# An Analysis of Environmental Conditions and Driver Behavior on the Severity of Accidents in California

Ashwin Bardhwaj

April 2022

Modeling the Future Challenge 2021-22

# Contents

# 1. Executive Summary

Vehicle collisions have been one of the leading causes of accidental death in the US. Specifically, we notice that a disproportionate number of national fatal collisions (10%) occur in California. We also notice that this trend continues for various severities of collisions ranging from “property damage only” to “severe injury” and “fatal”. To understand the cause of different types of collisions, we seek to create a mathematical model to analyze the extent to which various climate, behavioral, and location factors impact the severity of vehicle crashes in California and generate subsequent insights that provide a comprehensive analysis of this issue.

Using data from the SWITRS database collected by the California Highway Patrol and associates, we selected several factors about the accident related to environmental conditions such as climate information, road condition, lighting, time of day, road surface as well as driver information/behavior like age, gender, cellphone in use, drug influence, and alcohol influence among others. After conducting data validation and cleaning, we conducted the density-based algorithm DBSCAN to remove geographic outliers and focus on regions with high concentrations of accidents. Using this data, we characterize the frequency of different collision severity and trends over time. As a result, we predicted that collision frequency will rise as the economy improves in 2022 and beyond. We used the factors of these cases to construct a Random Forest, Decision Tree, and Naïve Bayes models to identify the most important factors affecting accident severity and have confidence in those results by providing different insights into the data. Exploratory Factor Analysis showed that our factors could be combined into five unrelated components and could account for 71.0% of the models’ variation.

Subsequently, we explored the three most influential variables indicated by these models to conduct a frequency analysis to determine the specific aspects of these variables that cause certain severities to occur. Coupling these results with a geospatial analysis of fatal accidents gave information on the specific regions where severe accidents commonly occur. We also identified high risk groups for e.g., young drivers, impaired drivers etc. and large private insurers, impoverished individuals, and local businesses.

Based on our analysis, we provided recommendations to both insurers and policymakers in California. We suggest the continuation of insurance programs targeted towards low-income families who would be unable to pay the premiums of liability insurance. Additionally, the use of improved public transportation practices in the “accident hotspots” described can bring down high collision periods. Also, the extension of graduated licensing practices in California can help more safely introduce driving to inexperienced learners and decrease the risk for these drivers. As severities and collisions increase in the coming years, we hope to make Californian roads a safer place for all drivers.

# 2. Background Information

Automotive transportation has always been the backbone of the American economy since the mass production of Henry Ford's Model T in 1908. In 2019, there are 276 million registered vehicles in the United States and 91% of American families have access to a vehicle. (Borrelli, 2021). As a result, it has altered the American lifestyle and economy by improving accessibility to goods and services as well as creating a larger market for manufacturers and producers. In 2016, $2.8 trillion in goods were shipped to and from sites in California, mostly by truck (TRIP, 2016). The utility of the automobile created a boom for the industry. For example, the growth of the automobile industry created 3 million jobs in the US (USBLS, 2022) and accounts for 3% of Americas GDP (Sky, 2021).

However, the industry has also brought with it a higher risk of injury and fatality. In 2019, there was an estimated 6.8 million police reported crashes across the United States, where 2.7 million people were reported injured and 36,096 reported fatal (NCSA, NHTSA , US-DOT, 2019). This trend is especially true in California, known to be the state with the highest collision and fatality rates in the country. In 2019 alone, there were 3,606 fatal and 74,000 injurious crashes in California. In addition, there were 174,000 non-injurious crashes leading to property damage only (PDO) (CHP, California, 2022). In total, California accounted for almost 9% of total number of crashes in America. It is 4th leading cause of death in the country and the 10th leading cause of death in the world (NCHS,CDC, 2022).

In a collision, the primary persons at risk are the drivers and others directly involved. The cost of collision increases exponentially depending on the severity of the crash. Individual medical bills, for instance, can range from $3,100 for non-injurious crash to $26,000 for severe injuries. In 2010, US economy spent around $23.4 Billion in medical cost for these injuries (Lawrence, Miller, Zaloshnja, & Lawrence, 2015). This could be even more expensive and sometimes financially disastrous for the uninsured driver (NCIPC, CDC, 2017). In addition, a severe injury can lead to disability, loss of income, and reduced quality of life. In 2019 Americans spent over one million days in the hospital each year from crash injuries. In 2012, more than 2.5 million Americans went to emergency and nearly 200,000 were then hospitalized for the same. Lifetime work loss due to 2012 crash injuries cost an estimated $33 billion. (NCIPC, CDC, 2017)

Additionally, the effect of a severe crash is felt not only by the involved parties, but also other vehicles on the road. In 2010, congestion cost due to vehicle crash including travel delay, fuel usage, and environmental impact totaled $28 Billion (Lawrence, Miller, Zaloshnja, & Lawrence, 2015). Employers and corporations are also at risk, in the form of work disruption, the cost of supporting employees, and the disruption the supply chain. Total economic cost of accidents in 2010 is estimated to be $242 billion (Lawrence, Miller, Zaloshnja, & Lawrence, 2015).

As mentioned before, fatality is not the only outcome of a collision. Instead, collisions have severities that indicate the seriousness of an accident and help to quantify different risks posed to a driver. While collisions are naturally randomly occurring events, the severities of those collisions are influenced by several factors surrounding the crash such as weather conditions, the state of the road, traffic congestion, and the geographic location of the accident. Another factor is driver characteristics such as age, gender, maturity etc. and his/her behavior. Additionally, the risk for drivers can be magnified in some areas of California over others. This is evident in certain cities such as Los Angeles containing 55,350 car accidents (20% of CA accidents in 2016) (Citywide Law Group, 2022). Hence, the necessity of safe transportation and the danger of accidents in California cannot be understated.

To protect these drivers from the devastating physical and fiscal effects of collisions, auto insurance has been mandated by the state of California to ensure every driver has at least basic liability protection. For those who cannot afford the premiums for liability can purchase them from the state at a lower cost. Automotive insurance is vital towards the well-being of the drivers because it provides them with financial security while driving. Drivers may choose to opt for additional insurance plans with different coverages based on personal need and risk appetite. This includes collision insurance, which covers damages towards the vehicle, med-pay, which covers the medical costs of the driver, and non/underinsured coverage, which covers the cost of repair of medical bills when the driver at-fault does not have the insurance to pay for the victim’s damages. These plans are written and insured by private companies who are licensed to operate in California. In 2019, the California automotive insurance industry wrote more than $33 billion in premium with a loss ratio of 65.50% (CDI, 2019). Hence, the risk on Californian roads does not only belong with the drivers, but also to the companies that insure them.

In summary, California experiences the largest number of traffic accidents and highest number of fatal accidents anywhere in the country. The risk of accidents affects not just the drivers and other involved, but also others stuck in traffic, insurance companies, and the economy at large. As a result, being able to predict the severity of collisions based on factors surrounding the accident is useful for all stakeholders. The paper attempts to analyze the extent to which various climate, behavioral, and location factors affect the severity of vehicle crashes. Also, knowing the factors that are most impactful in controlling severity can help guide appropriate government policies and inform local council to judicially apply the money allocated for infrastructure development.

## 3. Data Methodology

Effective analysis of traffic accidents in California and the impact of conditions leading to it, requires a comprehensive accident dataset covering the whole state with severity classification, along with location and other environmental factors. We used the Statewide Integrated Traffic Records System - SWITRS (CHP, California, 2022) database which collects, and processes data gathered from a collision scene. It is created and maintained by the California Highway

Patrol (CHP) and the Allied Agencies. Since it contains fields such as the exact location of the collision, as well as collision severity on a case-by-case basis; and covers collisions from 2001 to 2021, we use this as our primary dataset (Gude, 2021).

We also considered “Fatality Analysis and Reporting System” (FARS) (NHTSA, US-DOT, 2022) database from Nation Highway Traffic and Safety Administration (NHTSA), which is a nationwide census providing yearly data regarding fatal injuries suffered in motor vehicle traffic crashes from 1975 to present. While having a very diverse database of many factors, it only applied to fatal accidents which did not follow the goal of classifying multiple severity of accidents. However, we this data to conduct a regional risk analysis for fatal accidents.

We also considered NHTSA’s Crash Report Sampling System database (NHTSA, 2022) which, while containing distinct severities of accidents, did not include the precise location of the crash, a crucial factor of our model. As a result, we concluded that the SWITRS database would best suit our goals.

## 3.1. Data Identification and Categories

The California Highway Patrol has recorded a large dataset regarding the state and causes of accidents in the SWITRS database. We selected the collision severity classification from the dataset as the dependent variable. It is the quantification of the impact intensity and an indicator of the probability of injury and potential loss. Although many different metrics are used

*Table 1: Collision Severity on KABCO scale*

| ***Values*** | ***Definition*** | ***Description*** |
|---|---|---|
| ***1*** | Property Damage only (PDO) | There were no apparent injuries involved in the crash. If a party is transported and is subsequently examined and found to have no injuries. |
| ***2*** | Complaint of Pain | This classification could contain authentic internal or other non-visible injuries, as well as fraudulent claims of injury. This also includes persons who are dazed, confused, incoherent, or have been unconscious but recovered |
| ***3*** | Other Visible Injury | Injuries to victims were evident to officers at the scene, but they were non-disabling lacerations, scrapes, places where the body has received a blow (black eyes and bloody noses), or minor bruises. |
| ***4*** | Severe Injury | An injury other than a fatal injury which results in broken bones, dislocated or distorted limbs, severe lacerations, or unconsciousness at or when taken from the collision scene. |
| ***5*** | Fatal | Death because of injury sustained in a collision or an injury resulting in death within 30 days of the collision. This includes death of fetus |

to classify the severity of accidents across the US, the SWITRS database categorizes the severity of each accident on a California KABCO scale (FHWA, US-DOT, 2017) of property damage only (1) to a fatal crash (5) as shown in Table 1.

The SWITRS database contains large number of factors. For the purposes of this paper, we select only the driver of the party at fault for every crash because only the circumstances and situation around that driver is relevant to the analysis on the cause of the crash. Hence, we selected only the factors related to the environmental conditions, the driver behavior, and the location of the crash. These factors are listed in Table 2 and explained Appendix 9.1 in detail.

*Table 2: Factors affecting collision severity*

| Factor | Description | Data Type |
|---|---|---|
| **Age** | Age of the driver at the time of collision | Ratio |
| **Alcohol involved** | Indicates collision involved a party that had been drinking | Comparative |
| **Cellphone in use** | Classification based on if the party is using cell phone | Comparative |
| **Collision Time** | The time when the collision occurred (24 hr. time) | Ratio |
| **Financial responsibility** | Classification based on whether the party showed proof of insurance at the time of collision | Comparative |
| **Intersection** | Classification based if collision occurred in an intersection | Comparative |
| **Intersection Type** | Classification based on the type of intersection the collision occurred | Comparative |
| **Latitude** | Y- coordinate of the geocoded location of the collision | Comparative |
| **Longitude** | X- coordinate of the geocoded location of the collision | Comparative |
| **Lighting Condition** | Classification based on how bright location is at the time of collision | Comparative |
| **Party drug Impairment** | Classification based on physical or drug induced impairment | Comparative |
| **Party gender** | Primary party’s gender Classification | Nominal |
| **Population** | Population size classification at the collision zip code | Comparative |
| **Road Condition** | Classification based on road condition | Nominal |
| **Road Surface** | Classification based on slipperiness of the road at the time of the collision | Comparative |
| **Vehicle make** | Classification based on the vehicle make of the primary party’s vehicle | Nominal |
| **Vehicle year** | Model year of the party’s vehicle | Interval |
| **Weather** | Classification based weather condition at the time of collision | Comparative |

## 3.2. Data Reliability Evaluation

As observed before, SWITRS database has many fields that describe the crash, parties involved, and individual persons. We started by collating the relevant fields into a single table containing about nine million unique cases of accident data from the years 2000 – 2021 as explained in Appendix 2. We then removed all cases with fields containing nulls. By directly removing, instead of imputing missing data, we were able to conduct under-sampling after data cleanup, without any risk of data leakage between training and testing datasets. After that, the remaining fields containing unknown values were recoded appropriately.

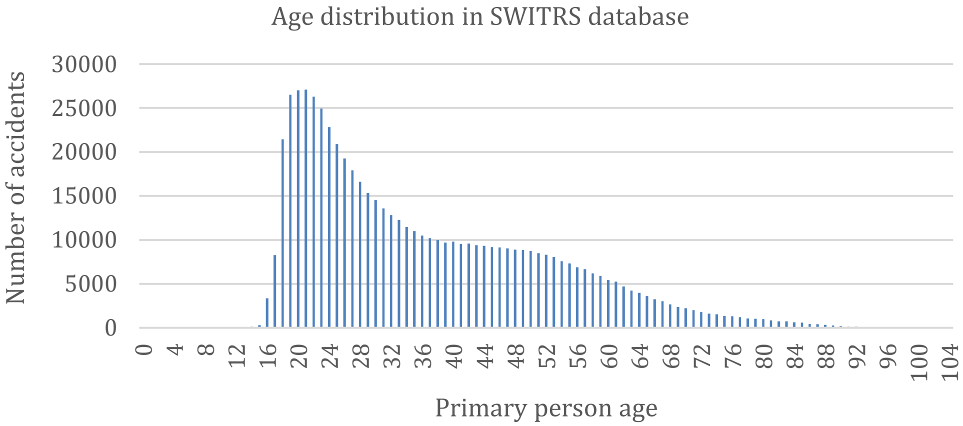


*Figure 1:Age Distribution in SWITRS database*

Other factors, such as age, were checked for any possible correlations between extreme values and accident severity. We removed the outliers determined by a 95% confidence level or a z-score over 1.96. This resulted in a range of 16 to 90 years as shown in Figure 1. Similarly, the SWITRS database holds 123 different vehicle makes. However, a frequency distribution of collision vs vehicle-make (as shown in Figure 2) exhibited a right skewed normal distribution. Using a Pareto chart, we found the top 40 vehicle make categories are responsible for 99% of total collisions in California and used these when executing our model.

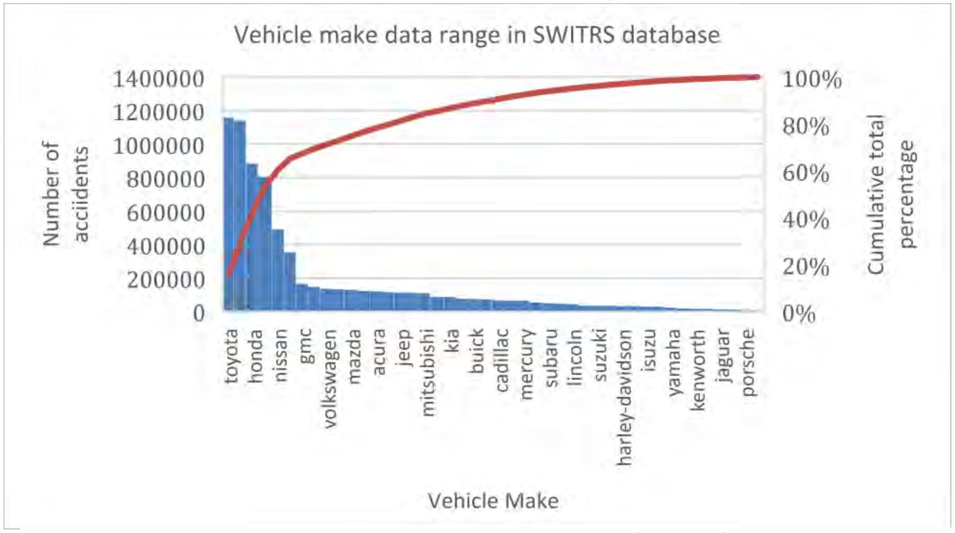


*Figure 2: Vehicle make distribution in SWITRS database*

## 3.3. Data Processing and Clustering

*Balancing*: We recognized that there was an incredibly high number of the accident severity classification of “property damage only” compared to other severities and this imbalance would cause the model to only favor this classification. As a result, random under sampling was utilized to balance the data to the least frequent classification (Fatal). (2u.Inc, 2021)

*Clustering:* To exclude outliers that might be caused by disparate factors, we utilize the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. This allows us to group accidents by their relative geographic location, thus minimizing the effects of external causes of variation during classification and assuring that we are observing similar cases.

The utility of DBSCAN compared to other cluster algorithms like K-Means or Agglomerative Clustering is both its ability to discover clusters of any shape (as opposed to gaussian-ball shaped) as well as being able to exclude noise which is important in our context. By running the DBSCAN algorithm at an epsilon value of 0.15, which was optimized by a k-nearest-neighbor search, we were able to create a cluster graph of all clusters as shown in Figure 13. The DBSCAN algorithm returned around 23 clusters, however, almost 80% of over 10,000 data points plotted are contained in 3 clusters localized in Los Angeles County, San Francisco Bay Area, and along central California. As a result, we focus on these high-density regions and sample data from each of them. (Appendix 3 & 4). We also ran the DBSCAN algorithm on the dataset derived from FARS database to do geo-spatial analysis.

*Trends Over Time*: We analyzed the monthly collision severity from SWITRS database from 2001-2020 and tried to characterize historical and future trends.

# 4. Mathematics Methodology

In this section, we aim to construct a model that can classify accident severities based on environmental conditions, driver behavior, and geographic location as well as analyze the most influential factors affecting them.

## 4.1. Assumptions and Justifications

1. *Minimal variation in the location of hotspots and the severity/frequency its data over period 2017-2019.* This assumption allows us to conduct geospatial analysis using data from more than one year.
2. *Geographic accident clusters can be defined as dense regions of accidents that are separated by non or less-dense regions of accidents*. This assumption is necessary to define clusters for the DBSCAN clustering algorithm.
3. *The data collected is correct and applicable to the goals of this paper*

4. *Statistics tool and models are calculated correctly and provide the correct result.* The "IBM SPSS Statistics" (IBM Inc, 2022) software from IBM has been widely used for data analytics and multivariate analysis." Math Works Mapping Toolbox" (MathWorks Inc, 2022) and "MathWorks Statistics and Machine Learning Toolbox" (MathWorks Inc, 2022) are used for modeling and clustering. As such, this assumption is made for confidence in the output of all involved models
5. *Only police reported crashes will be implemented.* As the data is collected by the State Highway Police and affiliates, accidents that are not reported are not considered in this paper
6. *Factors impact the collision frequency and severity in all clusters similarly.* We did analysis to determine factors affecting collision for LA cluster and applied the lessons learned for all clusters in California.

## 4.2. Model Development

Literature has historically recommended three types of intelligent classification techniques for modeling accident severity prediction (Buket & Kara, 2020). These are Random Forest, Decision Tree, and Bayesian models. Utilizing multiple models provides different insights into analyzing the trends within the data. In this paper, we will investigate the strength of all three of these models and determine the optimal strategy in which to classify accident severity.

### 4.2.1. Decision Trees

We use a Decision Tree model to determine what combinations of factors create a high risk for crashes, which is vital towards the creation of better risk mitigation strategies. Decision Tree is a technique that uses specified factors to make a series of conditional statements that results in the severity of an accident. A focus of the Decision Tree model is the tree itself, which shows the exact decision process that the model takes to make its conclusion. (Sharma, 2020)

We use the SPSS tool to form a Decision Tree using a CHAID growth method with a maximum of 3 levels. A minimum of 150 cases per node has also been specified to help limit overfitting of the model. We then run the model on a 60-40 training-testing split of the data. The CHAID growth method uses Chi-Squared testing to be the metric which decides what factor each node in the tree should be. (Sharma, 2020). It determines the "goodness of the fit" between the actual and expected values.

$$x^2 = \sum_{i=1}^{n} \frac{(O - E)^2}{E}$$

Where O is the actual value of a class, E is the expected value of the class, and n is the total number of nodes of the split.

### 4.2.2. Random Forest

We implemented a Random Forest model (Miaomiao & Yindong , 2022)), to not only predict future collision severity but also to receive the most relevant factors affecting collisions. Random Forest operates by creating an ensemble of decision trees and taking the highest frequency of classification. By collecting results from multiple decision trees, random forest more often has a higher accuracy of classification due to its ability to inform its answer through many different paths. It is a common model used by data scientists mainly for its robustness to outliers, quick training and testing speed, and usefulness with high dimensional data, similar to our selected factors.

We use the random forest algorithm provided by the SPSS tool, with 300 trees and 60-40 split for the training and testing datasets respectively. The model utilizes the Gini index to compare the impurities of different factors. The calculation for the Gini impurity index (G) is stated below where $p_i$ is the probability of the classification "i" occurring and $n_c$ is the total number of classifications (or in our case collision severities) which is 5.

$$\mathrm{G} = \sum_{i=1}^{n_c} p_i(1-p_i) = 1 - \sum_{i=1}^{n_c} p_i^2 = 1 - \sum_{i=1}^{4} p_i^2$$

Our model returns the mean decrease in impurity, which is the average of the decrease of Gini impurity in all trees. The calculation for the decrease in impurity (I) is as stated below.

$$\mathrm{I} = \mathrm{G}_{\mathrm{parent}} - \mathrm{P}_{\mathrm{split}t_1}\mathrm{G}_{\mathrm{split}t_1} - \mathrm{P}_{\mathrm{split}t_2}\mathrm{G}_{\mathrm{split}t_2} \ldots - \mathrm{P}_{\mathrm{split}t_n}\mathrm{G}_{\mathrm{split}t_n}$$

The decrease in Gini impurity is found by subtracting the Gini impurity (G) of every split (in n splits) multiplied by the proportion of cases in that split (P) from the Gini impurity of the parent node (Sharma, 2020)

### 4.2.3. Bayesian Models

We implemented a Bayesian model as well, to classify collision severity outcomes based on conditional probability and predict the most relevant factors. We choose the Naïve Bayes model provided in the SPSS tool, as it is among the best suited Bayesian models (Vadapalli, 2021) for working with categorical data. It uses Bayesian probability to find variable importance where P(A|B) is the probability that A occurs given that B occurred. The formula for this known as Bayes' Theorem is shown below

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A)\,P(B|A)}}{\mathrm{P(B)}}$$

P(A) is the probability that A occurred. P(B) is the probability that B occurred. And P(B|A) is the probability that B occurred given that A occurred.

The advantages of this model are that it returns the probabilities of the factors with respect to the dependent variable. (Vadapalli, 2021). Knowing the factor with the strongest probability correlation would help determine the most influential factor impacting collision severity. This would also help verify the factor importance results from the Random Forest model.

However, a significant assumption of the model makes is that all factors are independent. While this constraint has prevented Naïve Bayes from making accurate predictions in other applications, we hope to implement the unassociated components provided by Exploratory Factor Analysis to mitigate this risk.

# 5. Risk Characterization and Analysis

## 5.1. Collision Severity Characterization

Before we start analyzing the impact of various climate, behavioral, and geo-location factors on collisions, we will characterize collision severity based on its severity, frequency, and historical / future trends.

### 5.1.1. Collision Severity Distribution

In 2019, there was a total of 470K accidents across California. However, there is a significant variation between the frequencies of severities themselves as evident by Figure 3. For example, in 2019, “property damage only” accounts for approx. 60% of accidents, while complaint of pain is 24% and other injury accounts for 12%. Fatal (0.73%) and severe injury (3%) of accidents comprise only a total of 3.7% of accidents. While this demonstrates that the likelihood of being

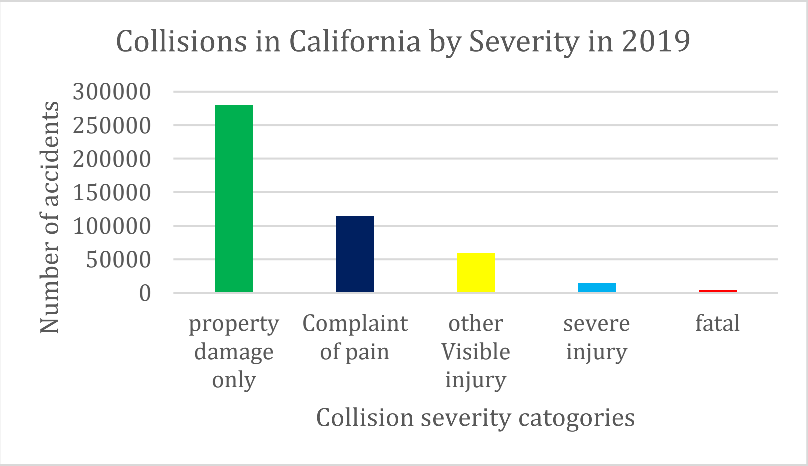


*Figure 3: Collision distribution in CA by severity in 2019*

involved in a dangerous and harmful collision in California is low, almost 18,000 people experienced life-threatening injuries or death as a result.

### 5.1.2. Severity, Frequency and Expected Values

As shown in Table 3, from 2001-2020 we see total mean collision of 38.6K (+/- 5.2K) per month in California, of which” Property Damage Only (PDO)” type collision (23K +/- 3.5K) is twice as much as “Pain” collisions (9K +/-1.2K), while "fatal” collision is about 272 (+/- 257) per month.

*Table 3: Mean, STD and Range of Collision severities per month from 2000- 2021*

| | PDO | Pain | Other | Severe | Fatal | Total |
|---|---|---|---|---|---|---|
| **Mean** | 23,356 | 9,295 | 4,762 | 905 | 273 | 38,591 |
| **Standard Dev.** | 3,483 | 1,198 | 804 | 151 | 43 | 5,243 |
| **Range** | 18,368 | 7,819 | 3,900 | 740 | 257 | 29,534 |

### 5.1.3. Collison Trends

We also see a long-term declining trend in collisions classified “PDO”, “Pain”, and “other injury”. This holds true even after discounting the unusual traffic conditions resulting from the COVID pandemic. However, “Fatal” and “severe-injury“ collisions show a stable and increasing trend respectively. Figure 4 shows the historical trends of average monthly collisions in the years 2001- 2020 by accident severity. Using the line of best fit for each severity to evaluate the trends, we see a decline in the number of collisions per month classified “property damage only”

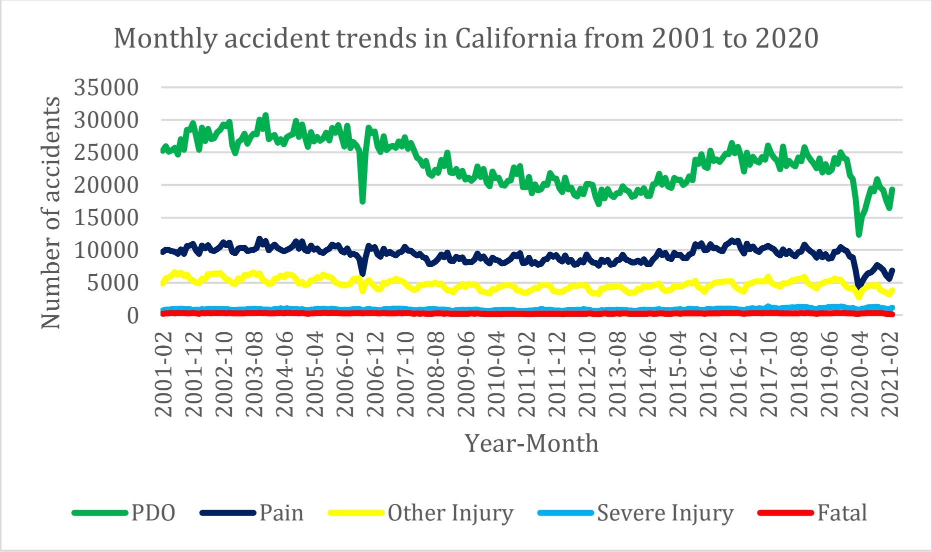


*Figure 4: Historical trend of monthly average Collision from 2000 – 2021*

(-0.11%), “pain” (-0.06%), “other injury” (-0.12%), and “fatal” crashes (-0.04%), while severe injury gained 0.13% of accidents per month. (See Appendix 9.5) This demonstrates that while the number of collisions classified pain or fatal have not changed significantly over time, there has been considerable variation in other severities. In particular, the number of “severe injury” collisions have been rising in contrast to other classifications and will be expected to do so in future years

By analyzing the historical frequency of the total monthly accidents recorded in California from 2001 to 2020 as shown in Figure 5, we observe a close relation between accidents and economy, expressed as a percentage of the employed population (USBLS,, 2022) as shown in red. This correlation even accounted for the sharp decline in 2020 (the emergence of the COVID pandemic) (See appendix 9.5). The Bureau of Labor Statistics projects that employment would increase by 10.8% in 2022 alone (USBLS,, 2022). This, along with the graphs recently rising trend, demonstrates that as the employed population rises, the number of collisions in California will also increase dramatically.

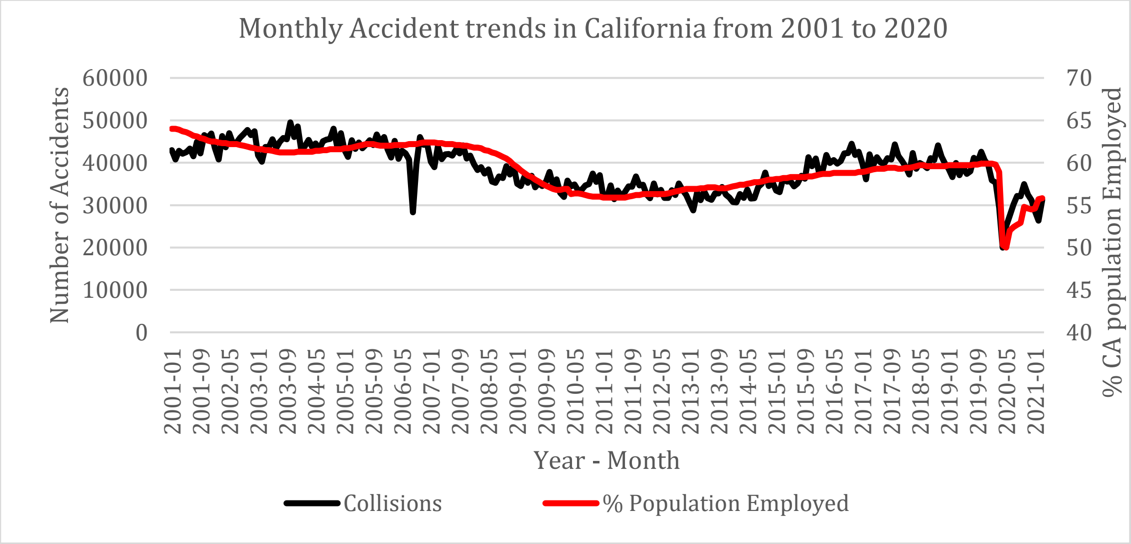


*Figure 5 Historical trends of monthly average Collision by severity type from 2000 - 2021*

## 5.2. Factor Analysis

The next goal of our model was to determine the impact of various climate, behavioral, and location factors, that contribute to varying degrees of accident severity, based on the mathematical model explained in Section 4.

### 5.2.1. Exploratory Factor Analysis of accident model

Before running the models, we conducted “Exploratory Factor Analysis” (EFA) in the SPSS tool for “dimensional reduction”, wherein we investigate the correlations in the independent factors themselves and determine if the variation in the data could be explained by a fewer number of hidden variables. (Fabrigar, 2011) These hidden variables should be independent with each other. While Random Forest and Decision Tree, do not depend on uncorrelated factors to make accurate predictions, Naïve Bayes depends on the assumption of independent data to correctly calculate probabilities. When running EFA with Principal Components Analysis (PCA) over our clustered dataset, the program returned a Kaiser-Meyer-Olkin (KMO) value of 0.59, demonstrating an adequate reliability of the factor analysis results. Also, the result of Bartlett’s test of sphericity shows a significance level (below 0.0001). meaning that the data reduction technique can compress the data in a meaningful way (Carlson, June 2010).

First, we removed the factors that have low communality loading (below 0.3). Then, after creating a covariance matrix to analyze the correlations between the factors, we computed the eigenvalues for our principal components (PC). By using the Kaiser criterion, we selected only the PCs with an eigenvalue greater than 1, as shown in the Scree plot Figure 6 (Costello, 2005)

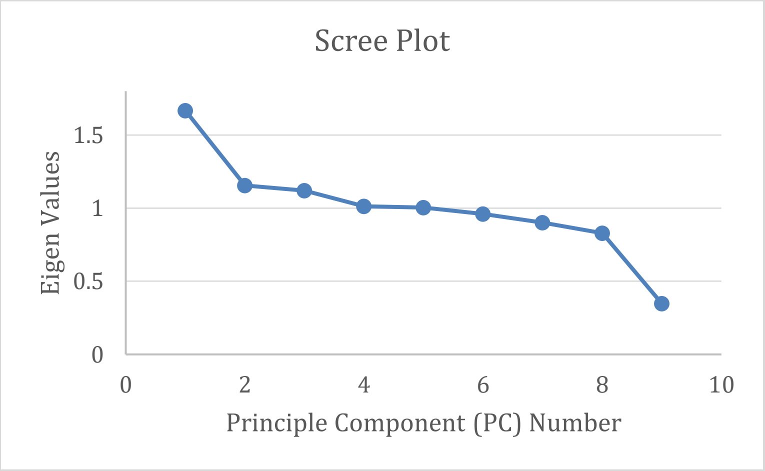


*Figure 6: Scree Plot using Principal Component Analysis*

EFA returns five components as shown in Table 4, with a total of 71% variance explained. The values in each cell of the component matrix reveal the correlations of each factor with a PC. For example, PC 1, explains the highest variance (18%) and contains a high correlation between weather and road surface. This is intuitively correct since type and amount of precipitation will correlate with the slipperiness of the road. In general, PC1 describes the atmospheric conditions (18%), PC2 the luminosity during the accident (12%), PC3 driver related information (12%), PC4 describes the type of location (11%), and PC5 the vehicle/driver conditions (11%).

*Table 4: Component matrix using Principal Component Analysis*

| | Component | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Weather | 0.908 | | | | |
| Road surface | 0.905 | | | | |
| lighting | | 0.785 | | | |
| Collision time | | -0.664 | | | |
| Gender | | | 0.759 | | |
| Location type | | | | 0.744 | |
| Population | | | | 0.684 | |
| Party age | | | 0.468 | | 0.653 |
| Vehicle year | | | -0.518 | | 0.611 |

### 5.2.2. Analysis of factors from Decision Tree

After running our models, we analyze the output from the Decision Tree model, which is the tree itself as shown in Figure 7. The first split in the Decision Tree is often considered the most influential factor. In this instance, the most influential factor is driving under the influence of alcohol (Chi-squared 1,127), which shows a strong separation between low severities (PDO or Pain) and high severities (Severe or Fatal). Another key take-away from the tree is the path where the driver is drunk and under the influence of drugs (Chi-Squared 150 to 300), resulting in a 95.5% chance of accident. This means that the impairment value is also a significant factor in characterizing severity.

We also observed a significant split due to vehicle makes (Chi-Squared of 725). While car makes such as Toyota, Honda, Chevrolet, etc. have shown to be often involved in safer collisions, motorcycles such as Suzuki, Yamaha, or Harley-Davidson have a much higher injury and fatality rate. In fact, there is almost three times the risk of severe collisions for motorcycles than cars and the risk of a fatal crash is doubled.

Additionally, the population classification (Chi-Squared 45) proves to be an efficient divider between mild and serious collisions. Among collisions involving non-DUI and cars, severe collisions are more than twice as likely to be found in an unincorporated area rather than in areas with a population of or greater than 10,000.

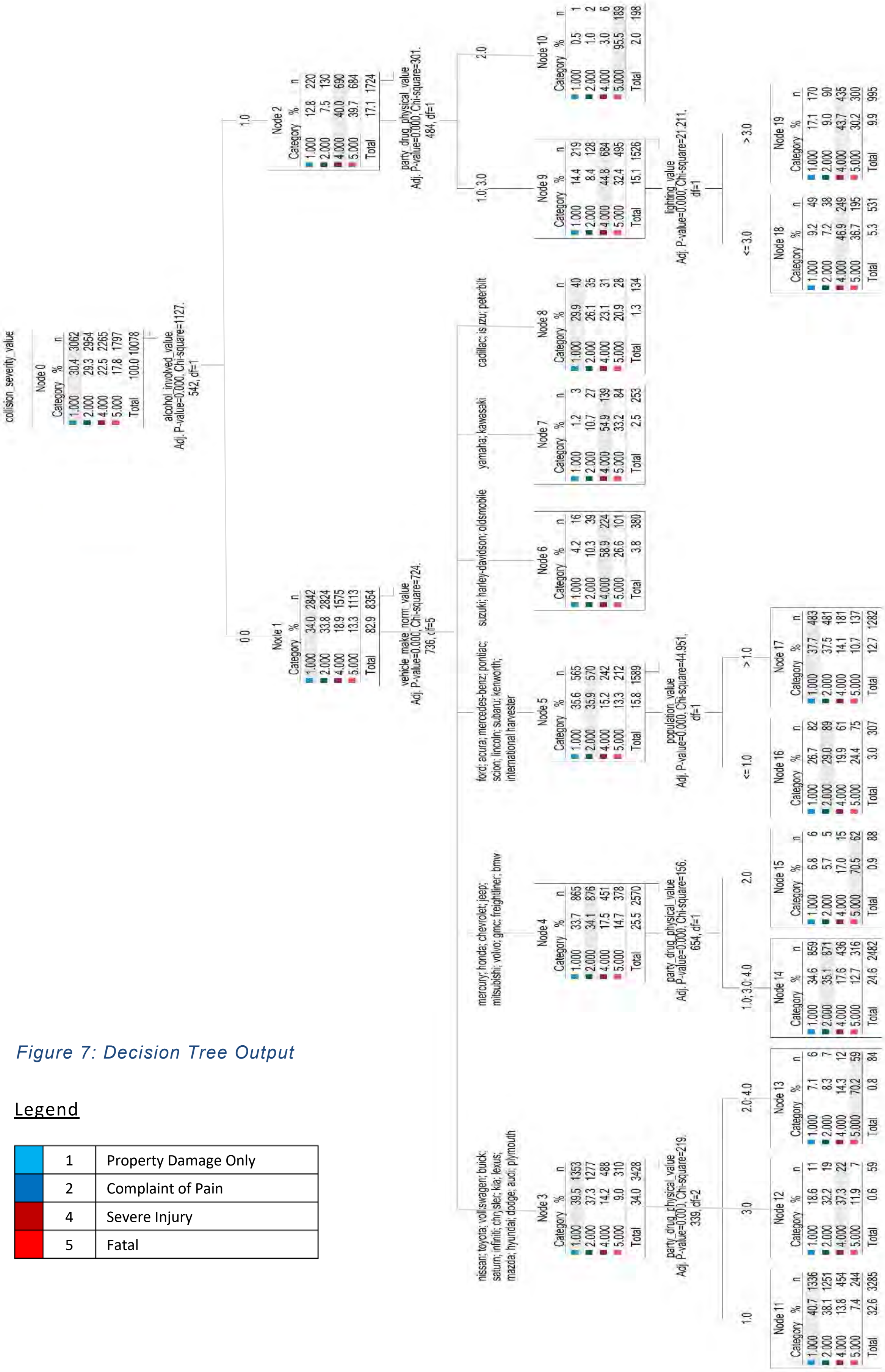


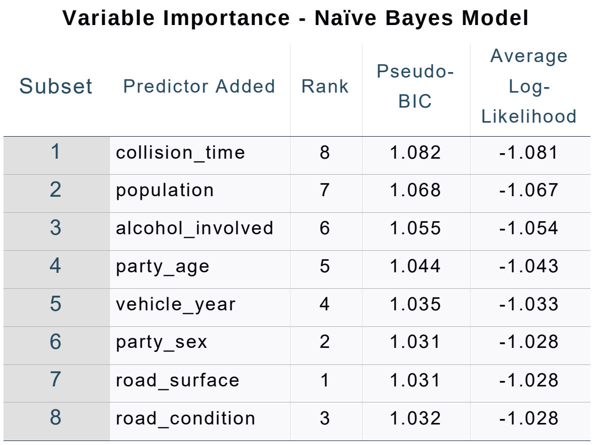
*Figure 7: Decision Tree Output*

## Legend

| | | |
|---|---|---|
| 🟦 | 1 | Property Damage Only |
| 🟦 | 2 | Complaint of Pain |
| 🟥 | 4 | Severe Injury |
| 🟥 | 5 | Fatal |

### 5.2.3. Analysis of factors from Bayesian and Forest Model

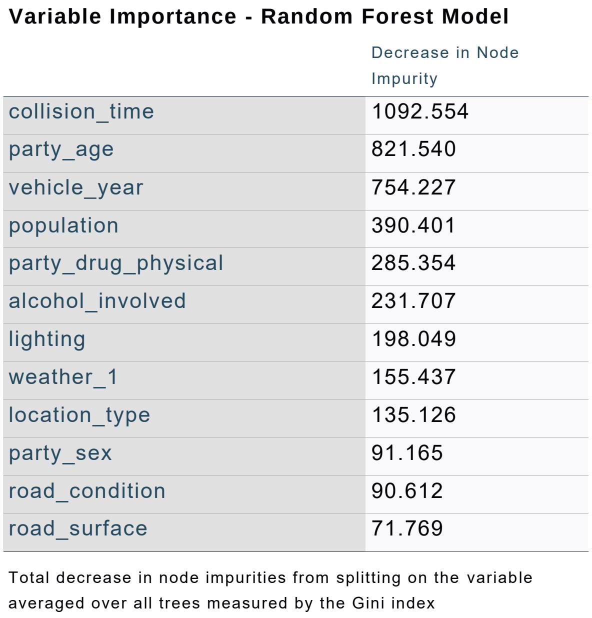
In addition to the Decision Tree, we obtained factor importance results from the Bayesian and Forest models as well. We ran both a probability analysis and a Gini index method to determine the primary factors that contribute to collision, as shown in the Figure 8 and Figure 9. Both methods agree that collision time has the highest rank, followed by party age, vehicle year, and population. Since vehicle year correlates with party age based on our EFA, we did not characterize vehicle year separately. We conducted a frequency analysis of the other three variables, to determine how they impact the severity of accidents. Variable Importance from Forest Model



**Variable Importance - Naïve Bayes Model**

| Subset | Predictor Added | Rank | Pseudo-BIC | Average Log-Likelihood |
|---|---|---|---|---|
| 1 | collision_time | 8 | 1.082 | -1.081 |
| 2 | population | 7 | 1.068 | -1.067 |
| 3 | alcohol_involved | 6 | 1.055 | -1.054 |
| 4 | party_age | 5 | 1.044 | -1.043 |
| 5 | vehicle_year | 4 | 1.035 | -1.033 |
| 6 | party_sex | 2 | 1.031 | -1.028 |
| 7 | road_surface | 1 | 1.031 | -1.028 |
| 8 | road_condition | 3 | 1.032 | -1.028 |

*Figure 9: Variable importance from Naive Bayes model*



**Variable Importance - Random Forest Model**

| | Decrease in Node Impurity |
|---|---|
| collision_time | 1092.554 |
| party_age | 821.540 |
| vehicle_year | 754.227 |
| population | 390.401 |
| party_drug_physical | 285.354 |
| alcohol_involved | 231.707 |
| lighting | 198.049 |
| weather_1 | 155.437 |
| location_type | 135.126 |
| party_sex | 91.165 |
| road_condition | 90.612 |
| road_surface | 71.769 |

Total decrease in node impurities from splitting on the variable averaged over all trees measured by the Gini index

*Figure 8: Variable Importance from Forest Model*

#### 5.2.3.1. Collision Time

A frequency analysis of the collision time factor is constructed as the frequency of collisions per hour in a day as a percentage of the total accidents, for every collision severity as shown in Figure 10.(see Appendix 9.7 for detail). There appears to be two peaks in overall collision frequency: the hours of 7 – 8 AM and 2 – 6 PM. This correlates well with traffic congestion (TomTom Inc, 2022) data as shown in the Figure 10. We also see significant number of severe Injury and Fatal crashes from 6 PM to midnight. This suggests that the lighting may have a correlation with severe car crashes which is consistent with EFA PC2 that found a significant correlation between lighting and time of the accident.

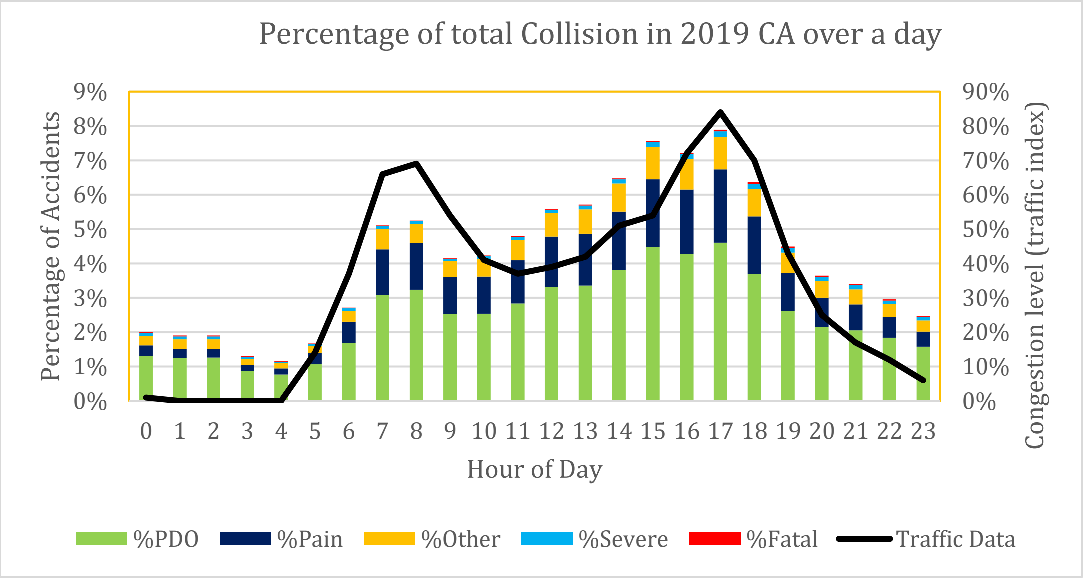


*Figure 10: Average collision per hour as % of total collision over a day*

### 5.2.3.2. Age

At-risk age groups for various collision severities could be evident through a frequency analysis of the age of drivers involved in a collision (see Appendix 9.6 for detail). To get a more accurate impact of age on the collision severity, we should normalize the collision distribution against the demographic size of registered drivers. Due to unavailability of this data, we use the demographic size of general population in 2019 obtained from “US Census Bureau” (US Census Bureau, 2022) (Sub Urban Stats, 2020). As is evident from Figure 11, we see that the biggest

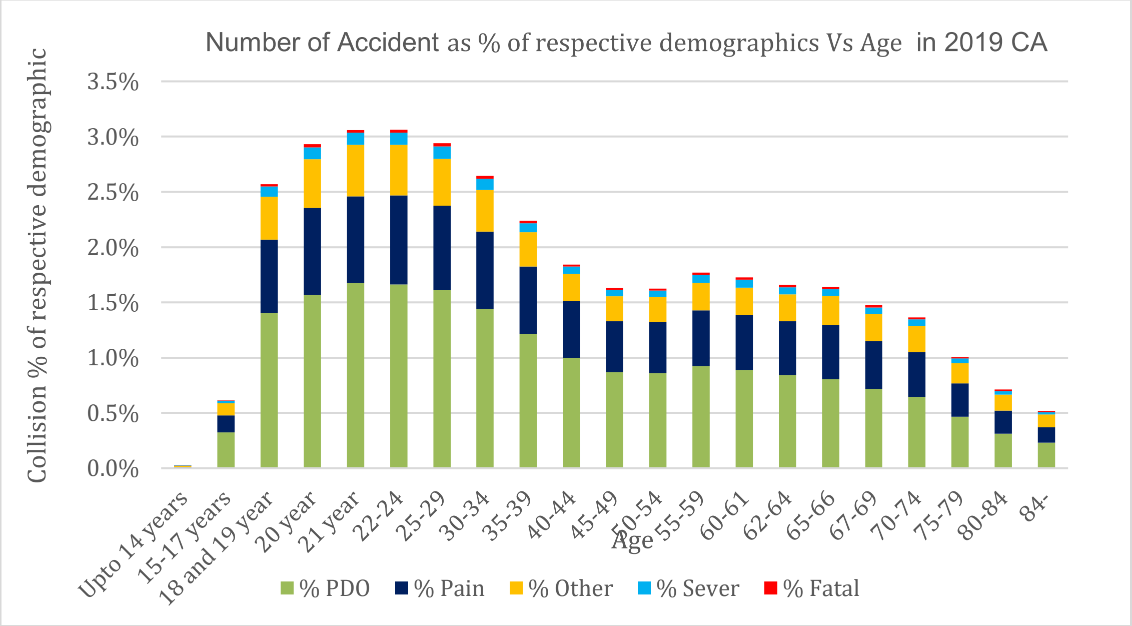


*Figure 11 Number of accidents as a percentage of demographic in 2019 California*

at-risk group is teens (15 – 21 years). The decrease in collision percentage in ages 15-17 may be because not all people in the age group drive. As driving becomes more common as people age (21-24 years), 2.5% - 3% of the demographic were involved in collision. As people mature, they drive more responsibly as is evident drop in collision percentage to 1.5% by age 40, and further decline from 1.5% - 1.2% for people in age group 40 -64 years. We see a continues drop of collision by senior drivers (0.5%). We see similar trends in all collisions, irrespective of the collision severity.

#### 5.2.3.3. Population

The third most important factor from our models is the population size. We created histogram of the frequencies of collision severities (as a percent of total collisions in 2019) over population size classification as shown in Figure 12. (See Appendix 9.8 for detail). We see that 25% of accidents occur in unincorporated areas for almost every severity classification (excluding pain). This might be because these areas may lack basic services such as active police force, proper infrastructure, etc. We also see a rising trend in the number of collisions as the population size grows. This holds true for all severities, demonstrating that large, densely populated areas are most at risk for a serious accident.

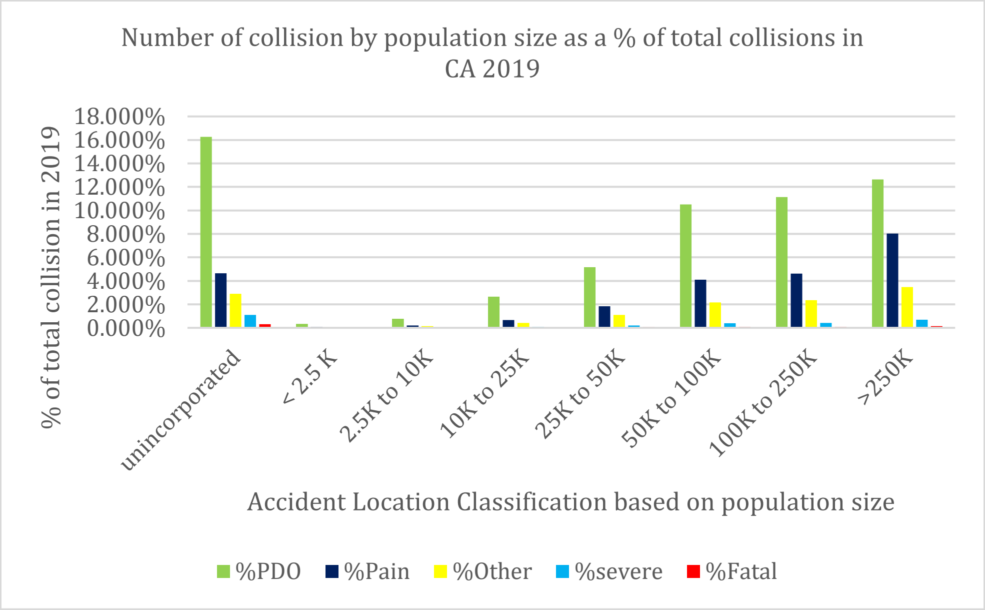


*Figure 12: Number of collisions by Population size classification as a % of total Collision in CA 2019*

### 5.2.4. Model Design and Accuracy

A common way to evaluate a multi-classification model such as ours is the confusion matrix. It's a matrix in which the number of correct and incorrect predictions are summarized with count values and broken down by collision severities. The metric we derive from the matrix is the accuracy (number of all correct predictions/ total dataset), precision, and recall. Precision is the number of predicted cases of a severity that were actually correct. Recall is defined to be the number of true cases of a severity that were predicted correctly (Shung, 2018)

$$Precision_i = \frac{True\ positive}{Actual\ Results} = \frac{True\ positive}{True\ Positive + False\ positive} = \frac{M_{ii}}{\sum_j M_{ij}}$$

$$Recall_i = \frac{True\ positive}{Predicted\ Results} = \frac{True\ positive}{True\ Positive + False\ Negative} = \frac{M_{ii}}{\sum_j M_{ji}}$$

where M is the confusion matrix; i and j are the index of row and columns. A metric to understand both the precision and recall is the F1 Score (Baeldung, 2020 ), which is the harmonic mean of the two as shown below.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + \ Recall}$$

F1 Score and all other metrics for the models are listed in Table 5. The other injury classification performed poorly (0.3% accuracy) and interfered with the accuracies of other severities, hence we removed other injury for the classification purpose.

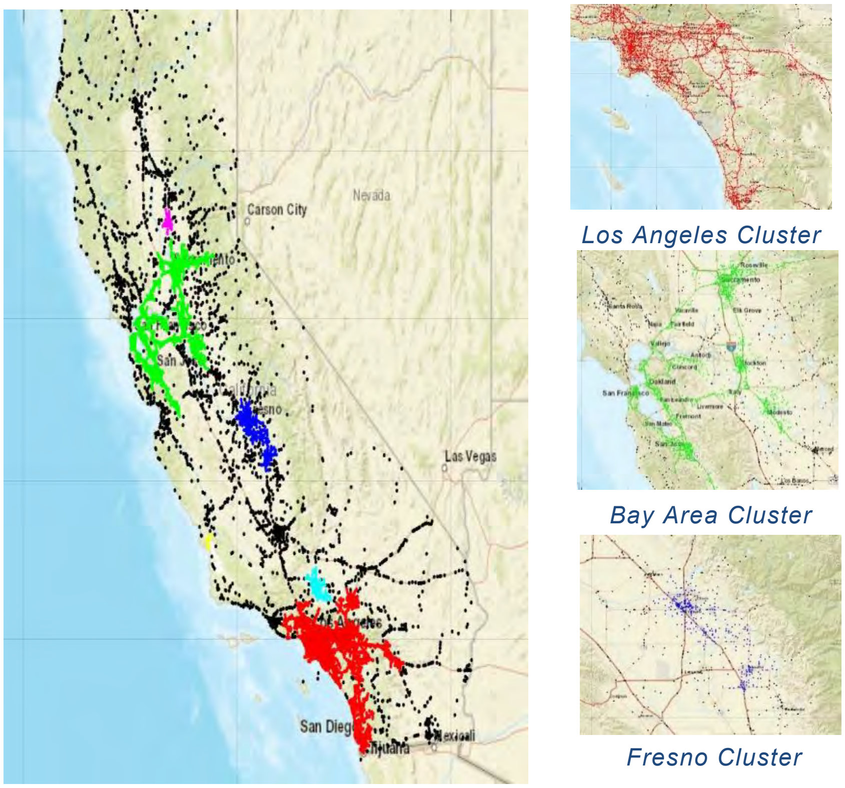
*Table 5: Accuracy, Precision, Recall and F1 Score for the models*

| Model Name | Accuracy | Collision Severity | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Decision Tree | 42.70% | Property Damage Only | 60.70% | 39.50% | 47.90% |
| | | Pain | 32.50% | 34.40% | 33.42% |
| | | Severe Injury | 48.00% | 48.20% | 48.10% |
| | | Fatal | 20.30% | 83.80% | 32.70% |
| Random Forest | 44.30% | Property Damage Only | 40.70% | 40.40% | 40.40% |
| | | Pain | 40.60% | 38.80% | 39.70% |
| | | Severe Injury | 44.40% | 45.30% | 44.80% |
| | | Fatal | 33.60% | 55.50% | 42.00% |
| Naïve Bayes | 55.70% | Property Damage Only | 60.20% | 52.20% | 55.90% |
| | | Pain | 55.50% | 52.20% | 53.80% |
| | | Severe Injury | 53.30% | 59.80% | 56.40% |
| | | Fatal | 51.10% | 66.50% | 57.80% |

## 5.3. Risk Analysis

### 5.3.1. Regional Risk Analysis

We then conduct geolocation analysis, using MATLAB's DBSCAN algorithm, on the data derived from FARS database to determine areas of highest accident density clusters called "accident hotspots". We use FARS dataset because it is better at providing data from highways and intersections, even though it only covers fatal accidents. We found that collisions in California are concentrated largely around three ”accident hotspots” located in Southern LA, Central California, and the Bay Area (Figure 13).



Los Angeles Cluster

Bay Area Cluster

Fresno Cluster

Figure 13: Three Most significant accident Clusters in California

The Los Angeles cluster is the largest and most dense region of accidents in California. To find specific dense areas within the LA Cluster, we created a heat map where denser areas are more brightly colored. This heat map shown in Figure 14 describes the region by aggregating

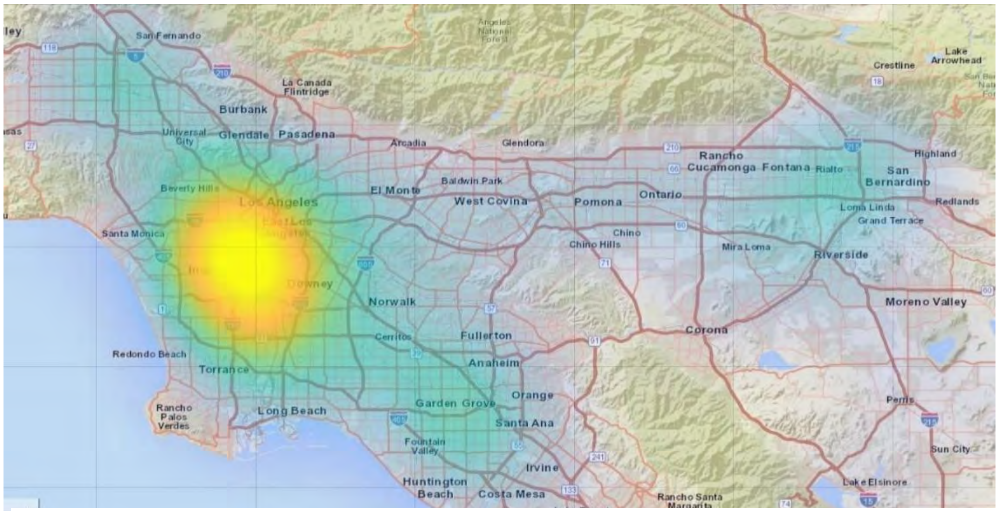


Figure 14: Heatmap for LA Area

the collisions per square mile. As apparent by the figure, the density of collisions in LA is unevenly distributed and largely focused on Southern LA, (defined as the region east of Inglewood and west of Highway 110), followed by Downtown LA, Beverly Hills, Torrance, Garden Grove, Pomona, San Bernadino, Long Beach, and Downtown San Diego. The Bay Area cluster mainly surrounds the coast of the San Francisco Bay and in the San Francisco city center. Here, collisions are centered around Downtown Sacramento, Stockton, Modesto, San Jose, and San Francisco. In central California, accidents are largely clustered around Downtown Fresno and Visalia.

### 5.3.2. Direct Risk to Automotive Insurance industry

From both our analysis of the influential factors controlling accident severity and our geo-spatial analysis, several risk groups were identified. While the primary risk group in a collision are the drivers and passengers involved, a large part of the risk of those insured is passed on to their insurance companies. Hence, it follows that a higher severity collision indicates a higher risk for both the drivers and insurance companies. As a result, we have determined that the insurance companies largely at risk are those that insure drivers exposed to the conditions and locations described in this past section.

In 2010 alone, insurance corporations lost 42 billion to covering medical expenses due to injuries caused by an accident. This means that the severity of the accident plays a significant role in the loss of the insurance provider. As calculated by Blincoe et all (Lawrence, Miller, Zaloshnja, & Lawrence, 2015) the average medical bill for a non-fatal accident after adjusting for inflation is as follows: for property damage has a unit cost of $3,148.30, a complaint of pain costs $5,379.42, other injury costs $6,099.45, a severe injury costs around $25,946.87. Automotive insurance (the primary insurance provider) may have to pay more if they are covering the at-fault driver.

Several private auto insurance brokers are active in California: State Farm, Farmers Insurance, Geico, Berkshire Hathaway, Allstate, Auto Club Exchange, Mercury Insurance, Kemper, and Progressive which accounts for 82.35% of the auto insurance market (EverQuote, 2019). As a result, we predict that these companies will take the largest hit as collisions rise during the employment rebound after COVID-19.

### 5.3.3. Ancillary Risk and at-risk Subgroups

The primary risk group are the drivers and passengers who were directly involved in the accident. The drivers most at risk to an accident or a fatal crash are in the “Accident hotspots” explained above and are exposed to the influential factors highlighted in section 5.2. For a driver involved in a collision there are multiple risk such as loss of life, the medical costs associated with injury, the loss to the quality of life and the property damage done to the driver or others involved.

When a victim is injured but survives, not only is there a medical and property cost, but also a loss of the quality of life. Depending on the degree to which the victim is affected, they can suffer pain and distress. In 2019, Americans spend over one million days in the hospital each year from crash injuries (NCIPC, CDC, 2017). More than 2.5 million Americans went to the emergency department (ED) and nearly 200,000 were then hospitalized for crash injuries in 2012 (NCIPC, CDC, 2017). Lifetime work lost because of 2012 crash injuries cost an estimated $33 billion· (NCIPC, CDC, 2017). Bryant et. Al (Bryant, 2021) states that settlement claims for pain and suffering can range anywhere from a few thousand dollars to an upwards of $250,000 or $500,000 dollars depending on the severity of the accident and its resulting impacts on the driver's emotional health.

There is also the risk of damage to the driver's car or other property, which has shown to be the most common instance of collision. While the collision liability insurance (required in California) of the at-fault driver often pays to repair or replace the victim's car, there is no coverage for an at-fault driver that is uninsured or underinsured to pay the victims damages. From our model, we see that in 2019 about 14% of Californians involved in an accident lacked insurance (CHP, California, 2022). Hence these group not only have the risk of economic ruin, if involved in crash, but also legal action since it's illegal to drive without basic insurance in CA. In these cases, having Underinsured/Uninsured Motorist Property Damage Insurance is a way for drivers to be sure that their property will always be covered in a collision.

Even if a driver is insured, they still might be subject to paying a deductible and/or pay out-of-pocket if the policy limit has been reached. Both cases are more likely/serious in a severe accident resulting in a greater risk for these drivers. As a result, having a plan with a lower deductible or a higher limit would reduce the risk. Additionally, drivers who can afford to buy additional insurance buy medical payment insurance, which covers the medical bill and usually does not have any deductibles. Having this form of insurance would significantly reduce the risk for the drivers themselves.

While the danger towards drivers and others on the roads is large focus of our model, the subsequent risk towards productivity is known to exceed $20 billion in California (Lawrence, Miller, Zaloshnja, & Lawrence, 2015). Not only the productivity of the driver/occupants is at risk, but also those affected by the subsequent congestion created by the crash, in terms of lost work time. Additionally, out of the many regions created by our clustering model, the Los Angeles Region held the highest frequency of all severities and was the densest. Since the Los Angeles also contains many traffic intensive highways and the highest level of traffic in the country (Mobility Division,, 2021), we consider it to be an especially high-risk cluster.

## 5.4. Strengths and Weakness

Our model provided practical results on the most influential factors that impact traffic severity. This information is useful in characterizing the risk that should be addressed through both insurance and public policy. One of the largest strengths of our model relates to our multiple-model approach. For example, this allows us to corroborate our Random Forest factor importance results with that from our Naïve Bayes model, confirming that there is no bias in the metric or model. Additionally, a multiple-model approach allows us to gain different insights into the data. For example, running the Decision Tree gives an understanding of the exact decision process whereas Random Forest does not.

Additionally, by implementing DBSCAN and geo-analysis using our latitude-longitude datasets, we were able to find general and specific areas of collisions that should be addressed by state and local officials. However, a disadvantage of DBSCAN is that it cannot create clusters with varying densities which is more often the case with traffic accident location.

Also, as demonstrated by the poor accuracy and other metric scores of our model, we did not have enough factors to fully describe the causes of collision severity. If provided with more relevant factors, we would be able to improve the performance of our model. Examples of such factors are the accident-avoidance technology of the car, autonomous driving level classification, more granular car model classification, safety equipment, distance from hospital, and the vehicle's crash test score, registered vehicle driver demographics to name a few. Moreover, we assumed that the factors impact collision frequency and severity in all clusters similarly. We did analysis to determine factors affecting collision for LA cluster and applied the lessons learned for all clusters in California. It would be better to apply the model for Bay Area and Fresno clusters individually and determine any special factors affecting them.

# 6. Recommendation

## 6.1. Insurance Recommendation

Established in 1983, California legislature enacted the Compulsory Financial Responsibility Law requiring all vehicles on Californian roadways to have a form of liability insurance. Based on the future trend in Fig 5, we predict that as the number of accidents will increase in the aftermath of the COVID pandemic, the premiums of auto insurance companies will also likely increase. This will disadvantage low-income individuals who would be unable to pay these premiums and be unable to drive legally as a result and put themselves in danger of losing their financial stability. In response, we recommend not only to enforce a stricter regulation of the insurance mandate, but also increased funding of "California's Low-Cost Automobile Program" (CLCA). (Insurance, 2022) CLCA is a California insurance program introduced in 1999 as a method to provide financially disadvantaged people and families with affordable insurance rates and drive legally.

The CAARP program (Insurance, 2022) caters towards high-risk individuals who have been involved in multiple accidents/tickets and are not able to find a standard insurance company that will insure them. All insurance companies in California are required to accept CAARP and we recommend that legislature funds and improve this program.

Google aims to prevent 100 million (Bradshaw, 2021) accidents each year, with a new "safe route" option in Google Maps. It uses similar factors used in our model, to suggest a "safer" alternative route. We recommend insurance companies to encourage drivers to proactively adopt similar accident-avoidance technologies.

## 6.2. Policy Recommendation

Addressing the root causes of vehicle collision in California, requires an in-depth analysis of government policies. As our model determines, time of collision, one of the most impactful factors affecting the intensity of collisions, correlates strongly with traffic congestion as shown in Figure 10, especially in California hotspots (section 5.3.1). As California is expected to receive $45.5 billion from the $1.2 trillion infrastructure bill recently introduced by President Biden, (Walters, 2021) we recommend that we use part of the $9.45 million that's allocated for public transport to fund Los Angeles Department of Transportations' (DOT) NextGen Bus Plan. We suggest focusing on Southern Los Angeles around the Inglewood and Hawthorne area to simultaneously tackle both congestion and driver safety. Similarly, in the Bay Area, we recommend using the funding to enable Caltrain's commuter rail service (Caltrain, 2022) to complete electrification and to extend Bay Area Rapid Transport (BART) further into the Silicon Valley (BART, 2022).

We also see that unincorporated areas faced the highest risk of collision in almost every severity (Section 5.2.3.3). We hypothesized that this might be due to a lack of police activity or traffic regulation in these areas, leading to more reckless driving behavior and more dangerous collisions (McCarthy, 1999). Hence, we recommend to use the infrastructure funding to improve rural infrastructure such as more active traffic police involvement, DUI enforcement, traffic lights, improved lighting, and setup of traffic cameras for improved vehicle regulation.

Additionally, our frequency analysis found that drivers from the age of 17-22 were most at risk (Figure 11) of not only being involved in a crash but also being involved in a serious accident. As a result, State of California has "Graduated driver licensing (GDL)" through a learner's permit which requires a graduated granting of driving privileges for a learning driver. Specifically, we recommend the passing of Assembly Bill AB-2388 (Villapudua, 2022) which will require graduated licensing from age 18 currently to the age 21. This bill, if passed, will help many young, learning drivers gain driving experience safely.

From our Decision Tree, we found that the influence of alcohol on a driver is more likely to increase the severity of a collision as shown in section 5.2.2. In fact, 95.5% of cases involving a combination of drunk driving and the use of drugs resulted in a fatal accident. California's Office of Traffic Safety (OTS) has created the California Impaired Driving Plan (CIDP) (OTS, 2020) that implements DUI / DUID treatment programs and advocate for stricter penalties as perceived risk of arrest has proven to be the strongest deterrent to impaired driving. We recommend continuing and strengthening these programs through additional funding. We also advocate the use of Ignition Interlock devices (IIDs) for all convicted drunk driving offense. (NCIPC, CDC, 2016) This device keeps the vehicle from starting unless the driver has a BAC below a pre-set limit. The International Council on Alcohol, Drugs and Traffic Safety maintains that IIDs, when combined with proper monitoring program, leads to a 40–95% reduction in the rate of repeat drunk driving offenses (DMV - CA, 2022).We also would recommend to fund and strengthen publicized sobriety checkpoints where police can check for DUI. Lastly, we encourage to continue enforcement of minimum drinking age (NCIPC, CDC, 2017).

The frequency analysis of the time-of-day factor (section 5.2.3.1) demonstrated collision peaks at morning (7-8 AM) and evening (3-7 PM) which correlates with work commute traffic as corroborated by the congestion report (TomTom Inc, 2022). As evident in Figure 4, collision accident frequency follows economy trends closely. As a result, we believe that work commute is one of the major causes of these traffic peaks. The state of California already encourages employees to use other forms of work commute through its Bicycle Commuter and “Mass Transit and Vanpool” programs. (CAL-HR, 2022). We recommend that such programs be encouraged for a broader workforce as well.

## 6.3. Concluding remarks

In total, our model and analysis demonstrated that several factors, especially the time of collision, the age of the driver, and the population of the area contributed to the severity of vehicle accidents. Not only the drivers and victims are at risk, but also the insurance companies, other commuters, employers, and the economy are affected, especially people living in hotspots such as LA, Bay Area, Fresno, and Visalia. We recommend the improvement of public transportation within these areas by implementing Los Angeles's NextGen Bus Plan, and improving Caltrain and BART, as well as improving infrastructure and increased traffic enforcement in unincorporated areas. We recommend the passing of Assembly Bill AB-2388 that will require the graduated licensing up to the age of 21. To reduce the risk due to impaired driving, we suggest the California State expands CIDP treatment programs, advocate the use of IIDs for all convicted drunk driving offense, fund and strengthen publicized sobriety checkpoints, and continue enforcement of minimum legal drinking age. We suggest that the state encourage work commute programs to a broader workforce. We also recommend the funding and continuation of the CAACP and the CAARP programs to increase insurance affordability to underrepresented or impoverished individuals.

# 7. References

1. 2u.Inc. (2021, June). *What Is Undersampling?* From Matser's in Data Sciences: https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/#:~:text=Undersampling%20is%20a%20technique%20to,information%20from%20originally%20imbalanced%20datasets.

2. Baeldung. (2020 , October 19). *F-1 Score for Multi-Class Classification*. Retrieved March 5, 2022 from Baeldung: https://www.baeldung.com/cs/multi-class-f1-score

3. BART. (2022). *About*. From Bay Area Rapir Transit: https://www.bart.gov/about

4. Borrelli, L. (2021, June 26). *Car ownership statistics*. From Bankrate.com: https://www.bankrate.com/insurance/car/car-ownership-statistic

5. Bradshaw, K. (2021, May 18). *Google Maps gaining 'safe routing' option to help avoid accidents.* From 9To5Google: https://9to5google.com/2021/05/18/google-maps-safe-routing-avoid-accidents/

6. Bryant, D. (2021, December 22). *How Pain And Suffering Damages Are Calculated In Car Accident Settlements*. From David Bryant Law, PLLC: https://davidbryantlaw.com/blog/car-accident-settlement-pain-and-suffering/

7. Buket, G., & Kara, M. (2020, June 26). Severity Prediction with Machine Learning Methods. *2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 1-7. doi:10.1109/HORA49412.2020.9152601

8. CAL-HR. (2022). *Commute Programs*. (California Department of Human Resources) From CA.Gov, CALHR: https://www.calhr.ca.gov/employees/Pages/Commute-Program.aspx

9. Caltrain. (2022). *About*. From Caltrain: https://www.caltrain.com/about.html

10. Carlson, S. T. (June 2010). BRIEF REPORT: BARTLETT'S TEST OF SPHERICITY AND CHANCE FINDINGS IN FACTOR ANALYSIS. *Multivariate Behavioural Research*, 375-377.

11. CDI. (2019). *2019 CALIFORNIA P&C PREMIUM AND LOSS SUMMARY.* Retrieved March 5, 2022 from California Department of Insurance: http://www.insurance.ca.gov/01-consumers/120-company/04-mrktshare/2019/upload/MktShrSummary2019wa.pdf

12. CHP, California. (2022). *SWITRS - Statewide Integrated Traffic Records System.* (California Highway Patrol, State of California) Retrieved March 20, 2022 from CA.Gov,

California Highway Patrol: https://www.chp.ca.gov/programs-services/services-information/switrs-internet-statewide-integrated-traffic-records-system

13. Citywide Law Group. (2022). *Los Angeles Car Accident Statistics.* From Citywide Law Group: https://www.citywidelaw.com/los-angeles-car-accident-attorney/los-angeles-car-accident-statistics/

14. Costello, A. &. (2005, January). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation. 10.*, 1-9.

15. DMV - CA. (2022). *Ignition Interlock Devices California DMV*. (State of California) Retrieved March 6, 2022 from CA.Gov, Department of Motor Vehicle: https://www.dmv.ca.gov/portal/driver-education-and-safety/educational-materials/fast-facts/ignition-interlock-devices-ffdl-31/

16. EverQuote. (2019, November 6). *Largest Auto Insurance Companies in California*. From EverQuote: https://www.everquote.com/blog/car-insurance/largest-companies-california/

17. Fabrigar, L. R. (2011). Exploratory factor analysis. In *Exploratory factor analysis.* Oxford University Press.

18. FHWA, US-DOT. (2017, January 19). *KABCO Injury Classification Scale and Definitions*. (US Department of Transportation) Retrieved March 5, 2022 from Federal Highway Administration, Safety: https://www.safety.fhwa.dot.gov/hsip/spm/conversion_tbl/pdfs/kabco_ctable_by_state.pdf

19. Gude, A. (2021, August 2021). *California Traffic Collision Data from SWITRS*. (Kaggle) Retrieved January 15, 2022 from Kaggle.com: https://www.kaggle.com/alexgude/california-traffic-collision-data-from-switrs

20. IBM Inc. (2022). *IBM SPSS Statistics*. Retrieved January 15, 2022 from IBM: https://www.ibm.com/products/spss-statistics

21. Insurance, C. D. (2022). *California's Low Cost Auto Insurance Program*. (California Department of Insurance, California Automobile Assigned Risk Plan) From California Department of Insurance: http://www.insurance.ca.gov/01-consumers/105-type/95-guides/01-auto/lca/

22. Lawrence, B., Miller, T., Zaloshnja, E., & Lawrence, B. (2015). *The Economic and Societal Impact Of Motor Vehicle Crashes, 2010 (Revised).* National Highway Traffic Safety Administration, US Department of Transportation. Washington, DC: National

Highway Traffic Safety Administration, US Department of Transportation. doi:DOT HS 812 013

23. MathWorks Inc. (2022). *Mapping Toolbox - MATLAB*. Retrieved February 10, 2022 from MathWorks: https://www.mathworks.com/products/mapping.html.

24. MathWorks Inc. (2022). *Statistics and Machine Learning Toolbox, Analyze and model data using statistics and machine learning*. Retrieved February 10, 2022 from MathWorks: https://www.mathworks.com/products/statistics.html

25. McCarthy, P. C. (1999). Public policy and highway safety: a city-wide perspective. *Regional Science and Urban Economics*, 231-244.

26. Miaomiao, Y., & Yindong , S. (2022). Traffic Accident Severity Prediction Based on Random Forest. *Sustainability*, 14. doi:10.3390/su14031729

27. Mobility Division,. (2021, June). *Urban Mobility Report — Mobility.* Texas A&M Transportation Institute, Mobility Division. College Station, Texas: Texas A&M University. Retrieved March 6, 2022 from ,Texas A&M Transportation Institute: https://mobility.tamu.edu/umr/

28. NCHS,CDC. (2022, January 13). *Leading Causes of Death*. (National Center for Health Statistics, Center for Disease Control and Prevention, US Department of Health & Human Services) Retrieved March 6, 2022 from Center for Disease Control and Prevention: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

29. NCIPC, CDC. (2016, July 6). *Motor Vehicle Crash Deaths, How is the US doing?* ( National Center for Injury Prevention and Control, US Department of Health & Human Services) Retrieved March 6, 2022 from Center for Disease Control and Prevention,: https://www.cdc.gov/vitalsigns/motor-vehicle-safety/index.html

30. NCIPC, CDC. (2017, October 2014). *Motor Vehicle Crash Injuries, Costly but preventable.* ( National Center for Injury Prevention and Control, Center for Disease Control and Prevention, US Department of Health & Human Services) Retrieved March 6, 2022 from Center for Disease Control and Prevention: https://www.cdc.gov/vitalsigns/crash-injuries/index.html

31. NCSA, NHTSA , US-DOT. (2019, September). *Summary of Motor Vehicle Crashes*. (National Center for Statistics and Analysis, National Highway Traffic Safety Administration, US Department of Transportation) doi:DOT HS 813 209

32. NHTSA, U.-D. (2022). *Crash Report Sampling System*. (National Highway Traffic Safety Administration, US Department of Transportation) Retrieved February 15, 2022 from

National Highway Traffic Safety Administration: https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system

33. NHTSA, US-DOT. (2022). *Fatality Analysis Reporting System (FARS)*. (National Highway Traffic Safety Administration, US Department of Transportation) Retrieved March 6, 2022 from National Highway Traffic Safety Administration: https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

34. OTS. (2020). *California Impaired Driving Plan 2020.* From California Office of Traffic Safety: https://www.ots.ca.gov/wp-content/uploads/sites/67/2021/03/California-Impaired-Driving-Plan_Remediation.pdf

35. Sharma, A. (2020, June 30). *Four Simple Ways to Split a Decision Tree in Machine Learning*. Retrieved March 5, 2022 from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/

36. Shung, K. P. (2018, March 15). *Accuracy, Precision, Recall or F1?* From Towards Data Science: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

37. Sky, A. (2021, December 7). *36 Important US Automotive Industry Statistics [2022]: Facts, Trends, And Projections*. Retrieved March 5, 2022 from Zippia: https://www.zippia.com/advice/automotive-industry-statistics/.

38. Sub Urban Stats. (2020). *California Population demographics 2020, 2019*. Retrieved March 5, 2022 from Suburban Stats Inc: https://suburbanstats.org/population/how-many-people-live-in-california

39. TomTom Inc. (2022). *Los Angeles traffic report*. Retrieved March 5, 2022 from TomTom Inc: https://www.tomtom.com/en_gb/traffic-index/los-angeles-traffic

40. TRIP. (2016). *California Transportation by the Numbers: Meeting the State's Need for Safe, Smooth and Efficient Mobility." California Transportation by the Numbers.* Washington DC: TRIP - National Transportation Research Nonprofit. doi:202-466-6706

41. US Census Bureau. (2022, March 4). *Census.gov.* (US Census Bureau, US Department of Commerce) Retrieved March 6, 2022 from US Census Bureau: https://www.census.gov

42. USBLS. (2022, March 4). *Automotive Industry: Employment, Earnings, and Hours*. (US Department of Labor) Retrieved March 6, 2022 from US Bureau of Labor Statistics: https://www.bls.gov/iag/tgs/iagauto.htm

43. USBLS,. (2022, March 2). *Local area Unemployment statistics – seasonally adjusted*. (US Department of Labor) From US. Bureau of Labor Statistics, US Department of Labor.: https://www.bls.gov/data/

44. Vadapalli, P. (2021, January 5). *Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2022*. (upGrad) Retrieved March 5, 2022 from upGrad: https://www.upgrad.com/blog/naive-bayes-explained/

45. Villapudua. (2022). *Assembly Bill 2388.* California Legislature 2021-22 Session. Legislative Counsel's Digest. CA. doi:2021AB2388_99

46. Walters, D. (2021, November 9). *California gets small share of infrastructure bill*. Retrieved March 6, 2022 from Cal Matters: https://calmatters.org/commentary/2021/11/california-gets-small-share-of-infrastructure-bill/

## 8. Acknowledgements

Our heartfelt thanks to the following people, who gave us the resources and opportunities to make this wonderful experience happen.

1. Our team Coach, Mr. Suresh Bardhwaj

2. Our Mentors for Modeling the Future Challenge – Mr. Alexander Bai, ASA, MAAA, and Mr. Zachary Johantges

3. All the Members and supporters of the Actuarial Foundation

We would like to thank the Modeling the Future Competition, The Institute of Competition Sciences, The Actuarial Foundation, and its sponsors for organizing this competition and giving us a chance to conduct educational and illuminating research. Taking part in the MTFC challenge has opened a new world of data science and has given us a greater appreciation for the importance and impact of actuarial work. This experience has changed the way we view society’s issues and wish that this program grows so that more students are exposed to the same valuable experiences as we had.

# 9. Appendix

## 9.1. Appendix 1: Data Factors

This section describes the relevant factors of the environmental conditions of the accident, the driver behavior and crash location used in our model.

| Factor | Description | Data Type | Value | Definition |
|---|---|---|---|---|
| Age | Age of the driver at the time of collision | Ratio | | The age of the driver in years |
| Alcohol involved | Indicates collision involved a party that had been drinking | Comparative | 0 | Alcohol not involved |
| | | | 1 | Alcohol involved |
| Cellphone in use | Classification based on if the party is using cell phone | Comparative | 0 | No |
| | | | 1 | Yes |
| Collision Time | The time when the collision occurred (24 hr. time) | Ratio | | Time when the accident occurred in military time |
| Drunk Driving | Classification based on whether the primary party is intoxicated | Comparative | 1 | Not due to drunk driving |
| | | | 2 | Due to drunk driving |
| Financial responsibility | Classification based on whether the party showed proof of insurance at the time of collision | Comparative | 0 | no proof of insurance obtained |
| | | | 1 | proof of insurance obtained |
| Intersection | Classification based if collision occurred in an intersection | Comparative | 0 | Not an intersection |
| | | | 1 | It is an intersection |
| Intersection Type | Classification based on intersection type | Comparative | 1 | Not an intersection |
| | | | 2 | Four way intersection |
| | | | 3 | T intersection |
| Latitude | Y- coordinate of the geocoded location of the collision | Comparative | degree | |
| Longitude | X- coordinate of the geocoded location of the collision | Comparative | degree | |
| Lighting Condition | Classification based on how bright location is at the time of collision | Comparative | 1 | Dusk |
| | | | 2 | DARK-Not Lighted |
| | | | 3 | Dark-Unknown |
| | | | 4 | DARK-Lighted |
| | | | 5 | Dawn |
| | | | 6 | Day-Light |

| Factor | Description | Data Type | Value | Definition |
|---|---|---|---|---|
| Party drug Impairment | Classification based on physical or drug induced impairment | Comparative | 1 | not applicable or 'G' |
| | | | 2 | under drug influence |
| | | | 3 | sleepy/fatigued |
| | | | 4 | impairment - physical |
| Party gender | Primary party's gender Classification | Nominal | 1 | Male |
| | | | 2 | Female |
| Population | Population size classification at the collision zip code | Comparative | 1 | unincorporated |
| | | | 2 | <2500 |
| | | | 3 | 2500 to 10000 |
| | | | 4 | 10000 to 25000 |
| | | | 5 | 25000 to 50000 |
| | | | 6 | 50000 to 100000 |
| | | | 7 | 100000 to 250000 |
| | | | 8 | >250000 |
| Road Condition | Classification based on road condition | Nominal | 1 | Normal |
| | | | 2 | Construction’ |
| | | | 3 | Flooded |
| | | | 4 | Holes |
| | | | 5 | Loose Material’ |
| | | | 6 | Obstruction’ |
| | | | 7 | Reduced width’ |
| | | | 8 | Other |
| Road Surface | Classification based on slipperiness of the road at the time of the collision | Comparative | 1 | Dry |
| | | | 2 | Wet |
| | | | 3 | Snowy or Icy |
| | | | 4 | Slippery (Muddy, Oily, etc,) |
| Vehicle make | Classification based on the vehicle make of the primary party's vehicle | Nominal | | Make of the primary party's vehicle |
| Vehicle year | Model year of the party's vehicle | Interval | | Model year of the primary vehicle model |
| Weather | Classification based weather condition at the time of collision | Comparative | 1 | Severe Crosswinds |
| | | | 2 | Snow |
| | | | 3 | Rain |
| | | | 4 | Cloudy |
| | | | 5 | Clear |

## 9.2. Appendix 2: Create Table.sql

Data derived from SWITRS database for the year 2000 - 2021.

```
CREATE TABLE collisiontable1M(case_id PRIMARY KEY, collision_severity TEXT, latitude REAL,
longitude REAL, population TEXT, lighting TEXT, collision_time TEXT, intersection INT,
location_type TEXT, road_surface TEXT, road_condition_1 TEXT, weather_1 TEXT,
alcohol_involved INT, at_fault INT, financial_responsibility TEXT, party_age INT, party_sex
TEXT, party_drug_physical TEXT, cellphone_in_use TEXT, vehicle_make TEXT, vehicle_year
INTEGER);


INSERT INTO collisiontable1M select
     collisions.case_id, collisions.collision_severity,
        collisions.latitude, collisions.longitude,
        collisions.population, collisions.lighting,
        collisions.collision_time, collisions.intersection,
        collisions.location_type, collisions.road_surface,
        collisions.road_condition_1, collisions.weather_1,
        collisions.alcohol_involved, parties.at_fault,
        parties.financial_responsibility,
        parties.party_age,
        parties.party_sex,
        parties.party_drug_physical,
        parties.cellphone_in_use,
        parties.vehicle_make,
        parties.vehicle_year
from collisions
inner join parties on collisions.case_id = parties.case_id
where parties.at_fault = 1 AND
         collisions.latitude IS NOT NULL AND
         collisions.longitude IS NOT NULL AND
         collisions.population IS NOT NULL AND
         collisions.lighting IS NOT NULL AND
         collisions.collision_time IS NOT NULL AND
         collisions.intersection IS NOT NULL AND
         collisions.location_type IS NOT NULL AND
         collisions.road_surface IS NOT NULL AND
         collisions.road_condition_1 IS NOT NULL AND
         collisions.weather_1 IS NOT NULL AND
         parties.financial_responsibility IS NOT NULL AND
         parties.party_age IS NOT NULL AND
         parties.party_sex IS NOT NULL AND
         parties.party_drug_physical IS NOT NULL AND
         p.cellphone_in_use IS NOT NULL AND
         parties.vehicle_make IS NOT NULL AND
         parties.vehicle_year IS NOT NULL
```

## 9.3. Appendix 3: POCDataClusters.m

```
%==========================================================================
% File      : POCDataClusters.m
% Project   : Modelling the future Challange
% Detail    : This is the main code.
%             Analying the eefct of light conditions, factor 2 on the
%             number of accidents in the Saceremento and LA Clusters
%==========================================================================

% Constant declaration ----------------------------------------------------
baseProjectDataDir = "Y:\59_MTFC Math Competition\0_Project Data\";
accDatafileName = baseProjectDataDir + "ashba_LocationDataWithFactorsZipCode_Active";
LightconditionFreqDataFilename = baseProjectDataDir +
"Factor_LightCondition_FreqData.xlsx";
WeatherFreqDataFilename = baseProjectDataDir + "Factor_Weather_FreqData.xlsx";
DrunkDrivingDataFilename = baseProjectDataDir + "Factor_DrunkDriving_FreqData.xlsx";
IntersectionDataFilename = baseProjectDataDir + "Factor_Intersection_FreqData.xlsx";
SeasonDataFilename = baseProjectDataDir + "Factor_Season_FreqData.xlsx";

% Cleanup -----------------------------------------------------------------
Utility.DeleteFile(LightconditionFreqDataFilename);
Utility.DeleteFile(WeatherFreqDataFilename);
Utility.DeleteFile(DrunkDrivingDataFilename);
Utility.DeleteFile(IntersectionDataFilename);
Utility.DeleteFile(SeasonDataFilename);
%Utility.DeleteFile(RandomForestInputTableFilename);

% Read the accident data from file to a matrix-----------------------------
data = readcell(accDatafileName);
header = data(1,:);
data(1,:) = [];

% Instantiate the cluster algorithm ---------------------------------------
 dbScanCluster = DBScanCluster(data, 2, 3);
 ClusterIndexList = dbScanCluster.Execute(5);
 dbScanCluster.Plot();

% Get the Data for biggest cluster ----------------------------------------
 %biggestClusterIndex = dbScanCluster.GetBiggestClusterIndex();
 biggestClusterData = dbScanCluster.GetCluster(3);
 %indexList = biggestClusterIndex * ones(size(biggestClusterData, 1),1);
 %dbScanCluster.PlotClusterData(biggestClusterData, indexList);
```

## 9.4. Appendix 4: DBScanCluster.m

```
%==========================================================================
% File      : DBScanCluster.m
% Project   : Math Challange
% Detail    : Create cluster based on the given data using density based
%             clustering algorithm
%==========================================================================
classdef DBScanCluster < handle
    % Predictive model using density based clustering algorithm

    properties
        Data;                  % accident data
        LatLngData;            % accident data
        ClusterIndexList;      % Cluster Index List
        ClusterSizeList;       % Cluster Size List
        OptimalEpsilon = 0.15; % optimal Epsilon
    end

    methods
        function obj = DBScanCluster(data, minPoints, maxPoints)
             obj.Data = data;
             obj.LatLngData = cell2mat(data(:,1:2));

             %Calculate Optimal epsilon for the given data
             obj.OptimalEpsilon = clusterDBSCAN.estimateEpsilon(obj.LatLngData,
minPoints, maxPoints);
        end

        function ClusterIndexList = Execute(obj, minNumOfPoints)
             % Performing density base clustering
             % input data - data, min number of Points

             ClusterIndexList = dbscan(obj.LatLngData,
                                       obj.OptimalEpsilon, minNumOfPoints);

             obj.ClusterIndexList = ClusterIndexList;
             obj.ClusterSizeList = UpdateClusterSize(obj, obj.ClusterIndexList);
        end
```

```
function clusterSizeList = UpdateClusterSize(obj, clusterIndexList)
    % Calulate size of each cluster
    % input - array of cluster numbers
    % output - array of size of each cluster
    clusterSizeList = zeros(max(clusterIndexList),1);
    for i = 1:size(obj.LatLngData)
         clusternumber = clusterIndexList(i);
        if(clusternumber ~= -1)
           clusterSizeList(clusternumber) = clusterSizeList(clusternumber) + 1;
        end
    end
end

function Plot(obj)
    % Create a Scatter Plot
    geoscatter(obj.LatLngData(:,1),
                obj.LatLngData(:,2), 5,
                obj.ClusterIndexList, "filled");

    geodensityplot(obj.LatLngData(:,1),
                   obj.LatLngData(:,2),[],"FaceColor","interp");

    geobasemap streets
    hold on
    %gscatter(obj.LatLngData(:,2), obj.LatLngData(:,1), obj.ClusterIndexList)
end

function biggestClusterIndex = GetBiggestClusterIndex(obj)
    % return the index of the biggest Cluster

    % Get the index of the BiggestcCluster
    biggestClusterIndexList =
     find(obj.ClusterSizeList == max(obj.ClusterSizeList));

    biggestClusterIndex = biggestClusterIndexList(1,1);
end

function clusterData = GetCluster(obj, index)
    % return the Data for a particular cluster index
    clusterData = cell(obj.ClusterSizeList(index), 13);
    j = 1;
    for i = 1:size(obj.ClusterIndexList)
        if obj.ClusterIndexList(i) == index
           clusterData(j,:) = obj.Data(i,:);
           j = j +1;
        end
    end
end
```

```
function PlotClusterData(~, data, indexList)
            % Create a Scatter Plot
            latlngData = cell2mat(data(2:end,3:4));
            gscatter(latlngData(:,1), latlngData(:,2), indexList(2:end,:))
        end
    end
end
```

## 9.5. Appendix 5: Data - Collision Severity

Monthly data derived from the SWITRS database. The percentage population is derived from “Labor static of California” from US bureau of California. Employment population ratio is defined as percentage of total population employed.

```
SELECT collisionYearMonth, Count(collisionYearMonth) AS Frequency
FROM collisions
where collision_severity = 'property damage only'
GROUP BY collisionYearMonth order By collisionYearMonth ASC

SELECT collisionYearMonth, Count(collisionYearMonth) AS Frequency
FROM collisions
where collision_severity = 'pain'
GROUP BY collisionYearMonth order By collisionYearMonth ASC

SELECT collisionYearMonth, Count(collisionYearMonth) AS Frequency
FROM collisions
where collision_severity = 'other injury'
GROUP BY collisionYearMonth order By collisionYearMonth ASC

SELECT collisionYearMonth, Count(collisionYearMonth) AS Frequency
FROM collisions
where collision_severity = 'severe injury'
GROUP BY collisionYearMonth order By collisionYearMonth ASC

SELECT collisionYearMonth, Count(collisionYearMonth) AS Frequency
FROM collisions
where collision_severity = 'fatal'
GROUP BY collisionYearMonth order By collisionYearMonth ASC
```

| YEAR | PDO | PAIN | OTHER INJURY | SEVERE INJURY | FATAL | ALL | CA EMPLOYMENT-POPULATION RATIO |
|---|---|---|---|---|---|---|---|
| **2001** | 26,464 | 10,062 | 5,838 | 890 | 293 | 43,547 | 63 |
| **2002** | 27,989 | 10,042 | 5,830 | 895 | 293 | 43,593 | 63 |
| **2003** | 27,588 | 10,360 | 5,761 | 894 | 311 | 44,913 | 61 |
| **2004** | 27,601 | 10,431 | 5,582 | 936 | 308 | 44,858 | 61 |
| **2005** | 27,516 | 10,206 | 5,439 | 913 | 319 | 44,394 | 62 |
| **2006** | 26,090 | 9,396 | 4,914 | 870 | 301 | 41,571 | 62 |
| **2007** | 26,113 | 9,666 | 4,830 | 920 | 296 | 41,826 | 62 |
| **2008** | 23,249 | 8,847 | 4,515 | 847 | 259 | 37,716 | 61 |
| **2009** | 21,658 | 8,606 | 4,242 | 779 | 234 | 35,519 | 58 |
| **2010** | 21,073 | 8,573 | 4,107 | 745 | 210 | 34,708 | 56 |
| **2011** | 20,157 | 8,448 | 4,048 | 764 | 219 | 33,636 | 56 |
| **2012** | 19,454 | 8,397 | 4,128 | 783 | 230 | 32,992 | 56 |
| **2013** | 18,673 | 8,305 | 4,006 | 764 | 238 | 31,987 | 57 |
| **2014** | 19,495 | 8,596 | 4,171 | 794 | 240 | 33,297 | 58 |
| **2015** | 21,352 | 9,533 | 4,499 | 867 | 264 | 36,514 | 58 |
| **2016** | 24,391 | 10,603 | 4,724 | 954 | 297 | 40,968 | 59 |
| **2017** | 24,107 | 10,237 | 4,874 | 1,019 | 299 | 40,536 | 59 |
| **2018** | 23,904 | 9,734 | 5,106 | 1,158 | 290 | 40,191 | 60 |
| **2019** | 23,316 | 9,479 | 4,945 | 1,170 | 286 | 39,196 | 60 |
| **2020** | 18,323 | 6,808 | 4,024 | 1,095 | 285 | 30,535 | 54 |

## 9.6. Appendix 6: Data – Collision Vs Age

Age data derived from the SWITRS database. The demographic data is derived from “US Census bureau” for the year 2019- 2020. The gross number of accidents may not reflect the

```
SELECT parties.party_age, Count(parties.party_age) AS Frequency
FROM collisions
INNER JOIN parties on parties.case_id == collisions.case_id
where (strftime('%Y', collisions.collision_date) == '2019') AND
      (collisions.collision_severity == 'property damage only')
GROUP BY parties.party_age
order By parties.party_age ASC;

SELECT parties.party_age, Count(parties.party_age) AS Frequency
FROM collisions
INNER JOIN parties on parties.case_id == collisions.case_id
where (strftime('%Y', collisions.collision_date) == '2019') AND
      (collisions.collision_severity == 'pain')
GROUP BY parties.party_age
order By parties.party_age ASC;

SELECT parties.party_age, Count(parties.party_age) AS Frequency
FROM collisions
INNER JOIN parties on parties.case_id == collisions.case_id
where (strftime('%Y', collisions.collision_date) == '2019') AND
      (collisions.collision_severity == 'other injury')
GROUP BY parties.party_age
order By parties.party_age ASC;

SELECT parties.party_age, Count(parties.party_age) AS Frequency
FROM collisions
INNER JOIN parties on parties.case_id == collisions.case_id
where (strftime('%Y', collisions.collision_date) == '2019') AND
      (collisions.collision_severity == 'severe injury')
GROUP BY parties.party_age
order By parties.party_age ASC;

SELECT parties.party_age, Count(parties.party_age) AS Frequency
FROM collisions
INNER JOIN parties on parties.case_id == collisions.case_id
where (strftime('%Y', collisions.collision_date) == '2019') AND
      (collisions.collision_severity == 'fatal')
GROUP BY parties.party_age
order By parties.party_age ASC;
```

impact of age on collision, as the size of this demographic vary. Hence, we normalize the accident as a percentage of the demographic size for the respective age groups in 2019

| AGE GROUP | 2019 POPULATION | TOTAL ACCIDENTS IN 2019 | PDO | PAIN | OTHER | SEVERE | FATAL |
|---|---|---|---|---|---|---|---|
| **Up to 14 years** | 7,612,696 | 2,038 | 385 | 556 | 843 | 216 | 38 |
| **15-17 years** | 1,651,013 | 10,155 | 5,347 | 2,554 | 1,791 | 383 | 80 |
| **18 and 19 years** | 1,035,338 | 26,597 | 14,536 | 6,885 | 4,026 | 940 | 210 |
| **20 years** | 523,576 | 15,342 | 8,198 | 4,139 | 2,299 | 559 | 147 |
| **21 year** | 517,694 | 15,836 | 8,662 | 4,074 | 2,413 | 562 | 125 |
| **22-24** | 1,572,440 | 48,171 | 26,157 | 12,633 | 7,201 | 1,768 | 412 |
| **25-29** | 2,675,956 | 78,673 | 43,125 | 20,435 | 11,356 | 2,987 | 770 |
| **30-34** | 2,515,804 | 66,554 | 36,297 | 17,559 | 9,446 | 2,572 | 680 |
| **35-39** | 2,521,794 | 56,446 | 30,658 | 15,360 | 7,813 | 2,042 | 573 |
| **40-44** | 2,556,100 | 47,109 | 25,543 | 13,106 | 6,261 | 1,743 | 456 |
| **45-49** | 2,636,048 | 42,981 | 22,868 | 12,203 | 5,911 | 1,555 | 444 |
| **50-54** | 2,516,572 | 40,903 | 21,620 | 11,718 | 5,624 | 1,516 | 425 |
| **55-59** | 2,170,407 | 38,423 | 20,056 | 10,918 | 5,447 | 1,529 | 473 |
| **60-61** | 775,227 | 13,384 | 6,900 | 3,865 | 1,896 | 557 | 166 |
| **62-64** | 1,032,955 | 17,155 | 8,695 | 5,039 | 2,498 | 670 | 253 |
| **65-66** | 561,670 | 9,202 | 4,523 | 2,768 | 1,457 | 350 | 104 |
| **67-69** | 725,771 | 10,716 | 5,213 | 3,126 | 1,778 | 442 | 157 |
| **70-74** | 957,430 | 13,064 | 6,189 | 3,874 | 2,281 | 538 | 182 |
| **75-79** | 750,040 | 7,546 | 3,495 | 2,266 | 1,362 | 333 | 90 |
| **80-84** | 579,704 | 4,123 | 1,807 | 1,206 | 850 | 187 | 73 |
| **84-** | 545,905 | 2,836 | 1,255 | 764 | 628 | 121 | 68 |

## 9.7. Appendix 7: Data - Collision Vs Time of Day

Derived from collision time filed of collision table of the SWITRS database. The collision is presented as % of total collisions from 2000 – 2021

```
SELECT strftime('%H', collision_time) AS collisionhour, count(*) AS Frequency
FROM collisions
GROUP BY collisionhour order By collisionhour ASC
```

| COLLISION HOUR | TOTAL | PDO | PAIN | OTHER | SEVERE | FATAL |
|---|---|---|---|---|---|---|
| 0 | 187,093 | 122,396 | 28,349 | 26,520 | 7,118 | 2,710 |
| 1 | 177,900 | 117,224 | 24,077 | 26,307 | 7,473 | 2,819 |
| 2 | 178,216 | 118,408 | 22,994 | 26,484 | 7,466 | 2,864 |
| 3 | 121,560 | 81,666 | 15,788 | 17,417 | 4,741 | 1,948 |
| 4 | 108,619 | 72,308 | 16,096 | 14,474 | 3,965 | 1,776 |
| 5 | 156,444 | 99,780 | 30,609 | 18,835 | 4,889 | 2,331 |
| 6 | 254,028 | 157,919 | 57,186 | 30,306 | 6,247 | 2,370 |
| 7 | 477,472 | 288,753 | 123,319 | 55,251 | 8,160 | 1,989 |
| 8 | 490,115 | 302,077 | 127,399 | 51,703 | 7,219 | 1,717 |
| 9 | 388,352 | 236,153 | 99,842 | 44,124 | 6,505 | 1,728 |
| 10 | 395,126 | 236,768 | 101,447 | 47,618 | 7,428 | 1,865 |
| 11 | 448,501 | 265,150 | 117,213 | 55,228 | 8,674 | 2,236 |
| 12 | 522,233 | 308,925 | 137,438 | 63,927 | 9,562 | 2,381 |
| 13 | 533,597 | 313,602 | 140,790 | 66,486 | 10,196 | 2,523 |
| 14 | 605,100 | 355,922 | 158,590 | 76,153 | 11,660 | 2,775 |
| 15 | 706,654 | 418,738 | 184,012 | 87,518 | 13,371 | 3,015 |
| 16 | 674,095 | 399,611 | 175,025 | 83,000 | 13,409 | 3,050 |
| 17 | 736,805 | 430,295 | 199,445 | 87,456 | 15,797 | 3,812 |
| 18 | 594,568 | 345,039 | 156,287 | 74,412 | 14,909 | 3,921 |
| 19 | 419,691 | 244,273 | 104,191 | 55,071 | 12,488 | 3,668 |
| 20 | 340,893 | 200,790 | 79,779 | 45,089 | 11,387 | 3,848 |
| 21 | 317,977 | 192,188 | 70,272 | 41,098 | 10,686 | 3,733 |
| 22 | 276,340 | 171,984 | 55,433 | 36,097 | 9,600 | 3,226 |
| 23 | 230,540 | 148,018 | 40,296 | 31,013 | 8,239 | 2,974 |

## 9.8. Appendix 8: Data - Collision Vs Population size

Location classification is derived from the SWITRS database. The collision is presented as % of total collisions from 2000 – 2021

```
SELECT population, Count(population) AS Frequency
FROM collisions
where strfTime('%Y', collisions.collision_date) == '2019'
GROUP BY population
order By population ASC
```

| Location classification based on population in 2019 | POD | PAIN | OTHER | SEVERE | FATAL | TOTAL |
|---|---|---|---|---|---|---|
| **Un- Incorporated** | 76,544 | 21,838 | 13,644 | 5,206 | 1,438 | 118,670 |
| **< 2.5 K** | 1,591 | 437 | 188 | 45 | 11 | 2,272 |
| **2.5K to 10K** | 3,697 | 980 | 722 | 154 | 40 | 5,593 |
| **10K to 25K** | 12,507 | 3,121 | 1,991 | 473 | 139 | 18,231 |
| **25K to 50K** | 24,301 | 8,678 | 5,177 | 990 | 259 | 39,405 |
| **50K to 100K** | 49,383 | 19,268 | 10,122 | 1,904 | 429 | 81,106 |
| **100K to 250K** | 52,374 | 21,705 | 11,105 | 2,000 | 441 | 87,625 |
| **>250K** | 59,389 | 37,717 | 16,395 | 3,269 | 680 | 117,450 |