DAIS FIRST DATA ANALYTICS COMPETITION

Shan Jiang

Department of Industrial and System Engineering

Rutgers University – New Brunswick

Email: sj576@soe.rutgers.edu

Zhifan Xu

Department of Industrial and System Engineering

Rutgers University – New Brunswick

Email: xu.zhifan.tony@gmail.com

Introduction

Identifying crash patterns through the historical crash database and qualitative measures has been commonly practiced by traffic authorities and researchers for many years [1-4]. For every crash, there are numerous factors such as human errors, vehicle defects, severe weather, and road conditions, contributing to its severity [4,5]. People are particularly interested in studying how fatal crashes occurred. Many reports analyze the contributing factors to fatal accidents [6–10]. In this report, we will address the following problems using the FARS dataset [11]:

1. What are the safest and most unsafe states in the United States?
2. What are the categories of factors that contribute to fatal crashes? And what are the most common factors in each class?
3. Is there any crash pattern that can be identified to help avoid the crash?
4. What are the factors that drivers have in common in fatal crashes?
5. What types of vehicles are more vulnerable in the crash?

To analyze the data and visualize the result, we will use Python/Jupyter Notebook and Tableau, and scripts/codes are attached in the Appendix. The organization of the report is as follows:

In section 1, the top 5 safe and unsafe states are listed based on our defined metrics: Death/ State Population and Death/State VMT (vehicle miles traveled). And an Autoencoder-LSTM model is utilized to predict future fatal crashes. In section 2, we list top candidates that lead to the majority of fatal crashes in Physical Mental Condition, Risky Events, and Vehicle Makes. In section 3, crash patterns that result in fatal crashes are examined, and suggestions for different types of vehicles are provided. In section 4, we inspect attributes of drivers in fatal crashes and provide insights to old/young, male/female drivers, respectively. In the last section, vehicle attributes are analyzed to find the most vulnerable vehicles, and a Multi-class Classification Neural Network model is utilized to make predictions on the type of damage each vehicle is received in the fatal crash.

1. The safest and most unsafe states

In this section, we will first analyze the historical crashes in 2018 in the United States by FARS dataset, which reveals 33,654 fatal motor vehicle crashes and 36,560 deaths. Next, we find the top 5 safe and unsafe states by the proposed death rate. To eliminate the bias caused by the population base difference in each state, we come up with metrics with common denominators, state population [12], and vehicle miles traveled (VMT) [13], respectively, to standardize the performance. In the last part of the section, a time series forecasting tool, Autoencoder Long Short-Term Memory (LSTM) [14], is utilized to predict the future fatal crashes in each state.

1.1 Death rate per population (DRP)

Since every state has different sizes of population and traffic network, solely counting the number of crashes does not reveal facts. Therefore, normalization needs to be done before comparing the risks. We start by proposing the metric rate of death crashes per population:

$$DRP = \frac{Number\ of\ death}{population} \tag{1}$$

Table 1: Top 5 safe states by Death per Population

STATE	NAME	Crashes	Deaths	Population	VMT	Death per Pop	Death per VMT
11	District of Columbia	30	31	702455	3690.677971	0.000044	0.008400
36	New York	889	943	19542209	123510.398633	0.000048	0.007635
25	Massachusetts	343	360	6902149	66771.980727	0.000052	0.005391
44	Rhode Island	56	59	1057315	8008.582207	0.000056	0.007367
34	New Jersey	525	564	8908520	77538.911172	0.000063	0.007274

Table 2: Top 5 unsafe states by Death per Population

STATE	NAME	Crashes	Deaths	Population	VMT	Death per Pop	Death per VMT
28	Mississippi	597	664	2986530	40730.439490	0.000222	0.016302
45	South Carolina	970	1037	5084127	56800.682752	0.000204	0.018257
1	Alabama	876	953	4887871	71167.206754	0.000195	0.013391
56	Wyoming	100	111	577737	10438.442613	0.000192	0.010634
35	New Mexico	350	391	2095428	27288.306972	0.000187	0.014328

The tables reveal the DRP ranges from 222 in Mississippi to 44 in the District of Columbia per million population. To better highlight the hot spots, Fig. 1 shows the DRP by states.

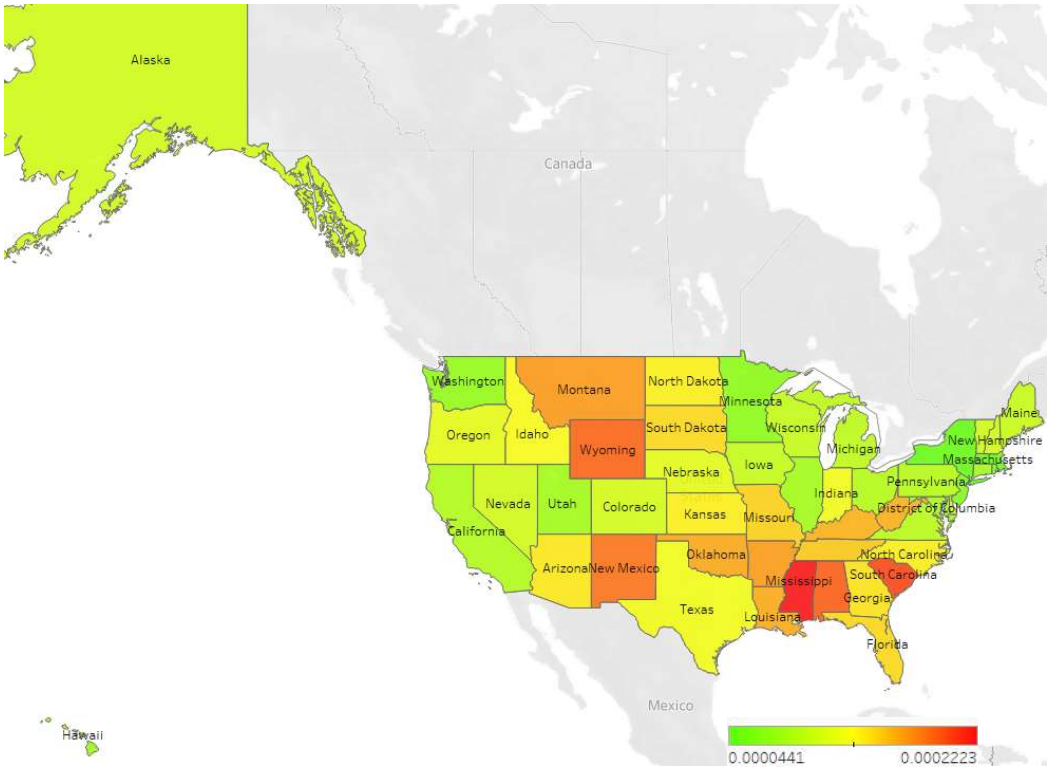


Figure 1: Heatmap of Death per Population

Light green indicates the safest, and dark red indicates the riskiest. In this case, the District of Columbia, New York, Massachusetts, Rhode Island, and New Jersey are the safest states in terms of DRP, while New Mexico, Wyoming, Alabama, South Carolina, and Mississippi are the worst.

1.2 Death rate per million VMT (DRV)

One may also argue that the amount of exposures (travels miles) for all vehicles in each state is very different. In light of this, we continue the investigation by introducing the metric by measuring the death per million vehicle mile traveled (VMT):

$$DRV = \frac{Number\ of\ death}{VMT} \tag{2}$$

Table 3: Top 5 safe states by Death per VMT

STATE	NAME	Crashes	Deaths	Population	VMT	Death per Pop	Death per VMT
25	Massachusetts	343	360	6902149	66771.980727	0.000052	0.005391
27	Minnesota	349	381	5611179	60438.314086	0.000068	0.006304
34	New Jersey	525	564	8908520	77538.911172	0.000063	0.007274
44	Rhode Island	56	59	1057315	8008.582207	0.000056	0.007367
36	New York	889	943	19542209	123510.398633	0.000048	0.007635

Table 4: Top 5 unsafe states by Death per VMT

STATE	NAME	Crashes	Deaths	Population	VMT	Death per Pop	Death per VMT
45	South Carolina	970	1037	5084127	56800.682752	0.000204	0.018257
28	Mississippi	597	664	2986530	40730.439490	0.000222	0.016302
22	Louisiana	716	768	4659978	50045.427449	0.000165	0.015346
4	Arizona	916	1010	7171646	66144.511205	0.000141	0.015270
54	West Virginia	265	294	1805832	19447.323211	0.000163	0.015118

The tables shows that the DRV ranges from 18.3 in South Carolina to 5.4 in Massachusetts per 1,000 million miles. To better reveal the hot spots, Fig. 2 shows the DRV by states.

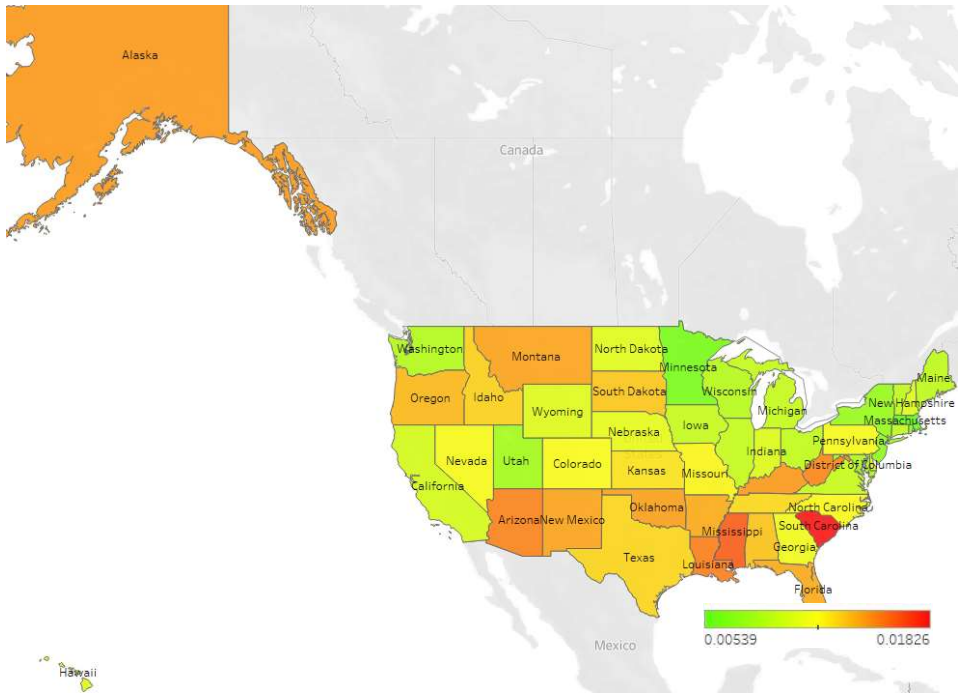


Figure 2: Heatmap of Death per VMT

In this use case, Massachusetts, Minnesota, New Jersey, Rhode Island, and New York are the safest states in terms of DRV, while West Virginia, Arizona, Louisiana, Mississippi, and South Carolina are the worst.

1.3 Fatal crash forecast

After analyzing historical crash data, we are also interested in possible future crashes in each state. With monthly fatal crashes being time-series, it is promising to utilize time-series tool to forecast future fatal crashes as head-ups for transportation authorities to take countermeasures in different states. Long Short-Term Memory (LSTM) is a type of recurrent neural network that can learn the order dependence between items in a sequence. LSTM is able to learn oscillation behavior such as seasonality and trend, and Autoencoder can make the computation more efficient by reducing the dimensions into some representation nodes [14]. The LSTM model will then learn the reduced representations instead of the whole thing. Since the LSTM maps a sequence of past observations as input to an output observation, the sequence of observations must be first transformed into multiple examples from which the LSTM can learn [14]. In our study, we divide the sequence into multiple input/output patterns called samples by a moving window. For example, we have crash data from 2018, which consists of 12 observations, and each observation contains a one-month fatal crash for 51 states. We would like to use a 6-time step moving window to generate input for the 1-step ahead prediction (predict the number of crashes in Jan 2019). Thus, the total number of observations is 12; moving window size is 6; time steps ahead is 1. With equation (3) for calculating the total number of samples,

$$No\ of\ Samples = No\ of\ Observations - Window\ Size - Time\ Step\ Ahead + 1 \quad (3)$$

six samples are obtained. Typically, researchers use the last sample for testing and the rest for training. Therefore, we have five training samples and one testing sample [14]. The final input shape of each sample is (6, 51), and the output shape of each sample is (1, 51). Fig.3 illustrates the training and testing procedures.

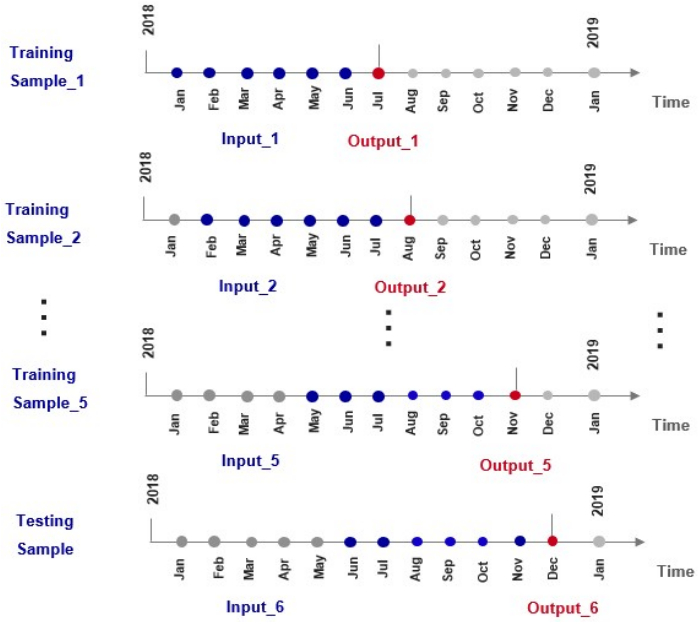


Figure 3: Illustration of training and testing procedures

In training, we first use the number of fatal crashes from Jan-Jun as inputs to predict the number of fatal crashes in Jul for 51 states. Then, we remove January data and add July data as inputs to predict August. Iteratively doing this with five training samples until reaching the maximum number of epochs. At this point, we get the final trained model and use it to test the performance of the testing sample. To evaluate the training and testing, we use the weighted mean absolute percent error (WMAPE) as the loss function and performance measure:

$$WMAPE = 100 * \sum_{i=1}^{51} \frac{|True_i - Forecast_i|}{True_i} \quad (4)$$

Where $True_i$ is the historical fatal crashes in the testing month and $Forecast_i$ is the estimate. With samples and defined loss function, we start building the Autoencoder LSTM model. Fig. 4 shows the structure of the model. Specifically, suggested by [15], rectified linear units (RELU) are utilized as the activation functions for the training.

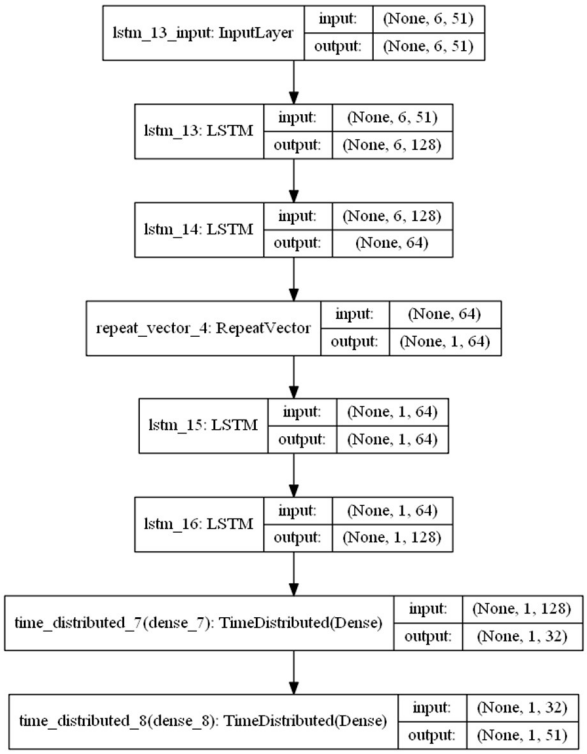


Figure 4: Structure of Autoencoder LSTM model

The first three layers represent LSTM layers, and the 4th layer RepeatVector acts as a bridge between the encoder and decoder modules. The dense layer is added at the end of the structure to get the output, where "51" is the number of states.

In training, most importantly, we need to figure out how long the training is and how many month's data we should use to predict future crashes. The corresponding hyperparameters, therefore, are training epochs and steps. To find the best parameter for the forecast, we trial and error on the number of epochs (1,000-6,000) and steps (1-8) and record the result. To reduce the randomness, each epoch and step combination will be repeated ten times, and the mean and

standard deviation of training and testing WMAPEs will be calculated to assist in finding the best model. Fig. 5 compares the training and testing WMAPEs concerning different epochs and steps.

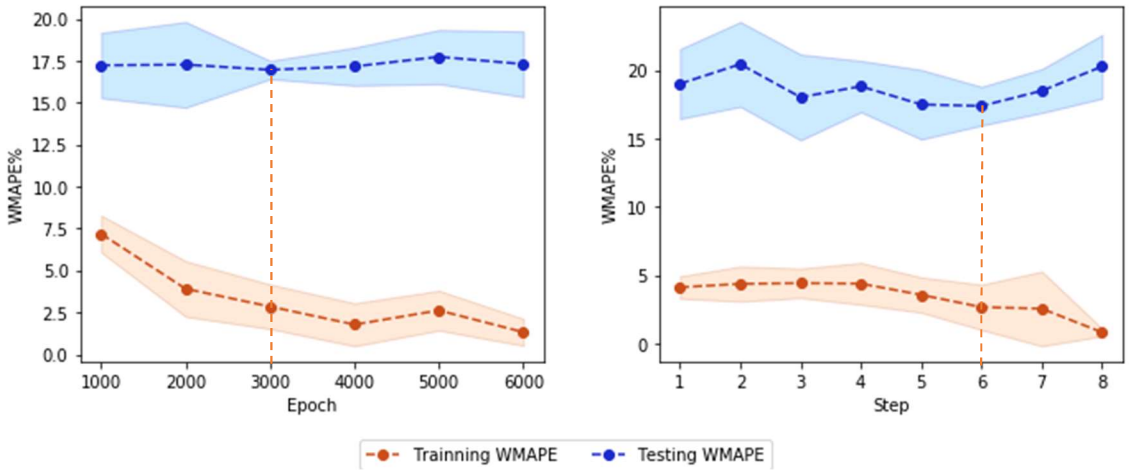


Figure 5: Training and testing WMAPE in different epochs (left) and steps (right).

It is not hard to see that the trend of training errors are going downwards while testing errors fluctuate. We pick the parameters where the mean and standard deviation of testing WMAPE sit lowest, which are 3,000 epochs and six steps. Finally, the trained model is obtained and ready for the case study. Fig. 6 shows the actual and predicted number of fatal crashes in sample states.

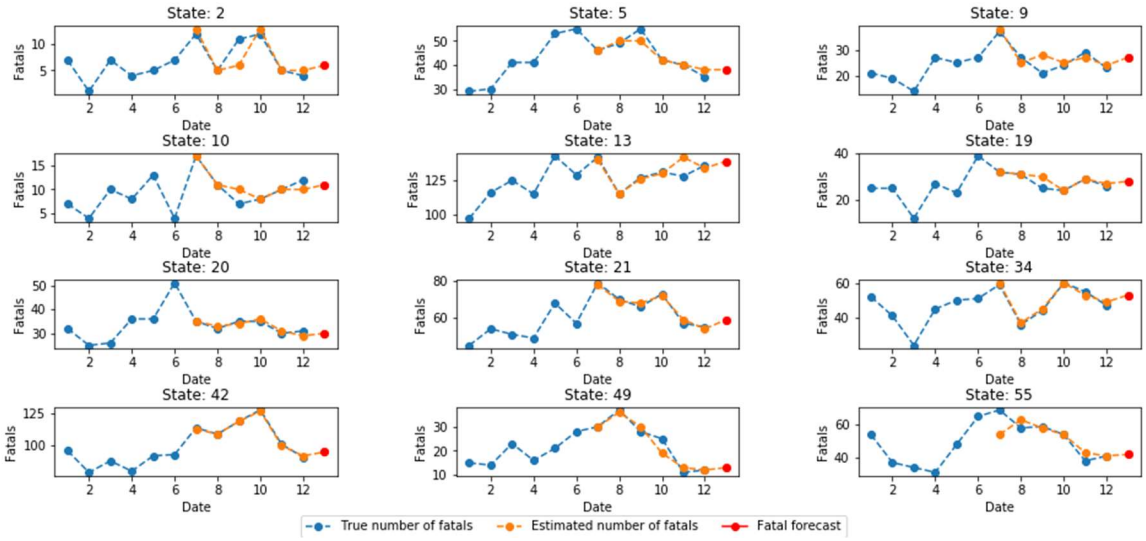


Figure 6: Fatal crash forecast for sample states

Blue circles are the number of actual fatal crashes in the states, and yellow circles represent the estimated results. Red circles represent the forecasted fatal crashes in January 2019, which is unknown as of 2018. As more data points available over time, we can continue this predicting process by adding more actual data points. To be noticed, this time series forecast can not only be done at the state level but also county level and city level. The result of this section can be used for transportation authorities as references to make plans to improve local safeties accordingly.

2. Contributing factors

Many factors contribute to fatal crashes in FARS data. In this section, we will focus on the events that lead to fatal crashes. Three representative categories Physical Mental Condition, Risky Events, and Vehicle Level Factors, are sorted out for this study. In Physical Mental Condition, we examine the influence of alcohol, drugs, and health issues to the fatal crash. In Risky Events, speeding, alcohol, drugs, driver distractions, drivers' vision obstruction, and vehicle level contributing circumstances are some examples. We perform a comprehensive review of the contributing factors, analyze the related tables in the dataset, and identify the most common contributing factors in each category.

Table 5: Top 5 factors associated with Physical Mental Condition

No.	Condition Description	No. of Fatal Crashes
1	DWI	5,177
2	Drowsy	694
3	Blackout	542
4	Physical Impairment	208
5	Emotional	172

DWI is the leading factor in the health category, and drowsiness and blackout follow.

Table 6: Top 5 factors associated with Risky Events

No.	Factor Description	No. of Fatal Crashes
1	Speeding	8,605
2	Positive BAC	6,970
3	No/Suspended/Revoked/Wrong/Expired License	4,079
4	Distracted	2,688
5	Visual Obstruction	1,942

Speeding is the most common event in fatal crashes, followed by drunk driving. Besides, people without a license or other types of license issues are likely to involve fatal crashes. Moreover, the number of fatal crashes caused by distraction and visual obstruction is not negligible. For the Vehicle Level Factors, we focused on analyzing the number of fatal crashes of each passenger vehicle make. The Makes of vehicles that have a large number of fatal crash records may indicate poor quality, more risks, and inadequate protection to drivers.

Table 7: Top 10 Vehicle Makes with most fatal crashes

No.	Makes	No. of Fatal Crashes	US Sales 2018	Fatal Crashes Per 10,000 Sales
1	Ford	7,234	2,386,588	30.31
2	Chevrolet	6,955	2,017,205	34.48
3	Toyota	4,304	2,224,156	19.35
4	Honda	3,704	1,445,627	25.62
5	Dodge	3,217	459,324	70.04
6	Nissan	2,636	1,344,597	19.60
7	GMC	1,554	556,451	27.93
8	Jeep	1,389	973,227	14.27
9	Hyundai	1,311	679,127	19.30
10	Kia	943	589,674	15.99

The top 10 Makes of vehicles on the list are popular makes in the U.S. that have more than 450,000 sales in 2018. We use the total sales of 2018 [16] as the normalization factor to represent the overall number of vehicles used by U.S. drivers for each makes. And we use the number of fatal crashes per 100,000 sales to check the risk level of each vehicle makes. The table shows that **Dodge** has an extremely high number of fatal crashes per 100,000 sales. Most U.S. brands such as Ford and Chevrolet, together with Japanese brand Honda, have moderate performance. Other Japanese and Korean brands, together with the U.S. brand Jeep, tend to have low risks. As a reference, two other make, **Subaru and Ram**, who also have more than 450,000 sales, are not even in the top 10 list for most fatal crashes.

3. Crash patterns

Fatal crashes can be related to many factors such as date, time, weather condition, type of road, type of vehicle, and so on. In this section, we identify such patterns by examining combinations of crash factors, including month, day of the week, hour index (every 4 hours belongs an index), road type, weather, and vehicle type to help the drivers avoid them. After initial data processing, the size of the generated table is 24,424. First, we check the frequency of different values in each crash factor. Fig. 7 compares the percentage of distinct values in each factor.

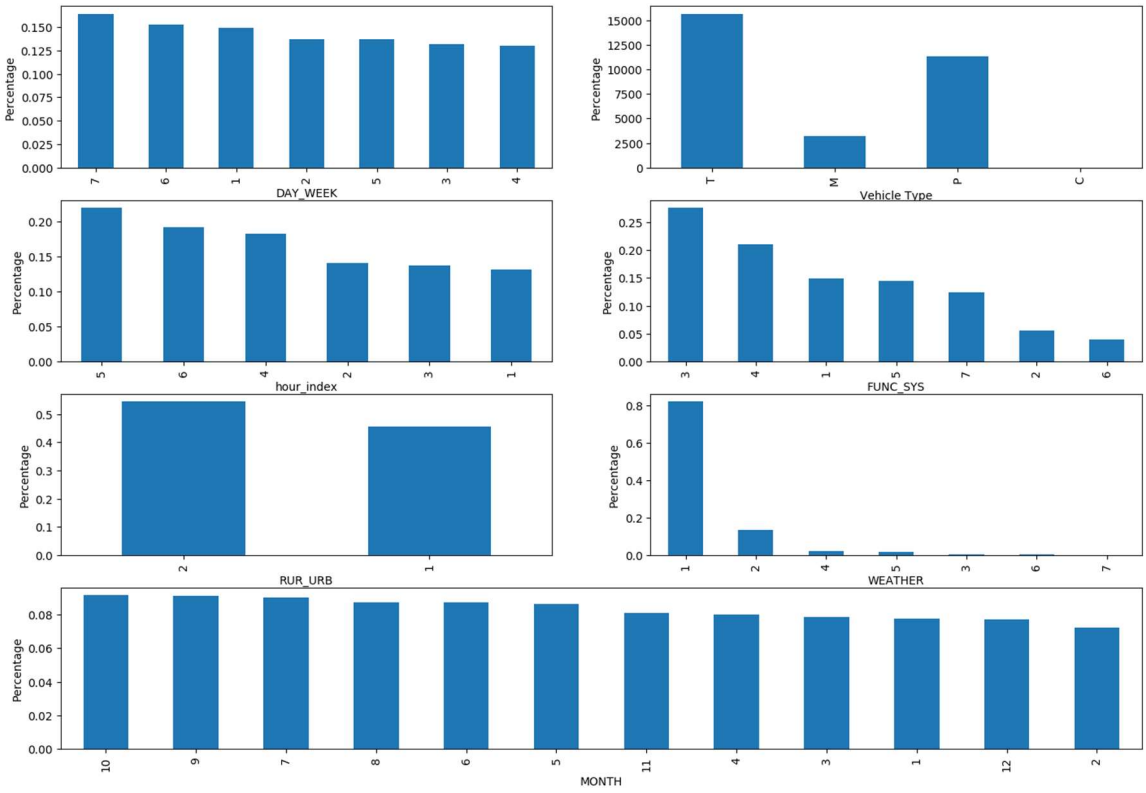


Figure 7: Comparison of the percentage of different values in each crash factor

Upon initial screening, the majority of crashes involved truck or passenger cars, on a clear weekend, within 4:00 pm – 8:00 pm, and on an arterial road. Next, we check the frequency of each unique combination of those crash factors. Table 8 lists the top 5 most unsafe unique combinations of crash patterns.

Table 8: Top 5 most unsafe unique combinations of crash patterns

	MONTH	DAY_WEEK	hour_index	FUNC_SYS	RUR_URB	WEATHER	C	M	P	T	cnt
0	9	7	6	3	2	1	0	0	1	0	12
1	5	6	6	4	2	1	0	0	1	0	10
2	12	4	5	3	2	1	0	0	0	1	10
3	8	6	6	3	2	1	0	0	1	0	10
4	11	5	5	3	2	1	0	0	0	1	9

From the above table, the most lethal combinations to the occurrence of fatal crashes are presented. However, it is not easy to interpret the result or arrive at any conclusions. To provide insights for each type of vehicle, Fig. 8 compares crash factors associated with different types of vehicles.

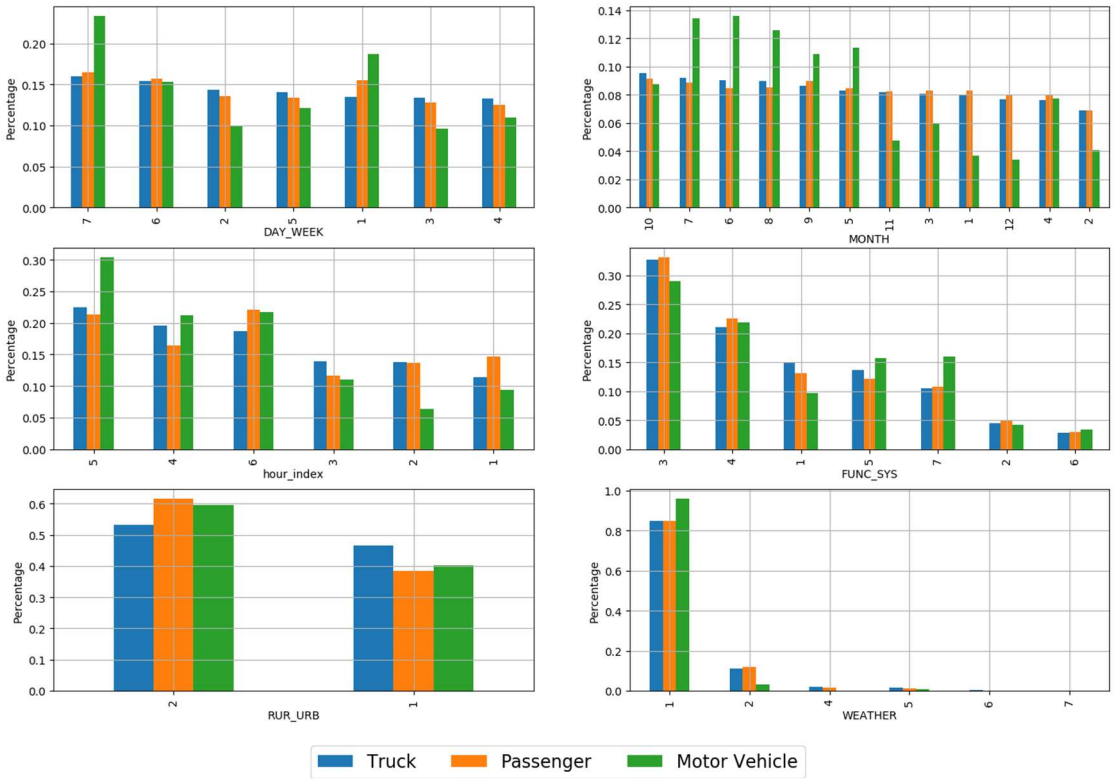


Figure 8: Comparison of crash factors associated with varying types of vehicle

After careful examination, we have the following findings and suggestions:

- Motorcyclists are not recommended to ride on Sunday and Monday due to high percentages of fatal crashes.
- Motorcyclists are more likely to involve in fatal crashes in the summertime since the number of fatal crashes in May through September is significantly higher than November through February. A possible reason is that motorcyclists ride less in cold winter.
- It is dangerous for all types of vehicles in the afternoon rush hours (16:00 – 20:00), especially for motorcyclists. And generally, nighttime is more dangerous than the daytime.

- It is safer to drive on Expressway and Minor collectors, and driving on principal arterial requires extra caution.
- Harsh weather does not necessarily lead to fatal crashes, especially for motorcyclists, as harsh weather is not as common as clear weather. Besides, drivers would pay more attention to bad weather or even avoid driving in bad weather.

4. Risky Drivers

Human error is another critical factor in Fatal crashes. In this section, we will investigate risky drivers according to their age and gender. We use the number of licensed drivers in 2018 [18] as the normalization factor to see how many accidents/million licensed drivers.

Table 9: Number of fatal crashes per million licensed drivers

	Male		
Age	No. of drivers involved in fatal	No. of licensed drivers (in million)	Per million driver
19 and under	1121	5.03	222.86
20 - 29	4651	17.15	271.20
30 - 39	3888	20.39	190.68
40 - 49	3433	20.16	170.29
50 - 59	3564	14.84	240.16
60 - 69	2411	9.15	263.50
70 - 79	1302	6.47	201.24
80 and over	734	2.62	280.15
	Female		
Age	No. of drivers involved in fatal	No. of licensed drivers (in million)	Per million driver
19 and under	609	4.71	129.30
20 - 29	1886	16.41	114.93
30 - 39	1447	19.82	73.01
40 - 49	1144	20.08	56.97
50 - 59	1094	14.79	73.97
60 - 69	894	9.65	92.64
70 - 79	567	6.91	82.05
80 and over	341	2.94	115.99
	Overall		
Age	No. of drivers involved in fatal	No. of licensed drivers (in million)	Per million driver
19 and under	1730	9.74	177.62
20 - 29	6537	33.56	194.79
30 - 39	5335	40.21	132.68
40 - 49	4577	40.24	113.74
50 - 59	4658	29.63	157.21
60 - 69	3305	18.8	175.80
70 - 79	1869	13.38	139.69
80 and over	1075	5.56	193.35

The table shows that drivers age 20-29, 60-69, and 80 or over are more likely to have fatal accidents. Besides, male drivers are more likely to have fatal accidents than females.

Following a similar fashion to crash pattern analysis, we also analyze driver data to identify attributes of high-risk drivers. In the analysis, we consider factors such as age group (every ten years as a group), sex, drunk, crash history, license issues, DWI, speeding, and handicapped. After initial data processing, the size of the generated table is 42,218. First, we check the frequency of different values in each driver's attribute. Fig. 9 compares the proportion of distinct values in each attribute.

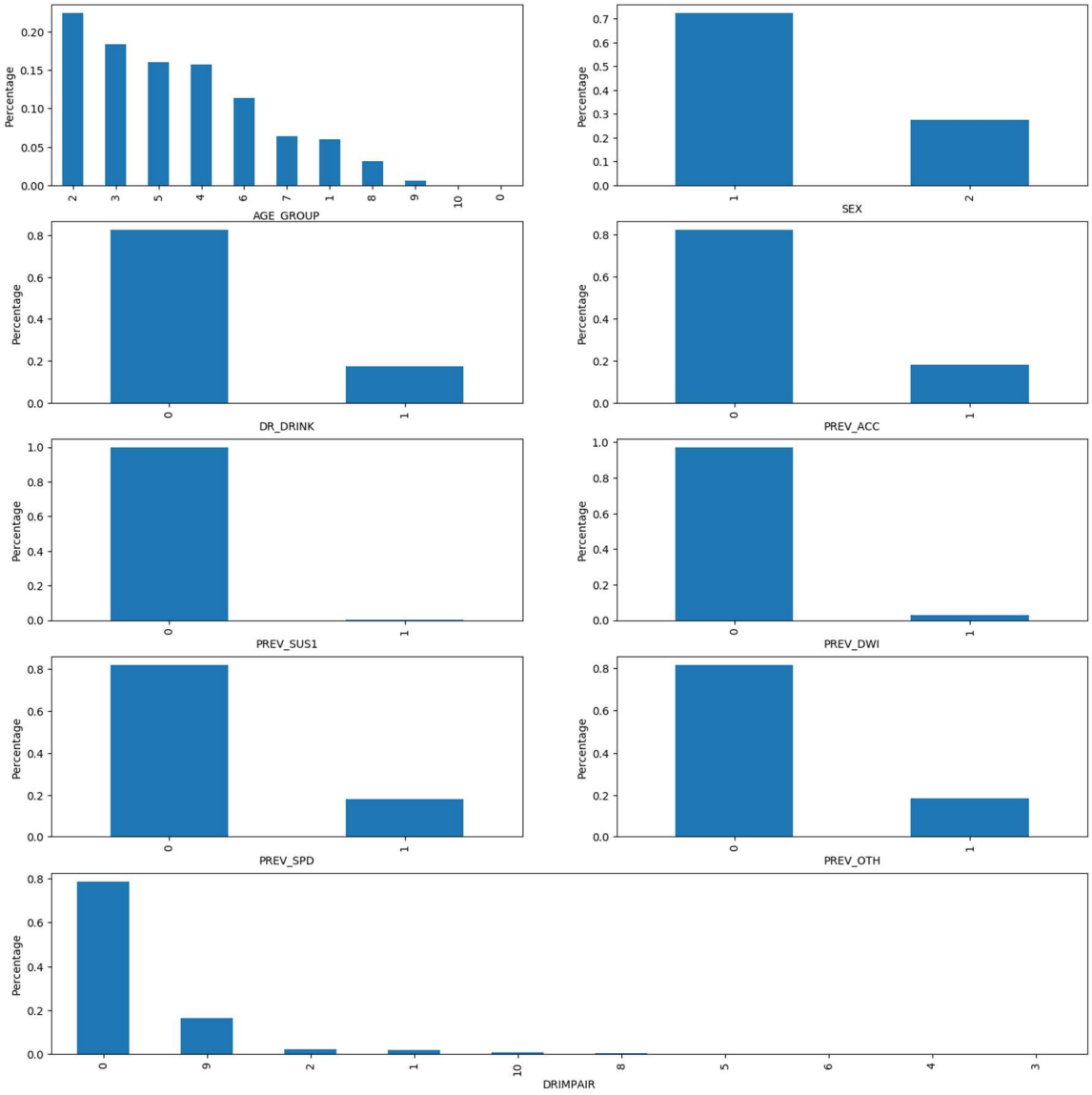


Figure 9: Comparison of the proportion of distinct values in each driver's attribute

At first glance, most drivers are young and have a clean driving history in terms of DWI, license issues, speeding, moving violations, and health issues. Next, we check the frequency of each unique combination of those risk factors. Table 9 lists the top 5 most unsafe unique combinations of driver's attributes.

Table 10: Top 5 most unsafe unique combinations of driver's attributes

	AGE_GROUP	SEX	DR_DRINK	PREV_ACC	PREV_SUS1	PREV_DWI	PREV_SPD	PREV_OTH	DRIMPAIR	cnt
0	5	1	0	0	0	0	0	0	0	1777
1	4	1	0	0	0	0	0	0	0	1506
2	3	1	0	0	0	0	0	0	0	1408
3	6	1	0	0	0	0	0	0	0	1300
4	2	1	0	0	0	0	0	0	0	1242

From the above table, we can barely interpret the result or arrive at any conclusions. To provide suggestions for male and female drivers at young ($age < 60$) and old ($age \geq 60$) ages, Fig. 10 compares the driver's attributes in different groups.

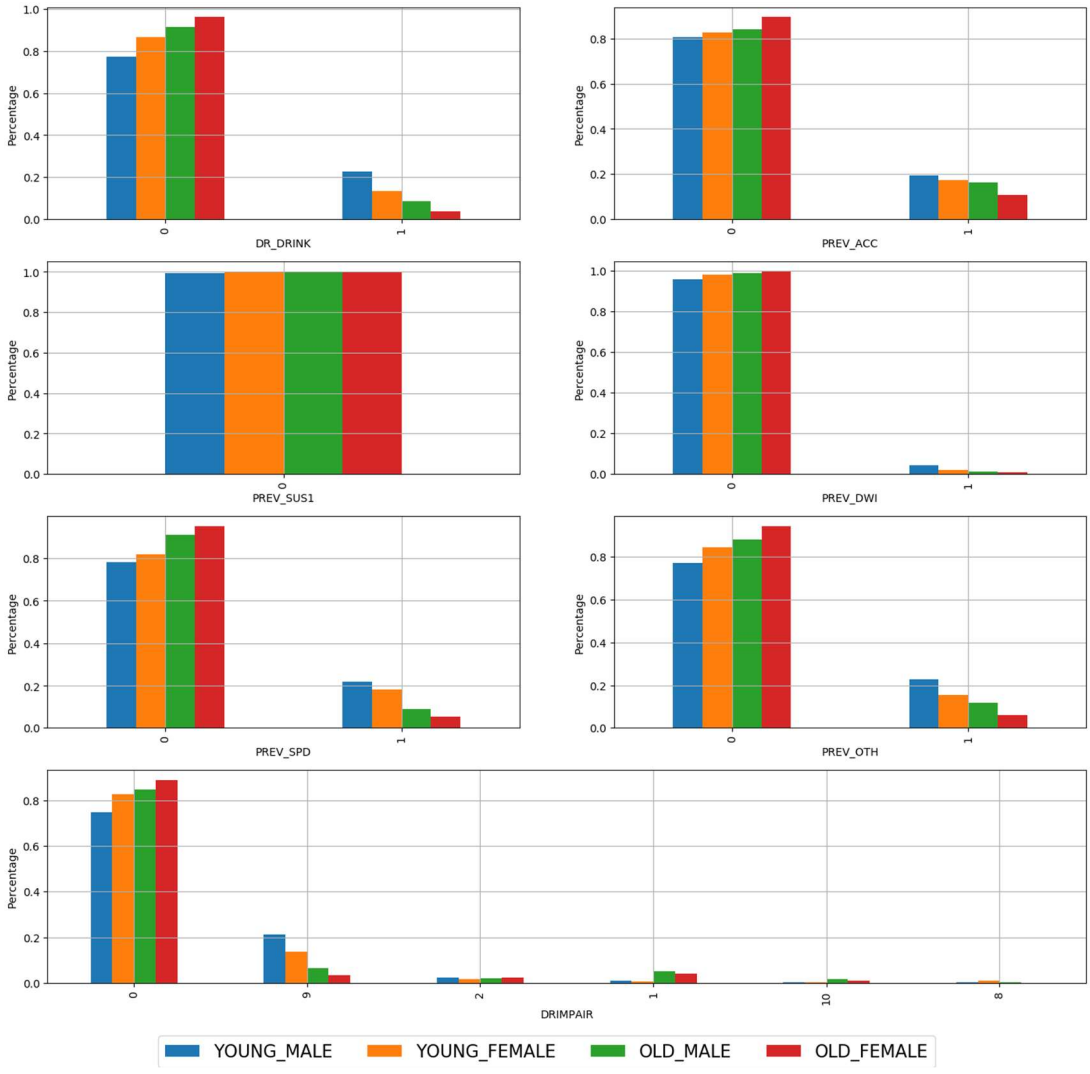


Figure 10: Comparison of driver's attributes in different groups

The above graph suggests that:

- If with clean driving history and without any impairment, old drivers are more likely to involve fatal crashes, especially old female drivers.

- If without a clean driving history, young drivers are more likely to get into fatal crashes, especially young male drivers.
- DWI is the leading risk factor in the impair category, and young drivers are more likely to involve in or get influenced by DWI.
- Fatal crashes due to drowsy and blackouts are more commonly seen among old drivers, especially old male drivers.

5. Vehicle Vulnerability

A sturdy vehicle can protect drivers and passengers in crashes, while a fragile one would fail in the crash. Therefore, investigating the vulnerability of vehicles is of vital importance for drivers. In this section, we will analyze patterns related to vehicle attributes such as type, make, model, body type, etc. combined with the kind of damage, rollover, fire/explosion, and frequency of incidents. And employ an exploratory approach similar to what is discussed in the previous question to hypothesize and validate vehicle vulnerabilities.

5.1 Exploratory analysis

This analysis considers the **deformation of a vehicle after an accident**. In the data set, possible deformations are:

Table 11: Deformation tag

Deformation	No damage	Minor Damage	Functional Damage	Disabling Damage	Unknown or not reported
Tag	0	2	4	6	8 or 9

For each accident, we consider a vehicle is more vulnerable when the Deformation Tag is greater. To evaluate the overall vulnerability of a vehicle, we use the expected deformation during the year 2018, and we exclude all accident records whose deformation information is unknown or not reported. Instead of using the expected deformation of a vehicle model among all types of accidents, we investigate **the expected deformation of a vehicle model per each collision type**. It is natural to assume that vehicles have different vulnerabilities, given another type of collision. The following table demonstrates a subset of data we used for our analysis.

Table 12: A subset of data used for analysis

Record No.	MAKE	MODEL	MAN_COLL	DEFORMED
0	82	881	0	6
28	20	431	0	4
37964	49	40	2	6

Note that there are more than 800 different models of vehicles in our dataset. However, most vehicle models don't have large samples. As shown in the following figure, more than 500 vehicle models have only very few accident records. We believe it is not possible to get persuasive analysis results for those vehicle models. Thus, **for each type of collision, we only focus on vehicle models with enough accident records**. We will use ten as the boundary number of records to separate the useful and not useful information for our analysis. Furthermore, we will focus on analyzing the vulnerability of passenger cars (including sedans, hatchbacks, SUVs, and pickups) only.

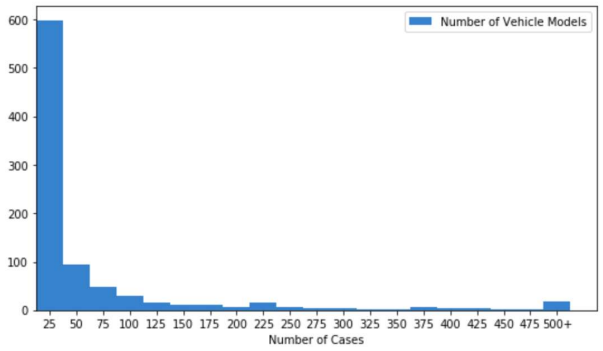


Figure 11: Frequency of different number of cases

In the original VEHICLE data, we have 51,872 records, covering 887 different vehicle models and all types of collisions. After removing vehicle models that don't have enough accident records for some types of collision and exclude non-passenger vehicles, we have **31,081 records, which belong to 100 vehicle models**. To calculate the expected deformation of a vehicle model per each collision type, we aggregate the data and count the number of cases on record of all degree of deformations for each vehicle model under each collision type. A sample subset of aggregated information is shown below.

Table 13: A sample subset of aggregated information

MAKE	MODEL	MAN_COLL	DEFORMED	COUNT	PERCENTAGE
2	403	0	0	7	4.76%
			2	14	9.52%
			4	17	11.56%
			6	109	74.15%
		1	0	0	0.00%
			2	1	3.23%
			4	6	19.35%
			6	24	77.42%
	404	0	0	3	1.28%
			2	35	14.96%
			4	20	8.55%
			6	176	75.21%
49	32	0	0	6	2.43%
			2	29	11.74%
			4	46	18.62%
			6	166	67.21%
		2	0	0	0.00%
			2	2	1.69%
			4	0	0.00%
			6	116	98.31%

Let $C_{i,j,k,h}$ be the number of accidents of Make i, Model j, Collision Type k that ends up with Deformation Level h. Let $P_{i,j,k,h}$ be the percentage of Deformation Level h under Collision Type k for Make i and Model j. We have

$$P_{i,j,k,h} = \frac{C_{i,j,k,h}}{\sum_{h'} C_{i,j,k,h'}} \quad (5)$$

The expected vulnerability of a car model us defined as

$$V_{i,j,k} = \sum_h D_h \times P_{i,j,k,h} \quad (6)$$

where D_k is the value of Deformation Level k. We believe $V_{i,j,k}$ accurately quantifies the vulnerability of passenger vehicles since $V_{i,j,k}$ is smaller when a car has minor deformation in most accidents given a specific collision type. Finally, we produce four top 10 lists that show the least vulnerable passenger vehicles for the four collision types.

Table 14: Top 10 least vulnerable passenger vehicles for the four collision types

Top 10 Least Vulnerable Passenger Vehicles for Single Vehicle Accidents (MAN_COLL=0)

MAKE	MODEL	Vulnerability
7	27	3.720930
12	461	3.777778
37	421	3.791045
59	403	3.884615
6	51	3.931034
12	880	3.967213
49	46	4.047619
12	24	4.076923
49	403	4.088235
12	881	4.108108

Top 10 Least Vulnerable Passenger Vehicles for Front-To-Rear Collisions (MAN_COLL=1)

MAKE	MODEL	Vulnerability
37	421	3.230769
20	431	3.818182
49	46	4.000000
19	18	4.200000
2	405	4.235294
35	401	4.285714
20	461	4.400000
51	881	4.428571
20	9	4.434783
12	881	4.450000

Top 10 Least Vulnerable Passenger Vehicles for Front-To-Front Collisions (MAN_COLL=2)

MAKE	MODEL	Vulnerability
49	46	4.695652
48	34	4.769231
18	7	4.800000
59	403	4.875000
41	50	4.888889
19	18	4.941176
59	31	4.941176
49	403	5.000000
12	881	5.030303
12	461	5.111111

Top 10 Least Vulnerable Passenger Vehicles for Opposite Direction Sideswipe Collisions (MAN_COLL=6)

MAKE	MODEL	Vulnerability
13	1	4.666667
49	402	4.753247
51	881	4.762887
12	24	4.800000
37	441	4.826087
20	482	4.833333
23	421	4.857143
7	27	4.909091
12	880	4.931034
12	881	4.939759

Here are some insights from the Top 10 Lists, given the makes and model information from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812828:

1. Since the 3-digits model number represents pickup trucks, it is clear that pickups tend to be the body type that can endure most kinds of accidents.
2. Ford Super Duty (model 12-881), appears on all Top 10 lists. We pick it as the most endurable automobile of 2018 that protects you from all unexpected accidents.

3. Toyota Prius (model 49-46), appears on 3 of the 4 Top 10 lists. We pick it as the most endurable non-pickup automobile of 2018 that protects you from most unexpected accidents.

5.2 Predictive modeling

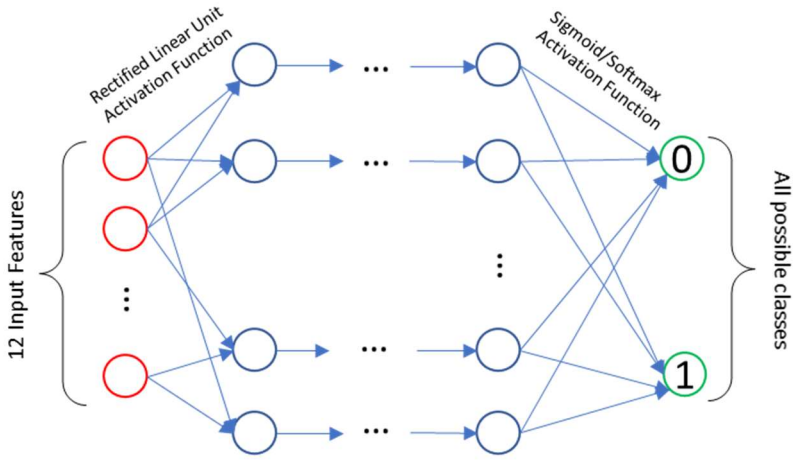
This section tries to use predictive models to answer the following questions:

1. Given a car's necessary information and the conditions when an accident happens, how severe the deformation will be in the crash?
2. Provided a car's necessary information and the conditions when an accident happens, whether there will be a death or not?

The input we use for our models include:

- The car's necessary information: make, model, body type, model year, and gross weight.
- Crash elements: travel speed, traffic flow, vehicle's activity, driver's action, underride or not, rollover or not, critical event, and collision type.

Apparently, the two questions we care about are classification problems. We build Sequential Neural Networks as our predictive models to answer the three questions. One of the advantages of Neural Networks we value a lot for our classification problems is that it can tell us the probability of falling into each class [17].



An illustrative structure of the Neural Network

Figure 12: Neural network model

Here are the setups to train the Neural Network:

- 75% of data records are used as training data, and 25% of data records are used testing data.
- We use the Sigmoid (SoftMax) activation function for the two classes (multiple-class classification) problem.
- We use the Linear Unit activation function for all other layers.

5.2.1 Predicting the Deformation

The predictive model built in this section tries to predict the deformation of a given car in an accident. Instead of using the four deformation levels as the labels, we simplify our problem to predict **whether a vehicle will have severe damage in an accident**. Cars that have no damage, minor damage, and moderate damage will be considered as not having severe damages.

The Neural Network we trained has 87.25% accuracy on the testing data. The table below shows that most cars in the testing data ended up (82%) with severe damage. Thus, our Neural Network classifier has better accuracy than claiming all vehicles will be severely damaged in accidents.

Table 15: Number of testing data

Testing Data	Not Sever Damage	Sever Damage	Total
Number of Records	1805	8263	10068

The confusion matrix shows that the True Negative rate is 53%, while the True Positive rate is 95%. It catches the fact that most cars will be severely damaged in the crash, while a few can end up with a better shape in accidents.

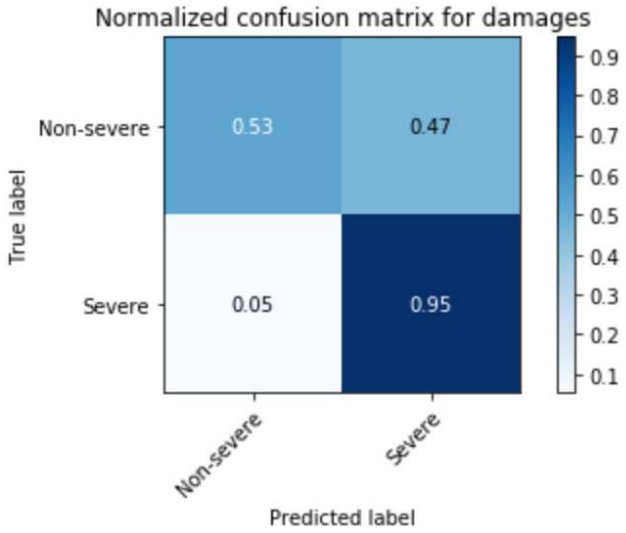


Figure 13: Confusion matrix

To use the model, we can input then vehicle's and the accident's information, and we get the probability that the vehicles will be severely damaged. For instance:

- Input: a 2017 Hyundai Elantra hatchback, traveling at 17mph on a two-way divided road, involves a rear-end accident in which the other vehicle is traveling in the same direction with higher speed. No override or rollover happened.
- Output: there is a 75.88% probability that this car will have severe damage.

5.2.2 Predicting Occurrence of Fatalities

When we discuss vehicles' vulnerability, the ultimate goal is to see how vehicles can protect drivers and passengers in an accident. This section presents a Neural Network model that predicts **whether there will be deaths in an accident**.

The Neural Network we trained has 79.3% accuracy on the testing data. The table below shows that about half of all accidents will have fatal events.

Table 16: Number of records associated with No Death, Have Death, and Total

Testing Data	No Death	Have Death	Total
Number of Records	5086	5654	10740

The confusion matrix shows that the True Negative rate is 73%, while the True Positive rate is 85%. It implies the Neural Network model has the following properties:

1. It has relatively better accuracy in both classes and outperforms random guesses.
2. It is more sensitive to situations where fatal events may happen since the True Positive rate is greater than the True Negative rate

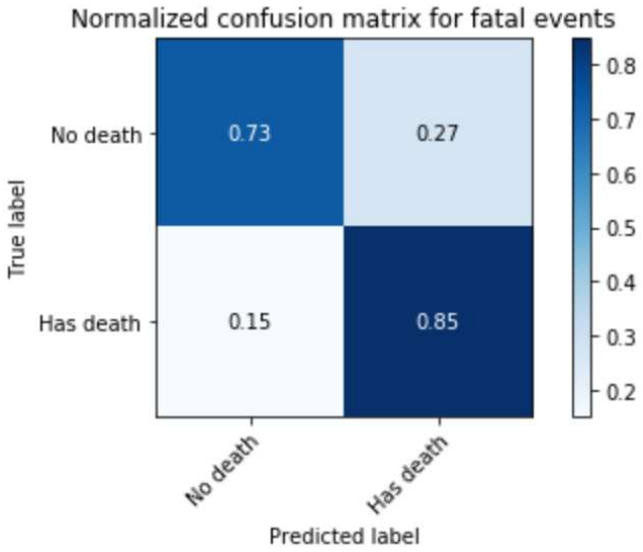


Figure 14: Confusion matrix

Property 1 proves that the Neural Network is effective. Property 2 shows that the model aligns with the risk-averse principle since people's life are priceless.

To make a prediction, we feed the model with the vehicle's and the accident's information, and we get the probability that there will be fatalities occurred to this vehicle. For instance:

- Input: a 2011 Chrysler 200 4-door sedan, traveling at 55 mph on a two-way not-divided road, involves in an opposite direction sideswipe accident in which the other vehicle from the opposite direction encroached into line. No override or rollover happened.
- Output: there is a 42.89% probability that fatalities will occur to this vehicle.

Conclusion

In this study, we started by exploring the number of historical crashes in different stats in the United States in 2018. We revealed 33,654 fatal motor vehicle crashes and 36,560 deaths and provided two lists of most unsafe and safest states by death per population and death per million vehicle miles traveled, respectively. The fatality rate per million people ranged from 222 in Mississippi to 44 in the District of Columbia. The death rate per 1,000 million miles traveled ranged from 5.4 in Massachusetts to 18.3 in South Carolina. To predict future risks, we utilized the Autoencoder-LSTM to forecast the one-step-ahead fatal crashes for all states parallelly.

We continued the investigation by exploring the contributing factors associated with driver's health/mental status, risk events, and the vehicle makes from the FARS datasets. To deep dive, we revealed factors that were related to fatal crashes such as date, time, weather condition, type of road, and type of vehicle. And point-to-point suggestions/findings for the motorcycle, truck, and commercial vehicle drivers were provided. To identify risky drivers, we analyzed driver attributes in terms of age, sex, and driving history. In the last part of this study, vehicle vulnerabilities were analyzed, and two NN models were proposed to predict the type of deformation a fatal crash could have caused and the likelihood of occurrence of fatalities to each vehicle based on the vehicle's attributes such as makes and model.

Acknowledgment

We would like to thank Dr. Saeed from Industrial and Systems Engineering at Ohio University, and Dr. Chen from Industrial Engineering, the University of Texas at Arlington to provide us with the opportunity to attend the first DAIS data competition. The challenge questions presented in this document are the intellectual property of the IISE Data Analytics and Information Systems (DAIS) board of directors. This report is exclusively shared with the competition entrants. Sharing this report in its entirety or parts of it with anyone other than the authorized competition entrants without official permission from the competition chairs are not allowed.

Reference

[1] Abdel-Aty, M., & Pande, A. (2005). Identifying crash propensity using specific traffic speed conditions. Journal of Safety Research, 36(1), 97-108.

[2] Polders, E., Daniels, S., Casters, W., & Brijs, T. (2015). Identifying crash patterns on roundabouts. Traffic injury prevention, 16(2), 202-207.

[3] Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proceedings of the National Academy of Sciences, 113(10), 2636-2641.

[4] D. Ascone, T. Lindsey, and C. Varghese, "Traffic safety factor: An examination of driver distraction as recorded in NHTSA databases," National Highway Traffic Safety Administration's National Center for Statistical and Analysis, 2009.

[5] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," IEEE Transactions on Intelligent Transportation Systems, vol. 12, no. 2, pp. 596–614, 2011.

[6] National Highway Traffic Safety Administration, "National motor vehicle crash causation survey: Report to congress," National Highway Traffic Safety Administration Technical Report DOT HS, vol. 811, p. 59, 2008.

[7] Federal Motor Carrier Safety Administration, Report to Congress on the Large Truck Crash Causation Study: Author Washington, DC.

[8] Preusser, D. F., Williams, A. F., Ferguson, S. A., Ulmer, R. G., & Weinstein, H. B. (1998). Fatal crash risk for older drivers at intersections. Accident Analysis & Prevention, 30(2), 151-159.

[9] Chihuri, S., Li, G., & Chen, Q. (2017). Interaction of marijuana and alcohol on fatal motor vehicle crash risk: a case–control study. Injury epidemiology, 4(1), 1-9.

[10] Kashani, A. T., & Besharati, M. M. (2017). Fatality rate of pedestrians and fatal crash involvement rate of drivers in pedestrian crashes: a case study of Iran. International journal of injury control and safety promotion, 24(2), 222-231.

[11] National Highway Traffic Safety Administration. (2018). Fatality analysis reporting system (FARS) encyclopedia. 2012.

[12] United census bureau. (2018). 2018 National and State Population Estimates. https://www.census.gov/newsroom/press-kits/2018/pop-estimates-national-state.html.

[13] Highway Statistics 2018. (2018). Office of Highway Policy Information. https://www.fhwa.dot.gov/policyinformation/statistics/2018/vm2.cfm.

[14] Gensler, A., Henze, J., Sick, B., & Raabe, N. (2016, October). Deep Learning for solar power forecasting—An approach using AutoEncoder and LSTM Neural Networks. In 2016 IEEE international conference on systems, man, and cybernetics (SMC) (pp. 002858-002865). IEEE.

[15] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[16] GoodCarBadCar.net. (2018). U.S. Auto Sales Brand Rankings – December 2018 YTD. https://www.goodcarbadcar.net/u-s-auto-sales-brand-rankings-december-2018-ytd/

[17] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y. and Alsaadi, F.E., 2017. A survey of deep neural network architectures and their applications. Neurocomputing, 234, pp.11-26.

[18] U.S. Department of Transportation. Licensed Drivers by Age and Sex (In Thousands). https://www.fhwa.dot.gov/ohim/onh00/bar7.htm