Final Paper

STOR 320.01 Group 3

May 01, 2024

INTRODUCTION

Using our data set, we wish to focus on examining the spatial patterns of severe traffic incidents in the United States using a data-centric approach. We hypothesize that severe accidents are geographically concentrated and that these concentrations are influenced by both environmental and infrastructural factors. More specifically, we believe that elevation and points of interest, such as cities and perhaps primary and secondary roads will have an impact on how accidents of different severities are concentrated. Our first question examines whether we can extract patterns by creating map plots to visualize accidents. In addition to perhaps aiding in predicting and preventing future incidents, the insights derived from our visualizations could be useful for traffic safety authorities and urban planners, enabling them to identify high-risk zones and allocate resources effectively. Additionally, we believe that by identifying the geographical distribution of severe accidents, we can better understand the dynamics at play, which can aid in the development of more effective preventive measures and policies.

In our second question, using models, we focused on whether we could predict the total time of the accident based on variables in our data set such as the state in which the accident occurred, if there was an exit nearby, and visibility. This question looks to examine what some key factors are that contribute to how long an accident affects the road it occurred on, as well as factors that have no impact on it at all. Having a formula that allows someone to know how long traffic will be impacted by an accident could be used by navigation apps to provide a more accurate arrival time or by urban planners to make improvements to roads that have a higher chance of being affected by an accident longer.

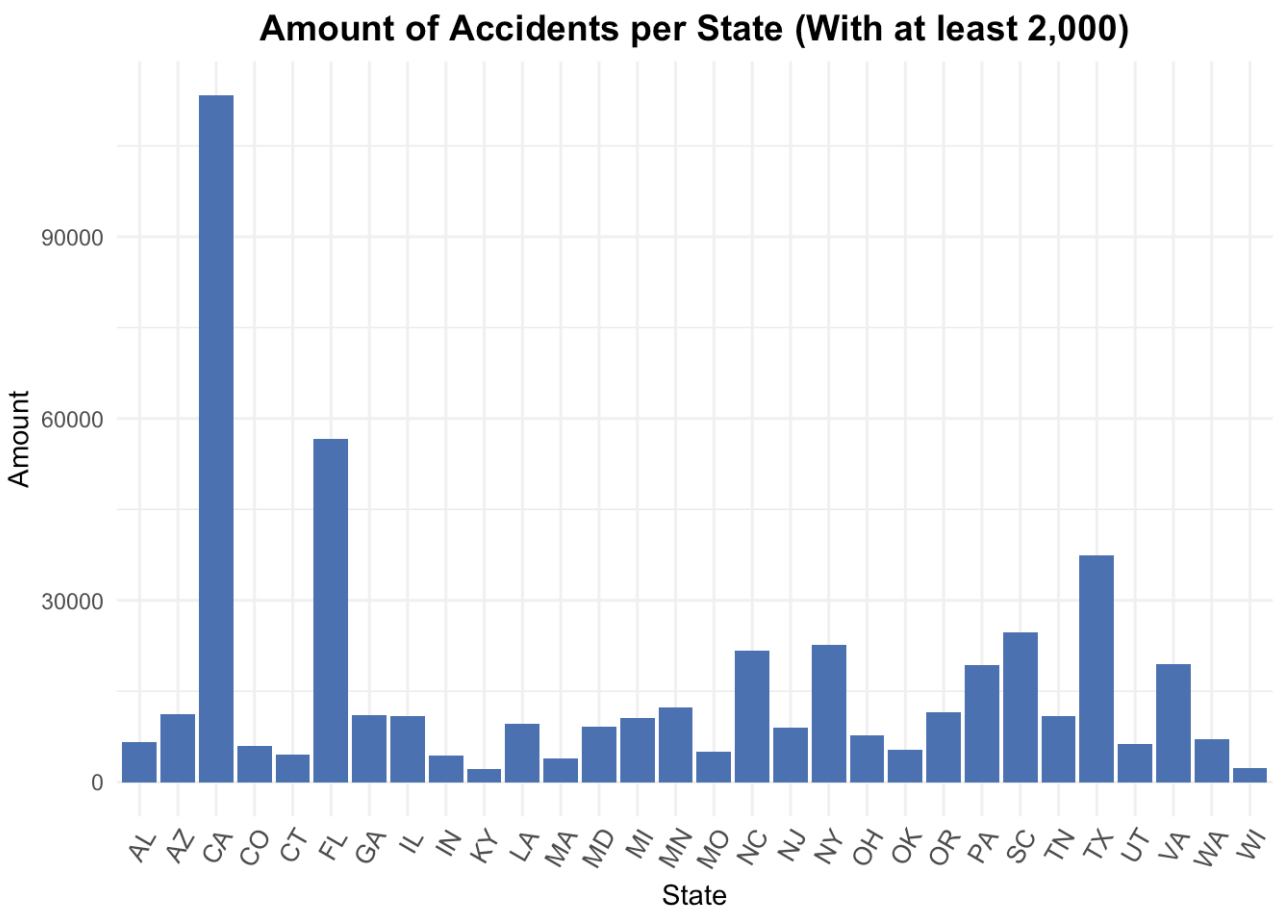
Questions:

1. Can we extract patterns from concentrations of different severity of accidents by creating map plots to visualize accidents?
2. Can we predict the total time of an accident based on a variety of variables in our data set?

DATA

We discovered this data set on Kaggle. A user by the name of Sobhan Moosavi recently updated this data set a year ago. The data was collected from a group of people and their papers as follows: Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019. Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019. The full dataset is comprised of 7.7 million car accidents in the United States, however, we used a subset of 500,000 random car accidents already created by the creator of the dataset. This allowed for R to process our code smoothly and actually be able to import the dataset. We believe this dataset is thorough enough for us to dive deeper into car accidents in the United States as well as explore the answers to our questions.

Below is a bar chart representing the states where the most accidents occurred. This image shows states that had at least 2,000 accidents. This image allowed us to see what states in particular we should investigate further.



This data set initially had 46 variables; however, we only used 38 through our analysis. We removed *ID* as it served as a unique identifier for each accident and we did not need this. We also removed *Civil_Twilight*, *Nautical_Twilight*, *Astronomical_Twilight*, *Source*, *Description*, *End_Lat*, *End_Lng*, *Country*, *Street*, and *Weather_Timestamp* as we believed they were irrelevant to our analysis.

Below are descriptions of the variables:

- *Severity* shows a number 1, 2, 3, or 4 where 1 represents the least impact on traffic and 4 the most.
- *Year*, *Month*, and *Day* shows the date on which the accident occurred.
- *City*, *County*, *State*, and *Zipcode* show the geographical location in which the accident occurred.
- *Distance.mi.* shows the length of the road affected by the accident in miles.
- *Start_Time* and *End_Time* show when the accident began initially and when the traffic flow returned to normal.
- *Sunrise_Sunset* shows whether the sun was up or down when the accident occurred.
- *Weather_Condition* shows the weather condition (rain, fog, cloudy, snow, etc).
- *Amenity*, *Bump*, *Crossing*, *Give_Way*, *Junction*, *No_Exit*, *Railway*, *Roundabout*, *Station*, *Stop*, *Traffic_Calming*, *Traffic_Signal*, *Turning_Loop* shows True or False values indicating each point of interest's presence nearby where the accident occurred.
- *Temperature.F.* and *Wind_Chill.F.* show the temperature and wind chill in Fahrenheit where the accident occurred.
- *Humidity…* shows the humidity around the accident as a percentage.

- *Pressure.in.* shows the air pressure in inches.
- *Visibility.mi.* shows the visibility in miles.

We also created another variable called *Total_Time_Min*, which represents how long the accident occurred in minutes. In other words, this is the total time from when the accident initially occurred to when the roads returned to normal. The following table represents accidents that occurred throughout the United States:

Example of Accidents

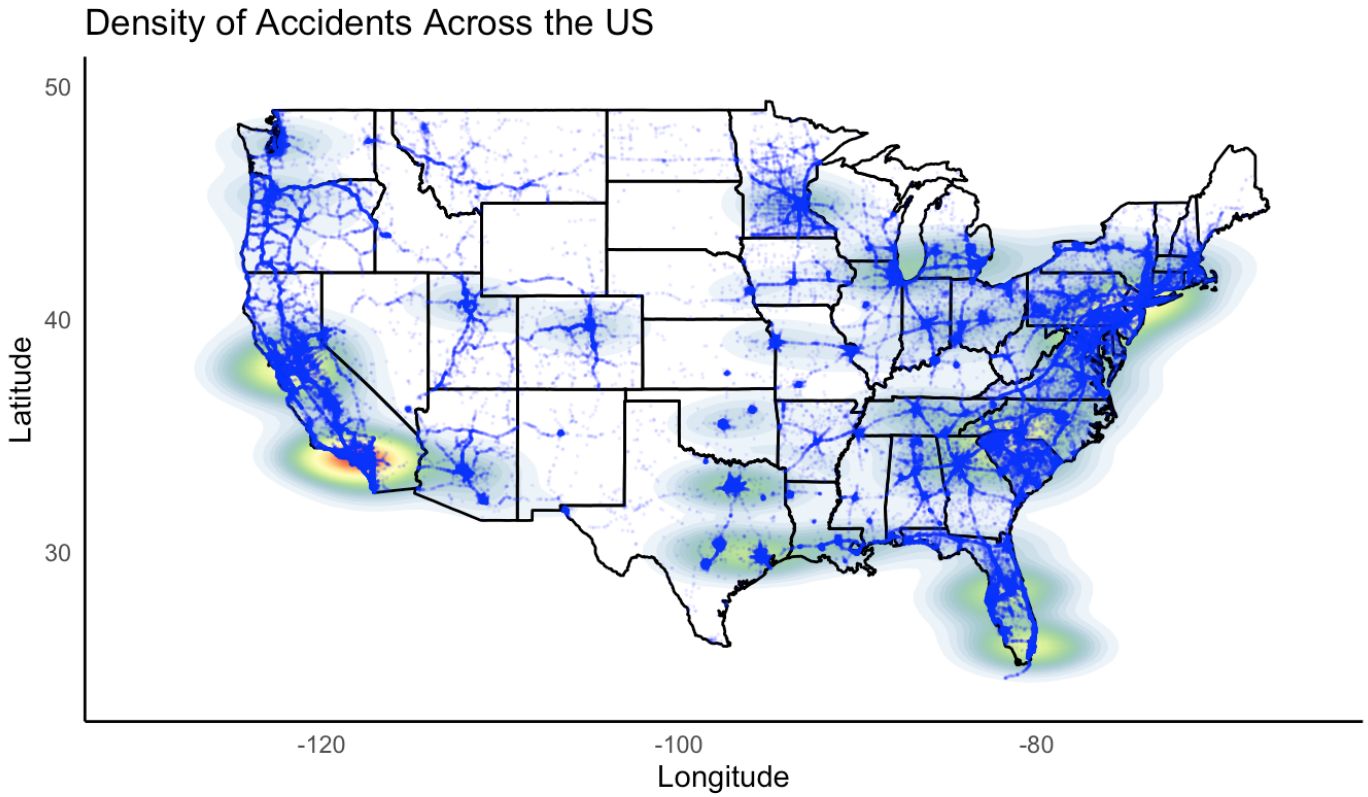
Severity	Year	Month	Day	Distance.mi.	City	County	State	Temperature.F.	Weather_Condition
2	2019	6	12	0.000	Zachary	East Baton Rouge	LA	77	Fair
2	2022	12	3	0.056	Sterling	Loudoun	VA	45	Fair
2	2022	8	20	0.022	Lompoc	Santa Barbara	CA	68	Fair
2	2022	2	21	1.054	Austin	Mower	MN	27	Wintry Mix
2	2020	12	4	0.046	Bakersfield	Kern	CA	42	Fair
2	2021	3	29	0.000	Peabody	Essex	MA	42	Fair
2	2020	1	14	0.000	Gold Hill	Jackson	OR	35	Light Rain
2	2021	8	13	0.047	Panama City	Bay	FL	90	Fair
2	2022	10	12	0.038	Dallas	Dallas	TX	91	Fair
2	2021	10	21	1.301	Indianapolis	Marion	IN	63	Cloudy
2	2021	8	25	2.480	Indianapolis	Hamilton	IN	70	Cloudy
2	2022	2	1	2.091	Saint Regis	Mineral	MT	13	Cloudy
2	2020	7	16	0.000	Huntsville	Madison	AL	85	Fair
2	2022	2	18	2.845	San Diego	San Diego	CA	63	Fair
2	2019	5	28	0.000	Tempe	Maricopa	AZ	64	Fair

RESULTS

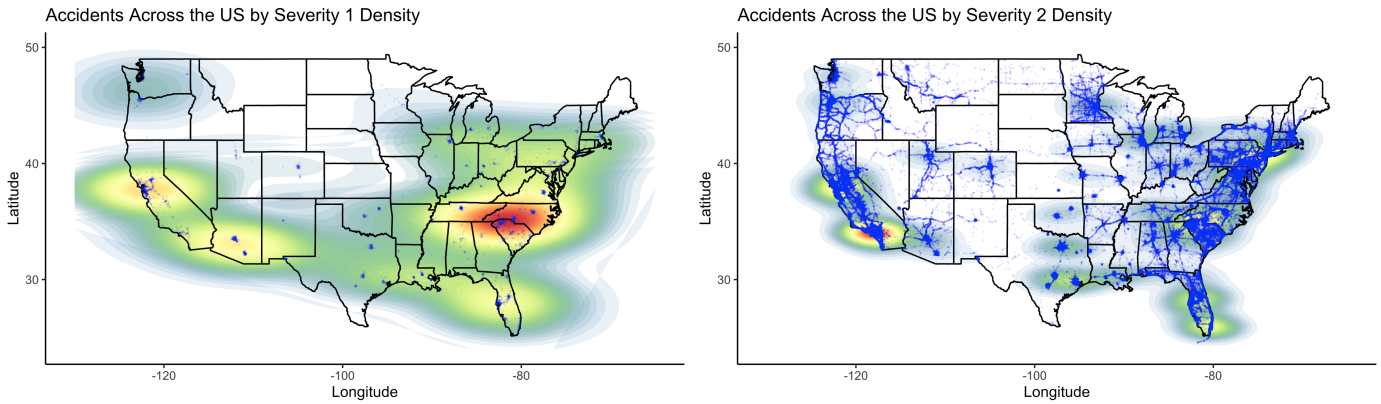
Question 1: Analyzing Density of Accidents Across the US

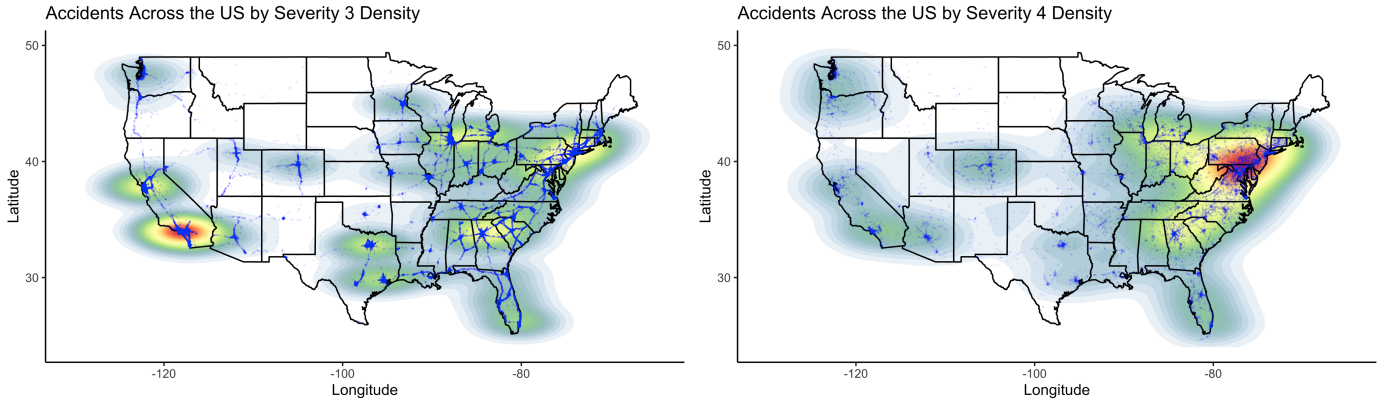
In our initial exploratory data analysis, we examined the geographic distribution of traffic accidents across the United States, with a particular focus on accidents classified as Severity 4 – a metric primarily defined based on the impact on traffic flow rather than the incident itself, as reported by the employed Traffic APIs such as Bing and MapQuest. By leveraging heat maps, we first visualized the overall density of accidents nationwide and,

subsequently, the concentration of accidents by Severity. While accidents in the most congested urban areas, such as Los Angeles, New York City, and Chicago, exhibit the highest density, the heat map also reveals clusters along major transportation corridors and intersections. This pattern suggests that not only do high traffic volumes contribute to the frequency of accidents, but also that the design and flow of traffic in these areas may also be factors. Furthermore, our findings reveal distinct geographic patterns in the distribution of accidents based on Severity, especially when comparing states near Washington, D.C. to California.



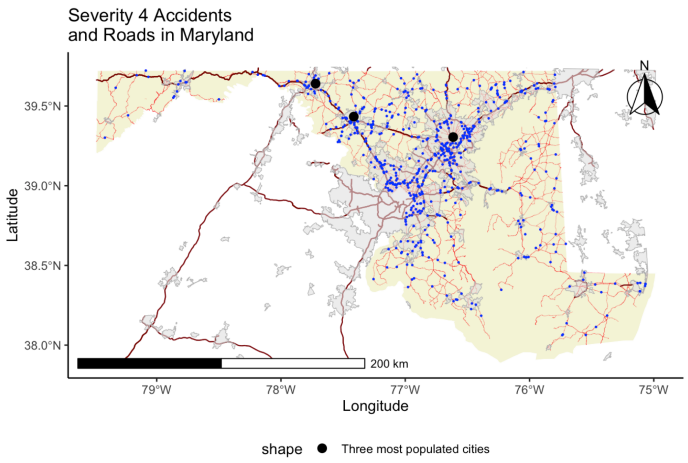
Accidents by Severity

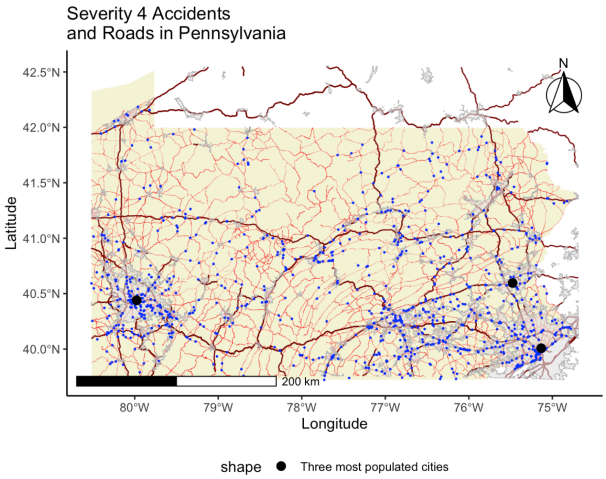


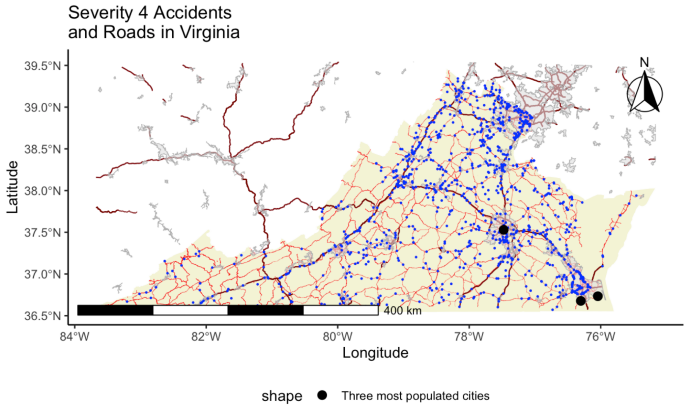


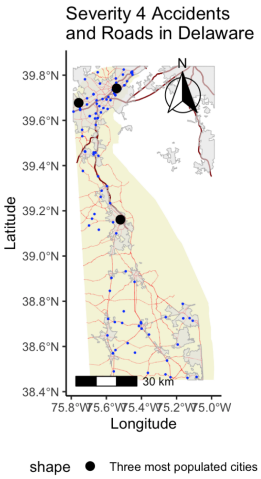
Looking at these plots, we were at first intrigued by how accidents of different severities were distributed. There was a much lower amount of Severity 1 accidents than expected, with the vast majority of accidents having a Severity of either 2 or 3. However, what was most striking was that despite the fact that most accidents within the data set occurred in California, we were surprised to discover with heat maps how the concentration of Severity 2, 3, and 4 accidents shifted from California to the East Coast, specifically towards the area around Washington, D.C. We found that the distribution of Accidents with a Severity of 4 seems to be concentrated near the states of Maryland, Virginia, Pennsylvania, and Delaware. We also found that these accidents were more prevalent in areas away from large concentrations of population. This might be due to Severity 4 accidents occurring more frequently on secondary roads rather than primary highways and interstates. Secondary roads often have lower traffic volumes and speeds, but their design and traffic patterns may contribute to more severe impacts on traffic flow when incidents occur. To further investigate this hypothesis, we decided to analyze the distribution of accidents by road type, focusing on the differences between primary and secondary roads across the highlighted states.

Primary and Secondary Roads

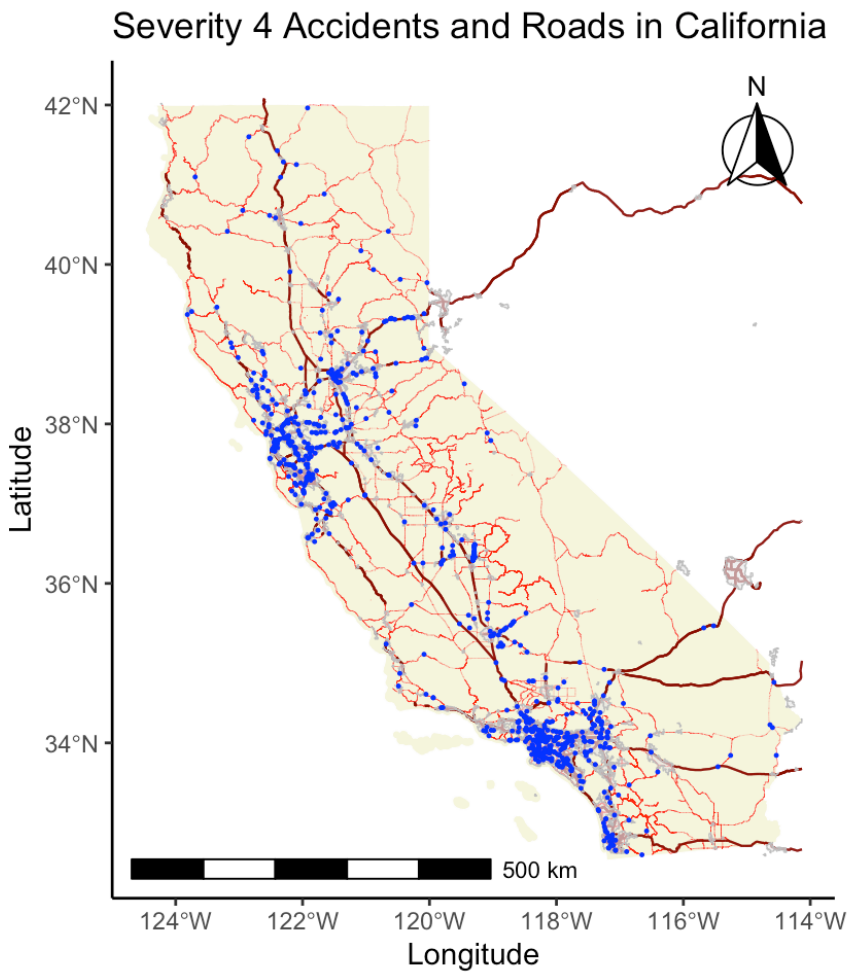








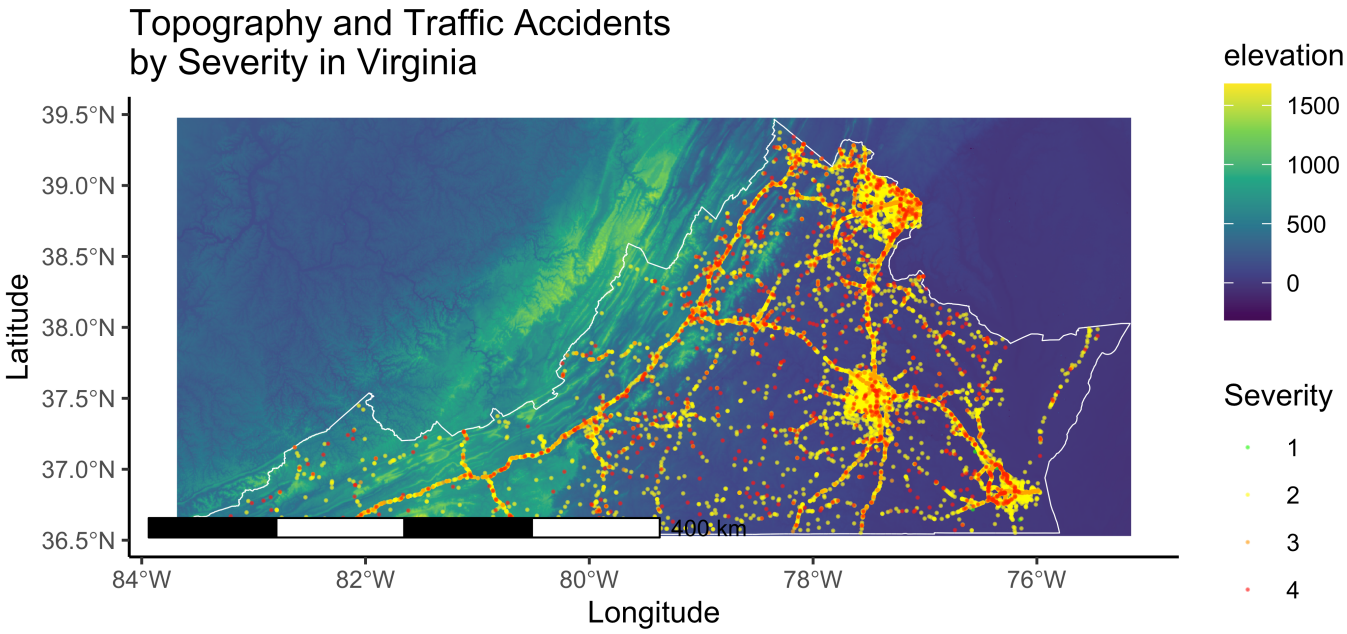
From previous observations, it would seem that while most Severity 4 accidents are concentrated around major urban areas and along primary roads. In this case, major urban areas, which are the areas in the maps highlighted in gray, stand for any area with at least 2,000 housing units or a population of 5,000. The pattern of Severity 4 accidents shifts when we look at the East Coast. In Maryland, Virginia, Pennsylvania, and Delaware, there is a notable spread of Severity 4 accidents across secondary roads. This spread is particularly evident in areas away from large urban populations, suggesting that the risk factors for severe accidents in these states extend into less populated regions. Secondary roads, which are often less equipped to handle emergencies and may lack robust safety features, appear to be significant contributors to the occurrence of high-severity accidents.

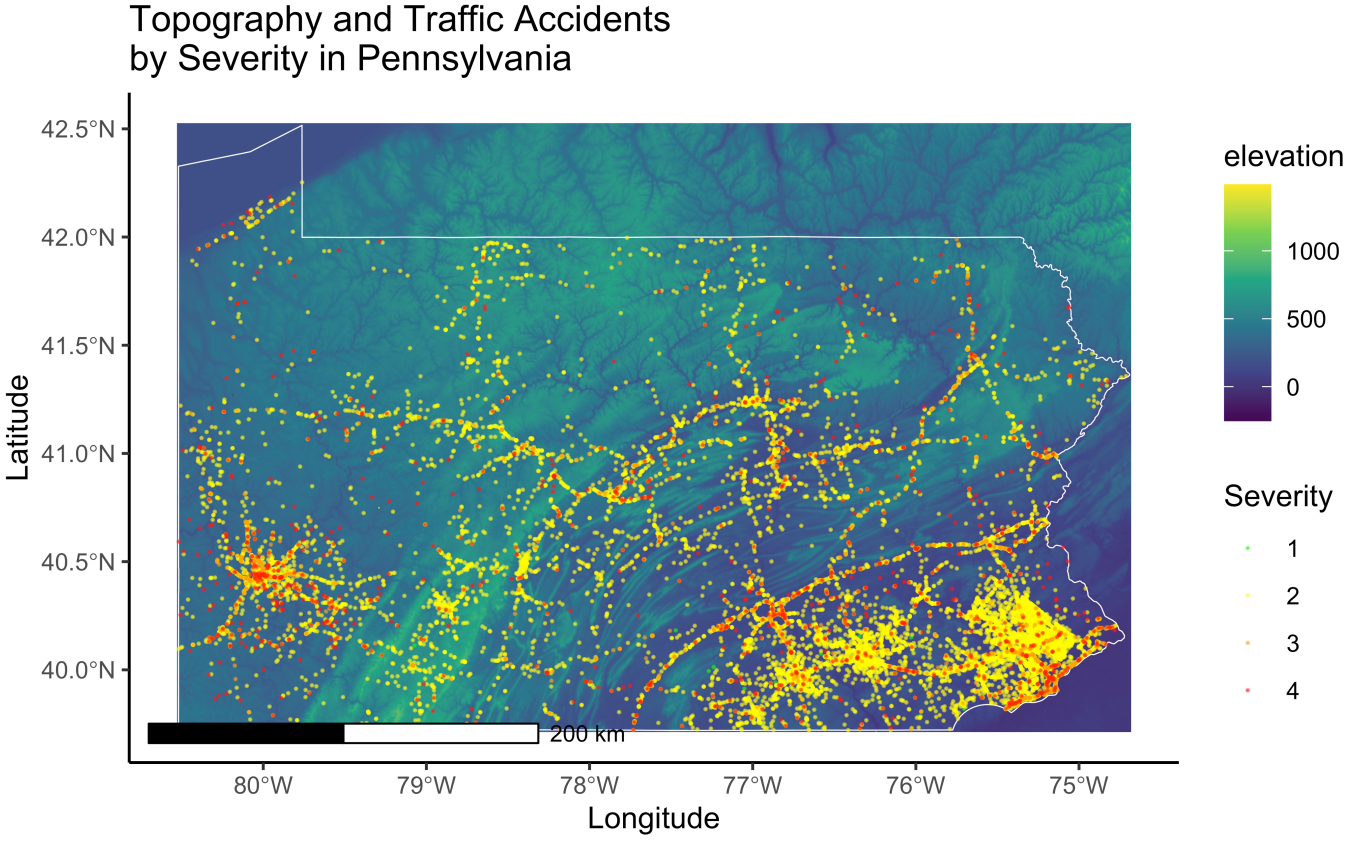


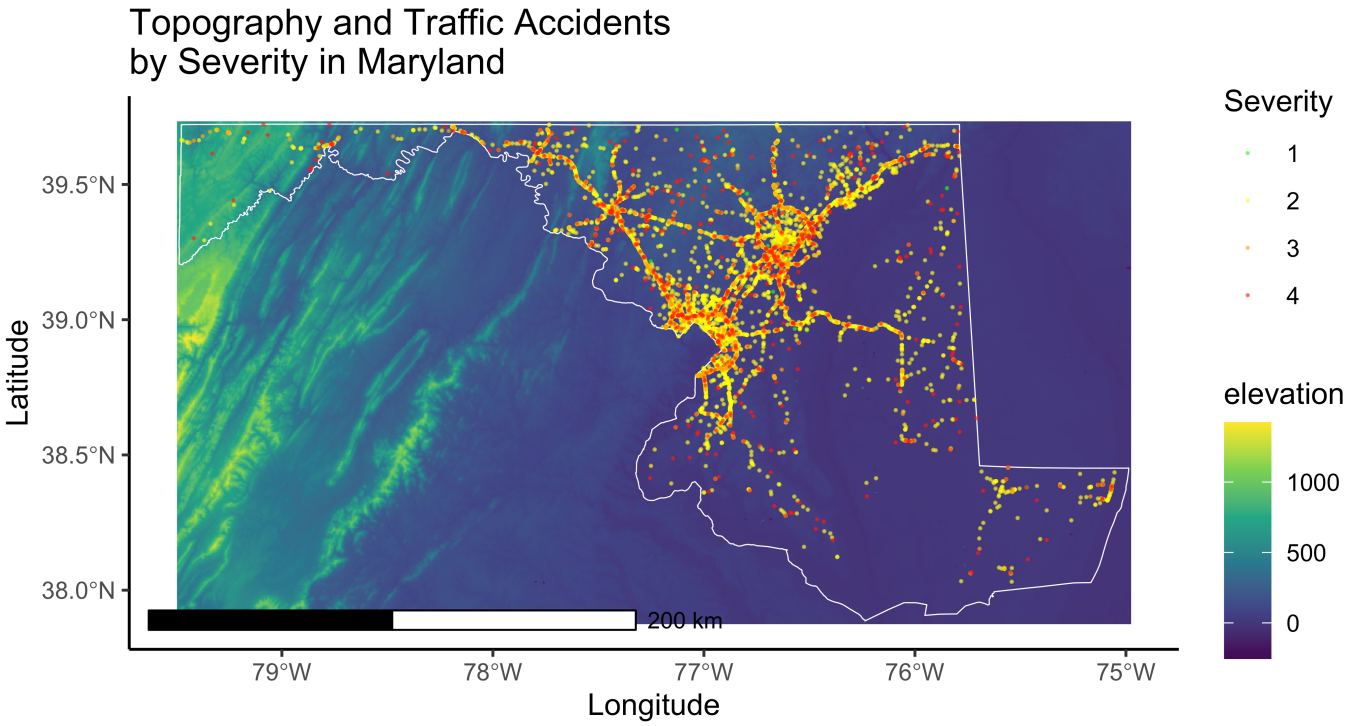
Comparing this to California, where Severity 4 accidents are densely concentrated along primary roadways, particularly in highly urbanized areas, the contrast becomes apparent. This wider dispersion, particularly in Maryland, Virginia, Pennsylvania, and Delaware, is indicative of varied contributing factors beyond traffic volume alone, including perhaps road conditions, maintenance frequency, and less stringent safety measures. One such factor that we want to consider is elevation.

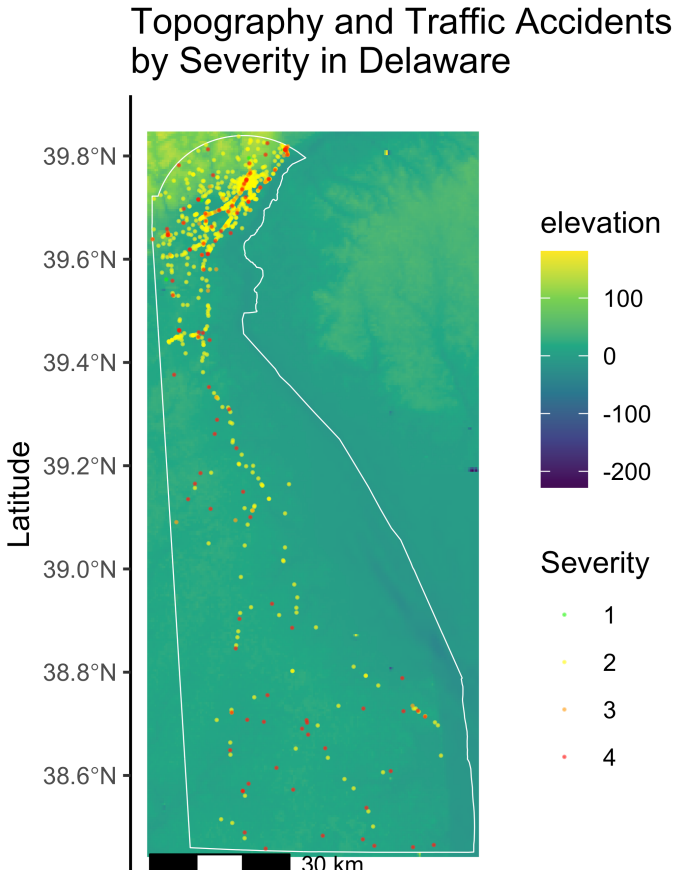
Elevation

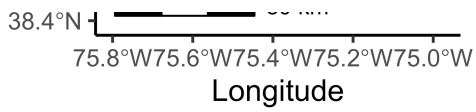
Moving into the dimension of elevation, it becomes essential to consider the topography as a potential contributor to accident severity. For example, Virginia's and Pennsylvania's varying topographies include mountainous regions, which could impact accident rates and severity differently than the relatively flat landscapes of Delaware or the varied terrain of California.


Topography and Traffic Accidents
by Severity in Virginia
39.5°N
39.0°N
38.5°N
38.0°N
37.5°N
37.0°N
36.5°N
Latitude
84°W
82°W
80°W
78°W
76°W
Longitude
400 km
elevation
1500
1000
500
0
Severity
1
2
3
4


Topography and Traffic Accidents
by Severity in Pennsylvania
42.5°N
42.0°N
41.5°N
41.0°N
40.5°N
40.0°N
Latitude
80°W
79°W
78°W
77°W
76°W
75°W
Longitude
200 km
elevation
1000
500
0
Severity
1
2
3
4


Topography and Traffic Accidents
by Severity in Maryland
Latitude
39.5°N
39.0°N
38.5°N
38.0°N
79°W
78°W
77°W
76°W
75°W
Longitude
200 km
Severity
1
2
3
4
elevation
1000
500
0


Topography and Traffic Accidents
by Severity in Delaware
Latitude
39.8°N
39.6°N
39.4°N
39.2°N
39.0°N
38.8°N
38.6°N
30 km
elevation
100
0
-100
-200
Severity
1
2
3
4



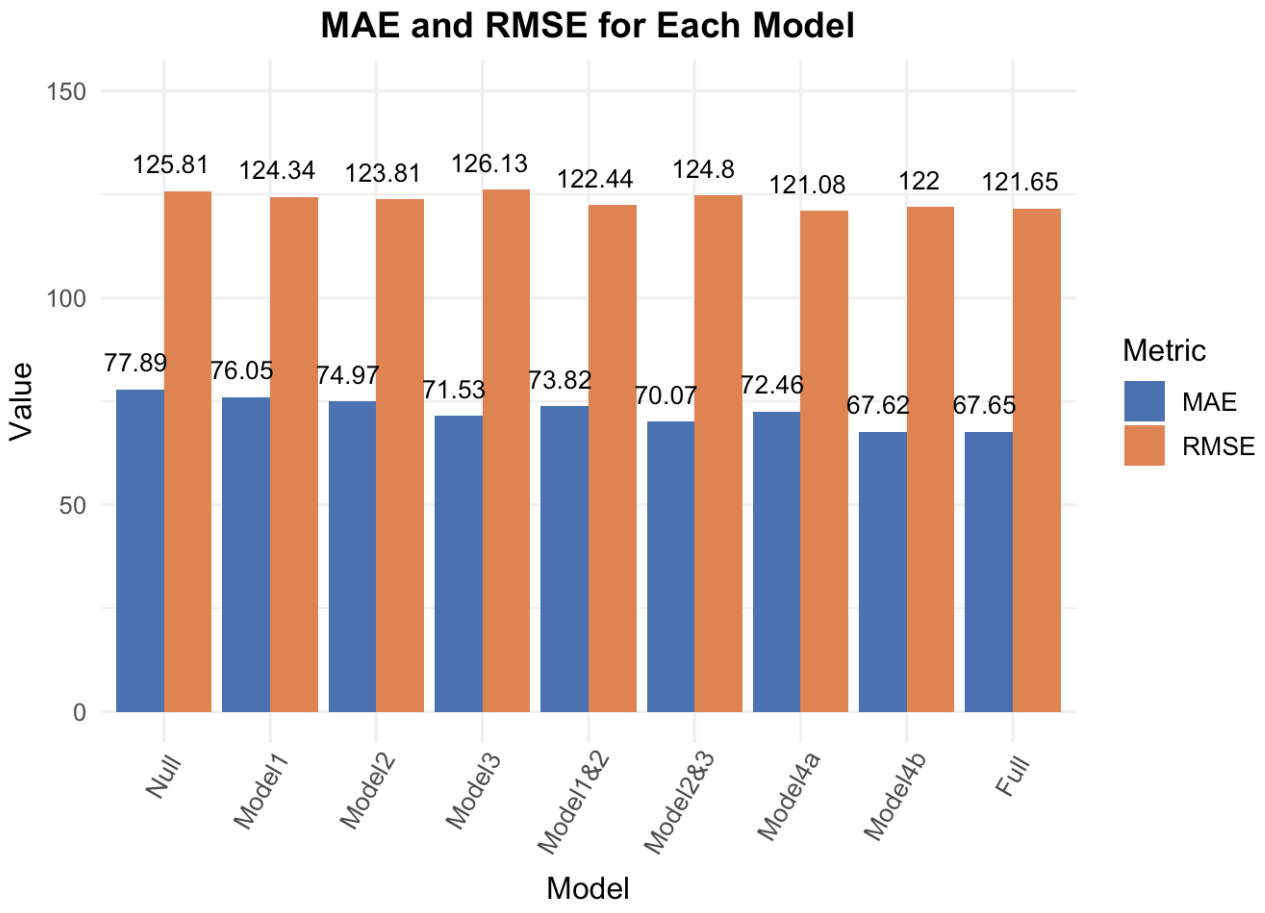
The visualization of traffic accidents by severity levels across the varying topographies of Virginia, Pennsylvania, Maryland, and Delaware presents insights into how geography may influence the impact on traffic flow. While accidents occurred across diverse elevations in Virginia and Pennsylvania, there was no clear correlation between elevation and severity level. The prevalence of high severity accidents impacting traffic flow in both low-lying coastal plains and along the Appalachian Mountains in Virginia indicates elevation alone is likely not a key determinant of severity level. The flat terrains of Delaware and Maryland also saw a spread of accident severity levels, with slightly more high-severity, traffic-disrupting incidents near urban areas despite the overall lower elevations. These preliminary observations suggest that while elevation might impact driving conditions and accident occurrences, other factors such as road design, traffic density, and driver familiarity with the roads are likely playing more substantial roles in the severity of accidents. To better understand severity level predictors, a more comprehensive analysis incorporating road type, traffic data, weather, visibility, and other variables along with elevation could yield insights.

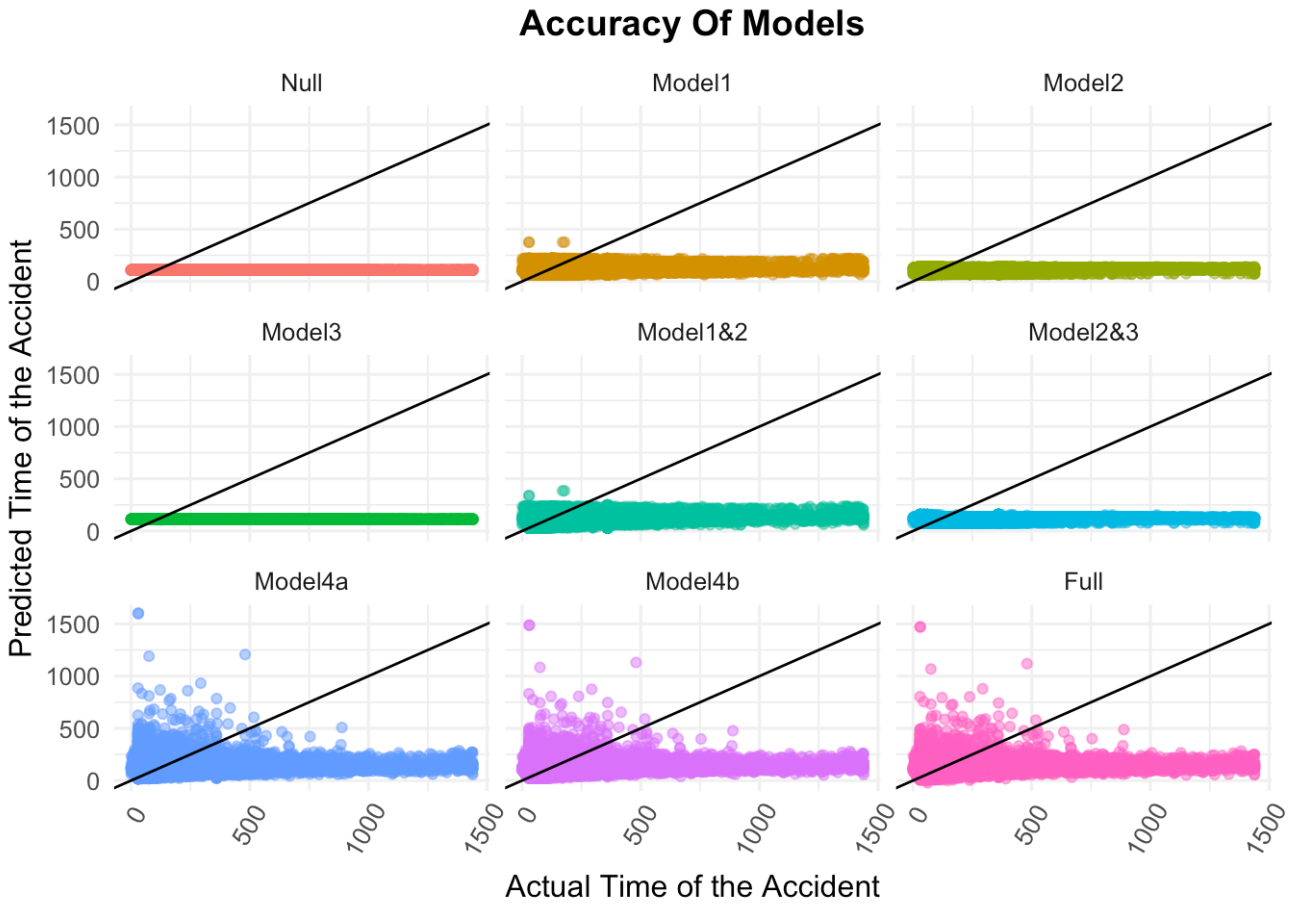
Question 2: Predicting the Total Time of an Accident

We created 9 different models in order to see which variables were the best at predicting the total time of an accident. We randomly split the data up and put half of it in a testing data set and the other half into a training data set. The training data was used to create the models and the testing was used to see how well these models predicted the total time of the accident. We chose the variables for the models based on what we thought would be the best predictors and excluded variables where many of them had a response that only appeared once throughout the data set, such as *Zipcode*. We also excluded *Severity*, *end_time_min*, and *start_time_min* to avoid multicollinearity and *Turning_Loop* as well as *Bump* because these variables only had one response. Below are the models we chose:

- Null: total_time_min = 1
- Model1: total_time_min = State
- Model2: total_time_min = Year
- Model3: total_time_min = Wind_Chill.F.
- Model1&2: total_time_min = State + Year
- Model2&3: total_time_min = Year + Wind_Chill.F.
- Model4a: total_time_min = State + Sunrise_Sunset + Junction+ factor(Year)+ Distance.mi.+ Visibility.mi. + No_Exit + Pressure.in. + Stop
- Model4b: total_time_min = State + Sunrise_Sunset + Junction+ factor(Year)+ Distance.mi.+ Visibility.mi. + No_Exit + Wind_Chill.F. + Pressure.in. + Stop
- Full_Model: total_time_min = factor(Month) + State + Sunrise_Sunset +Junction + factor(Year) + Distance.mi. + Visibility.mi. + No_Exit + Temperature.F. + Crossing + Traffic_Signal + factor(Day) +Timezone + Railway + Wind_Chill.F. + Humidity… + Pressure.in. + Wind_Speed.mph. + Weather_Condition + Amenity + Give_Way + Roundabout + Station + Stop + Traffic_Calming

We looked at MAE and RMSE, as well as the graphed predicted vs actual values, to determine which models predicted the best.


MAE and RMSE for Each Model
Value
150
100
50
0
125.81
124.34
123.81
126.13
122.44
124.8
121.08
122
121.65
77.89
76.05
74.97
71.53
73.82
70.07
72.46
67.62
67.65
Null
Model1
Model2
Model3
Model1&2
Model2&3
Model4a
Model4b
Full
Model
Metric
MAE
RMSE


Accuracy Of Models
Null
Model1
Model2
Model3
Model1&2
Model2&3
Model4a
Model4b
Full
Predicted Time of the Accident
1500
1000
500
0
0
500
1000
1500
Actual Time of the Accident

Model 4a, 4b, and the full model all predict very similarly. The full model is much more complicated than models 4a and 4b even though they all have almost the same MAE and RMSE, which makes models 4a and 4b the better choices.

CONCLUSION

In the first part of our analysis, we focused on understanding the geographic distribution and severity of traffic accidents across the United States, employing heat maps to highlight areas of high accident density and severity. Our findings indicate a significant concentration of accidents in major urban areas like Los Angeles, New York City, and Chicago, as well as along major transportation corridors. Interestingly, while most accidents occurred in California, the pattern of severe (Severity 4) accidents notably shifted towards the East Coast, particularly around Maryland, Virginia, Pennsylvania, and Delaware. These Severity 4 accidents were more prevalent on secondary roads away from large population centers, suggesting that factors such as road design, maintenance frequency, and safety measures might play a crucial role in these high-severity incidents. This pattern points to unique regional characteristics that influence the severity and distribution of traffic accidents.

Moreover, when considering the role of elevation, our findings indicate that while topography varies—ranging from the mountainous terrains of Virginia and Pennsylvania to the flatter landscapes of Delaware and Maryland—there is no clear correlation between elevation and accident severity. The data showed high-severity accidents occurring at various elevations, which suggests that elevation alone does not determine the severity of traffic accidents. Instead, other elements such as road conditions, traffic volume, and possibly less stringent safety measures on secondary roads seem to contribute more significantly to the occurrence of high-severity accidents.

In the second part of our analysis, we were trying to predict the duration of an accident. Knowing how long the specific road will be affected after the accident occurs can greatly improve the traffic flow. Navigation software can redirect traffic onto surrounding roads based on more accurate data. This information could also improve state resource allocations so that they can be more precisely distributed around areas where they are needed the most. After fitting multiple models to our data set, we were unable to generate accurate predictions for the total time of an accident. The reason behind this could be that we did not have sufficient data because the variables available in our data set were not able to be predicted correctly.

An approach for future analysis could be adding more data sets with variables on the types of vehicles involved in the accidents, response times from the local authorities, if alcohol was a factor, and if injuries or deaths occurred to improve predictions. Future analysis could also investigate the connection between specific road types, like interstates, US highways, state highways, country roads, and the length of an accident. With the number of accidents increasing continuously and large population growth, it is especially important to develop models that will aid in traffic management. Newly improved systems could not only make travel quicker but also save many lives by preventing collisions.