Understanding Fatalities Amongst US Drivers

WYATT BABB* and MAR LONSWAY*, University of Colorado at Boulder, USA

Driver action plays a substantial role in the survivability of an accident. Utilizing the National Highway Traffic Safety Administration's Fatality Analysis Reporting System data, key characteristics of accidents present themselves. Some key characteristics include, but are not limited to, travel speed, BMI, vehicle weight, and time. Utilizing clustering analysis, we aim to understand what features present in an accident can predict the presence (or absence) of a fatality in an accident.

Additional Key Words and Phrases: Clustering, Traffic Data, Vehicle Safety, Traffic Fatalities

ACM Reference Format:

Wyatt Babb and Mar Lonsway. 2024. Understanding Fatalities Amongst US Drivers. 1, 1 (April 2024), 10 pages.

1 INTRODUCTION

Car culture is a defining aspect of the United States that has slowly developed since the rapid development of American society, headed by Gerald R. Ford and the Model T. With such a rapid introduction and involvement in American society, cars, trucks, and motorcycles slowly came to dominate transportation, even bleeding into infrastructure and development. As early as 1925, futurists in the United States were predicting cars to be dominant in society and transform the "modern" city landscape as seen by the cover of *Popular Science Monthly*'s August 1925 issue [6]. However, as time marched on, people like Robert Moses began to appear countrywide influencing and impacting millions of people. As cars became more important on the streets than people, city infrastructure expanded extensively at the expense of the average pedestrian[2]. Such a great displacement of persons was often disproportionate, too, impacting people's of color more frequently than their white counterparts, as was the way with Moses' Lincoln Center in New York [7].

With the introduction of such character's as Robert Moses, cities continue to develop in such a way that cars are often put first. This, combined with the considerable growth in technological achievements, cars became faster, stronger, and, to a simple pedestrian, more lethal. With this, manufacturing companies continue to develop key safety features necessary for improving the survivability of accidents beyond a simple prayer. Understanding what the successes and

*Both authors contributed equally to this research.

Authors' address: Wyatt Babb, wyba3752@colorado.edu; Mar Lonsway, malo6737@colorado.edu, University of Colorado at Boulder, Boulder, Colorado, USA.

Manuscript submitted to ACM

failures of recent technological advancements have been direct the key question of this paper: Which features present in accidents predict the presence or absence of a fatality?

This paper first introduces related works in the topic area. It then describes the data used for research and analysis and elaborates on the usefulness and collection of such data. The methods of the project are then defined to include key features assessed and evaluated in conjunction with data exploration and modelling. Finally, this paper moves into the results of such modelling and conclusion to be made.

2 RELATED WORKS

Feature exploration is key in understanding traffic data and traffic safety implications. Many key vehicle features and personable action are reviewed for the sake of understanding what causes or influences accidents across the US. On a larger scale, many studies focus on risk aversion and human contributing factors. One such study by [4] looks at the efficacy and determinants of seat belt use amongst US drivers. Similarly, [3] examines drug presence in deceased drivers in the state of Maryland. Looking to features this way provides a better understanding and explanation as to why these accidents occur, and provide implications as to what legislation should be put in place to prevent such tragedies.

3 DATA

3.1 Source

The data for the following research was collected from the National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis Reporting System (FARS). The specific dataset used references all vehicles present in fatal incidents and their key characteristics. Specific fatal incidents are only included if they occur within the United States and her territories and a fatality occurred within 30 days of said incident, and have been collected since FARS inception in 1975 [1]. With over 200 unique columns present from years 2020 and 2021, much of the data included corresponds to the key specifics of the vehicle present as well as notable actions taken by the driver or passenger that relate to the outcome of the incident. Some of such data includes, but is not limited to, car make/model, the presence of alcohol in the driver, damage to the vehicle, and injuries/deaths reported.

3.2 Bias

Immediate exploratory analysis of the data presents that the majority of fatal incidents occur in 3 key states: Texas, California, and Florida as presented below in Figure [1]. While data collection did not appear to be impacted by location, potential cultural factors have the potential to influence presence of fatalities or lack thereof. These cultural factors include county and state level jurisdiction general perspectives to the law, and education around motor transportation. As this paper does not aim to measure the efficacy of the education or legislation of certain states, this is recognized as a potential influence to deaths, however no transformations were made to normalize cultural influence to better understand overall impact of specific influences rather than compare between states.

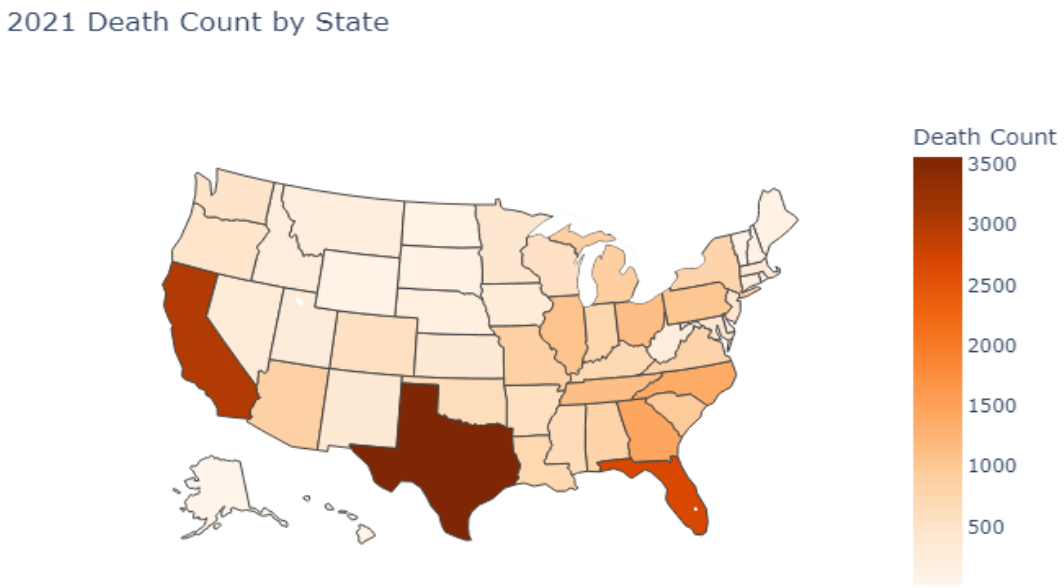


Fig. 1. Total Vehicular Fatalities in 2021 by State

In conjunction with this, due to the nature of how data is collected within this dataset, there is some possibility of survivorship bias present. It is entirely possible that human contributing factor to these accidents is over represented among a dataset. This comes from the fact that, on the average, the individuals represented in this data set are inherently more risky and take action that could be skewed from the population's average individual. Such a bias and its trade-offs are further examined by Mannering et al and their analysis of research methods in highway safety analysis [5]. As the data used exhibits exclusively fatal accidents, those incidents that lack any sort of fatality are can provide some insight into causality and more of the human contributing factor that more closely represents the population as a whole rather than the sample. It is the goal of this project to provide insight into the survivorship, so this is recognized as potential skew within the data.

Lastly, for sake of data analysis many of the codes and categorical data provided needed to be converted to ordinal data for the sake of machine learning applications and model development. While attempts were made to evaluate impacts and neutrally apply ordinal values according to values and categories of codes provided by FARS, The nature of ordinal data provides some innate levels of bias that may impact accuracy of the models presented below.

4 METHODS

4.1 Data Cleaning

Immediate inspection of the original dataset shows high levels of dimensionality through the sheer quantity of features present. A first pass of the data was used to determine redundancies and unnecessary features based unofficially on practicality and good sense. One such example of this includes a set of 10 columns that were eliminated. One of the ten columns was a vehicles license plate and the succeeding 9 columns were each individual license plate characters in order. While this information may be pertinent to specific case data, it does not necessarily apply to predicting deaths present in an accident. As such these 10 columns were entirely eliminated. Following primary elimination, the dataset was reduced by over 90 unique columns. This larger dataset was used for exploratory data analysis prior to cleaning data for modelling.

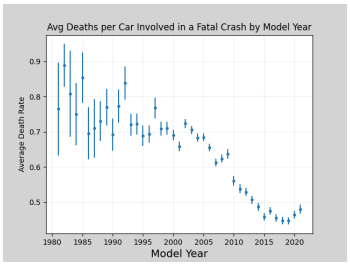
Subsequent cleaning was guided by exploratory data analysis and the driving research question of this paper: which features are prominent in predicting the presence of fatalities amongst car accidents? As many of the succeeding methods used for model implementation involved machine algorithms as described in the Experiment Design section of this paper, quantitative data was required that pertained to the topic. As such more features were eliminated directly, in the case they would be unhelpful in modelling fatality prediction, or were modified in such a way they could be used for such algorithms. Qualitative data was often remapped and translated to either ordinal data or binary dummy variables on a case by case basis. For example, in cases where the NHTSA used multiple different codes/characteristics as values for a single feature, such data was converted to ordinal data. Furthermore, the deaths column of the dataset prior to remapping, contained a integer value of the total quantity of deaths present. This was changed to a boolean value to represent whether or not a single death was present for prediction purposes.

Finally, further cleaning was necessary for use in modelling algorithms. As many of the features used units of substantially different magnitudes across the dataset, each of the features was normalized for scaling purposes, this assisted in preventing any further feature bias as described above. In order to prevent any data leakage, scaling occurred after splitting the dataset into training and testing sets. Following refitting, remapping, and elimination of many of the original variables, the final dataset to be used in data analysis maintained fourteen features.

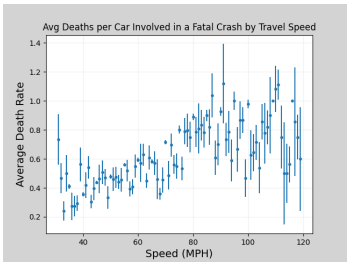
4.2 Exploratory Data Analysis

Initial exploratory data analysis assisted heavily in developing and designing experimentation and modelling by providing insight into the FARS dataset. Immediate analysis involved looking into the what was predicted to be the more impactful features of each incident. Initial impressions pushed further examination into 3 main features anticipated to be impactful in predicting the presence of a death: time of day, travel speed, and model year. Figure [2] demonstrates the evaluations done on each of these variables, and allowed curation in data cleaning to better understand the clear patterns present in the graphs below.

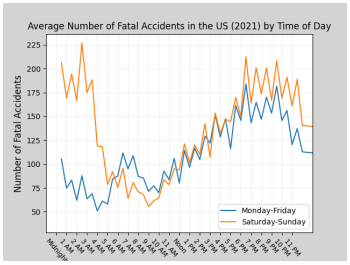
As seen in Figure [2a], newer model years are less likely to account for the deaths present in a fatality. This is in part due to the fact that they have had less time on the roads and are therefore less prevalent, but also are likely to exhibit more technologically advanced safety features than their older counterparts. Figure [2b] describes how travel speed impacts the prevalence of deaths in an accident. While it is true that there is noticeable non constant variance as speeds increase, there is still a positive relationship between speed and fatalities present. Finally, there is a distinct cyclical pattern in fatalities and time of day as seen in figure [2c]. With initial observable patterns amongst these features and a relationship with fatalities in accidents confirmed, experimentation and model design could be better curated around present suspicions.



(a) Death Rate by Model Years



(b) Average Death Rate by Travel Speed



(c) Number of Fatal Incidents by Time and Day

Fig. 2. Exploratory Data Analysis Key Figures

4.3 Model Implementation and Feature Engineering

We implemented different types of support vector machines (SVM), random forest, and k-means clustering to classify our data points. We first implemented different variations of SVM to classify vehicles involved in fatal accidents based on their body type (i.e, sedan, SUV, pick-up truck, etc.). SVM is a reasonable method to apply to the dataset given its ability to handle outliers and large quantities of attributes. While we had only selected between 10 and 20 attributes for a given analysis, one-hot encoding categorical values like vehicle make or model quickly expanded the amount of data the model would have to process. Moreover, the slack parameter in the SVM models allowed for a soft margin to account for tricky overlaps in the body type classification (e.g, large sedans sometimes weigh as much as the smaller SUVs, thus rendering weight a noisy attribute for that vehicle instance). Even with the encoded data, the SVM models yielded poor accuracy (less than 60%) in body type classification. We decided to proceed with two other main approaches, random forest and k-means clustering (two methods that would prove more accurate given the categorical, boolean, and ordinal nature of the data) and one last implementation of SVM given a dataset with reduced dimensionality. But before we moved onto the other models, we took time to feature engineer attributes and explore the data with more depth.

After the inaccuracy of the SVM models, we took time to feature engineer our dataset. One example of feature engineering the data involved the weight attribute. Surprisingly, the vehicle weight was not included in our original

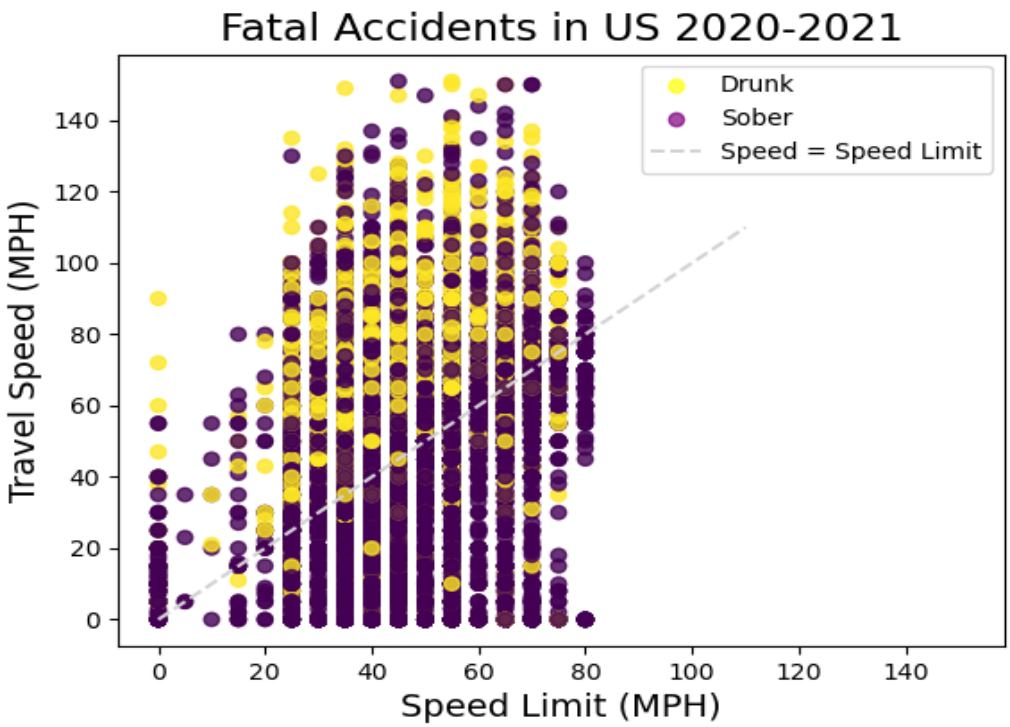


Fig. 3. Drunk drivers seem to drive faster

data set; rather, we were provided number codes that correlate to large bins of vehicle weight. For example, the first number code by weight included all vehicles under 6000 pounds. Given that most commuter vehicles fall below this measurement, the number codes did not provide much use in our modeling. Rather, we decided to create a vehicle weight attribute based off of body type. For a given vehicle body type, we sampled a weight from a normal distribution centered at the mean of vehicles in that class. This allowed us to attribute a reasonable weight measurement per vehicle in the data without knowing its exact curb weight. While this feature engineering abstracted the body class type, we also simplified attributes too. For instance, instead of looking at the number of deaths in a given vehicle, we simply reduced the fatality attribute to a boolean: did any occupant in the vehicle perish? This allowed for a binary target in the proceeding analyses. After cleaning out data for the next set of models, we decided to explore the data further.

In an effort to reduce one of the boolean attributes before implementing the random forests (which are less well suited than SVM to data with high dimensionalality), we decided to look at features of drunk drivers. Did they drive differently than their sober and more responsible counterparts? The answer is yes; they drove faster. Figure 3 shows the travel speed in miles per hour versus the recorded speed limit for a vehicle involved in a fatal accident. The axes have been scaled equally and the grey dashed line set through the middle of the plot to quickly identify the travel speeds

consistent with going the speed limit. Lastly, the data was separated into two data frames, those where the drivers were drunk and those were they were not. The yellow points, indicating driver inebriation, appear to be in the speeding region at a higher percentage than purple points. While no statistical tests were performed to distinguish whether the speeds come from two different distributions, the yellow points intensity above the grey line showed us enough evidence to treat the drunk group as a smaller subset of data points to which we could apply our models.

From here, we took all of the vehicle instances with drunk drivers and implemented a random forest. We targeted the feature attribute we had made boolean: did anyone in the vehicle perish? Implementing the random forest yielded our most accurate models in the entire project. When only considering drunk drivers, the random forest predicted fatality with about an 84% accuracy. Moreover, the feature importances, conveniently pulled out of the scikit-learn decision tree classifier object, revealed the attributes that were most important in classification. To clarify, the features that are deemed the most important are the ones that regularly lead to less impurity in the data. Mathematically, these features are the ones that limit the variance in the outcomes of their decisions. It is often the case that the important features are or help reveal the most compelling attributes for clarification. Luckily for us, the most important feature helped us produced a succinct result by showing us that weight is the most important predictor of vehicular fatality. 4 Shows us the distribution of vehicle weights for the target classification. Similar to the two classifications of data visualized in 3, we divided the data in 4 by the target attribute instead of an exploratory attribute (i.e, drunk versus sober). The values in the orange histogram (correlating to at least one fatality) overtake those in the blue histogram (no fatalities) for bins of lower vehicle weight. Notably, motorcyclists die at a frequency similar to occupats of sedans, hatchbacks, and coupes despite the latter's vastly more abundant road presence. However, beyond the dotted line representing about 3900 pounds, the frequency of vehicles involved in fatal accidents that did not endure an occupant fatality becomes the more prominent population. Lastly, this graph shows that occupants of semi trucks have the highest percentage of occupant safety relative to their body type classification, which makes sense given that the average vehicular weight of an unloaded semi is greater than four times that of the average sedan.

Lastly, in an attempt to get higher accuracy, we decided to use principal component analysis before implementing K-means and one more look at SVM. To reduce the dimensionality of our dataset, we programmed the code to make three principal components (i.e, linear combinations of the attributes) to describe the data. Total separation into clusters did not occur, however, 5 shows that the purple and yellow points seem to have separate loci. While clustering did not occur, it seems that the data were, by eye, separated well enough to apply other models to the values of the principal components. Both K-means clustering and SVM, with the RBF kernal and a small (0.025) regularization variable fared worse than the random forest, but still better than the original implementation on SVM when classifying body type. When these two methods were implemented on the principal components, both models yielded about 70% accuracy, over 10% more than the first implementation of SVM.

Note that the SVM was improved by using the RBF instead of the linear kernal and hard coding in the regularization value that generated the highest accuracy as shown in 6a. For k-means, while there is much data overlap shown in 5

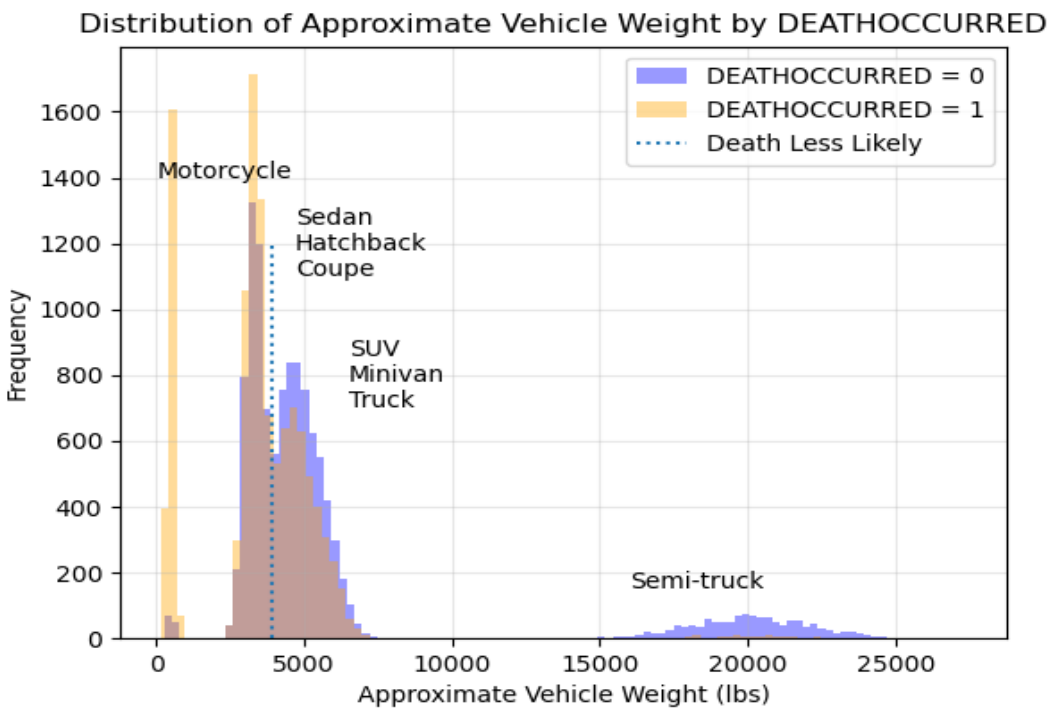


Fig. 4. Those in larger vehicles have better survival odds.

of the two classifications, the binary clusters shown by the k-means analysis in 6b shows that the epicenter of each classification seem to be centered in the correct position.

5 SUMMARY AND CONCLUSION

The primary results from our analysis is that heavier vehicles keep their occupants safer, as shown by 4. While this is the most concrete conclusion from our analysis, this is not new information. This conclusion aligns with common sense, known research, and the physical laws of the universe that mandate that when two objects collide, the forces acting on each object are identical in magnitude. It follows that a small sedan will be harmed more greatly from a force of the same magnitude than a large SUV. While the concrete conclusions of our research end here, the data from the FARS database is ripe for further investigation, albeit noting the challenges.

5.1 Future Work

The primary challenges of this dataset came from the categorical, boolean, and ordinal attributes collected by the Department of Transportation (the entity that oversees FARS data collection). Because many machine learning algorithms

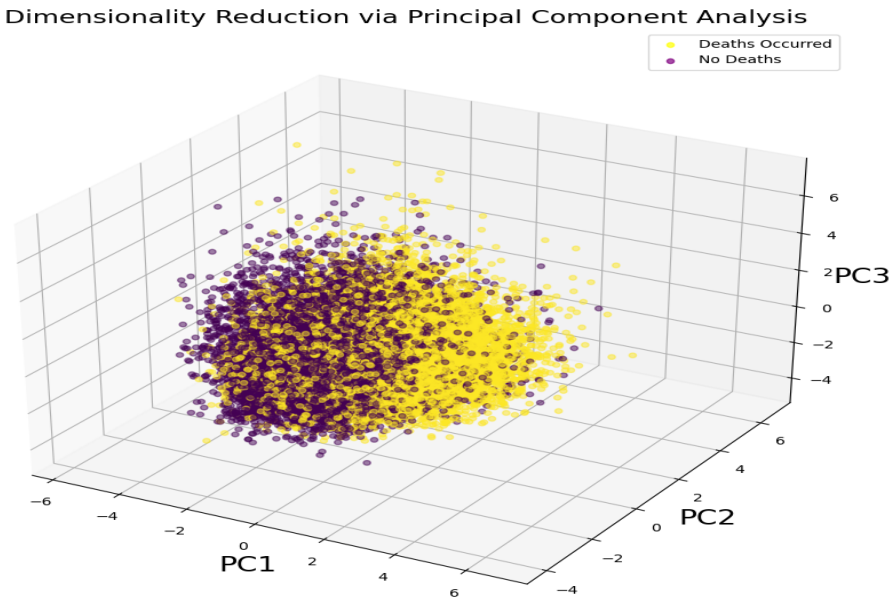
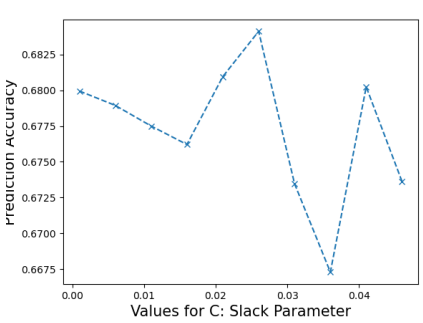
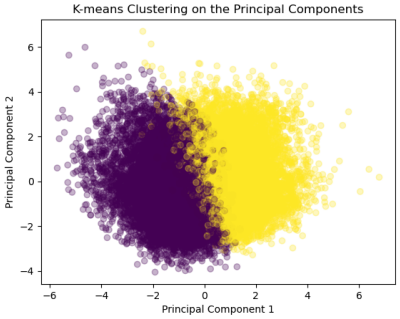


Fig. 5. Dimensionality reduction through PCA allows model implementation on only three attributes.



(a) Tuning of the slack parameter.



(b) K-means classification of principal components.

rely on normally distributed data, careful feature engineering must be done that doesn't abstract the original data while increasing model performance. An alternative to feature engineering is to source more data. For example, the result shown in 4 is the product of the nicely sampled data on the vehicle class weight means. Instead of sampling the data from a distribution, it would be more accurate to write a helper program that uses vehicle identification numbers to scrape the accurate weight of the vehicle from another source. In doing so, one could more confidently determine a point

(if it exists) in which a vehicle's weight generates a higher probability of survival than fatality. With this information ascertained, more questions and opportunities arise. One could, for example, provide suggestions to legislators on what weight and engineering standards should be required for manufacture. Researchers could also explore what sets of modern safety features boost a small car's survival rate to that of the heavier vehicles (i.e, can a comprehensive suite of safety features account for the disparity of survival rates in small cars versus large?) In conclusion, automobiles are an ordinary part of life and remain one of the highest causes of non-illness death across all ages in the United States. Research in the realm of vehicle safety should continue to be conducted to inform the public about the risks of driving and to better inform decision makers and vehicle occupants on how to ensure safety while driving.

REFERENCES

[1] National Highway Traffic Safety Administration. 1975-2021. *Fatality Analysis Reporting System.* Technical Report Version 1.0.0. National Highway Traffic Safety Administration. https://crashviewer.nhtsa.dot.gov/CrashAPI

[2] Christopher Dunn. 2011. The Rise Of Robert Moses And The Fall Of New York Constitutional Protections Against Eminent Domain. *Alb. Gov't L. Rev.* 4 (2011), 270.

[3] Johnathon P. Ehsani, Jeffrey P. Michael, Michelle Duren, Wendy C. Shields, Richard P. Compton, David Fowler, and Gordon Smith. 2021. Drug presence in driving deaths in Maryland: Comparing trends and prevalence in medical examiner and FARS data. *Accident Analysis Prevention* 154 (2021), 106066. https://doi.org/10.1016/j.aap.2021.106066

[4] Frank Goetzke and Samia Islam. 2015. Determinants of seat belt use: A regression analysis with FARS data corrected for self-selection. *Journal of Safety Research* 55 (2015), 7–12. https://doi.org/10.1016/j.jsr.2015.07.004

[5] Fred Mannering, Chandra R. Bhat, Venky Shankar, and Mohamed Abdel-Aty. 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research* 25 (2020), 100113. https://doi.org/10.1016/j.amar.2020.100113

[6] Popular Science Monthly. 1925. May Live to See: May Solve Congestion Problems. *Popular Science Monthly* 107, 2 (1925), 41.

[7] Keith Williams. 2017. How Lincoln Center was built (It wasn't pretty). *New York Times* 21 (2017).