Team 6 - Sayuri Jayawardena (suj85), Nicole Wu (nw7688), AJ Lewis (ajl477), Annmarie Chang (ac77233), Adrian Barajas (Ab73559), Thang Truong (ttt2525), Sarah Tu (syt322)

Traffic Fatalities in Austin

1. Introduction

Since the inception of the automobile there has been a need to address safety issues for both the driver and the pedestrian. The earliest Ford automobile's only safety feature was the Triplex Shatter-Proof Glass (1). Many of the basic safety features we recognize on today's vehicles weren't mandatory until the late 1960's (2). While today's vehicles are some of the safest to have ever been manufactured, more than 2.5 million Americans are injured in vehicle collisions each year with over 42 thousand individuals in 2021 dying as a result of a collision (3). While vehicle safety standards are constantly increasing, vehicle deaths per 100 thousand people have stayed constant over the years (2). This stagnancy warrants further observation. Our vehicles are becoming exponentially safer with smart features like adaptive braking and smart cruise control and other safety features like side airbags and crumple zones, yet we haven't seen a dramatic decrease in vehicle deaths since the beginning of the *Click it or Ticket* campaign in the early 2000s (4). Did vehicle safety reach its peak in the 2006 Ford Taurus or is there some other reason for this sustained amount of death? Is there some other factor that predisposes individuals to be involved in a fatal crash?

Our mission is to delve deep into comprehensive datasets gathered by the city of Austin related to traffic collisions and fatalities from the last ten years to shed light on the factors that contribute the greatest to whether a crash is a fatal one. In addition, by harnessing the power of data analysis and predictive modeling, our goal is to develop a model that can predict the likelihood of an individual being involved in a fatal accident within the city of Austin given specific factors. We hope this project will help the city of Austin reduce traffic fatalities as we identify locations, times, days, and any other factors that are predictive of fatality. In analyzing this data, we will also look at other factors that are well-documented as being unsafe. The safest driver is a driver who follows the speed limit, wears their seatbelt, and follows traffic signals. Our data should not be used to conclude that unsafe driving behaviors are somehow as safe or safer than "safe" driving. By sharing our project on the publicly available source that is Github, we hope this project can initiate public discussion and awareness of risk factors that are within an individual's control and others that must be addressed by the government.

2. Data

This project used two different datasets. All datasets were gathered from the Official City of Austin Open Data Portal. First, we combined traffic fatality reports for each year from 2013-2019 into one dataset. This dataset had many interesting variables to investigate for our big question of what factors make an accident more fatal, like speeding, running a red light, driver's license status, location, time of day, etc. A limitation to this dataset was that it only had 531 observations, which is not ideal for training a generalizable classifying method. The other major limitation was that the traffic fatalities dataset only contained information on accidents that resulted in a fatality. Therefore, our predictive model would be missing information for all crashes and factors that did not result in a fatality. A separate dataset was obtained for all car crashes reported by the Austin Police Department from 2013-2023.

Each dataset contained variables for time, date, and x,y coordinates for the location of the crash within the city of Austin, along with crash IDs. This crash report dataset (2013-2023) had 148,417 observations, which was attractive for building a predictive model, but it did not have all the factors we were interested in exploring from the fatalities dataset, such as whether the drivers were speeding or impaired. The crash report dataset had information like the posted speed limit of the area with the crash, location, number and severity of injuries, and the types of vehicles involved. We used both the traffic fatalities dataset and the crash report dataset to investigate the factors that were the most intriguing and to create predictive models on traffic accidents involving fatalities so we could explore what factors were important in an unbiased way.

For the compiled fatality dataset, the data was cleaned by assigning normalized entries. For example, the speeding variable contained entries of yes, y, no, n, X, u, or blank. For the year with Xs and blanks, we assumed the Xs were “yes” entries and blank spaces were “no” entries. This assumption gave a similar distribution of speeding incidences as years with standard yes or no entries. A similar process was used for variables with similar entry styles. Time was standardized to HH:MM:SS and date was standardized to MM/DD/YYYY. When we visualized speeding and running a red light, we decided to not include “uncertain” observations because they didn’t provide useful information and grouping it under “no” would skew our estimations.

For the crash report dataset, we made similar assumptions about missing observations for fatalities (that the common pattern of missing observations was because there had been no fatality) and converted the “crash_date_time” column to a datetime object. We removed observations missing location information from analysis since that was a factor we were interested in, and we had so many observations the distribution would not be greatly affected.

3. Exploratory Analysis

Our ultimate goal is to figure out which factors contribute the most to whether or not a crash is fatal. To solve this problem, we first used the crash report dataset to see if there are any trends in crash frequency over different time scales. We wanted to see if the total number of crashes change with the hour, day, month, and year. We didn’t detect much change over days, with only the 31st day of the month having a lower number of crashes most likely because not all months have a 31st day (Fig 1b). Months also had a uniform distribution of crashes (Fig 1c). Looking over years we did see a decrease in the number of crashes from 2020-2023, likely due to travel restrictions in the pandemic and the increase in work from home (Fig 1d). We did see an interesting distribution of crashes looking over the hour of the day. What we see is that the number of crashes follows a left-skewed normal distribution with the lowest number of crashes at the earliest hours and then a peak in crashes at approximately 5:00 pm. (Fig 1a). Next, we wanted to see the distribution of fatal crashes from the crash report dataset over the hour of the day. There was a period with fewer fatal crashes from 7am-5pm (Fig 2). Interestingly, this distribution of fatal crashes over a day is not proportional to the number of all crashes seen from figure 1a. This made us wonder, what is it about the early and late hours of the day that makes fatal crashes more common? We hypothesize that due to a decrease in traffic, people will drive more recklessly and are more likely to be involved in a fatal accident if they crash.

We wanted to explore if there was any relationship between fatalities and moving violations. Previous studies conducted by the DOT and NHTSA have shown that vehicle speed correlates with an increased chance of a collision being fatal (5). From our visualization using the speeding variable recorded in the traffic fatalities dataset, we do not observe a clear relationship between speeding and fatalities. Vehicular fatalities that involved speeding represented <50% of the total number of fatalities (Fig 3). While speeding may not be a major contributor in our dataset, from this visualization we hypothesize that time of day is a primary contributor to the chance a vehicle collision is fatal. The trend observed from this visualization suggests that between 7:00 am and 5:00 pm, there is an overall reduction in the chance a crash is fatal with the chance of fatality reaching its peak at around 1:00 am (Fig 3).

Since we determined speeding wasn’t a major factor, we next looked at whether running a red light was a major contributing factor to vehicle fatalities. If you drive around Austin for over 30 minutes, you’ll likely see someone run a red light. Running red lights is a serious issue that plagues this city (6). Running a red light is incredibly dangerous for other drivers and pedestrians. With this in mind, we decided to look at how many fatalities over the years involved running red lights (Fig 4a). Again, we found no significant increase in the number of fatalities for individuals who ran a red light. We also found no trend between running a red light and the time of day, though there weren’t many cases of running a red light in the first place (Fig 4b). Based on these graphs, we hypothesize that running a red light is not an important feature of fatal crashes. However, there are likely a couple of reasons this relationship seems so weak. The incidents are likely underreported as an officer or bystander would have to witness the fatality. Additionally, not all roads have red lights like interstates. Furthermore, we don’t have any data on cases when a red light was run but there was no fatal accident. It’s possible the chance of being involved in a fatal crash is higher than average if you run a red light, but we can only see how many people already in a fatal accident ran a red light, and this behavior is likely limited to only the most unsafe drivers. Overall, we reject our initial hypothesis that running a red light would be a common feature in fatal crashes, and fail to reject the null hypothesis that running a red light is a major contributor to fatal crashes.

Next, we decided to look at where these fatal collisions were occurring. Previous studies have shown that there is a strong correlation between driver speed and fatality risk (5). Due to variability in speed limits, we wanted to explore if certain roads are more prone to fatal collisions. To analyze if this relationship might exist, we used latitude and longitude coordinate variables for each fatal collision and created a heat map of traffic fatalities for the City of Austin

(Fig 5). We noticed a large concentration of fatalities along a specific road. We then did a side-by-side comparison of our heat map to a map of Austin with a red line indicating I-35. We hypothesize that the large number of deaths along the highway is due to the high speeds and the large volume of vehicles that travel on I-35. Indeed, I-35 is one of the oldest freeways in Texas, and according to the City of Austin, has exceeded its functional lifespan. Our analysis supports a need to assess the safety of I-35 (7).

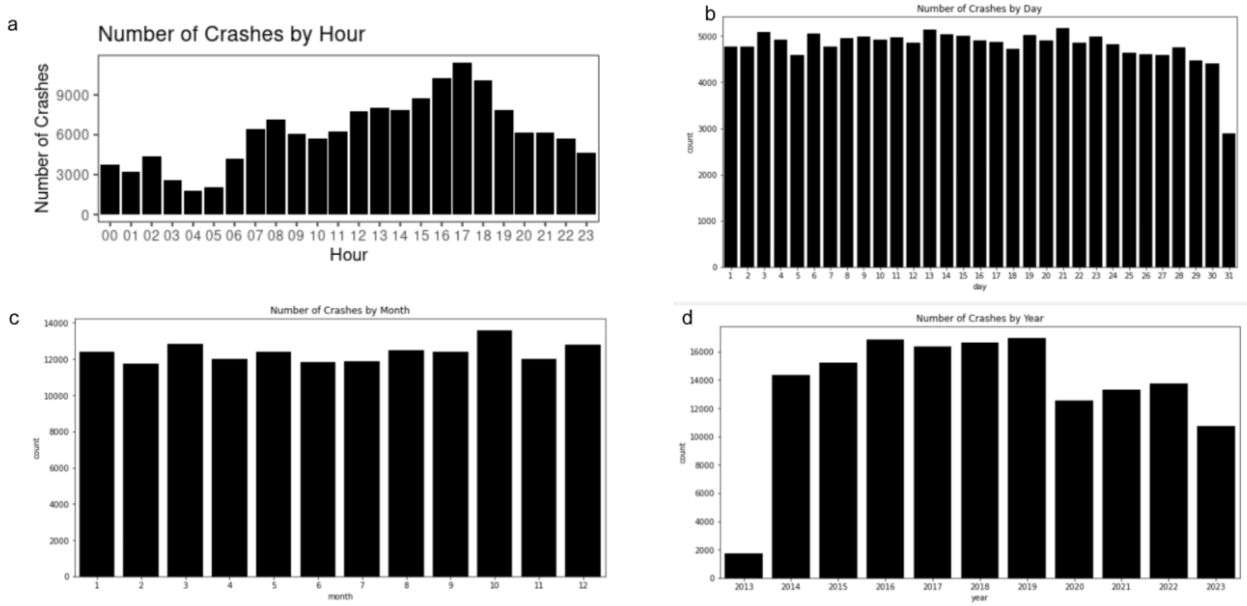


Figure 1: Number of collisions by hour (a), day (b), month (c), and year (d). Crash report data (2013-2023).

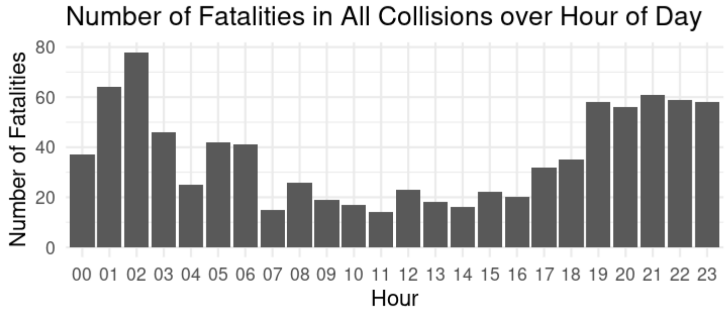


Figure 2: Total number of fatalities over hours in the day. Crash report data (2013-2023).

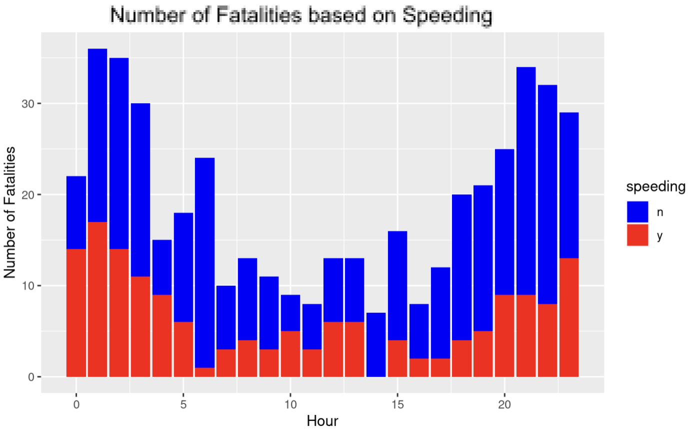


Figure 3: Total number of fatalities ordered by hour of occurrence and speeding status. Traffic fatalities data (2013-2019).

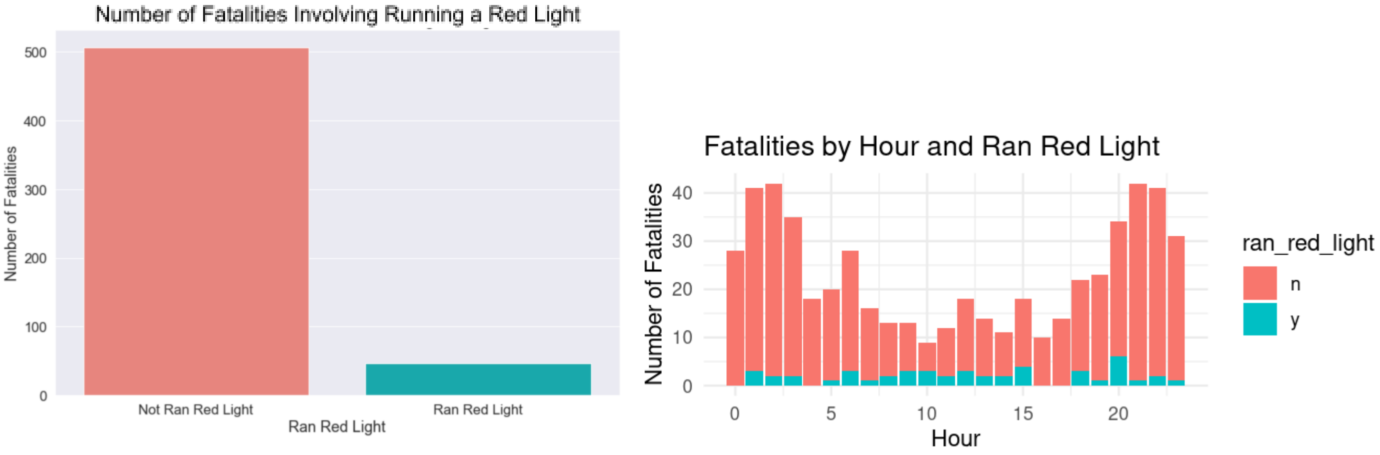


Figure 4: Number of cars that ran red light (a) and number of fatalities in accidents that involved running a red light or not (b). Traffic fatalities data (2013-2019).

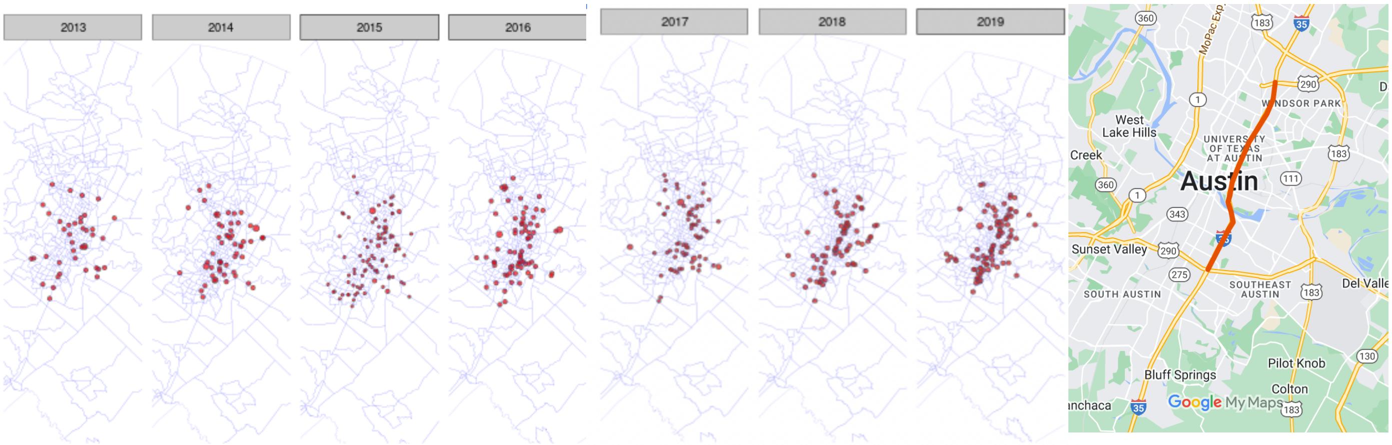


Figure 5: Heatmap of fatalities within the city of Austin. The smallest dots represent 1 death with the largest dots being 4. Traffic fatalities data (2013-2019).

4. Modeling

While the exploratory analysis shed some interesting insights, a machine learning model would be enlightening for investigating the individual factors that predispose individuals to fatal accidents. We used a decision tree model because of its highly interpretable nature behind the transparent decision-making process, which facilitates its use in downstream applications by city council or interested Austin residents. Moreover, we can visualize the structure of the decision tree graphically with important features near the root of the tree. A logistic regression model was also run to give probabilistic predictions when classifying the binary outcome response variable. The model provides interpretable results and is less prone to overfitting. The goal is to predict whether a collision will result in a fatality using the crash report 2013-2023 dataset. Ultimately, predictive analysis employing logistic regression and decision tree models was utilized to estimate the likelihood of fatal crashes in Austin, based on specific variables including time of day, latitude, and longitude. We used location in our decision model because, despite the fact that exact coordinates could be a source of noise, our heat map did not have a smattering of small locations with fatal crashes through the city and instead appeared to reveal trends (Fig 5). These models are designed to discern patterns within the data that correlate these conditions with the probability of a crash being fatal, providing valuable insights for understanding and potentially mitigating crash risks in different areas and times in the city.

Before training the logistic regression and decision tree models, we undertook rigorous data preparation to ensure the integrity and utility of our dataset. This process began with thorough data cleaning, crucially involving the removal of rows lacking vital latitude and longitude information, essential for accurate location-based predictions. Recognizing the importance of handling missing data, we also addressed gaps in other columns, particularly in our target variable crash_fatal_fl (which records whether a person died). Missing values here were imputed with a default 'N', signifying a non-fatal outcome. Moreover, to suit the requirements of our machine learning algorithms, we converted crash_fatal_fl from its categorical form ('Y' for fatal, 'N' for non-fatal) to a binary format, mapping 'Y' to 1 and 'N' to 0. This comprehensive data preparation was pivotal in setting a strong foundation for the effective training of our predictive models. Our classifiers were trained on 80% of the data, randomly selected, and tested on the remaining 20%.

Our logistic regression model proved somewhat capable of distinguishing between fatal and non-fatal crashes with an accuracy of 62%. After visualizing the predicted probabilities (Fig 6), confusion matrices were made to quantify the correct and incorrect classifications for each case. It tended towards over-predicting fatal crashes with many more high false positives (Fig 7). Delving into the decision tree model, the initial classification process starts with evaluating the hour_of_day, making its first split at 6.5 hours, suggesting a distinct pattern in crash outcomes before and after 6:30 AM (Fig 8). We also noticed a spike in fatal crashes before 7am when we explored our datasets with visualizations (Fig 2&3). Further branching based on rpt_latitude reinforces the significance of location in the occurrence of fatal crashes. Despite a more nuanced consideration of time and geographic factors, the decision tree model did not correctly predict any fatal crashes (zero true positives). However, there was more overlap between predicted and actual probabilities with the decision tree (Fig 6) and the decision tree reduced false positives by about 600 compared to the logistic regression model (Fig 7). This reduction is notable given the very few fatal crashes present in the dataset to begin with (Fig 7). Overall, the decision tree boasted an accuracy of 99% in its predictions, yet this figure is heavily influenced by the imbalance in the dataset towards non-fatal crashes.

The confusion matrices for both models highlighted a significant challenge – the difficulty in accurately predicting the minority class, which in this case are the fatal crashes. This suggests a strong class imbalance influence, where the models are biased towards the majority class of non-fatal crashes. As a result, both models, particularly the decision tree, displayed limited effectiveness in identifying the critical instances of fatal crashes, marked by a low number of true positives and high false negatives for the logistic regression model. The predictive models suffer from data imbalance due to a scarcity of fatal crash instances within the dataset, leading to a majority of non-fatal crash data. This skew causes the models to underperform in predicting the critical minority class – fatal crashes – with the decision tree failing to identify any fatalities and the logistic regression generating a high number of false positives. To correct this, resampling strategies such as oversampling the minority class or undersampling the majority class, potentially through methods like SMOTE, are essential. Resampling aims to provide a more balanced dataset for training, which is crucial for enhancing model sensitivity to fatal incidents and ensuring the models' utility in real-world safety applications.

We implemented an additional machine learning model using the traffic fatalities dataset aimed at predicting the impact severity of crashes, distinguishing between high-impact and low-impact scenarios. High-impact was defined as involving more than one person in a fatal crash. A decision tree model was used to uncover what criteria is most influential in predicting if an accident from the traffic fatalities data set would be high-impact. Before model training, we conducted data cleaning and engaged in feature engineering to incorporate meaningful variables into the model. The eleven selected variables were time, hour, day, month, type of road, speeding, running a red light, restraint type, suspected impairment, drivers license status, and failure to stop and render aid.

Logistic regression and decision tree algorithms were again employed for analysis. The outcomes were promising, revealing a 95% accuracy for the decision tree model and a 93% accuracy for logistic regression. Further assessment using the confusion matrix indicated similarities in true negatives between both models. The decision tree slightly outperformed the logistic regression model in correctly identifying high-impact cases, correctly classifying one more case, but both models had very similar results as seen in the confusion matrix (Fig 9). More important than the slightly better performance, the decision tree allows the user to easily understand what factors were the most important in predicting a high-impact fatality. The first three steps used by the decision tree to classify an accident as high or low impact involved time and driver’s license status (Fig 10). We recognized the importance of time before from our visualizations, and driver's license status seems like another factor that would be a priority to study in future works.

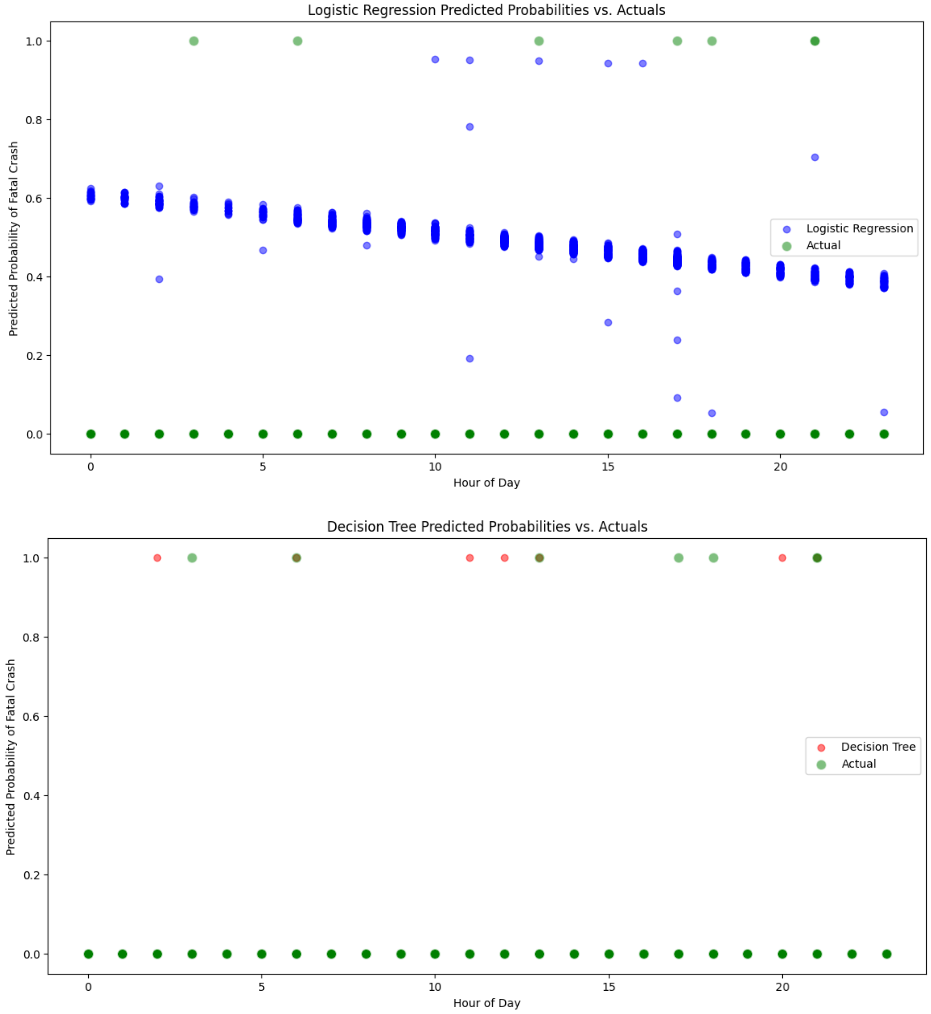


Figure 6. Scatter plots with predicted probabilities for Logistic Regression (above) and Decision Tree models (below) of the crash report 2013-2023 dataset.

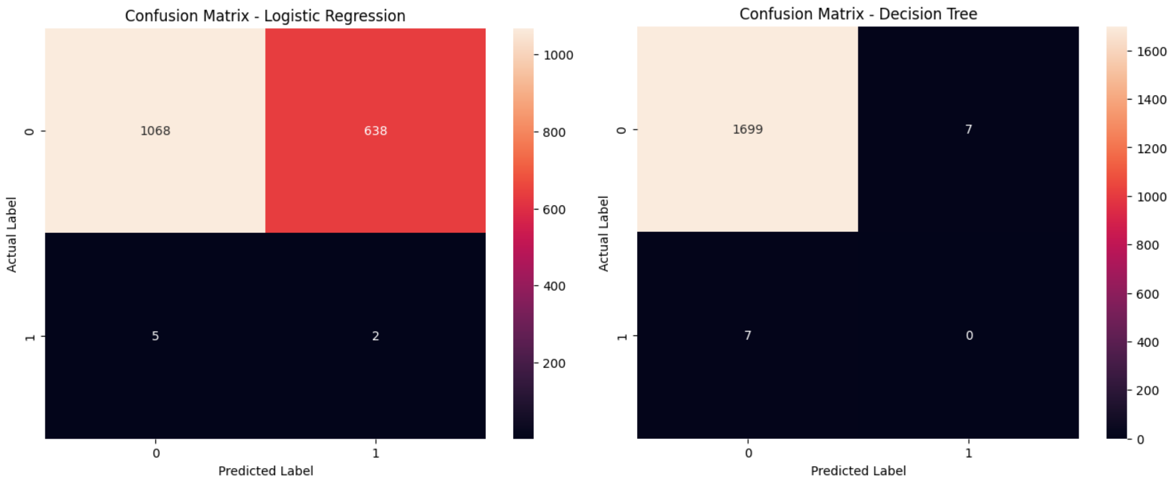


Figure 7. Confusion matrices for both predictive models (logistic regression, left, and decision tree, right). Crash report data (2013-2023).

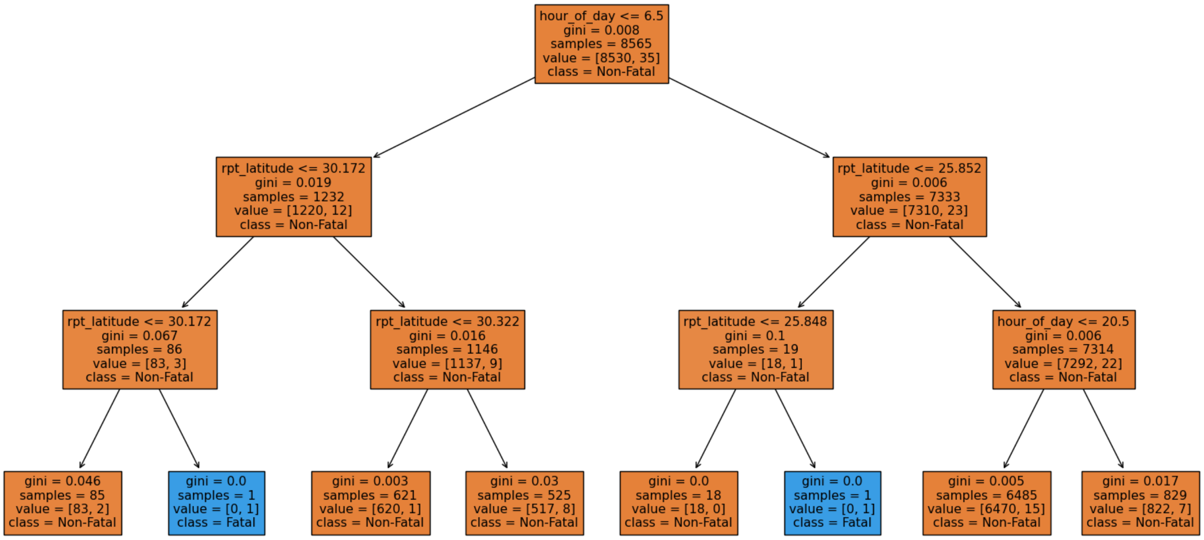


Figure 8: First four levels of the decision tree for predicting fatality. Crash report data (2013-2023).

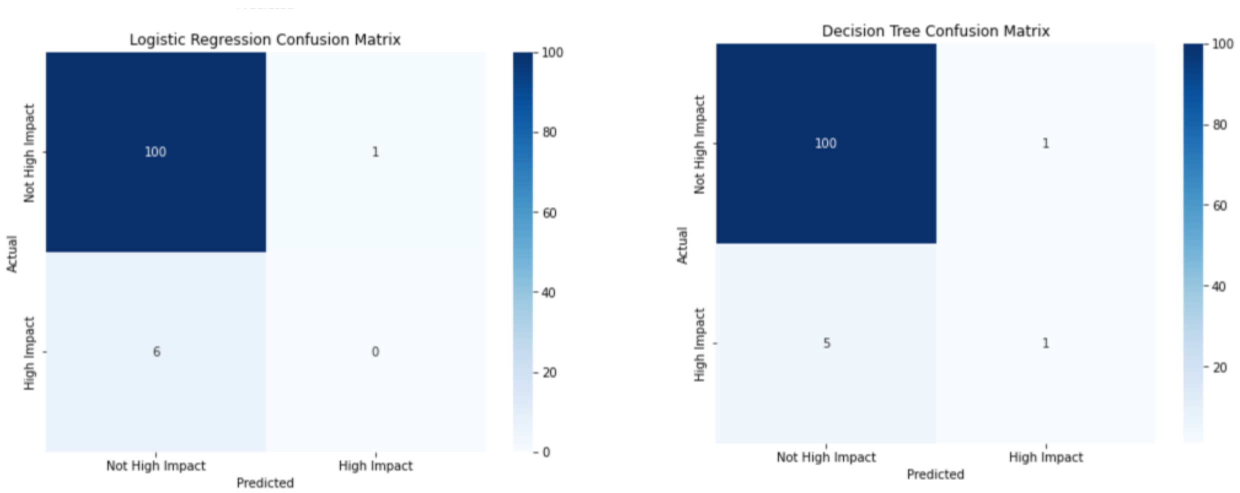


Figure 9: Confusion matrices for both predictive models (logistic regression, left, and decision tree, right). Traffic fatalities data (2013-2019).

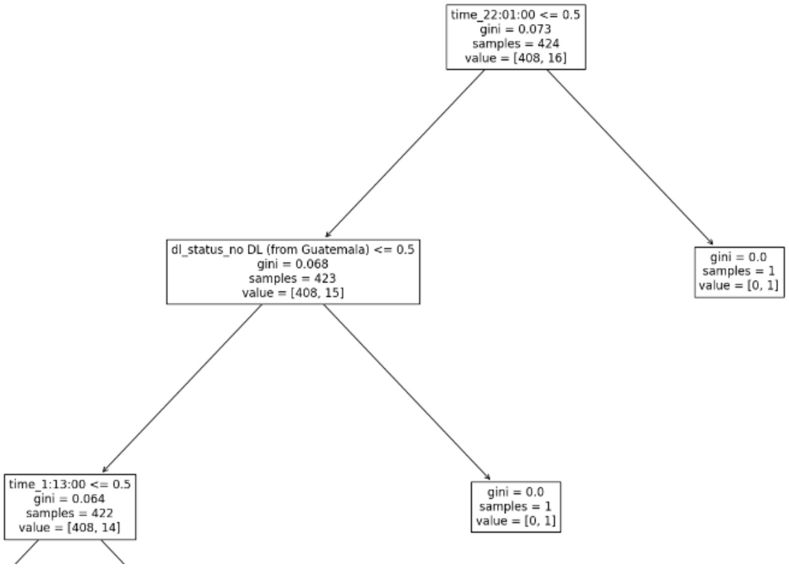


Figure 10: First three steps of the decision tree for predicting high-impact crashes with more than one fatality. First and third steps are time, the second step is driver's license status. Traffic fatalities data (2013-2019)

5. Discussion

Through our predictive models, we found that time of day and whether a person had a valid driver's license were important factors in distinguishing accidents with more than one fatality from accidents with one fatality. Overall, using the decision tree classifier and the crash report dataset, our model could predict whether a crash would have a fatality with 99% accuracy by considering the hour and location in Austin. Using another decision tree classifier with the traffic fatalities dataset, we could predict whether a fatal crash would have more than one death with 95% accuracy. A benefit of decision trees is the readability of the model, as it provides the exact criteria used to make the decision. This level of detail enables decision-makers, such as city planners and traffic safety authorities, to prioritize interventions based on what factors took priority in the decision-making process. City planners can use the knowledge of times that were used to recognize conditions leading to high-impact fatalities when deciding when to open and close parts of the highway and setting speed limits. The police department can choose to increase presence at these times in areas we noticed had a high density of fatal collisions from our heat map, to discourage unsafe driving practices. Additionally, drivers could benefit from knowledge of these factors, allowing them to make informed decisions about when they go out to drive, safety habits, routes and to adjust their driving behavior accordingly to enhance personal safety. This aims to bridge the gap between predictive analytics and actionable insights for stakeholders invested in enhancing road safety in Austin and beyond.

Our exploration of time on a larger scale (days, months, and years) and all crashes found no interesting connections. From our exploratory data analysis we also found that there were more fatal accidents before 8am and after 5pm. Publicizing this finding in Austin could promote safe driving behaviors during these higher-risk periods. We must also consider why we saw this trend. In future studies, we could explore whether this is because of increased intoxicated driving at night, if the areas where these fatal collisions occurred had insufficient street lighting, and finally if people are more likely to speed at these times of day.

Another important finding we had was that there was a trend of a growing density of fatal accidents along I-35. We did not normalize the number of accidents by the volume of cars on the road because that data was not provided in our data set, but doing so would be beneficial to see if it the frequency of accidents is proportional to other roads or if there is an increased risk that comes with driving on I-35. We hope efforts will be made to improve the safety of the many drivers who use I-35 and that understanding how traffic volume plays a big role in accidents could help narrow down what type of action should be taken. For now, visualizations like the heat map we created are powerful ways to send drivers the message to exercise more caution when on big interstate highways.

6. Ethics

Under the AREA-4Ps guidelines, we evaluated several aspects of ethics of our work, including potential conflicts, under-representation, and shaping the future.

We would like to **reflect** upon the **potential conflicts** of our findings. Our data analysis did not show a large correlation between speeding and the number of fatalities as we hypothesized. This may cause people who speed or run a red light to justify their actions to be not a public safety concern. However, we only investigated speeding incidents resulting in deaths and did not include other consequences such as serious or life-threatening injuries, irreversible vehicle damage, etc. Additionally, these incidents are likely underreported due to the need for a police officer to witness the speed the individual was traveling before the collision. Without these considerations, our results may be misleading and misinterpreted at first glance.

We hope to **engage** in recognizing **underrepresentation**. There is a possibility of disproportionate reporting. We noticed that most of the observations recorded in the dataset, 149 observations out of a total of 532 observations, were on or near large roads in Austin, such as I-35. It is possible that larger, busier roads garnered more attention in public transportation reporting and that smaller roads did not get as much spotlight and funding for surveillance. In this case, people who live near smaller roads in more rural areas may have less of a voice regarding public safety in their community. It is also possible that there are inherently more reports on roads like I-35 because there are more people who travel day to day and thus more chance of accidents. Additionally, several variables in our Traffic Fatalities dataset are likely underreported because, as stated earlier, they need a witness to accurately report the variable condition. Finally, the variables we explored could be more robust . We noticed that the peak number of fatalities occurred between 9 PM and 3 AM (Hour 21- Hour 3). With the variables contained in our dataset, we are limited in our analysis to explain this observation. One explanation we thought of, that there may be fewer cars on the road during the hours of peak fatalities, seems to be contradictory. However, we are missing information such as whether seat belts were fastened, or if the roadway had broken, malfunctioning, or a lack of adequate lighting. Even variables we have in our dataset are not always detailed. For example, we found no significant increase in the proportion of crashes involving speeding at the peak fatality times, but we don’t know how high over the speed limit anyone was going. Victims and their families may be faced with inaccurate, unsatisfactory explanations. There may also be underrepresentation in our attempts to call for public safety. Our data does not include demographic info on the individuals involved in the crashes, so we do not know if there may be biases in the reporting of crashes. Given this, it is hard to say whether our findings are truly applicable to everyone in Austin.

We hope to create **action** in the realm of public safety to **shape the future**. Our results highlight key variables that play an important role in road safety, such as hotspot locations and time of day. However, more data is needed to analyze additional factors that can impact likelihood of fatality, such as road infrastructure, possibility of drunk driving, etc. We hope to gain support from the government and APD to provide us with more data in our research. Additionally, we hope our findings will encourage more funding on surveillance and research in public safety. This report is available to the public through Github, and public conversation can be initiated and encouraged to address this issue.

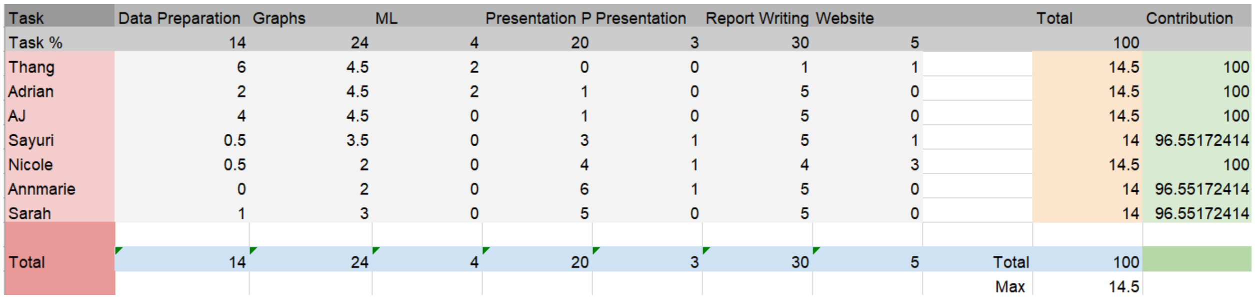
7. Conclusion

As a result of data analysis on information surrounding traffic fatalities in Austin, we have found that several factors remained influential in these outcomes. Namely, factors like the time of day and the validity of one's driver's license played a significant role in distinguishing between different instances of traffic fatalities. Surprisingly, factors like speeding and running a red light did not appear to be associated with times high in fatal crashes. However, further investigation is warranted to explore the relationships between these factors and potentially uncover nuanced patterns specific to the Austin region. Expanding on this, a deeper examination of socio-economic factors, such as income levels and access to public transportation, could provide valuable insights into disparities in traffic safety outcomes within the Austin community. Future research should also delve into the effectiveness of existing safety interventions and assess the impact of emerging technologies, such as autonomous vehicles and current traffic management systems, on overall road safety. In addition, the incorporation of traffic volume and activity would give further insight into determining the danger of certain roads and areas. These insights could contribute to the formulation of evidence-based policies and regulations aimed at reducing traffic fatalities in Austin as well as other similar urban environments.

8. Acknowledgment

We would like to express gratitude to the Austin Police Department for making the dataset available to use with detailed descriptions of each variable, enabling students and researchers to harness the power of predictive machine learning for academic purposes. Thank you to Alexander Marden from the UT Geography Department for supplying the shapefile for the city of Austin. Finally, we would like to give thanks to Dr. Arya Farahi for supporting us, listening to our concerns, and answering questions for how to best complete this project.

The table that lists the member's contributions is shown here:



Task	Data Preparation	Graphs	ML	Presentation P	Presentation	Report Writing	Website		Total	Contribution
Task %	14	24	4	20	3	30	5		100	
Thang	6	4.5	2	0	0	1	1		14.5	100
Adrian	2	4.5	2	1	0	5	0		14.5	100
AJ	4	4.5	0	1	0	5	0		14.5	100
Sayuri	0.5	3.5	0	3	1	5	1		14	96.55172414
Nicole	0.5	2	0	4	1	4	3		14.5	100
Annmarie	0	2	0	6	1	5	0		14	96.55172414
Sarah	1	3	0	5	0	5	0		14	96.55172414
Total	14	24	4	20	3	30	5	Total	100	
								Max	14.5	

Sources:

(1)

Appointments of the New Ford. Directory index: Ford/1930_ford/1930_ford_brochure_02. (1930). https://www.oldcarbrochures.com/static/NA/Ford/1930_Ford/1930_Ford_Brochure_02/1930%20Ford-14-15.html

(2)

Global Road Safety. (2023, January 10). Centers for Disease Control and Prevention. Retrieved December 4, 2023, from https://www.cdc.gov/injury/features/global-road-safety/index.html#:~:text=For%20example%2C%20the%20U.S.%20had,increased%20in%202020%20and%202021.

(3)

NHTSA. (2022). Overview of Motor Vehicle Traffic Crashes in 2021. Austin, TX; DOT.

(4)

NHTSA (2010). Analyzing the First Years Of the Ticket or Click It Mobilizations. D.C.; DOT

(5)

SWOV. (2012, April). *The relation between speed and crashes*. Institute for Road Safety Research. Retrieved December 4, 2023, from https://safety.fhwa.dot.gov/speedmgt/ref_mats/fhwasa1304/Resources3/08%20-%20The%20Relation%20Between%20Speed%20and%20Crashes.pdf

(6)

Philip Jankowski. (2019, June 4). It's official: Austin's red light cameras shuttered. *Austin American-Statesman*. https://www.statesman.com/story/news/local/flash-briefing/2019/06/04/its-official-austins-red-light-cameras-turned-off/4988719007/

(7)

City of Austin & I-35 Capital Express Projects. AustinTexas.gov. (2023, November 20). https://www.austintexas.gov/I-35Mobility