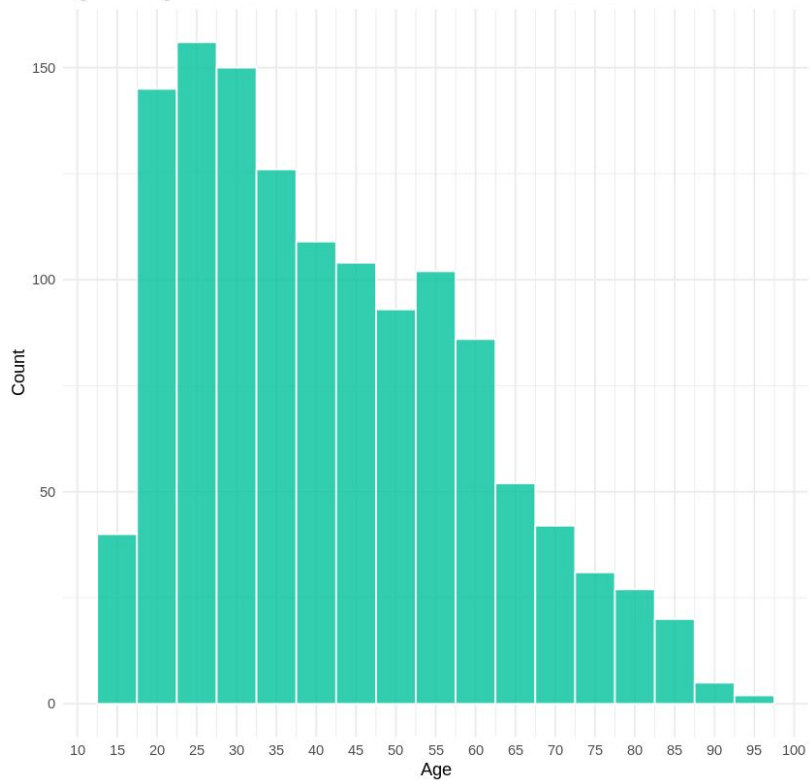
Population Demographics of High-Risk Driver-Producing ZIP Codes



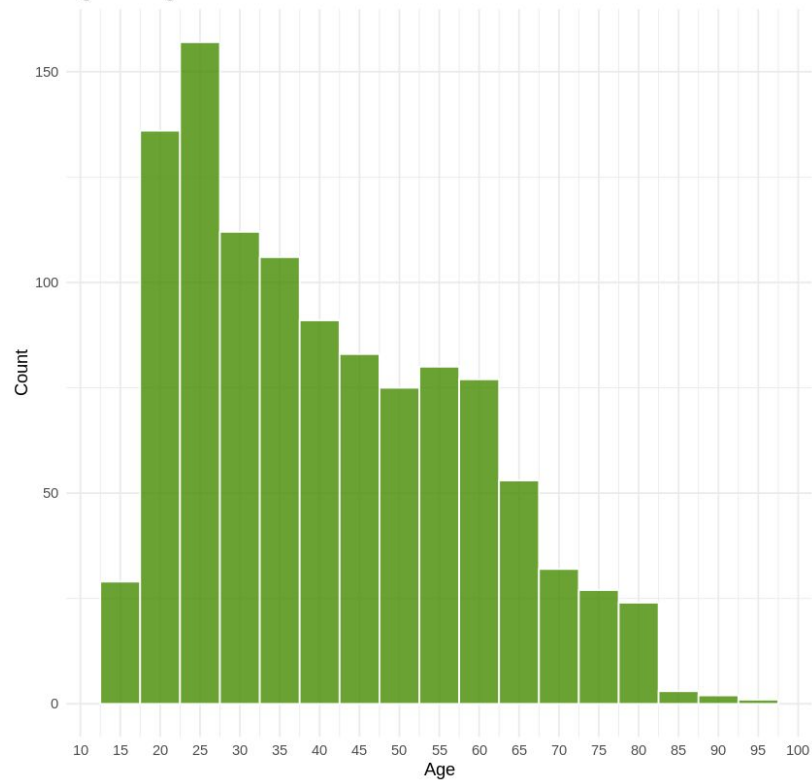


# Age

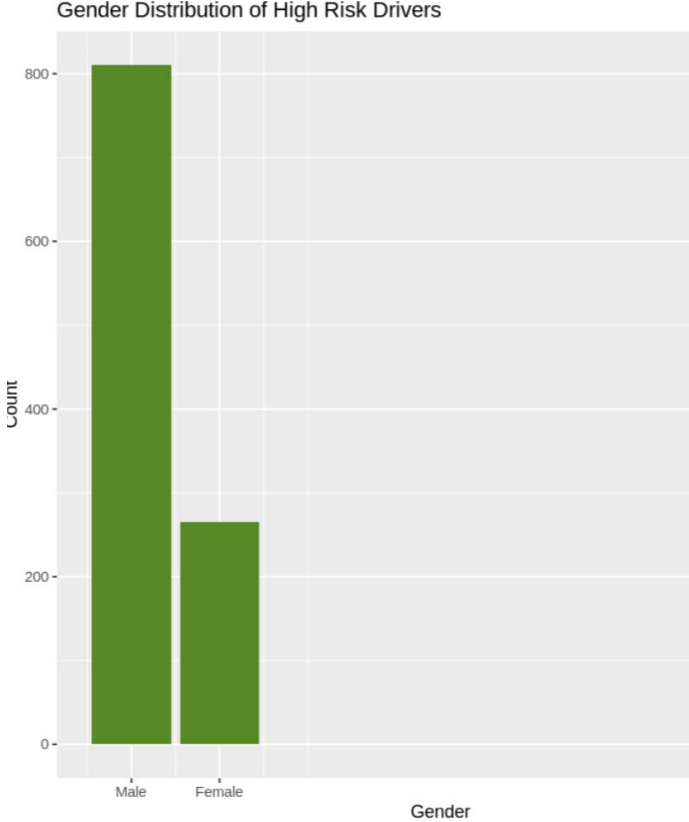
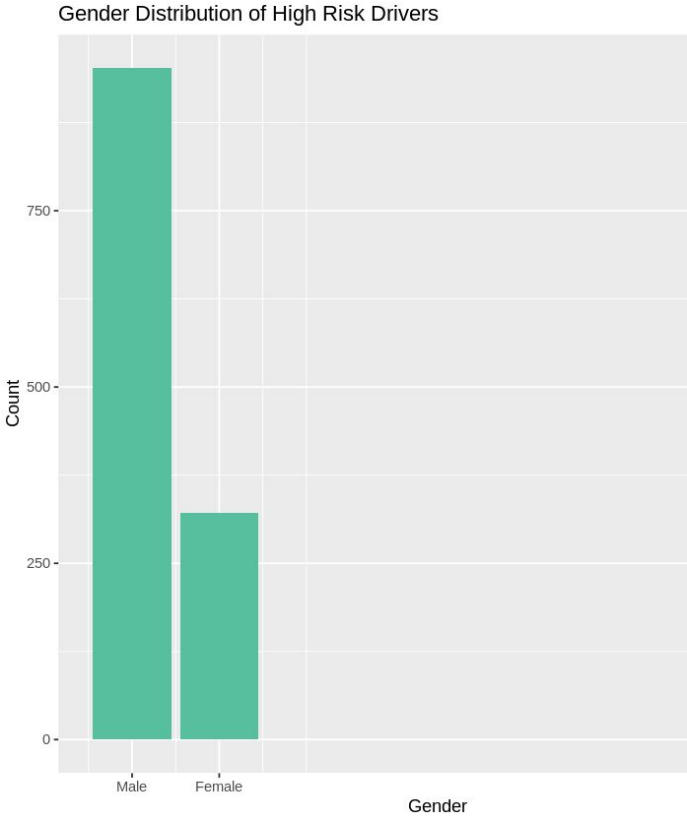
Ages of High-Risk Drivers



Ages of High-Risk Drivers

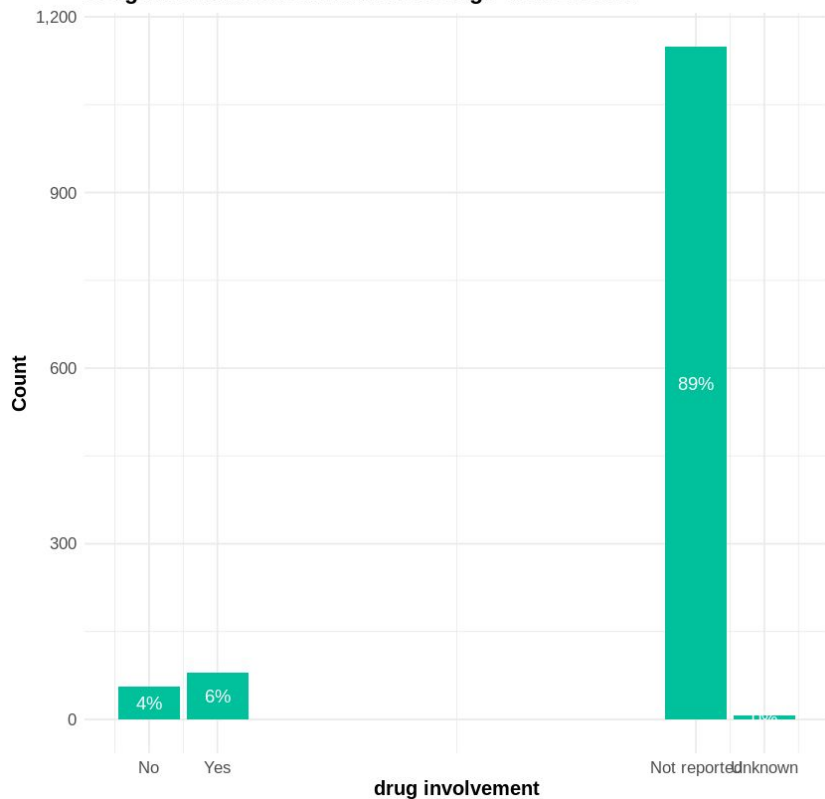


# Gender

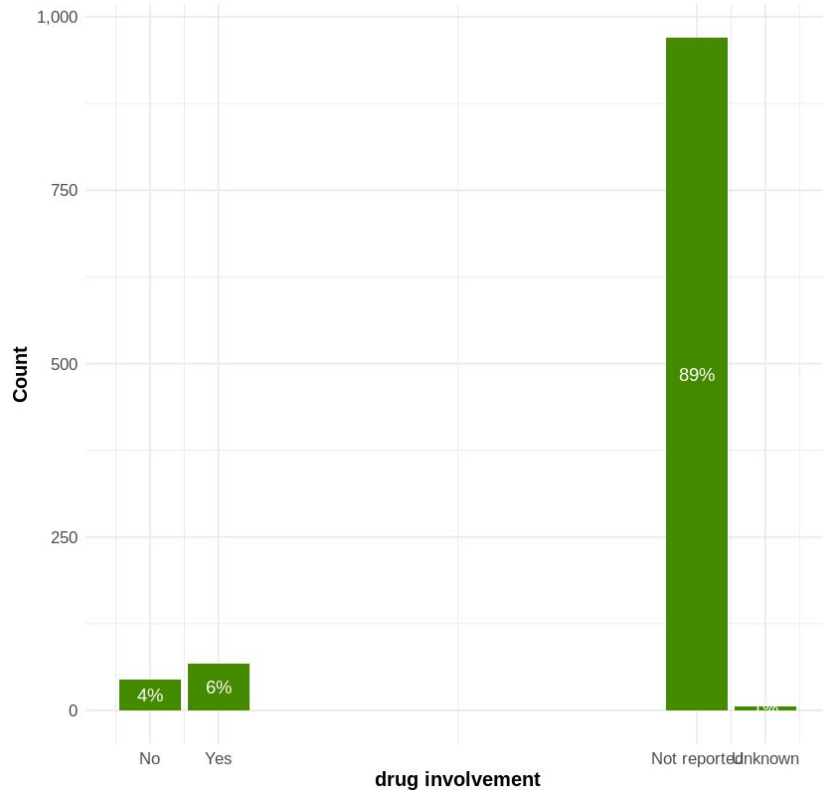


# Drug Involvement

Drug Involvement Distribution of High Risk Drivers

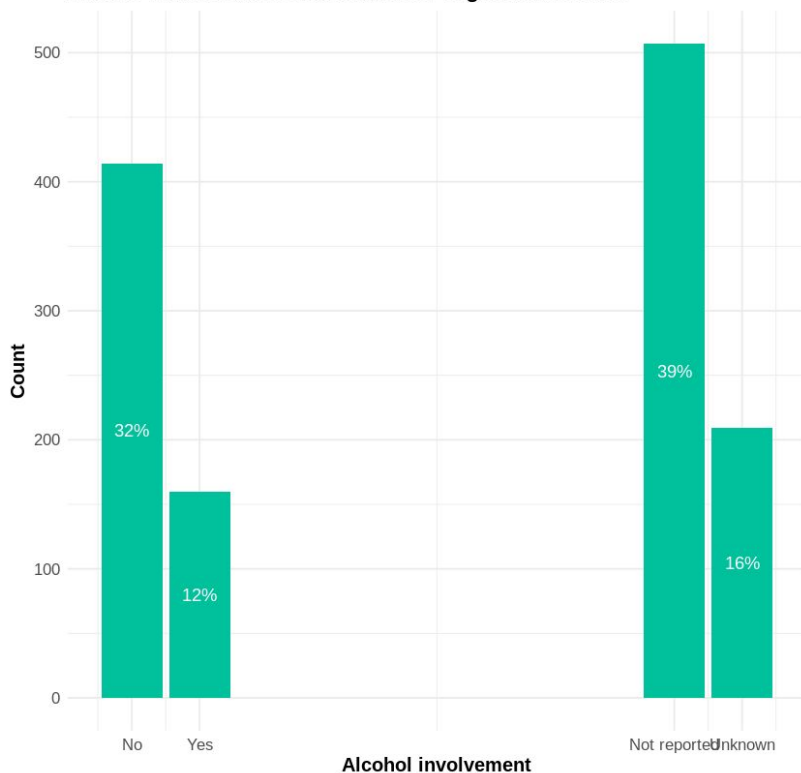


Drug Involvement Distribution of High Risk Drivers

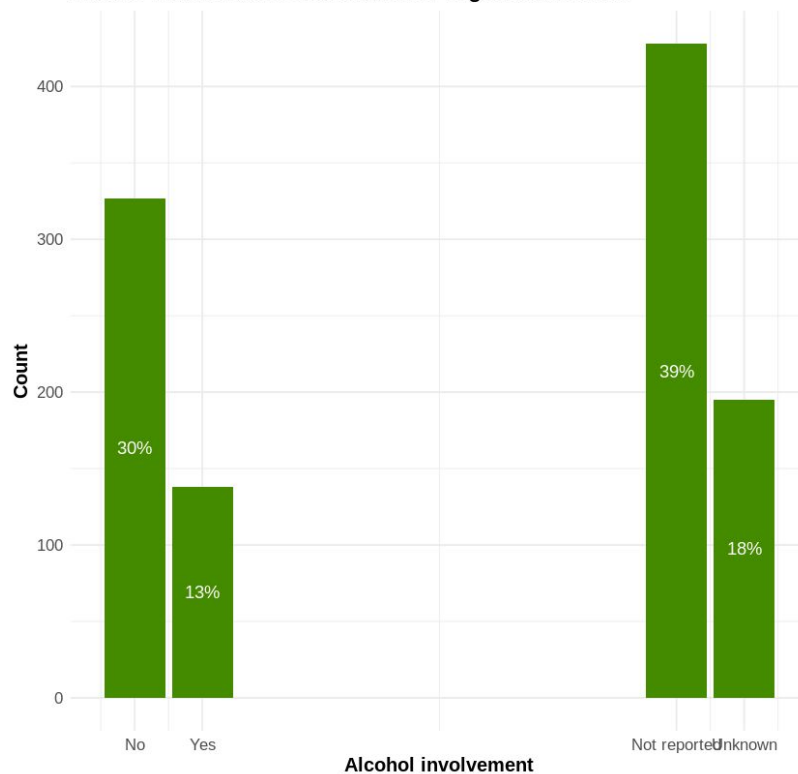


# Alcohol Involvement

Alcohol Involvement Distribution of High Risk Drivers



Alcohol Involvement Distribution of High Risk Drivers



---

# 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