Analysis of US Accidents

CSPB 4502 Data Mining

Matthew Scott
Group 12
University of Colorado-Boulder
Boulder, Colorado, USA
masc6977@colorado.edu

Daniel Simonson
Group 12
University of Colorado-Boulder
Boulder, Colorado, USA
daniel.simonson@colorado.edu

Paul Schneider
Group 12
University of Colorado-Boulder
Boulder, Colorado, USA
paul.schneider-1@colorado.edu

Kevin Qian
Group 12
University of Colorado-Boulder
Boulder, Colorado, USA
kevin.qian@colorado.edu

ABSTRACT

Accidents pose a serious concern for all motorists on the road. They can lead to property damage, severe injury, supply chain delays, and traffic congestion. The goal of our analysis is to find patterns and correlations between accident occurrences and external factors that may influence a higher rate of accidents in the US. By identifying supporting data for increased accidents we hope to inform motorists and the DMV of measures that could be taken both individually and as a community as a whole to reduce the number of accidents, thereby saving everyone's time, money, and lives. The results of our analysis suggest that time of day/year and high-density areas are the largest contributing factors to the occurrence of car accidents. While hazardous driving conditions are not the primary cause of most crashes, they do play a major role in predicting the severity of an accident. These results are displayed through pattern visualizations, heat maps, bar charts, and line graphs.

INTRODUCTION

We intend to explore and answer questions about how weather, time of day, geolocation, and other external factors lead to an increase in accidents in the US. Our particular interest is finding a correlation between why an accident might be more likely to occur near a certain location vs another, during extreme weather instances vs not, or during daytime rush hour. Particularly, our questions seek to understand,

- What driving conditions are associated with the majority of accidents?
- What driving conditions cause the most severe accidents?"
- How does traffic accident volume change over time?
- What locations have the most traffic accidents?

Answering these questions is important for making informed policy decisions on how to lower the frequency of traffic accidents across the country. By discovering frequency patterns and persistent external factors, policymakers can make changes to our transportation network that correct these issues. We predict our findings will support recommendations for upgrades to shared transportation methods, focusing on accessibility, affordability, and efficiency.

LITERATURE SURVEY

L. S. Boyagoda and L. S. Nawarathna, "Analysis and Prediction of Severity of United States Countrywide Car Accidents Based on Machine Learning Techniques," 2022 7th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2022,

pp. 1-5, doi: 10.1109/ICITR57877.2022.9993371.

keywords: {Road accidents; Machine learning algorithms; Biological system modeling; Computational modeling; Urban planning; Predictive models; Road safety; classification; decision tree; k-nearest neighbor; random forest},

RELATED WORK

The data covers 49 states in the United States from 2016 to 2020 and consists of 1.5 million accident records along with 45 attributes. Boyagoda and Nawarathna used classification-based methods discussed in class (Decision Tree, K-Nearest Neighbor Algorithm, Random Forest, and Model Validation) to predict which variable is the greatest indicator of severity. They found that using a random forest classification technique resulted in the highest accuracy for test and training sets, and could best extract the essential features used to predict the severity level of accidents. The results from their study overwhelmingly suggested that the distance variable was the most reliable means of predicting the severity level of an accident.

While our analysis uses similar attributes to the literature surveys (time, precipitation, distance, etc), we differentiate on the type of questions we seek to answer. Our goal is not to predict suitable indicators for the level of accident severity but to find interesting correlations between accident frequencies and variables that may have broader implications for dangerous traffic patterns. Results from this analysis may produce policy recommendations that could help prevent future accidents. Additionally, we greatly expand the previous work on US Traffic Accidents by using exact geographical locations in our analysis. Our goal is not only to achieve a higher level of understanding of the data through spatial depiction but also to draw additional hypotheses that suggest common themes among drivers in certain areas. Furthermore, if time permits, depicting our results geographically will add a layer of complexity and greater interpretability to our final evaluation of the data.

DATA SET

The data set,

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data,

is an accident data set that covers 49 states. The data was collected between February 2016 and March 2023. Data collection was done using several traffic APIs that provide streaming traffic data. The APIs broadcast traffic data collected from various sources including but not limited to, US and state departments of transportation, traffic cameras, law enforcement agencies, and traffic sensors within road networks. This data set contains 7.7 million accident records. The data contains several attributes that will be detailed below:

- ID : Nominal - A unique identifier for each accident record
- Source : Nominal - The origin of the accident data
- Severity : Ordinal - Ranked from 1-4 where 1 indicates least impact on traffic
- Start_Time : Interval - The date and time when the accident occurred
- End_Time : Interval - The date and time when the accident's impact ended
- Start_Lat : Ratio - The latitude coordinate of where the accident occurred
- End_Lat : Ratio - The latitude coordinate of where the accident ended
- Start_Lng : Ratio - The longitude coordinate of where the accident started
- End_Lng : Ratio - The longitude coordinate of where the accident ended
- Distance(mi) : Ratio - The length of the road extent affected by accident in miles
- Description : Nominal - A description of the accident

- Street : Nominal - The street where the accident occurred
- City : Nominal - The city where the accident occurred
- County : Nominal - The county where the accident occurred
- State : Nominal - The state where the accident occurred, Including Washington DC
- ZipCode : Nominal - The postal code where the accident occurred
- Timezone : Nominal - The time zone where the accident occurred
- Weather_Timestamp : Interval - The date and time of the weather observation
- Temperature : Ratio - The temperature at the time of the accident
- Wind Chill : Ratio - The wind chill in Fahrenheit
- Humidity : Ratio - The humidity in percentage
- Pressure : Ratio - The air pressure in inches
- Visibility : Ratio - The visibility in miles
- Wind Direction : Nominal - The wind direction
- Wind Speed : Ratio - The wind speed in miles per hour
- Precipitation : Ratio - The precipitation in inches if any
- Weather_Condition : Nominal - Shows the weather condition (rain, snow, thunder, fog, etc.)
- Amenity, Bump, Crossing, Give_way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_loop : Binary - Shows if the feature is present nearby
- Sunrise_Sunset : Nominal - Indicates whether the accident occurred in the day or night
- Civil_Twilight : Nominal - Shows the period of day based on civil twilight
- Nautical_Twilight : Nominal - Shows the period of day based on nautical twilight
- Astronomical Twilight : Nominal - Shows the period of day based on astronomical twilight

There are a total of 46 attributes in the data set, with 7728394 objects.

PROPOSED WORK

Overall the data set is incredibly clean as is, only two of the data points have any Null values. Those columns would be the Latitude and Longitude for the location of the accidents. Upon reviewing the sources involved in collecting these data points as well as the nature of the accident. Typical police reports do utilize location services to automatically record accident Longitude and Latitude when law enforcement is on site, this however does not get utilized with accidents such as “Hit and Runs” or less serious accidents. Rather than removing or excluding these data sets, the goal is to utilize average longitude and latitude data per zip code to fill out these null points. This does lead to an issue with inconsistencies within the zip code data set, most rows have the common 5-digit zip code. However, some do include the “+4” values which will have to be removed to allow for latitude and longitude data to be added. Having latitude and longitude data will allow for location analysis to be paired with time of day and weather data to dive deeper into trends related to the causes of the accidents. At this time there is no plan to remove any columns from the data set. Each works with the others to help paint a deeper picture of the situation that caused the accident. At this time there is no plan to utilize the three different “Twilight” sections and one will most likely be chosen over the others.

The current questions around this data set revolve mostly around location distribution paired with timing. Previous work has been done on the surface-level distribution simply based on the number of

accidents, the plan for this is to pair that with the time and weather information. The first question being evaluated would be; “Do we see a higher rate of accidents occurring during times with higher precipitation?” From there location and time can be taken into consideration to evaluate if rain or snow impacts certain areas more or if the combination of precipitation and low visibility causes an increase in accident frequency. The other use of time would be to evaluate the relationship between rush hour traffic and accident count. Is there an uptick in accidents between the hours of 7:00-9:00 AM and 4:00-6:00 PM? Finally, depending on findings with location trends, a secondary data set correlating public transportation use by zip code will be used to see if this impacts accident counts.

MAIN TECHNIQUES APPLIED

Data Cleaning

Utilizing the Pandas module in python, we removed any invalid, missing or not applicable data from the dataset. This process removed several million points of data that were incomplete or more importantly not relevant to our assessment. This way we were able to retain only the attribute features we planned to use. This involved us dropping several of the features present in the original dataset. This included dropping all features pertaining to time of day except sunset sunrise, which was what was ultimately used to determine time. We also removed all features regarding road objects such as roundabouts, crossings, bumps etc. Using this method of cleaning, we were able to heavily reduce the overall size of our dataset and focus on the attribute features we were interested in.

Data Preprocessing

In order to perform our frequency analysis, we needed to perform some data preprocessing. This involved utilizing a script to convert several weather-related attributes from quantitative data to categorical data. These attributes were Windspeed(mph), Temperature(F), Windchill(F), Humidity(%), Pressure(in), Visibility(mi), Precipitation(in), Sunrise vs Sunset. This required defining the numerical ranges that each category could fall into. For example, all values less than 12 miles per hour in the ‘Windspeed’ attribute were categorized as ‘Light _Breeze’. Values between 12 and 18 miles per hour were labeled ‘Moderate Breeze’, and any greater values fell under the ‘Strong Breeze’ category. This transformation of the data was necessary to provide the Apriori Algorithm with the required parameters to complete a compressive frequency analysis.

EVALUATION METHODS

In order to evaluate our results we plan to present a collection of visualizations that shape and inform our correlation analysis. We will apply the Apriori algorithm to conduct a pattern analysis that finds the most similar attribute values present in an accident. Through this method, we will identify and use these common attributes to conduct a Bayesian Classification to determine the probability of the most similar attribute values.

We will then conduct a reanalysis of our predictions and questions asked at the beginning of our research to determine if there are unforeseen or unexpected results that bear a reason to revisit and produce our final determinations based on the evidence discovered.

Finally, after presenting our evidence as potential answers to the questions we put forward, we will offer a proposal of solutions that fit the evidence and may help to reduce the likelihood of accidents in the future. Potential solutions that require evidence at this time are: implementation of additional public transportation, the subsidization of public transportation during certain weather/times of day at certain locales, and work from home office policies.

TOOLS

- Python
  - Pandas
  - Numpy
  - MatplotLib
- Excel
- Github

- Zoom
- Text Message
- Google Drive
- Tableau or Pow

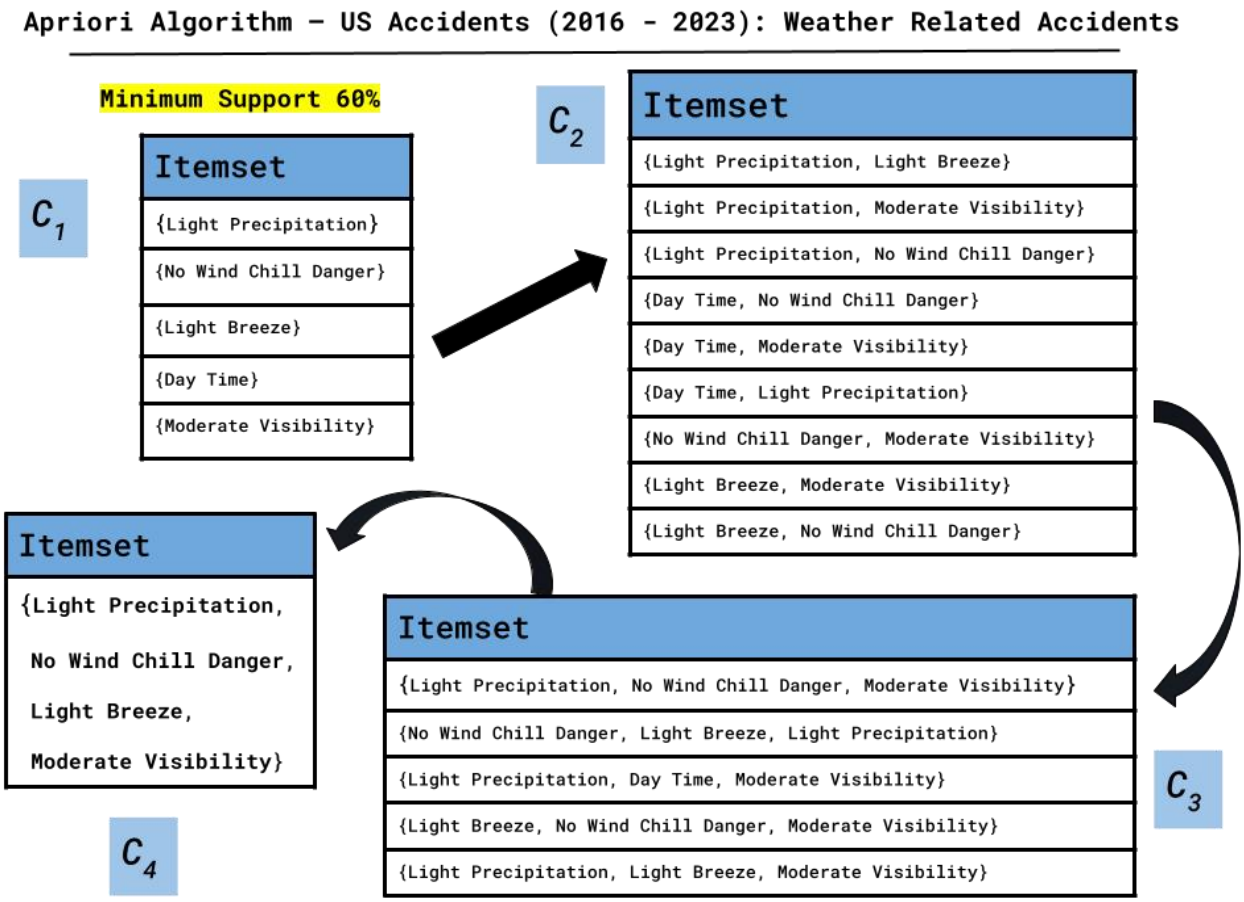
MILESTONES

Week	Item Completed
March 4 - 10	-Part 2: Proposal Paper **Draft**
March 11 - 17	-Part 2: Proposal Paper Final Draft **Completed** (due 03/18) -Clean data and integrate zipcode data with latitude and longitude (removing NULL issue) -Questions finalized
March 18 - 19	-Data scraped for our questions -Visualization decisions made -Heat Map of Accident locations -Flowchart of precipitation vs accident
March 25 - 31	-Break -Team reviews current visualization **Drafts**
April 1 - 7	-Part 3: Progress Report **Draft**
April 8 - 14	-Part 3: Progress Report Final **Completed** (due 04/22) -Visualizations **Completed** -Part 5: Project Code & Descriptions **Draft**
April 15 - 21	-Part 4: Project Final Report **Draft** -Part 5: Project Code and Descriptions **Complete** -Part 6: Project Presentation **Draft**
April 22 - 28	-Part 4: Project Final Report **Complete** -Part 6: Project Presentation **Complete** -Part 7: Peer Evaluation **Completed**
April 29 - May 5	-Part 4-7 due May 2nd -4: Project Final Report -5: Project Code and Descriptions -6: Project Presentation -7: Peer Evaluation

KEY RESULTS

Frequency Analysis

Attributes associated with changes in hazardous driving conditions were run through the Apriori Algorithm to reveal correlations between natural effects and accident frequencies. The attribute chosen for this analysis (Windspeed(mph), Temperature(F), Windchill(F), Humidity(%), Pressure(in), Visibility(mi), Precipitation(in), Sunrise vs Sunset) had to be manually converted from quantitative to categorical data. This required defining the numerical ranges that each category could fall into. For example, all values less than 12 miles per hour in the 'Windspeed' attribute were categorized as 'Light _Breeze'. Values between 12 and 18 miles per hour were labeled 'Moderate Breeze', and any greater values fell under the 'Strong Breeze' category. This transformation of the data was necessary to provide the Apriori Algorithm with the required parameters to complete a compressive frequency analysis.



Initial results showed that the four categories itemset, {Light Precipitation, No Wind Chill Danger, Light Breeze, Moderate Visibility}, were present in at least 60% of the accidents recorded from this database. These results are statically significant, as they suggest that a majority of accidents are not heavily influenced by hazardous driving conditions, specifically bad weather. This may suggest one of two things. First, hazardous driving conditions may increase people's awareness and caution while on the road. Second, there's likely a more external factor contributing to most accidents besides the influence of natural forces. However, these results do not suggest that hazardous conditions do not contribute to car accidents. Perhaps their effects are better correlated with an increase in severity rather than frequency.

Bayesian Classification

Most of the attributes used in the Apriori analysis were input into the Bayesian Classification process attempting to predict the severity level of a car accident given condition X. This analysis was run to identify a correlation between hazardous driving conditions and accident severity. We sought to confirm suspicions that while bad weather and nighttime driving may not directly increase the frequency of car accidents, it may increase severity. Our analysis used the attribute, 'Distance(mi)', to measure the severity of an accident, as personal guidance suggests the two are highly correlated. Again, the 'Distance(mi)' attribute required conversion from quantitative to categorical data. Based on subjective reasoning, we coded an accident causing less than .25 miles of affected roadway as a 'Minor Accident', a value greater than 0.25 miles, but less than 1 mile, was labeled a "Major Accident,

Bayesian Classification:
Hazardous Driving Conditions Predict Accident Severity

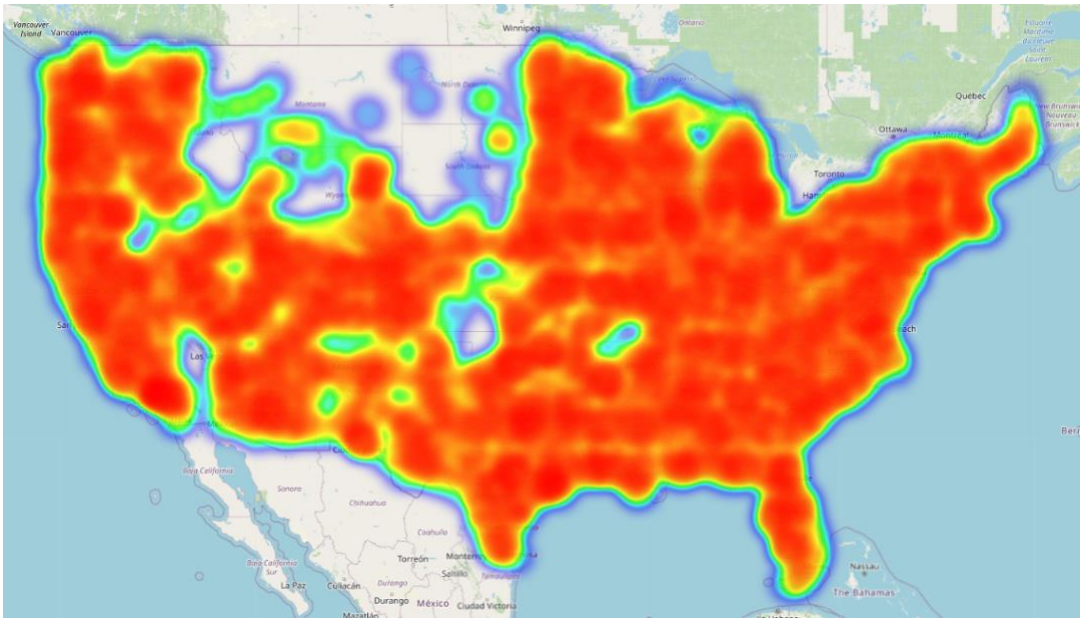
Conditions ⬇ Severity ➡	Minor Accident	Major Accident	Severe Accident
'Wind Speed' = Strong Breeze 'Visibility' = Poor	0.0840%	0.0347%	0.0485%
'Wind Speed' = Strong Breeze 'Visibility' = Poor 'Time of Day' = Night 'Temperature' = Freezing	0.001814%	0.001239%	0.002026%
'Wind Speed' = Strong Breeze 'Visibility' = Poor 'Temperature' = Freezing 'Time of Day' = Night 'Wind Chill' = Dangerous 'Humidity' = Wet	0.0001262%	0.0001067%	0.0002115%

and any value greater than 1 mile was categorized as a "Severe Accident".

The table created attempts to predict the severity of an accident based on three versions of conditional category item sets. The resulting computations show the predicted probabilities of a Minor, Major, or Severe accident occurring given a series of conditions present in a single accident. The table reveals accidents occurring alongside strong breezes and poor visibility are predicted to be minor. However, with the addition of two, then four, hazardous driving categories in the analysis, the Bayesian Classification predicts car accidents under these conditions will most likely be severe. Indeed, this confirms our hypothesis that weather conditions and nighttime driving have a positive correlation with an increase in the severity of accidents. Drawing inferences from these results, we can conclude that because most accidents are not weather-related, a majority are not as severe.

Heat Map

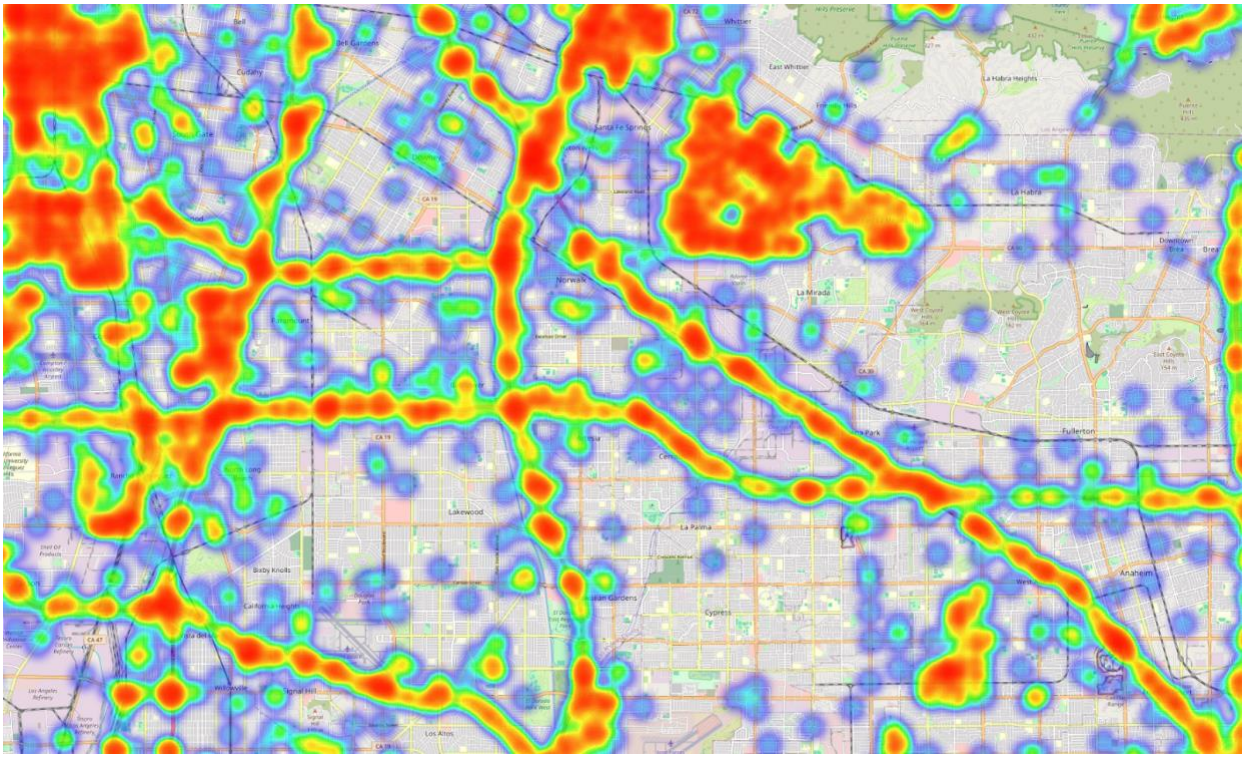
In our analysis of US Accidents data, we've utilized heat maps to visualize the spatial distribution of accident frequencies across the country. The initial step involved setting a minimum frequency threshold to identify the areas with the highest occurrence of accidents. By doing so, we aimed to pinpoint regions within states or nationwide that experience the most frequent accidents.



The subsequent phase of our analysis is focused on drilling down into specific geographical areas, including intersections, highways, and zip codes, where accidents are prevalent. By identifying these

hotspots, we can gain valuable insights into the transportation infrastructure's shortcomings and areas in need of improvement. This information is crucial for decision-makers involved in transit planning, commuter safety, and infrastructure development.

The second heat map presented depicts the situation in Los Angeles County, where highways and key city centers are prominently highlighted due to their high

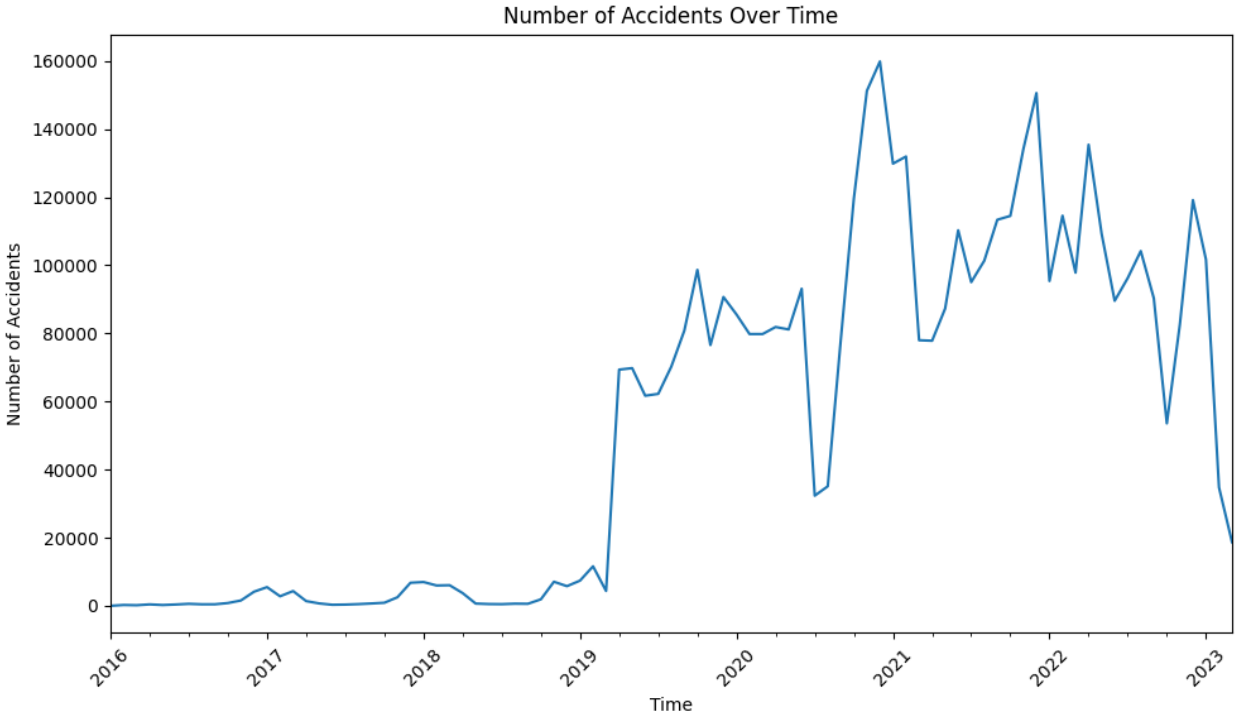


frequency of accidents. By visualizing accident patterns at a granular level, we can discern trends and patterns that inform targeted interventions and resource allocation to enhance road safety and mitigate accident risks.

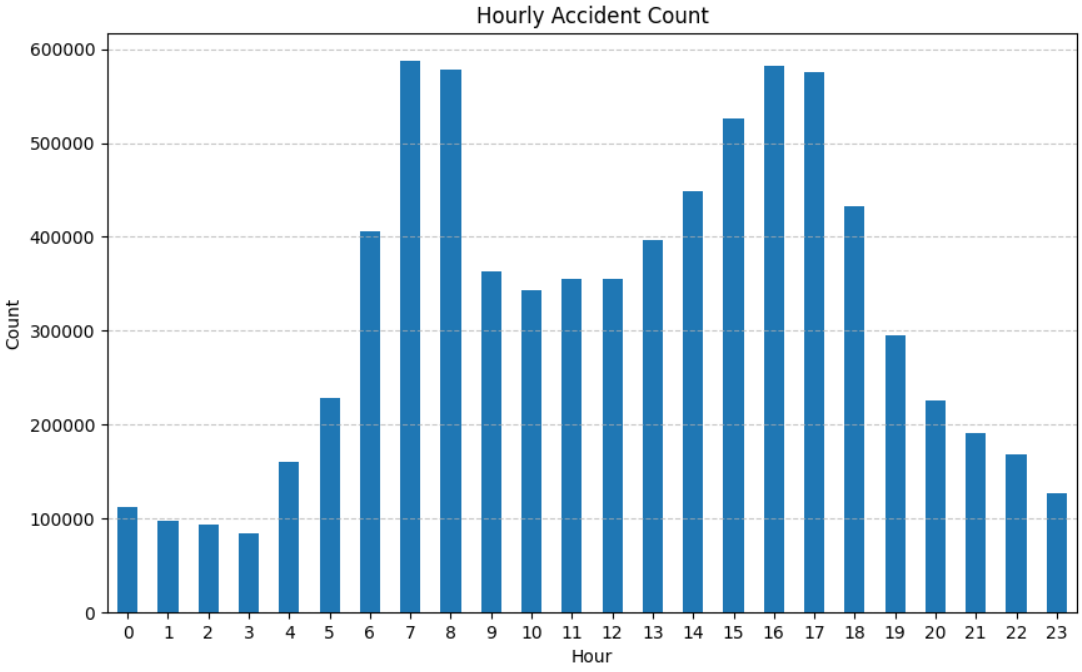
Bar Graph of Time Distribution

One of our original questions asked about frequency and distribution of accidents with relationship to time. This resulted in two scopes to look at, the first was a graph of the number of accidents by date, showing a progression year over year. The second scope was to analyze when accidents are most likely to happen day over day based on time.

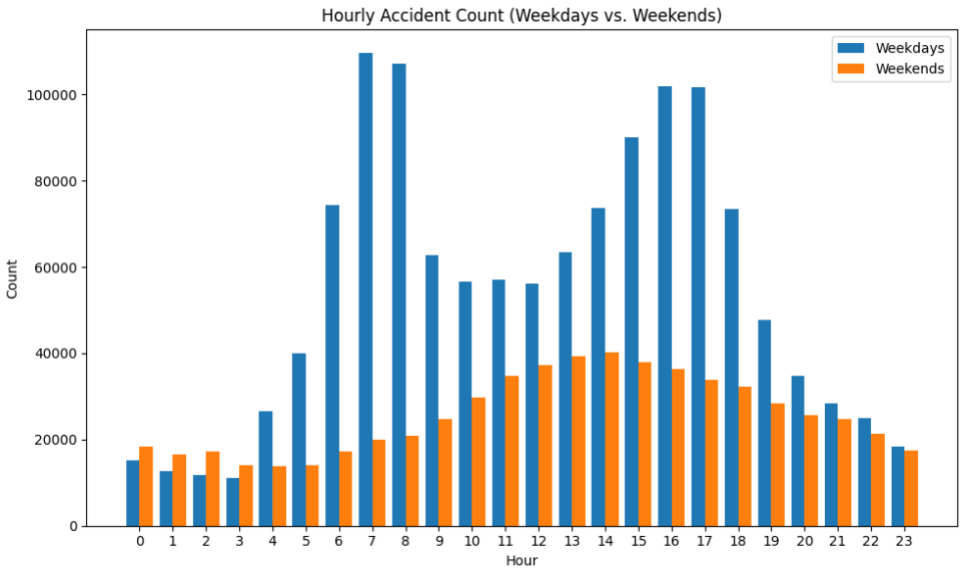
There are two main elements that stand out when analyzing the “Number of Accidents Over Time” would be the shockingly low number of accidents up to 2019, and the significant drop in March of 2020. After a deeper analysis of our data, our best guess as to why there was a discrepancy in the number of accidents pre 2019, would be because of the two different sources supplying the information. The second element that stood out was the extreme drop in number of accidents in the 2nd quarter of 2020, from March through June. There is no doubt that this is due to the global pandemic causing a majority of the population to shift to doing work and school remotely.



The second question we asked was how are accidents distributed over the day, specifically will we see a spike in accidents in what is commonly referred to as "rush hour." We chose to define these time periods as 7-9:00 AM and 4-6:00PM (or 16-18 on the graph). The graph clearly shows that there are significantly more accidents that occur during these time frames. This led us to ask a more detailed question: what is



the difference between the distribution when breaking up weekdays and weekends. To make sure we had an accurate view as to what was happening, we made the graph account for the number of days being examined, so this is average per day for each of these time slots.



This was a question that developed naturally after examining the original bar graph so we never made any official predictions. That being said, the results were not shocking to anyone. We still have a majority of accidents occurring on weekdays during the time most people are commuting to work. But we do see an interesting dip in early morning weekends, when people tend to sleep in. There is also a weekend spike in the early afternoon, when people tend to be out and about to enjoy their weekend. We also see that the only times an accident is more likely on a weekend is from the midnight to 4:00 am time slot.

Our data did confirm our predictions for each of our time based questions and led to some new interesting insights.

APPLICATIONS

Our findings indicate that good driving conditions (moderate visibility, low wind speeds, comfortable precipitation) are present in most traffic accidents recorded. Given our findings that suggest most accidents occur within peak hours of driving, (7 am - 9 am, 4pm - 6pm) we can infer a positive correlation between the frequency of accidents and the volume of traffic. Therefore, because we've established a correlation between the frequency of accidents and good driving conditions, and have predicted a positive relationship between the frequency of accidents and the volume of traffic, we can infer a positive correlation between good driving conditions and the volume of traffic. There could be many reasons for this relationship; commuters want to travel to an outdoor location or event, good driving conditions encourage commuters to get their errands done, etc. If commuters desire to drive during good weather conditions, they should be aware of the increased chance of hitting heavy traffic and getting into minor accidents.

Results from the Bayesian Classification predict a strong positive correlation between hazardous driving conditions and the severity of an accident. That is, as more hazardous driving attributes are present in an accident, there is a higher probability that the accident is severe. This knowledge can be applied in several ways. First, commuters may suspect based on the results from the frequency analysis that people drive more carefully in hazardous conditions, thus resulting in fewer overall accidents. However, the results from this classification analysis confirm that hazardous driving conditions do not make driving conditions safer overall. Even though fewer accidents will occur in hazardous conditions, the resulting damage from accidents is likely to be much worse than in normal conditions. This should inform commuters that they should remain off the road during harsh weather or nighttime.

Commuters in highly populated areas should consider using public transport due to its potential to reduce traffic congestion as well as accident rates. Heat maps of U.S. accident data show higher accident

frequencies in densely populated regions, particularly around major highways and city centers. These high-traffic zones indicate areas with increased vehicle volume, leading to a greater likelihood of accidents. By adopting public transport, carpooling (shared mobility), or alternative/shared transportation (bikes, skateboards, scooters, etc.), commuters can help alleviate traffic, resulting in fewer cars on the road, which can directly translate into a reduced number of accidents. Additionally, opting for shared mobility as a primary source of commuting offers a more sustainable alternative, reducing carbon emissions and promoting a greener environment.

Expanding on this idea, local governments should contemplate subsidizing public transportation not only to encourage its utilization among commuters but also to bolster the range of transportation options available. This multifaceted approach not only helps alleviate road congestion but also enhances accessibility and convenience for residents. The heat maps provide valuable insights into where accidents are most prevalent, suggesting that targeted efforts to increase public transportation accessibility to these areas could significantly impact accident reduction. Subsidies could make public transport more affordable and attractive, drawing more commuters away from personal vehicles and onto public transit. This shift would not only decrease the risk of accidents but also contribute to improved traffic flow and reduced stress on transportation infrastructure. By supporting public transportation through subsidies, local governments can play a crucial role in enhancing road safety, promoting a more efficient and sustainable commute, and ultimately saving money by reducing infrastructure wear and tear. This comprehensive approach benefits commuters and contributes to the long-term preservation of transportation infrastructure and associated costs.