Exploring the Dynamics of Traffic Crashes and Outcomes

INFO 523

Group 7

by Kapil Parab & Nicaise Irambona

Introduction

Introduction

- Our dataset has been derived from **Fatality Analysis Reporting System (FARS)** system and the **Crash Report Sampling System (CRSS)**, maintained by the **National Highway Traffic Safety Administration (NHTSA)** of the United States.

- These systems collect and analyze motor vehicle traffic crash data to enhance road safety, reduce injuries, and prevent fatalities on the trafficways.

- The raw dataset can be found on nhtsa.gov.

- Rows = 233,069

Introduction

- FARS was established in 1975 by NHTSA's National Center for Statistics and Analysis (NCSA) to:
    - Measure highway safety.
    - Identify traffic safety problems.
    - Propose solutions and evaluate motor vehicle safety standards and highway safety programs.
- CRSS is a nationally representative sample of police-reported motor vehicle crashes, including:
    - Property damage-only crashes.
    - Crashes resulting in injuries or fatalities.

Introduction

- NHTSA has adopted the term "crash" instead of "accident" to align with the American National Standard Institute (ANSI) recommendations. The term "crash" comprehensively includes:
    - Collision Events:
        - Crashes involving fixed objects like poles, walls, or barriers.
        - Crashes involving non-fixed objects like pedestrians, animals, or other vehicles.
    - Non-Collision Events:
        - Fires in moving vehicles.
        - Vehicles running off trafficways into water.
        - Injuries caused by shifting cargo or objects within vehicles.
        - Damage due to pavement irregularities like potholes or loose plates.

Dataset

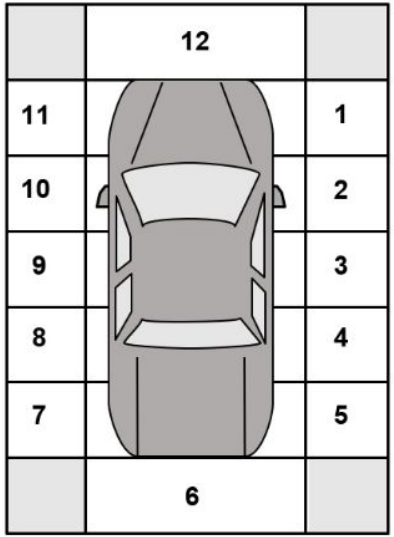
- For our analysis we are using 3 years of data i.e 2020 to 2023.
- The raw dataset is divided into **accident** data and **person-level** data.

STATE	STATENAME	ST_CASE	PEDS	PERNOTMVIT	VE_TOTAL	VE_FORMS	PVH_INVL	PERSONS	PERMVIT	COUNTY	COUNTYNAME	CITY	CITYNA
1	Alabama	10001	0	0	2	2	0	3	3	107	PICKENS (107)	0	NOT A
1	Alabama	10002	0	0	2	2	0	5	5	101	MONTGOMERY (101)	0	NOT A

STATE	STATENAME	ST_CASE	VEH_NO	PER_NO	VE_FORMS	COUNTY	MONTH	MONTHNAME	DAY	DAYNAME	HOUR	HOURNAME	MINUTE
1	Alabama	10001	1	1	2	107	1	January	1	1	12	12:00pm-12:59pm	30
1	Alabama	10001	2	1	2	107	1	January	1	1	12	12:00pm-12:59pm	30
1	Alabama	10001	2	2	2	107	1	January	1	1	12	12:00pm-12:59pm	30
1	Alabama	10002	1	1	2	101	1	January	1	1	16	4:00pm-4:59pm	40
1	Alabama	10002	2	1	2	101	1	January	1	1	16	4:00pm-4:59pm	40

Dataset

- Whenever a crash/accident occurs, the police have to follow certain guidelines to code the values.
- If the evidence is unavailable, the value is reported as “Not Reported” or “Unknown”.
- Example:
  - In a hit and run case, a vehicle’s “Make” or Manufacturer may not be reported due to lack of evidence.
  - Latitude and longitude values are also sometimes not reported.
  - Some counties report impact locations as clock points.
    Clock point 6 represents collision from behind.
    Clock point 1 represents front right.
- The dataset contains 6 different injury levels:
  - Death
  - Minor
  - None
  - Possible
  - Serious
  - Not Reported (Coded as Minor/Serious/Death)



Dataset

- To make visualization easier, we have derived new columns and encoded certain fields to better suit our needs.

- Example:

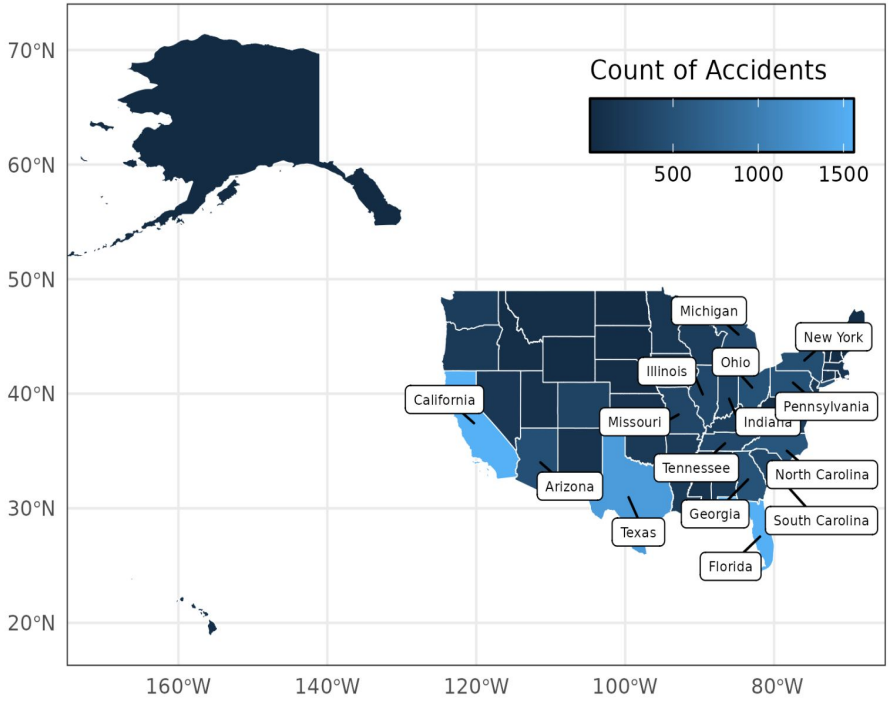
    - ACC_TOD (Accident Time of Day): This field is derived from the accident timestamp to indicate if the crash happened during the Morning, Afternoon, Evening or Night.

    - Invalid latitude and longitude values are replaced by the mean coordinates of each state.

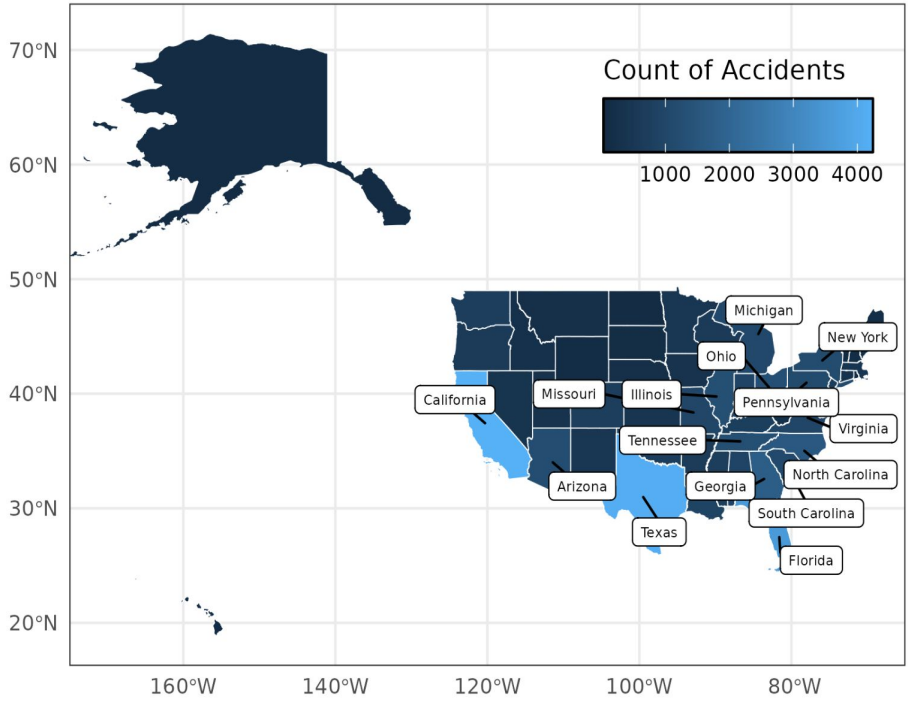
Visualizations

Accidents in the United States by State

Motorcycle Accidents across United States
2020-2022

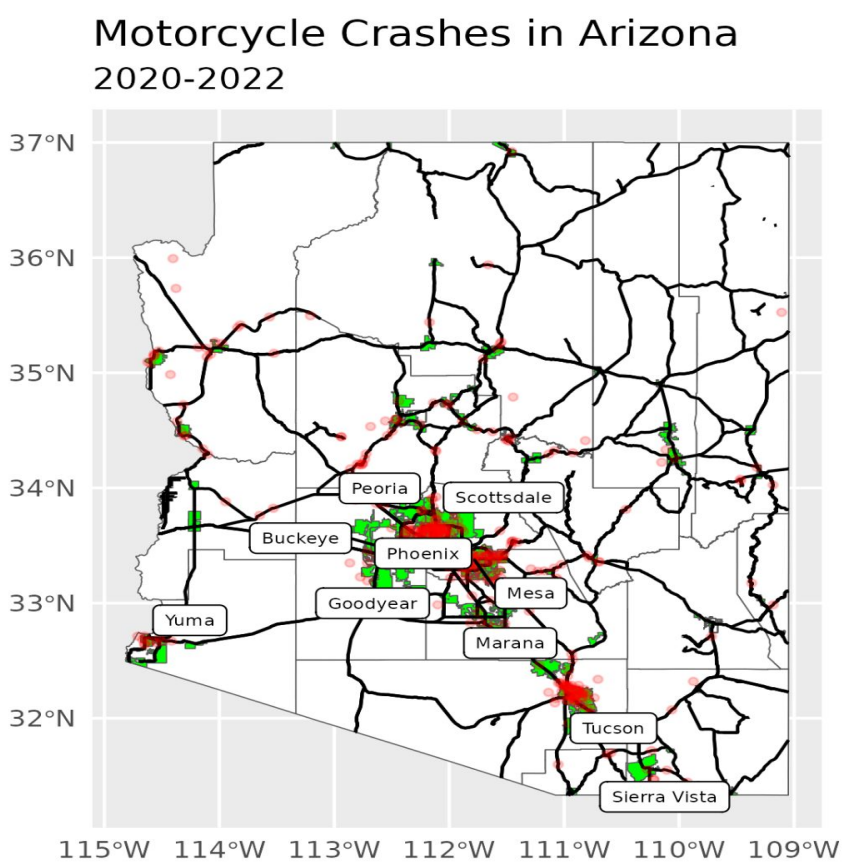


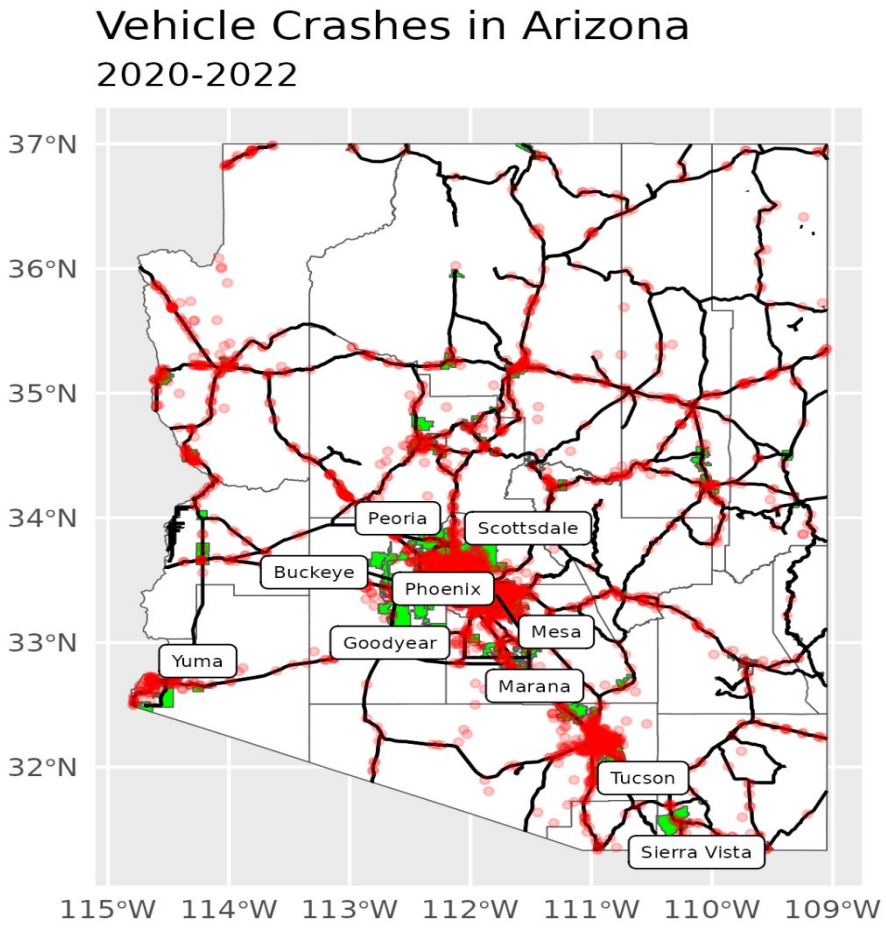
Vehicle Accidents across United States
2020-2022



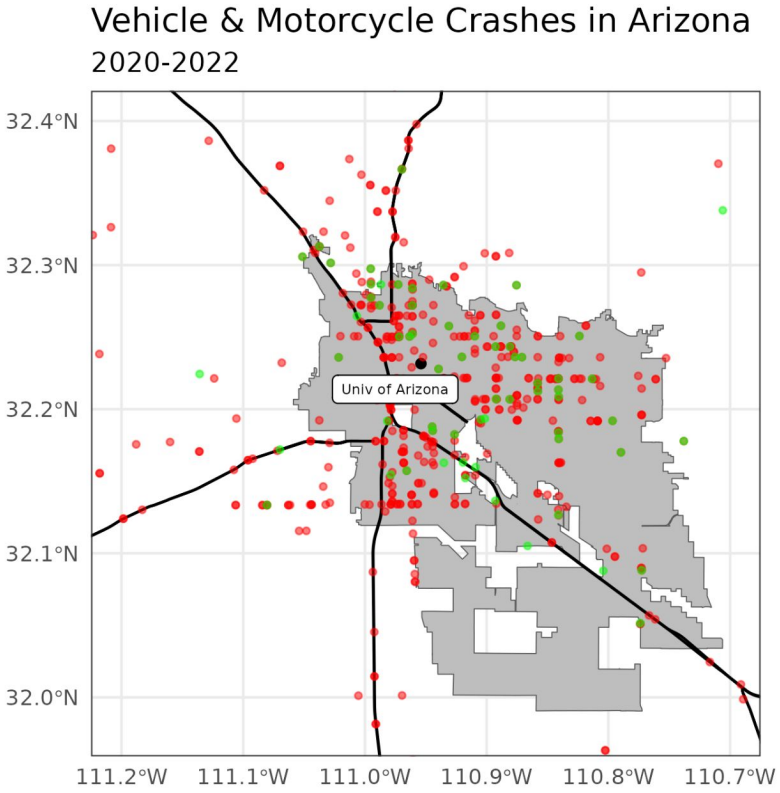
Cities and Major Roadways of Arizona

Accidents in Arizona

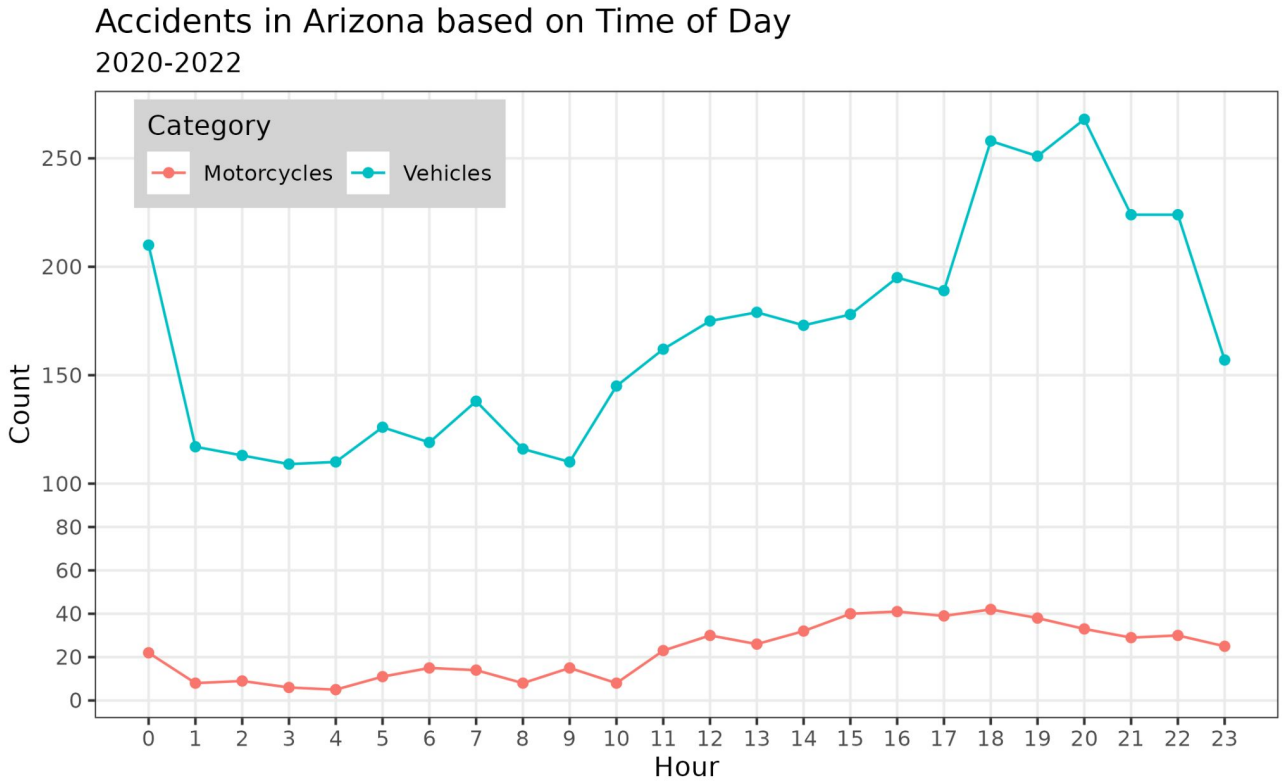




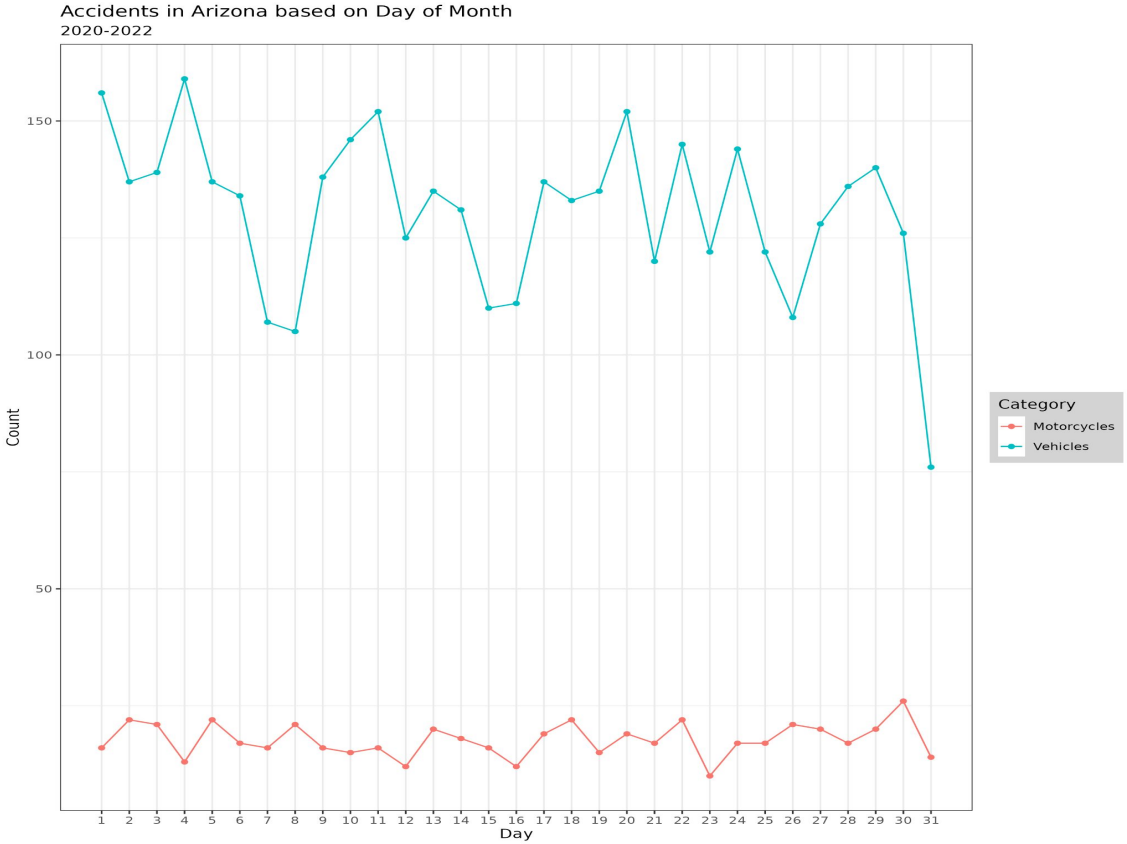
Accidents in Tucson



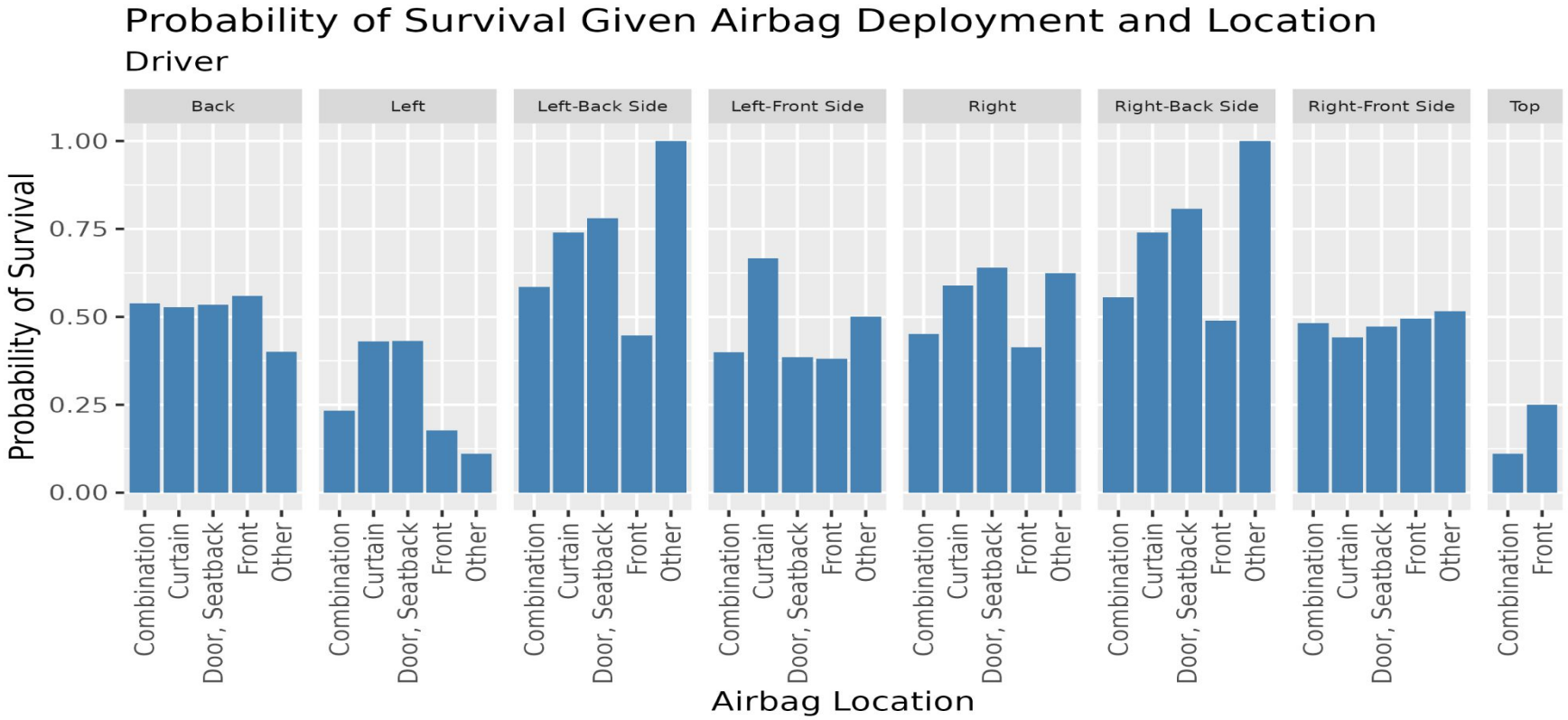
Accidents in AZ based on Time of Day



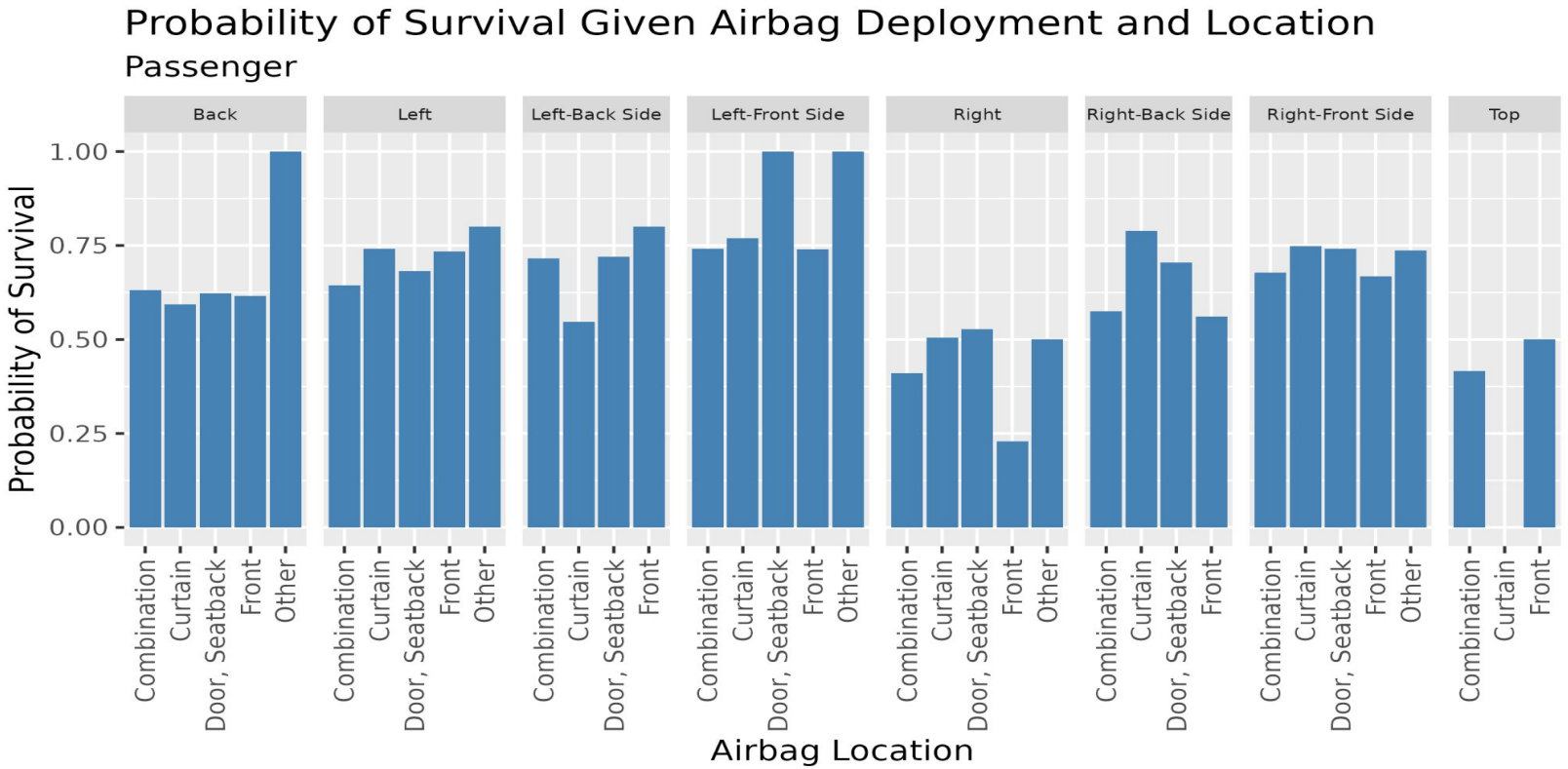
Accidents in AZ based on Day of Month



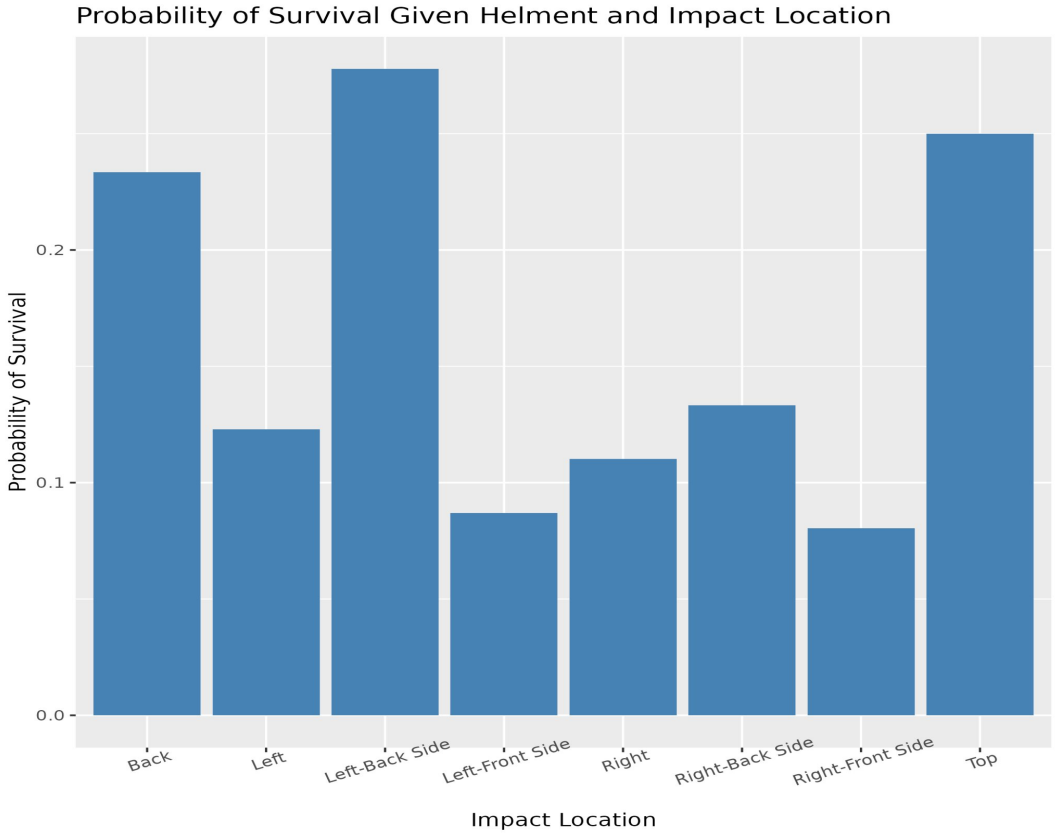
Probabilities - Vehicle



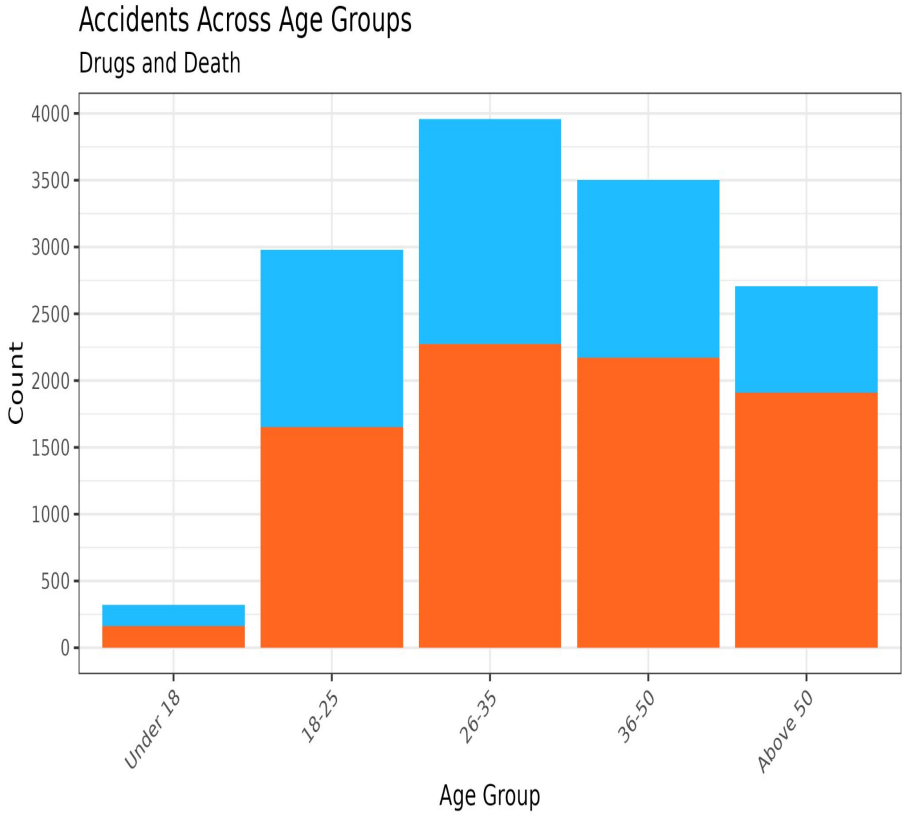
Probabilities - Vehicle

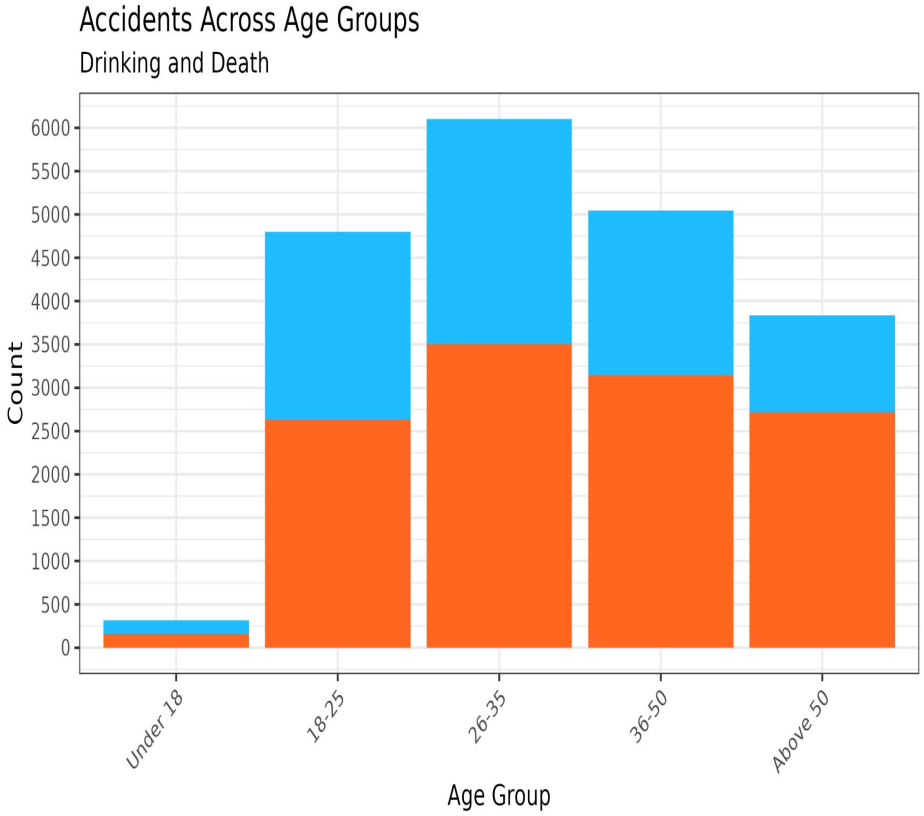


Probabilities - Motorcycle

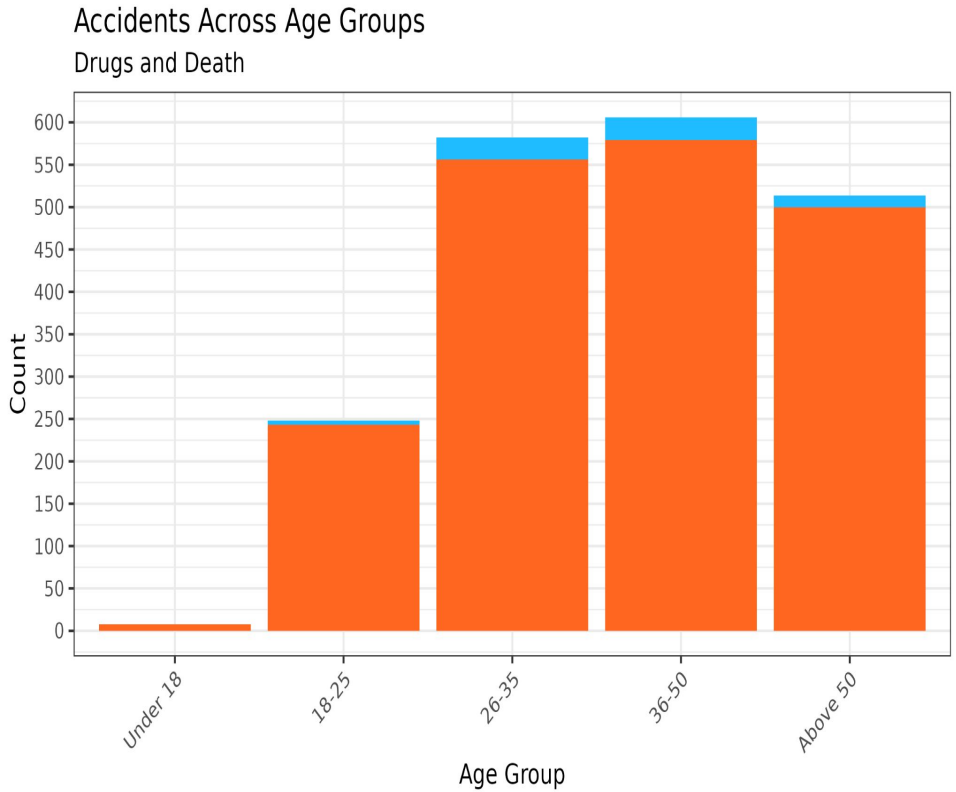


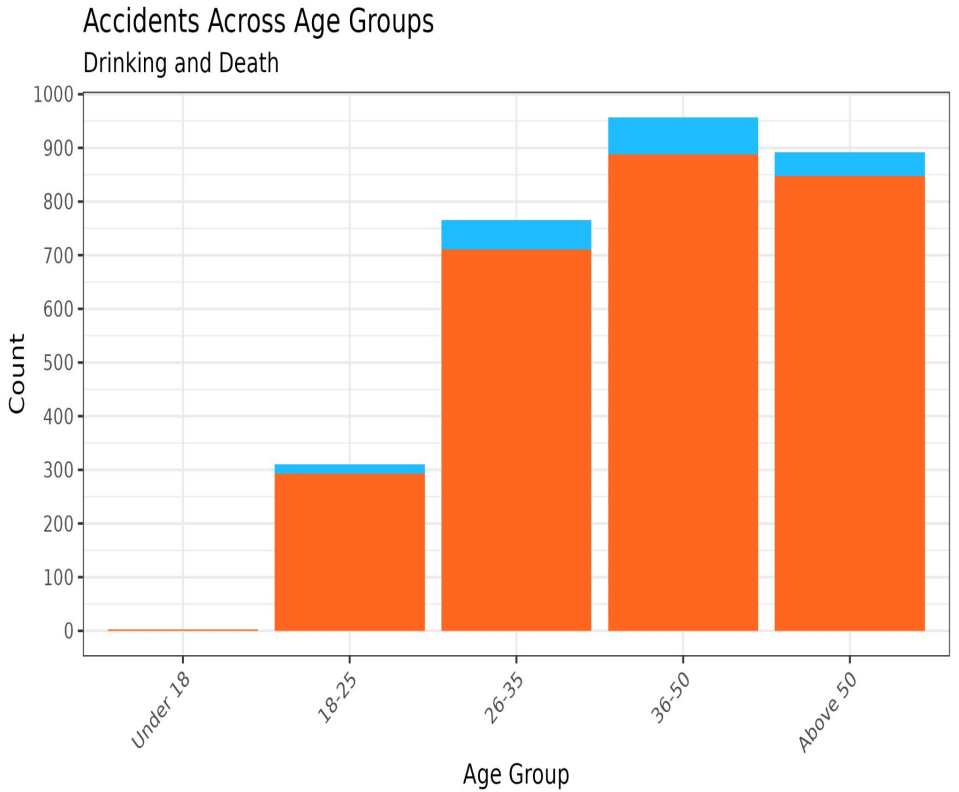
Probabilities - Vehicle





Probabilities - Motorcycle





Association Mining

Apriori

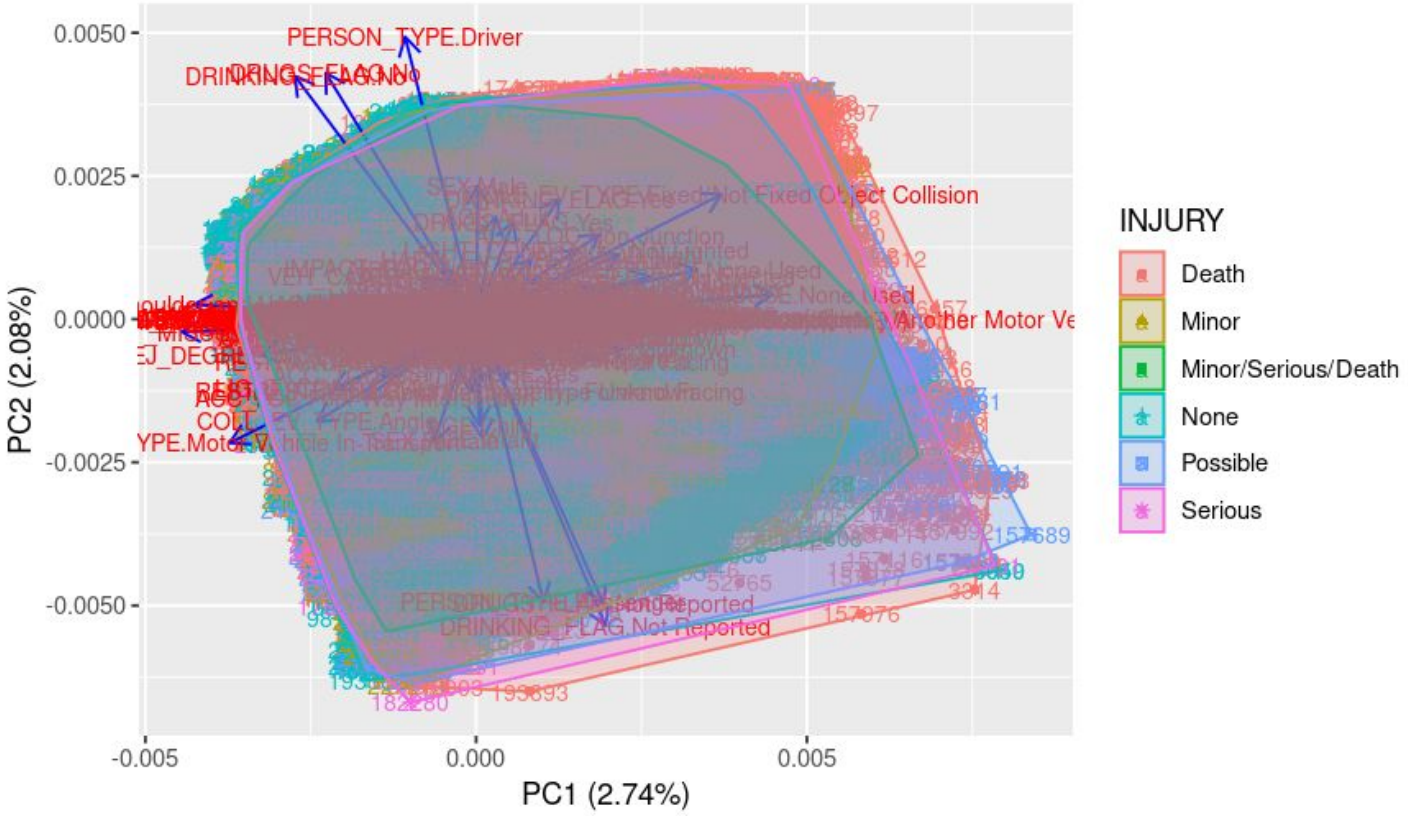
- Identifies frequent itemsets in a dataset and generates association rules based on these itemsets.
- We have filtered our dataset to include only drivers where their drinking status was reported (Yes/No). Records containing “Not Reported” have been excluded.
- Some columns such as STATE, LATITUDE etc have been removed for simplicity.
- The algorithm was applied to identify rules where the consequent (RHS) is INJURY=Death, with a minimum support of 1% and confidence of 80%.

Clustering

K Means

- Enables the algorithm to operate on the data without supervision.
- It assigns data points to one of the K clusters depending on their distance from the center of the clusters.
- Since we have 6 different injury levels, the plot contains 6 clusters.
- We have also reduced dimensionality using PCA.

K Means


PERSON_TYPE.Driver
DRINKING_FLAG.No
0.0050
0.0025
0.0000
-0.0025
-0.0050
PC2 (2.08%)
-0.005
0.000
0.005
PC1 (2.74%)
DRINKING_FLAG.Not Reported
INJURY
Death
Minor
Minor/Serious/Death
None
Possible
Serious

Classification

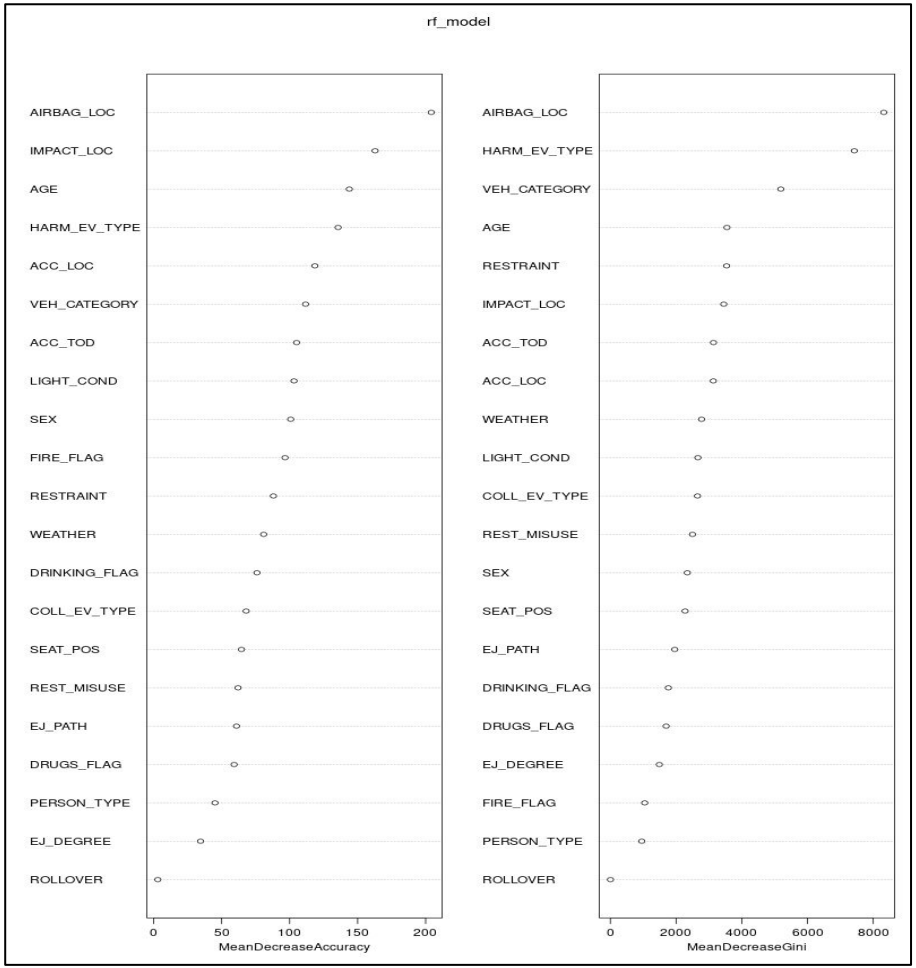
Random Forest

- Random Forest combines multiple decision trees to produce a single result.
- One of the major advantages is its avoids overfitting.
- By testing different combinations of parameters, we were able to create a model having 61% accuracy.
- Hyperparameter Values:
  - ntree = 500
  - mtry = 4
  - nodesize = 1
  - sampsize = floor(0.7 * nrow(train_data))
  - importance = TRUE

Random Forest

Overall Statistics:

Accuracy : 0.6101
95% CI : (0.6056, 0.6145)
No Information Rate : 0.3515
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.4441



Conclusion

Conclusion & Recommendations

- People above 50 are likely to die in an accident involving rollovers.
- A lot of people engage in DUI knowing well the consequences.
- Extra care has to be taken when driving/riding after dusk.
- Recommendations:
  - Improve road lighting and monitoring, especially on interstates and highways.
  - Enforce stricter helmet laws and educate motorcyclists about safety gear.
  - Deploy targeted traffic monitoring in high-crash urban areas.
  - Adjust traffic control measures during peak hours and weekends.