Analysis of Traffic Fatality Records

This manuscript (permalink) was automatically generated from uiceds/project-team-front-row@a001265 on December 13, 2024.

Authors

- **Justin Rebholz** ✉
  · jrebholz12
  Department of Civil and Environmental Engineering, University of Illinois

- **Cameron Kimber** ✉
  · ckimber2
  Department of Civil and Environmental Engineering, University of Illinois

- **Hannah Daggett** ✉
  · hed2
  Department of Civil and Environmental Engineering, University of Illinois

- **Riley Kelch** ✉
  · rileykelch
  Department of Civil and Environmental Engineering, University of Illinois

✉ — Correspondence possible via GitHub Issues or email to Justin Rebholz <rebholz4@illinois.edu>, Cameron Kimber <ckimber2@illinois.edu>, Hannah Daggett <hed2@illinois.edu>, Riley Kelch <rjkelch2@illinois.edu>.

Abstract

Introduction

Drunk driving fatalities are a significant concern to public safety, demanding targeted strategies to mitigate risks and save lives. This report analyzes traffic fatality data from the Fatality Analysis Reporting System (FARS), a nationwide census curated by the National Highway Traffic Safety Administration (NHTSA), with the goal of identifying patterns that can inform preventive measures. Using the 2015 FARS dataset, we focus on spatial and temporal trends in fatal motor vehicle accidents across the United States, with a particular emphasis on Illinois.

By examining factors such as drunk driving, time of day, and weather conditions, this report uncovers critical insights through exploratory data analysis. Additionally, predictive models, including k-means clustering and neural networks, are employed to forecast high-risk areas for fatal crashes, offering actionable guidance for resource allocation, such as the placement of DUI checkpoints. These findings aim to empower policymakers and law enforcement with data-driven solutions to reduce fatalities and enhance road safety.

Description

In this analysis, traffic data from the FARS created by the NNHTSA is used to predict where traffic accidents are most likely to occur. The FARS dataset for the year 2015 is available in the CSV format. The specific subset data that this project will be focused on is labeled “accidents.csv” and includes 32K+ instances and 52 columns. The columns are described in the FARS Analytical User’s Manual and in the below table:

Table 1: Abbreviation Legend.

Column	Description
ARR_HOUR	This data element records the hour when emergency services arrived at the scene.
ARR_MIN	This data element records the minute when emergency services arrived at the scene.
CF1, CF2, CF3	These data elements record contributing factors to the crash, such as driver behaviors or environmental conditions.
CITY	This data element identifies the city in which the crash occurred.
COUNTY	This data element identifies the county in which the crash occurred.
DAY	This data element records the day of the month on which the crash occurred.
DAY_WEEK	This data element identifies the day of the week on which the crash occurred.
DRUNK_DR	This data element records whether a driver involved in the crash was suspected of drinking alcohol.
FATALS	This data element records the number of fatalities resulting from the crash.

Column	Description
FUNC_SYS	This data element identifies the functional classification of the trafficway segment where the crash occurred.
HARM_EV	This data element records the first harmful event that occurred in the crash sequence.
HOSP_HR	This data element records the hour when the injured were admitted to the hospital.
HOSP_MN	This data element records the minute when the injured were admitted to the hospital.
HOUR	This data element records the hour when the crash occurred.
LATITUDE	This data element identifies the location of the crash using latitude coordinates.
LGT_COND	This data element identifies the light condition at the time of the crash, such as daylight, dark, or dusk.
LONGITUD	This data element identifies the location of the crash using longitude coordinates.
MAN_COLL	This data element identifies the manner of collision, such as rear-end, head-on, or angle.
MILEPT	This data element records the milepoint nearest to the crash location.
MINUTE	This data element records the minute when the crash occurred.
MONTH	This data element records the month in which the crash occurred.
NHS	This data element identifies whether the crash occurred on a National Highway System (NHS) route.
NOT_HOUR	This data element records the hour when the crash was reported to authorities.
NOT_MIN	This data element records the minute when the crash was reported to authorities.
PEDS	This data element records the number of pedestrians involved in the crash.
PERMVIT	This data element counts the number of persons in motor vehicles in transport (motorists) involved in the crash.
PERNOTMVIT	This data element counts the number of persons not in motor vehicles in transport (non-motorists) involved in the crash.
PERSONS	This data element is a count of the total number of persons involved in the crash.
PVH_INVL	This data element is the number of parked or working vehicles involved in the crash.
RAIL	This data element identifies if the crash involved a rail system or crossing.

Column	Description
RELJCT1	This data element identifies the relationship of the crash to a junction, such as intersection or non-intersection.
RELJCT2	This data element provides additional information about the crash's relationship to the junction.
REL_ROAD	This data element identifies the relationship of the crash to the road, such as on the roadway or off the roadway.
RD_OWNER	This data element identifies the entity responsible for the ownership of the road where the crash occurred.
ROUTE	This data element records the type of route where the crash occurred, such as Interstate, U.S. Highway, or State Highway.
RUR_URB	This data element identifies whether the crash occurred in a rural or urban area.
SCH_BUS	This data element identifies if a school bus was involved in the crash.
SP_JUR	This data element identifies if the crash occurred in a special jurisdiction, such as military or Indian reservations.
STATE	This data element identifies the state in which the crash occurred. The codes are from the General Services Administration's (GSA) publication of worldwide Geographic Location Codes (GLC).
ST_CASE	This data element is the unique case number assigned to each crash. It appears on each data file and is used to merge information from the data files together.
TWAY_ID	This data element identifies the primary trafficway on which the crash occurred.
TWAY_ID2	This data element identifies the secondary trafficway associated with the crash, if applicable.
TYP_INT	This data element identifies the type of intersection involved in the crash, if applicable.
VE_FORMS	This data element is a count of all vehicle forms applicable to this crash.
VE_TOTAL	This data element is the number of contact motor vehicles that the officer reported on the PAR as a unit involved in the crash.
WEATHER	This data element identifies additional weather factors at the time of the crash.
WEATHER1	This data element records the primary weather condition at the time of the crash.
WEATHER2	This data element records the secondary weather condition at the time of the crash.
WRK_ZONE	This data element identifies if the crash occurred in a work zone.

Column	Description
YEAR	This data element records the year in which the crash occurred.

Link: https://www.kaggle.com/datasets/nhtsa/2015-traffic-fatalities

Plan and Proposal

Using the FARS dataset from the year 2015, our team aims to understand the trends in traffic fatalities over the entire continental United States as well as just Illinois and what factors are affecting those trends. This analysis can then be used to implement safety factors that guard against drunk driving (traffic cameras, sensor systems, DUI checkpoints, etc.) The trends in traffic fatalities found through this data analysis can be used to inform policy makers and ultimately decrease the number of traffic fatalities.

PRJ2.1 Exploratory Data Analysis

Exploratory Data Analysis

The dataset that we have chosen describes the details surrounding motor vehicle crashes in the United States during the year 2015.

From the data, we have interpreted that 32,166 fatal crashes occured in 2015. Out of the total number of crashes, 26.78% of accidents involved an intoxicated driver. In the state of Illinois, 264 of 914 crashes involved a drunk driver (28.9%). More statistics are found in the table below.

Table 2: Misc Statistics.

Data Summary	Statistic
Crashes (US)	32,166
Total Fatalities (US)	35,092
Drunk Driver Crashes (US)	8,617
Drunk Driver Percentage (US)	26.78%
Crashes (IL)	914
Total Fatalities (IL)	998
Drunk Driver Crashes (IL)	264
Drunk Driver Percentage (IL)	28.88%
Most DD Crashes, Dates (US)	03May, 15Aug, 02Aug, 16Aug
Most Crashes, Dates (IL)	07Mar, 27Jun, 17Apr

To understand the trends in the data, we first analyzed the location of accidents and how the location relates to other variables. We looked at a map of the United states to plot the fatal accidents vs the drunk driving fatal accidents, as seen in Figure 1. This visualtion shows hotspots for both categories which are generally in more populous areas and in coastal regions.

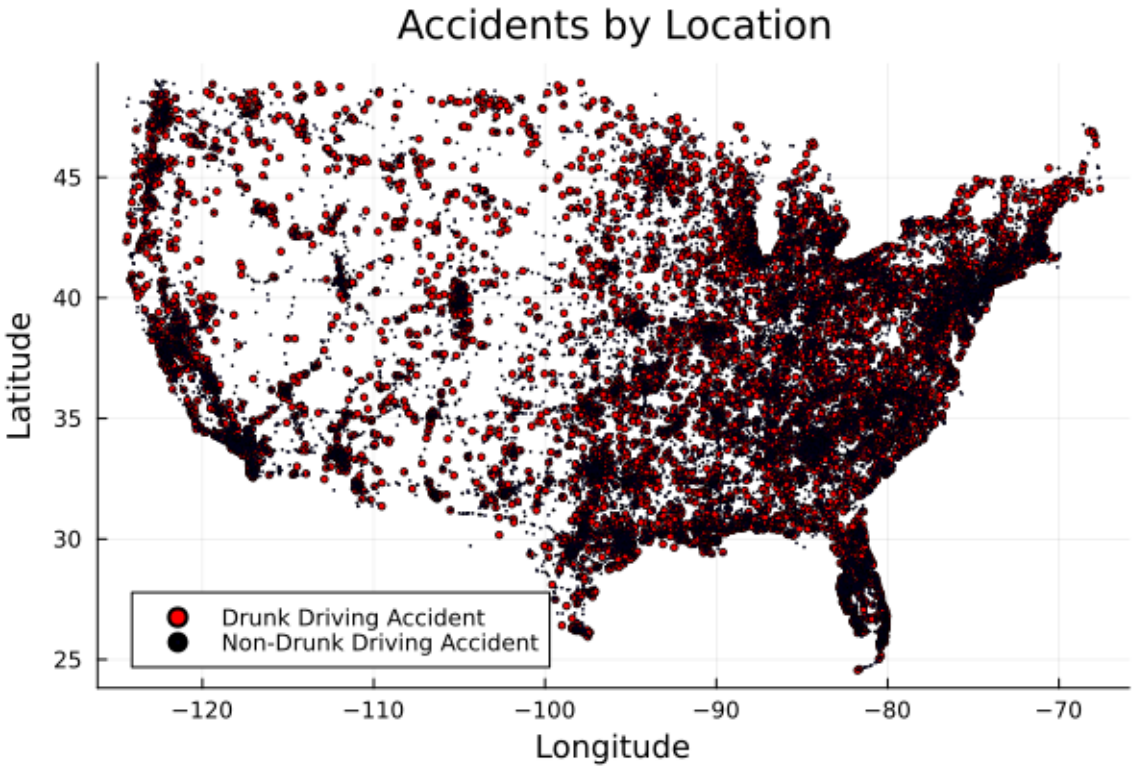

Figure 1: US Map

We then created a scatter plot to visualize the number of accidents and fatalities per accident for the United States. As seen in the figure, most of the crashes result in only one fatality, however, there are a handful of multiple fatality crashes. The scatter plot annotates the amount of crashes for drunk drivers and sober drivers next to the datapoint.

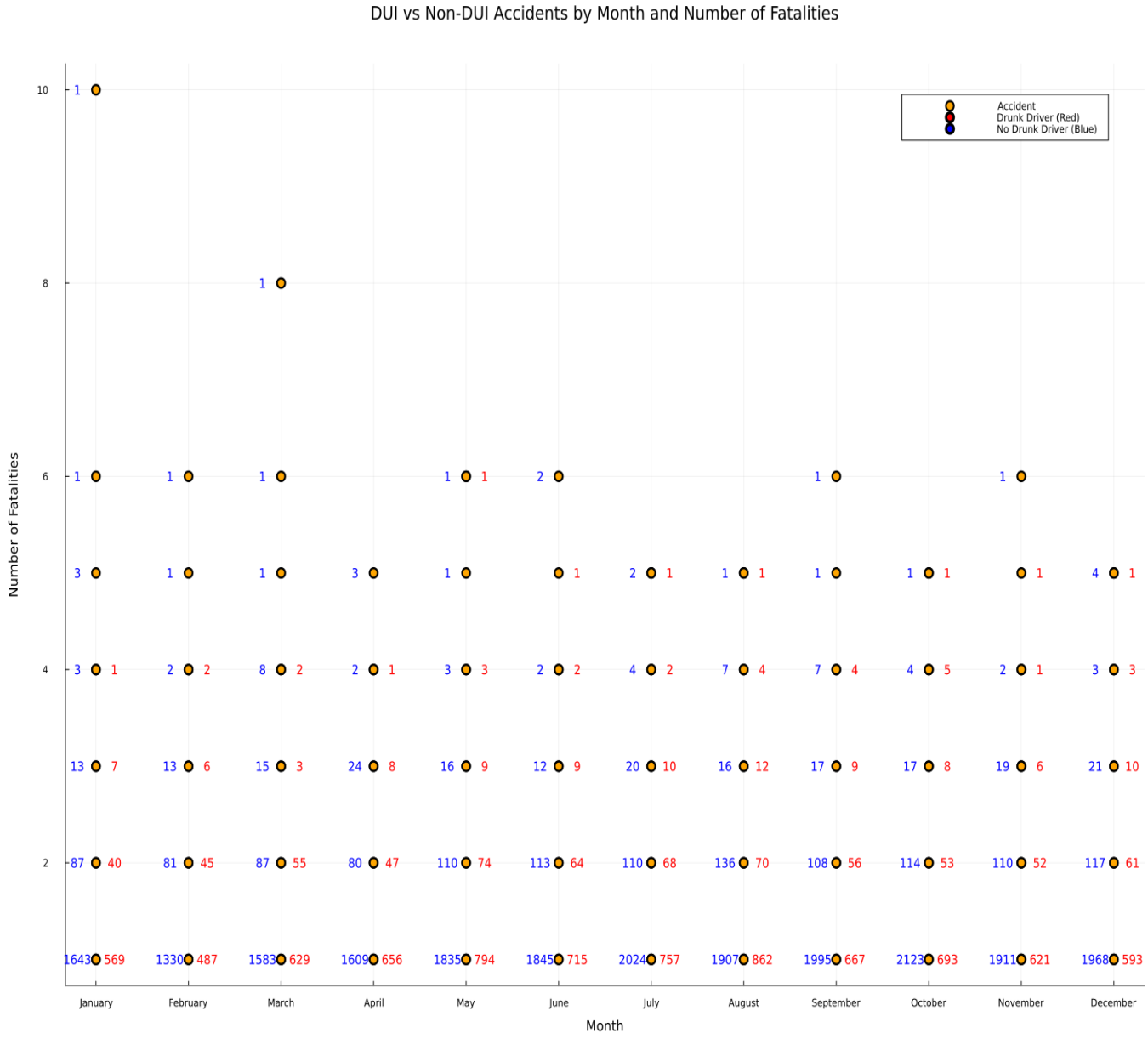

Figure 2: DUI vs Non-DUI Accidents by Month and Fatality Count

The next factor that we analyzed to understand the data was the specific day of accidents. As seen in the figure below, the highest number of accidents occurred on the weekends. On average, 104 accidents occurred on a given day of a weekend whereas 82 accidents occurred per weekday. Monday and Tuesday have the lowest number of accidents and as the week progresses, the number of accidents increases.

The portion of accidents due to drunk driving by the day of the week follows similar trends. The average ratio of drunk driving accidents to total number of accidents was 37.47% for the weekend and 21.19% for weekdays. This is most likely due to the fact that drinking is more popular on the weekends. Similarly, there is a figure for crashes vs time of day.

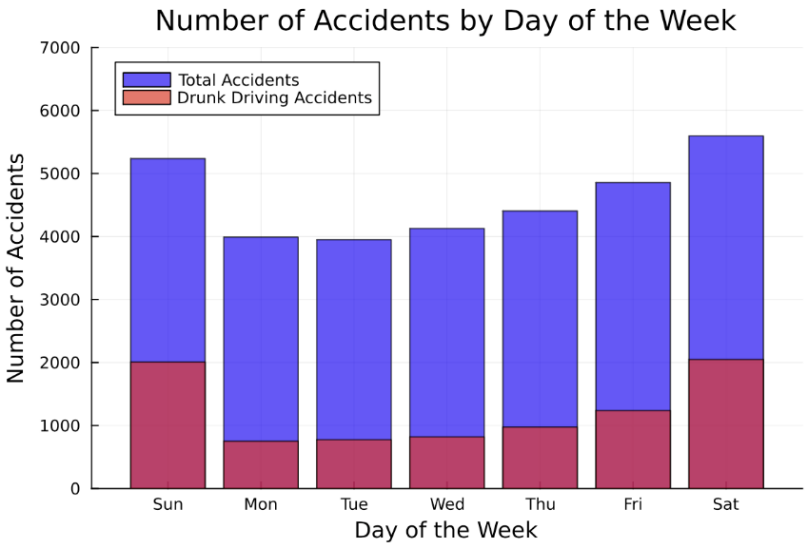

Figure 3: Accidents vs Day of Week

Our team also looked at the number of drunk driver accidents per day of the year. See Table 1 for the most popular days for drunk driver accidents. The total number of accidents by hour of the day is the highest at 3 AM and decreases until 8 AM. The number of accidents then gradually increases by the hour.

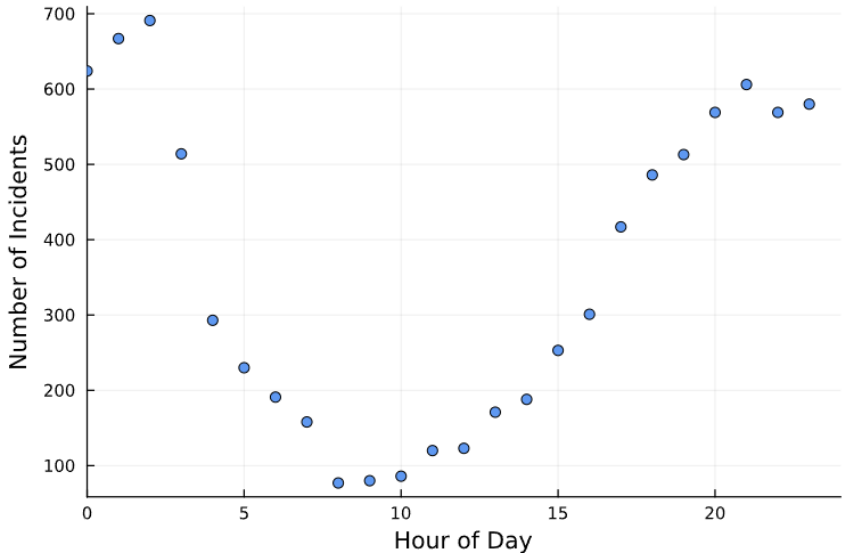

Figure 4: Time of Day

Looking at the total number of DUI related accidents throughout the year, the summer months see higher numbers of accidents.

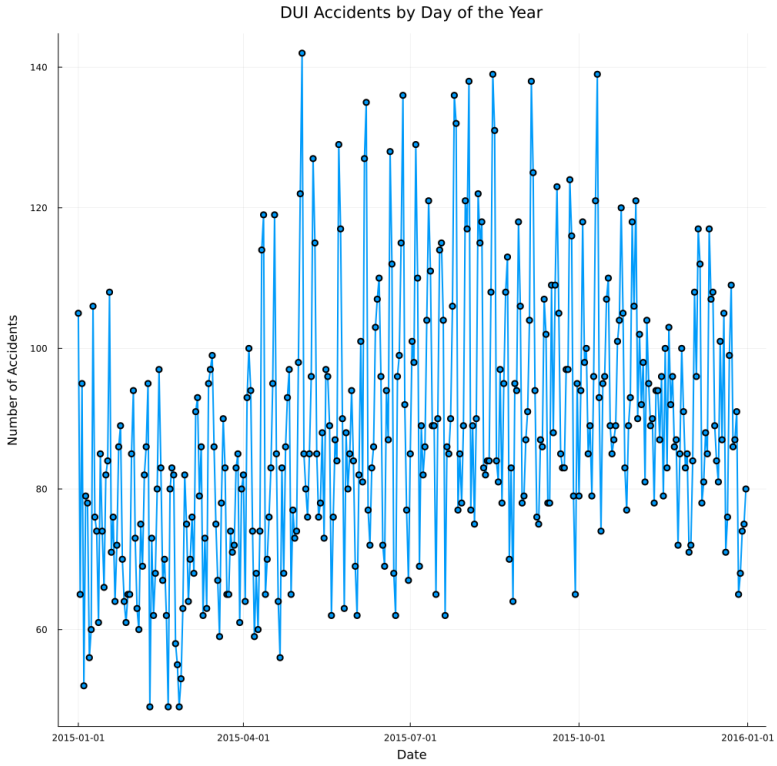

Figure 5: DUI Crashes by Day (US)

As we conducted our Exploratory Data Analysis, we aimed to focus in on the state of Illinois. Below is the same information as Figure 1, but specific to Illinois for 2015. From this visual, we can see that most crashes are in the areas with larger cities (i.e. Chicago).

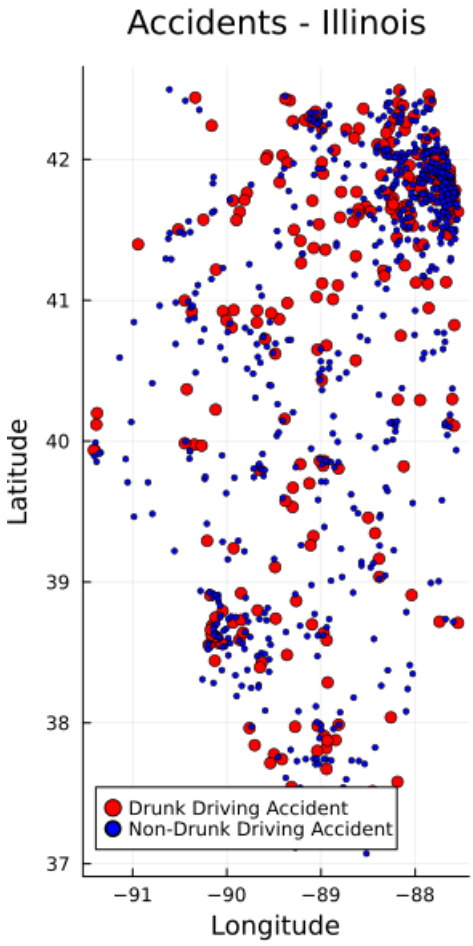

Figure 6: Illinois Map

PRJ3.1 Predictive Modeling

Predictive Modeling

K-means

Based on the analysis conducted on the provided data, a predictive model utilizing k-means clustering can assist in determining ideal locations for the implementation of DUI checkpoints based on crash sites. Adjustment of k in the clustering algorithm can be done using parameters such as resource allocation towards policing in distinct areas. The k-means clustering process will begin by taking crash data with drinking involved within Illinois and the United States. This data will be clustered based on a predefined k value, representative of a decision made based on resource allocation.

Figure 7 below depicts all of the drunk driving incidents that occurred in Illinois. Using k-means clustering, the instances were grouped into 20 groups based on their location. These groups were then snapped to the nearest major interstate in order to avoid centroid points being placed in unrealistic locations such as over agricultural land.

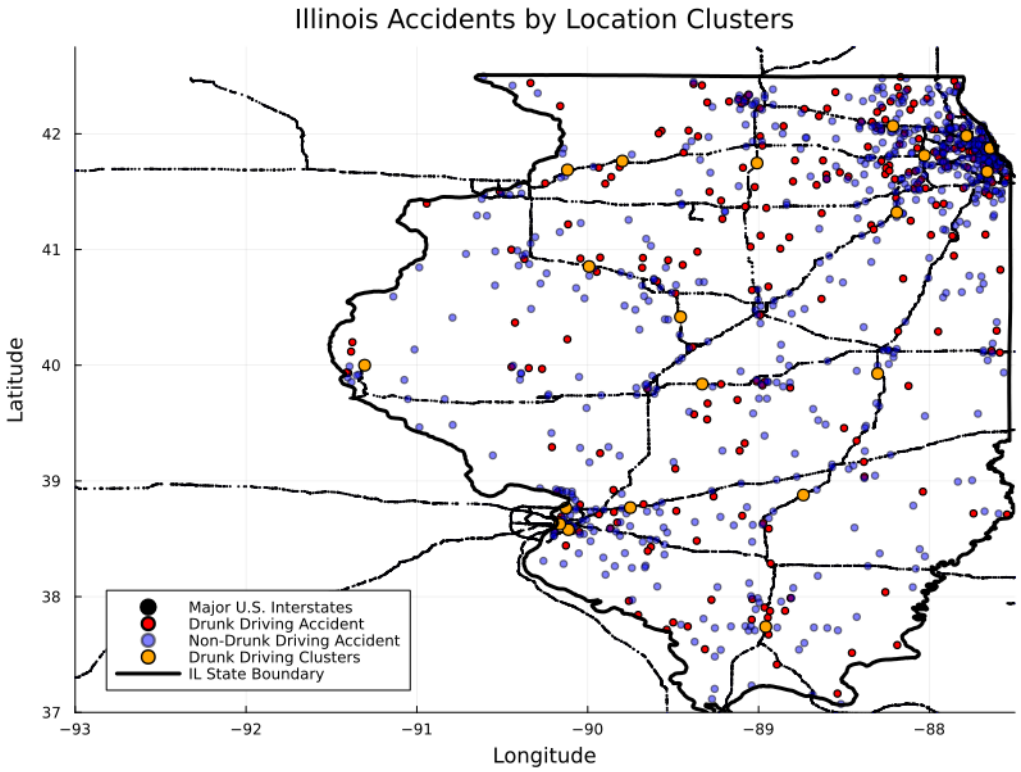

Figure 7: Illinois DD Clusters

Additional k-means clustering models were created to provide insights into potential high-risk zones in regards to drunk driving incidents. The seven plots below illustrate the distribution of drunk driving accidents on a daily basis, revealing notable trends throughout the week. Drunk driving incidents show a clear increase during the weekends, likely due to bars, restaurants, and clubs tending to have higher activity on Friday and Saturday nights. On weekdays, drunk driving incidents appear to be more clustered along commuter routes. However, on weekends, these clusters appear to shift to more urbanized areas where more recreational zones and leisure activities may be present. In other words, weekend incidents appear more concentrated on city centers. A k-value of 50 (i.e. 50 centroid points) is used for this analysis; however, this value is arbitrary and can be adjusted to suit the needs of policymakers as to where resources can be allocated to most efficiently drunk driving incidents.

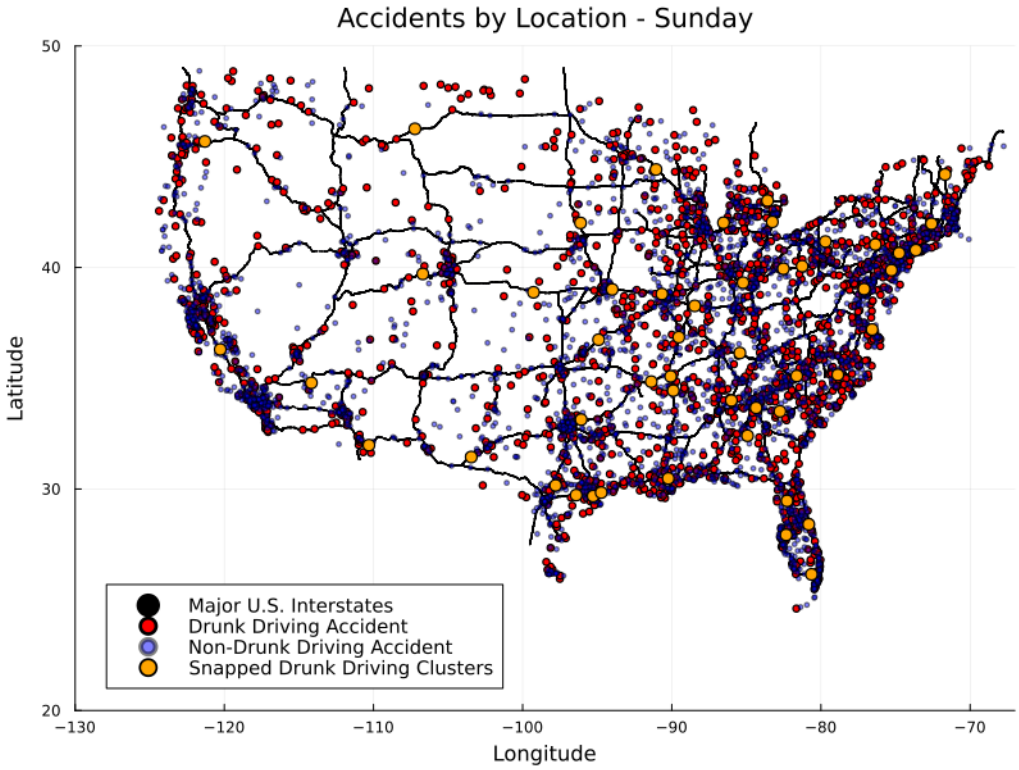

Figure 8: Sunday Cluster

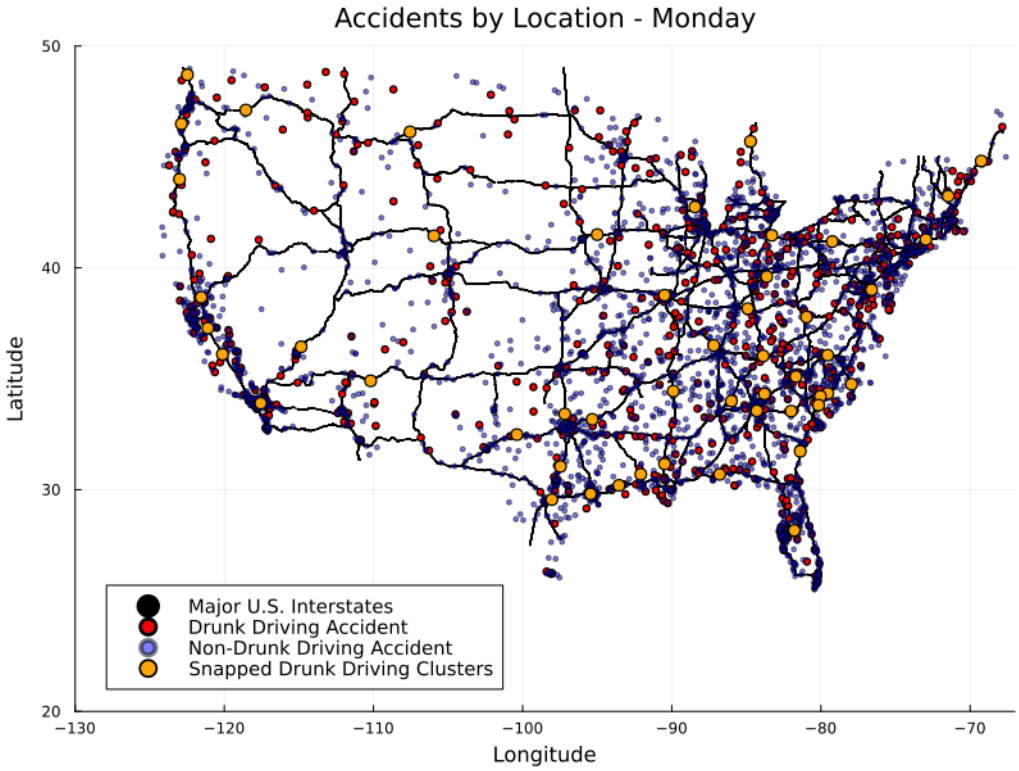

Figure 9: Monday Cluster

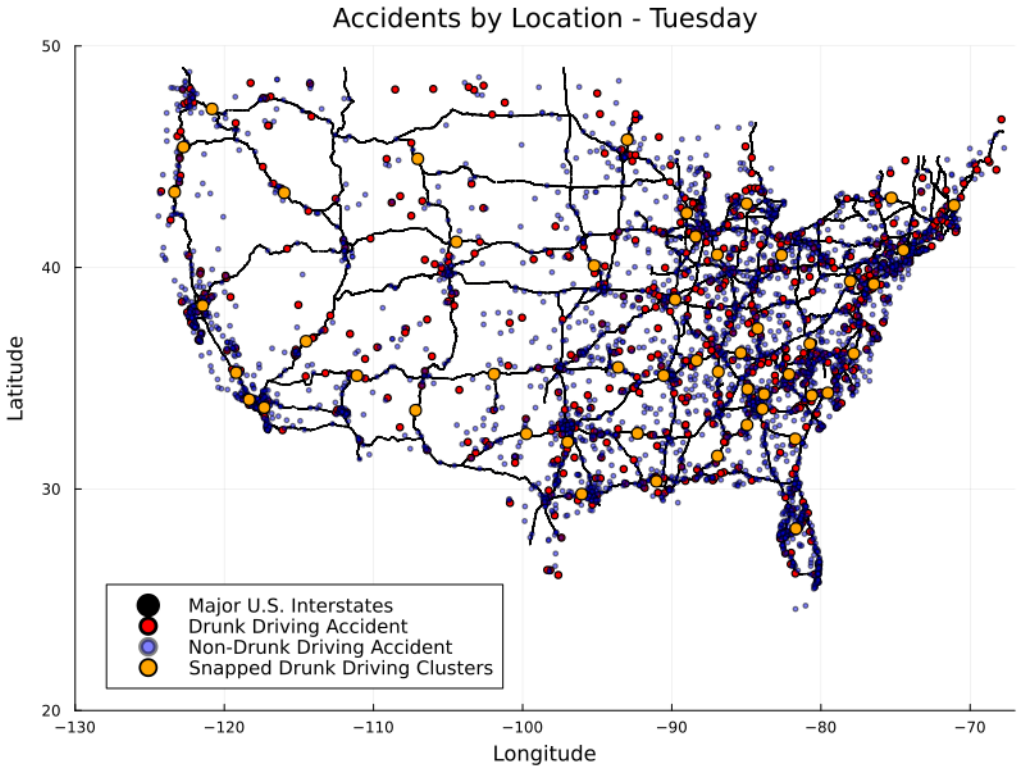

Figure 10: Tuesday Cluster

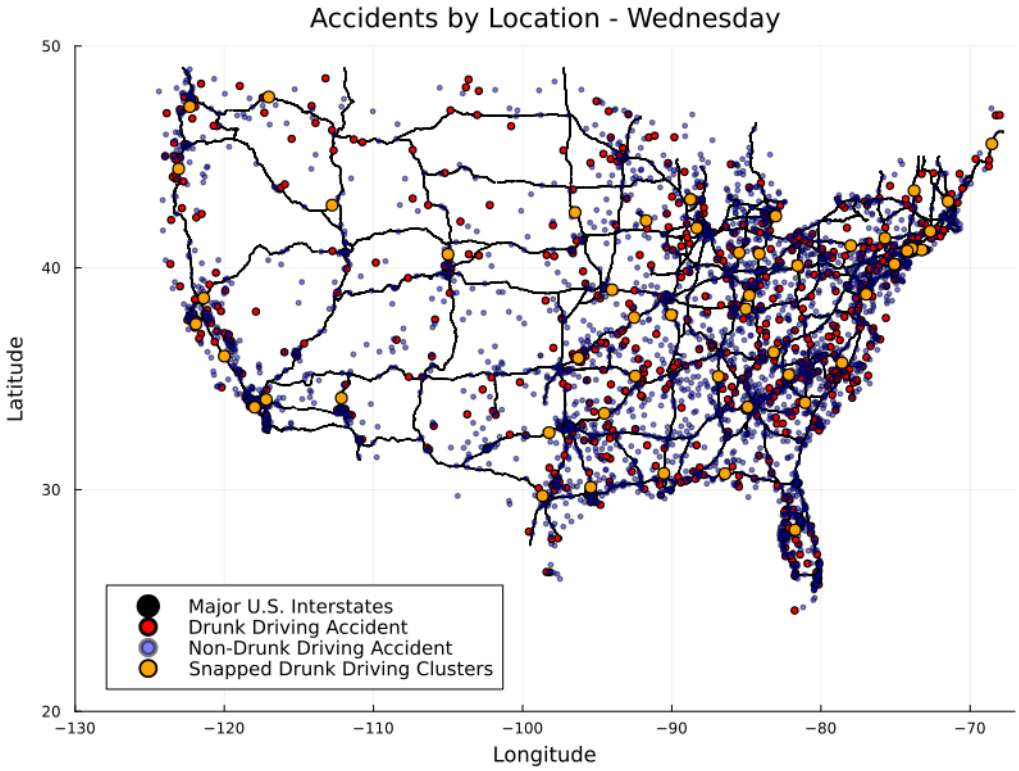

Figure 11: Wednesday Cluster

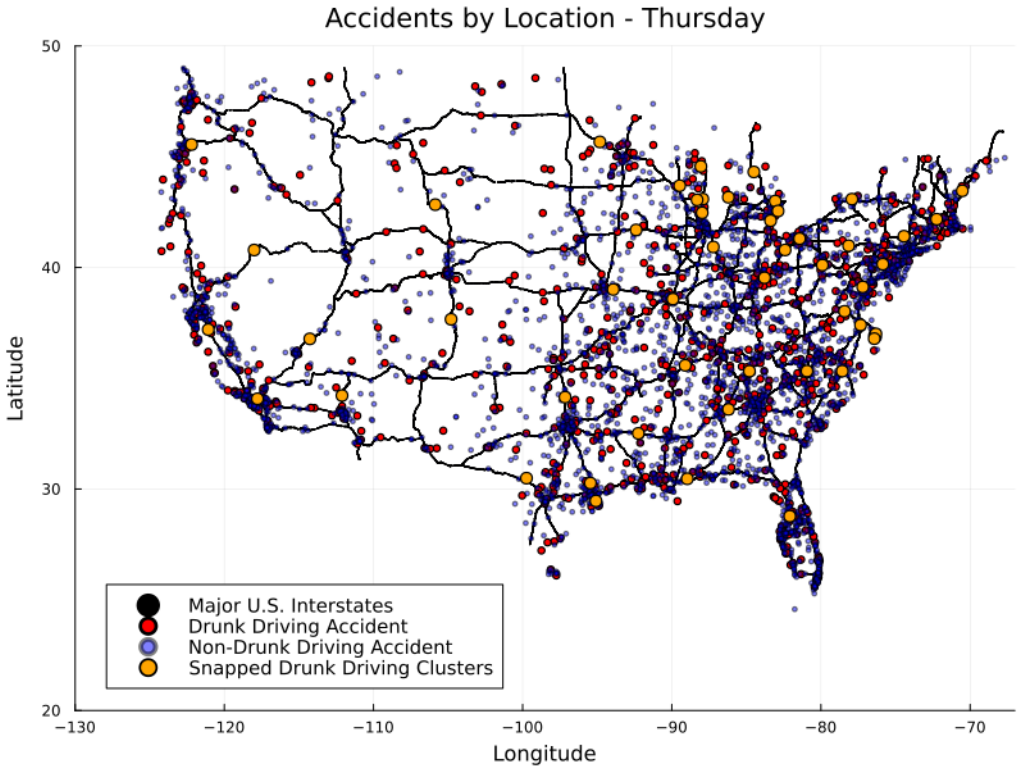

Figure 12: Thursday Cluster

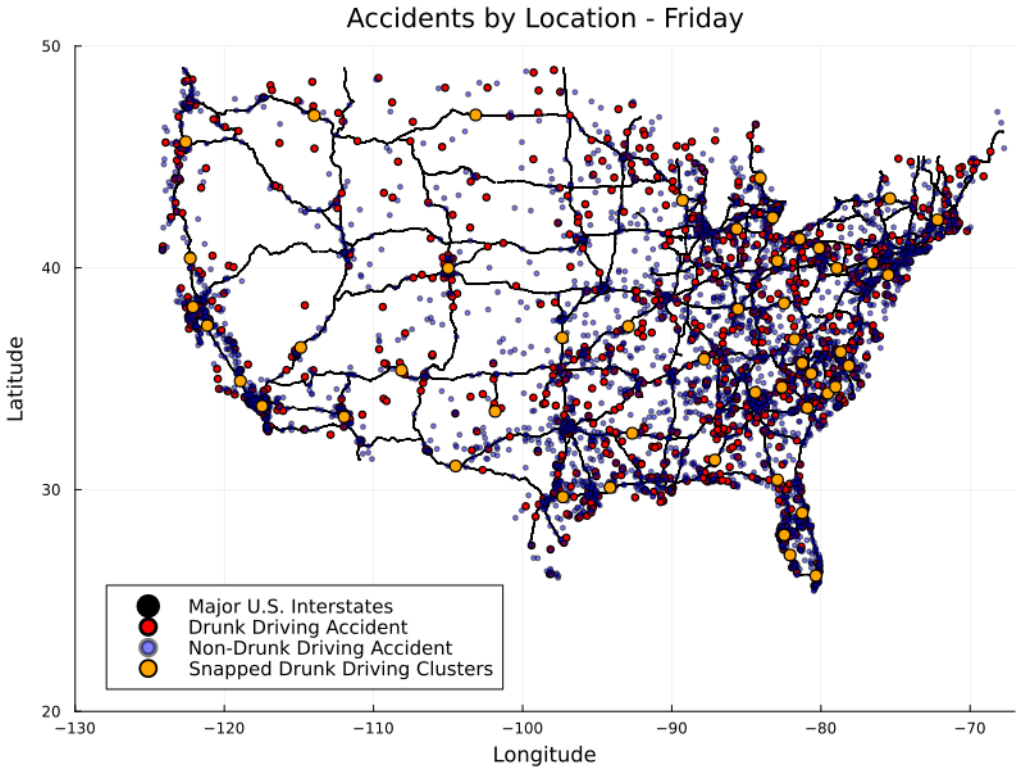

Figure 13: Friday Cluster

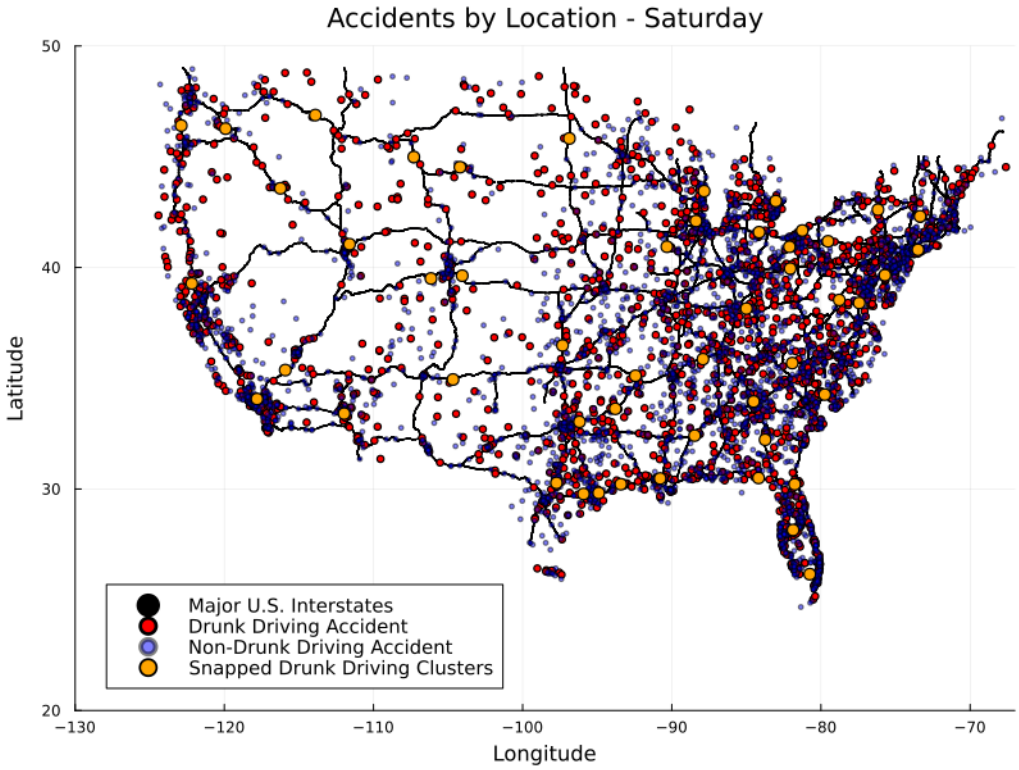

Figure 14: Saturday Cluster

Drunk driving incidents across the entire United States were further analyzed, as depicted by Figure 8 below. The visualization demonstrates the spatial distribution of all drunk driving accidents and the resulting clusters identified using k-means analysis, offering insight into nationwide trends. As expected, the amount of incidents are most dense in areas with higher populations. However, some major interstates contain higher amounts of centroid points than others, indicating to drivers areas they may want to avoid as well as informing law enforcement areas to increase measures to reduce drunk driving incidents.

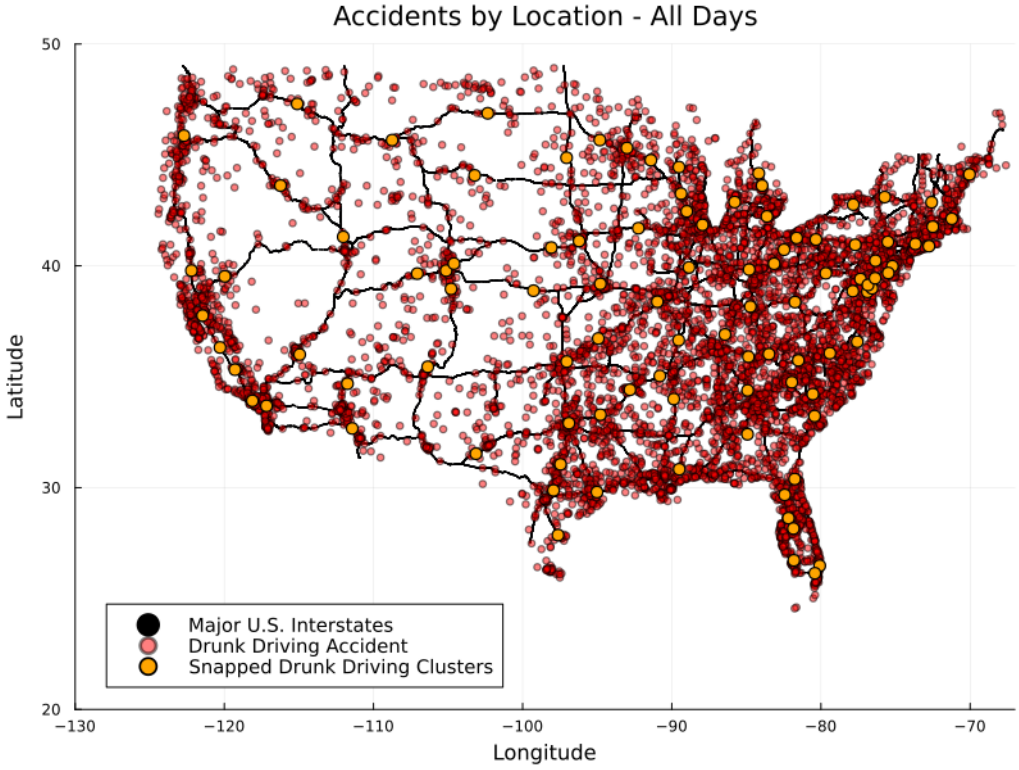

Figure 15: Drunk Driving Cluster

Some limitations exist with the k-means modeling process. Running the scripts multiple times often results in different centroid locations with each rerun. In the future, this could be corrected by incorporating a weight system into the clustering. Centroids could be weighted based on factors such as fatality count, whether the incident involved drunk driving, or both. This approach would add consistency to the clustering results and allow the model to better reflect the relative severity and risk associated with each cluster. Additionally, further refinement of the algorithm parameters, such as initializing centroids, could enhance the accuracy and reliability of the model.

Additional consideration was given towards dimensions of time regarding crash likelihood. According to Figure 4, noticeable variation occurs in the amount of accidents occurring at specific times of day, indicating a need to manage resource allocation with time consideration. The implementation of a neural network using the provided data can be applied towards a proposal regarding DUI checkpoint locations or patrol areas in designated areas based on available resources and shift time, therefore optimizing provision of safety from DUI-related incidents.

Neural Network

The model below represents a neural network designed to predict high-risk areas for drunk driving accidents in Illinois, potentially useful for the Illinois State Police. It is a multiple dense-layered, supervised, feedforward neural network. Standardization and component engineering were employed to improve model performance. The independent variables used in the model include the day of the week, the hour of the shift, weather conditions, and latitudinal and longitudinal bins corresponding to a twenty-mile radius. The dependent variable is the number of drunk driver crashes. Unlike simply calculating the average number of accidents per area, this model incorporates additional factors such as time and weather conditions, offering a more precise prediction of high-risk areas. By considering these variables, the model allows law enforcement to focus patrol efforts based on the conditions that increase the likelihood of drunk driving accidents, making it a more actionable tool than average-based approaches.

The figures below represent visualizations with varying input data for specific standard police patrol shifts. For instance, a police officer working on a Friday from 12 PM to 12 AM with current rainy weather conditions would input the day, time of shift, and weather into the model. The model would then generate predictions, highlighting the 20-mile radii areas with the highest probabilities of drunk driving accidents, as seen in the upper left of Figure 9.

The top right of Figure 9 displays the predicted high-risk areas for drunk driving accidents on Saturday from 6 PM to 12 AM under clear weather conditions, while the bottom of Figure 9 illustrates similar predictions for Sunday from 12 AM to 4 AM, but for all weather conditions. Saturday and Sunday have more areas of interests, due that more drunk driving fatalities commonly occur on late Saturday nights, as discussed in the Exploratory Data section.

After producing the output map, the law enforcement officer could focus patrols on the displayed areas.

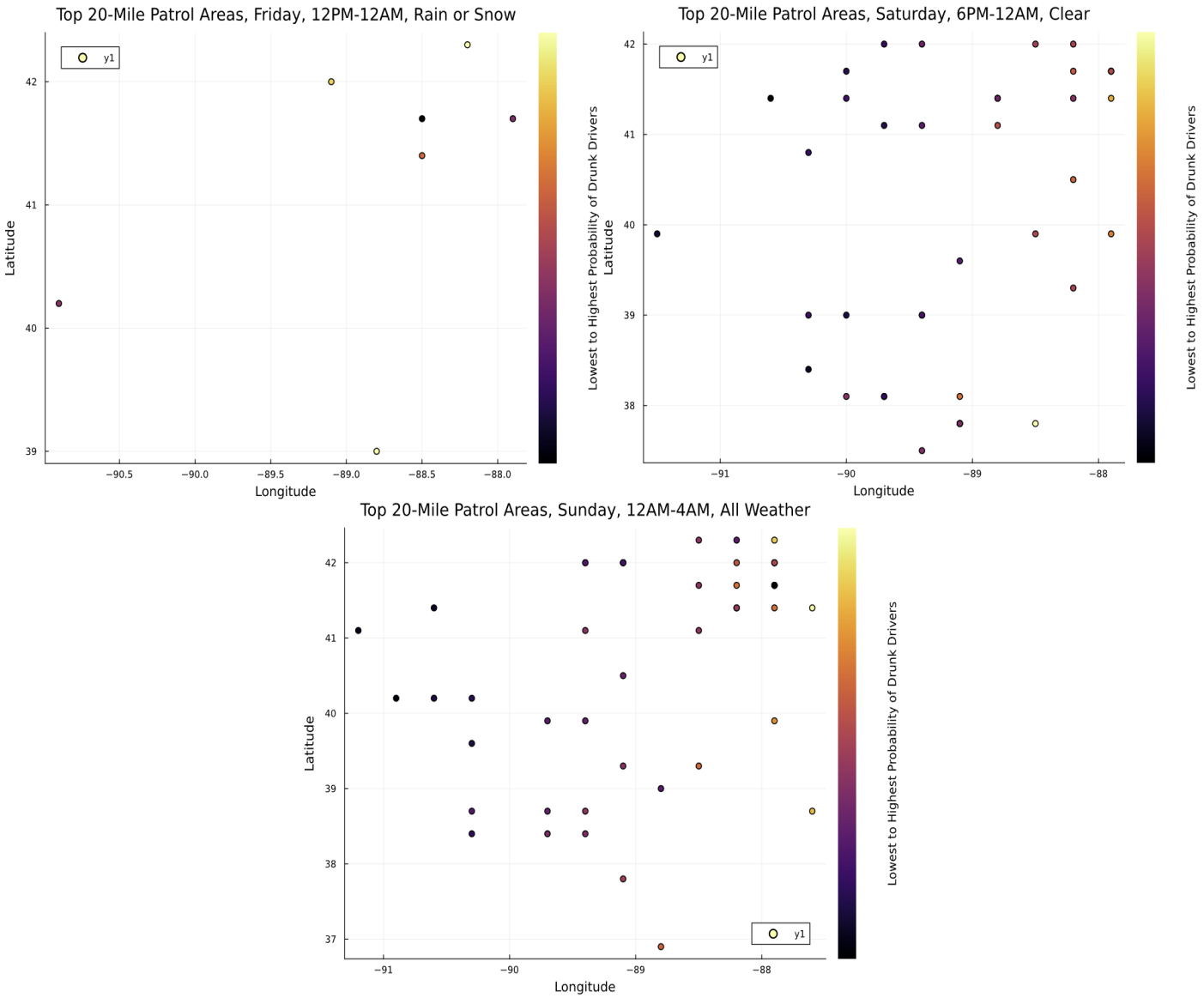

Figure 16: 3-Day Output from Neural Network Model

PRJ 4.1 Discussion

Modeling Use

The findings of this project provide insight into the temporal and spatial distribution of drunk driving incidents, both within Illinois and across the United States. By using k-means clustering and the neural network model to identify high risk areas, this analysis offers a starting point for policymakers and law enforcement to allocate resources efficiently in addressing drunk driving. Several key implications, limitations, and potential next steps emerge from the work done in this analysis.

The k-means model demonstrates its usefulness in identifying patterns in drunk driving incidents based on their spatial distribution. Nationwide, the analysis confirms the intuitive relationship between population density and the frequency of incidents, with denser regions and major interstates showing the highest concentration of incidents. However, some interstates exhibit higher numbers of centroid points, identifying potential targets for intervention, such as enhanced law enforcement presence.

The neural network model could enhance targeted intervention for standard police shifts. The model could be integrated into existing police vehicle computer systems for real-time targeting. Additionally,

it could be used by high-levels of the organization to plan future patrol areas. It shows it's effectiveness by consistently targeting more highly populated areas (i.e. zones within Chicago area).

Conclusion and Future Development

While the clusters provide valuable insights, the model's reliance on a fixed k-value introduces some subjectivity. Though the k-values were set in this analysis, they remain arbitrary and may need adjustment to better suit a specific region's needs or policy objectives. These k-values can be tailored to the desired scale of intervention or available resources, ensuring that high-risk zones are accurately identified. There are further limitations with the k-means clustering model. The k-means algorithm's randomness in its initialization of centroids can result in a variation of cluster locations across multiple runs of the model. This inconsistency undermines the reproducibility of the results. Future implementations should explore methods to stabilize cluster initialization. Furthermore, the model treats all incidents equally, without accounting for factors such as fatality counts. Incorporating a weighting system could improve the model's accuracy by emphasizing more severe incidents, leading to a more meaningful representation of risk analysis.

If this project were to continue beyond this current semester, the k-means model could be improved to take in additional factors such as weather, road conditions, and event data to enrich the clustering model and perhaps increase the reproducibility of the results. Additionally, more data could be used over a larger span of time, especially for smaller communities in which the amount of drunk driving incidents might not be large enough over the course of a single year to create meaningful clusters.

While the neural network model shows promise, its current accuracy does not make it usable for practical purposes. A negative R^2 value was calculated, which indicates a poor model fit. This is primarily due to the limitation of the dataset, which includes only one year of fatal accident data. Although the model's accuracy is not yet high enough to be considered fully reliable, any reduction in drunk driving accidents, even modest, could still have a significant positive impact on public safety. To improve the model's accuracy, one potential approach would be to extend the dataset to include all DUI-related incidents, not just fatal accidents. Additionally, using data from multiple years could help the model capture more diverse patterns and improve its overall performance.

References

National Highway Traffic Safety Administration. “2015 Traffic Fatalities.” Kaggle, https://www.kaggle.com/datasets/nhtsa/2015-traffic-fatalities. Accessed 24 Oct. 2024.

National Highway Traffic Safety Administration. Fatality Analysis Reporting System (FARS) Analytical User’s Manual 1975-2015. U.S. Department of Transportation, Aug. 2016.