VEHICLE COLLISIONS IN SEATTLE

IBM CAPSTONE | BY ARTEM SHEVCHUK

Contents

1. INTRODUCTION

Road accidents have become very common nowadays. As more people are buying automobiles, the incidences of road accidents are just increasing day by day. According to The Annual Global Road Crash Statistics report published by **www.asirt.org**, approximately 1.35 million people die in road crashes each year, and on average 3,700 people lose their lives every day on the roads. An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities.

Considering the importance of this problem, we have to try our best to understand what factors affect the probability of being injured during the car crash the most. To achieve this, we will use the modern tools data science has to offer.

This research is based on collisions data from Seattle; however, I believe that the derived conclusions can serve as a benchmark for all other cities and counties.

This report will be of interest to anyone who drives cars or works in traffic regulation authorities. Understanding the roots of this problem may bring us one step closer to making our cities a safer place to live and may save the lives of those readers who take this issue seriously.

2. A DESCRIPTION OF THE DATASET

Every data-driven solution requires one most important thing: data. In our research, we will use the Traffic Collisions Dataset compiled and maintained by the Seattle Department of Transportation (SDOT). This dataset has been referred to within the IBM capstone project, and therefore we have no doubts regarding its authenticity and quality.

Below is the **quick summary** of our dataset:

a. Range: 2004 to Present
b. 37 Attributes
c. Almost 195,000 rows (observations)

The key intention of our research will be an attempt to measure the **impact of various characteristics on the severity** of the accident. It will be helpful to mention what severity codes correspond to:

Code	Meaning
3	Fatality
2b	Serious injury
2	Injury
1	Property damage
0	Unknown

Our research will be focused only on the following severities:

- ✓ 1 (Property Damage, 136 485 observations)
- ✓ 2 (Injury, 58 188 observations)

For this research I decided to create 2 separate data frames: one will be used for exploratory analysis and another for the machine learning (ML) algorithm. The key motivation behind separating the dataset into 2 is the fact that data is labeled unevenly, which means that while trimming data for unbiased model creation we will lose almost half of observations, which can be useful for exploratory analysis. I hope after following the below steps it all will make sense to you.

Below is the description of our ML dataset:

	SEVERITYCODE	ADDRTYPE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	SPEEDING
0	2	Intersection	N	Overcast	Wet	Daylight	NaN
1	1	Block	0	Raining	Wet	Dark - Street Lights On	NaN
2	1	Block	0	Overcast	Dry	Daylight	NaN
3	1	Block	N	Clear	Dry	Daylight	NaN
4	2	Intersection	0	Raining	Wet	Daylight	NaN

Let us have a quick look at what each column represents and what actions we might need to take to prepare it for analysis:

Column	Description
SEVERITYCODE	Already described in the previous table
ADDRTYPE	Whether an incident occurred at an intersection or block. Requires 0/1 encoding
UNDERINFL	Whether a driver was under influence of drugs or alcohol. Requires 0/1 encoding. Contains NA values which will be filled with 0 by default
WEATHER	Weather conditions during the accident. The column includes 11 types of conditions. For efficiency, they will be converted to FAVORABLE (0) and UNFAVORABLE (1)
ROADCOND	Road conditions during the accident. The column includes 9 types of conditions. For simplicity, they will be encoded to FAVORABLE (0) and UNFAVORABLE (1)
LIGHTCOND	Light conditions during the accident. The column includes 9 types of conditions. For simplicity, they will be encoded to DAY (0), OTHER (1)
SPEEDING	Whether an incident occurred while the driver was obeying or disobeying the speed limit. Requires 0/1 encoding.

Other important points:

- ✓ The data has unbalanced labels (more property damage in comparison to injuries). To avoid bias, we will have to trim it and keep an equal amount of labeled observations for both categories.
- ✓ Some categories like weather, road, and light conditions contain “unknown” observations. To avoid guessing, we will just remove those observations entirely. I

believe this will not have a significant statistical impact on the quality of research, as the overall number of recorded rows is still very large.

I will not include the long list of all processing actions taken here; they will be available in my Git repository in a separate workbook. After processing the data this is the final table we arrived at:

	SEVERITYCODE	ADDRTYPE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	SPEEDING
0	2	1	0	1	1	0	0
1	2	0	0	0	0	0	0
2	2	1	0	1	0	0	0
3	2	0	0	0	0	1	0
4	2	1	0	0	0	0	0

- ✓ The shape of data: 7 columns (1 label, 6 categorical variables)
- ✓ All categorical variables are encoded in 1/0 format
- ✓ The total number of rows is 96 662
- ✓ Both Severity Codes 1 and 2 have an equal amount of observations (48 331 each)

Next, let us create an exploratory data frame, which we will use for trend investigation, statistical summary, and dynamic map plotting. Below is what it looks like:

	SEVERITYCODE	X	Y	PERSONCOUNT	INCDATE	ADDRTYPE
0	2	-122.323148	47.703140	2	2013/03/27 00:00:00+00	Intersection
1	1	-122.347294	47.647172	2	2006/12/20 00:00:00+00	Block
2	1	-122.334540	47.607871	4	2004/11/18 00:00:00+00	Block
3	1	-122.334803	47.604803	3	2013/03/29 00:00:00+00	Block
4	2	-122.306426	47.545739	2	2004/01/28 00:00:00+00	Intersection

Column	Description
SEVERITYCODE	1 for Property Damage, 2 for Injury
X,Y	Coordinates of occurred accident
PERSONCOUNT	How many people were involved in the accident
INCDATE	When did the accident happen
ADDRTYPE	Whether an incident occurred at an intersection or block.

- ✓ The shape of data: 6 columns (1 label, 5 exploratory variables)
- ✓ The total number of rows is 194 673

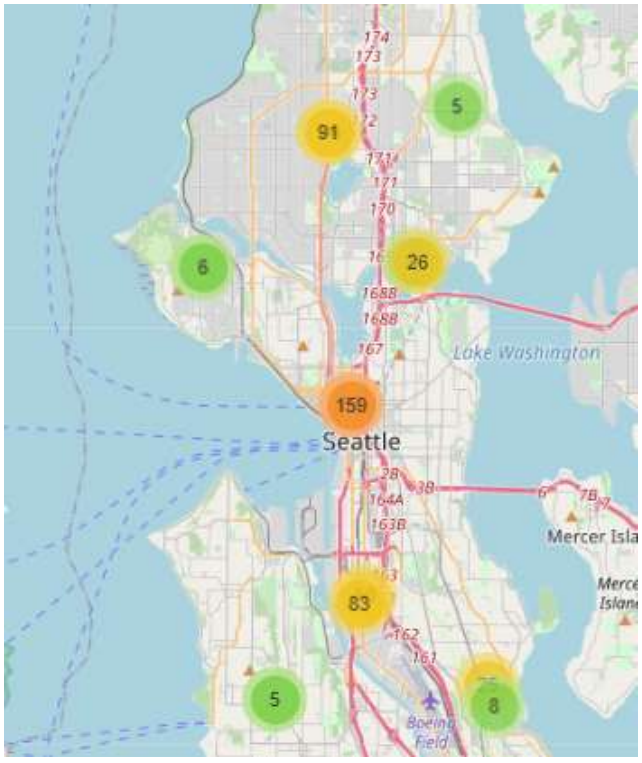
We can now move to our next “Methodology” section, where we will make a quick visual exploratory analysis, and then try to use some Machine Learning techniques to predict what conditions can be linked to severity codes.

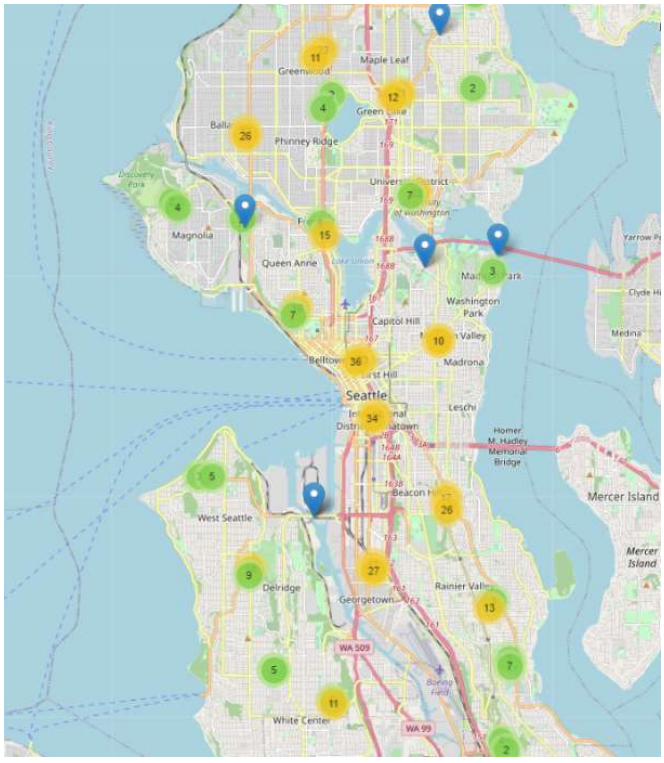
3. METHODOLOGY

This section is the core of our research. Here we will explore the data and try to come up with some scientific conclusions on what attributes impact the chance of injury during a car accident.

I assume that majority of my readers (including myself before I got the assignment) are not familiar with how Seattle looks like and what is the geographic layout of this town. It will be helpful to start our research with anchoring to something visual.

We will build a map of the city and include their dynamic accident markers. For that purpose, we will use the Folium library in Python and create another data frame, which will contain only X & Y coordinates alongside with ADDRTYPE label. To save the computational cost and time, we will use a random sample of 400 observations from the data set. This should be sufficient enough to provide us with the basic idea of the accident density based on the city area. All the details of the code are provided in the notebook. Below are the maps:

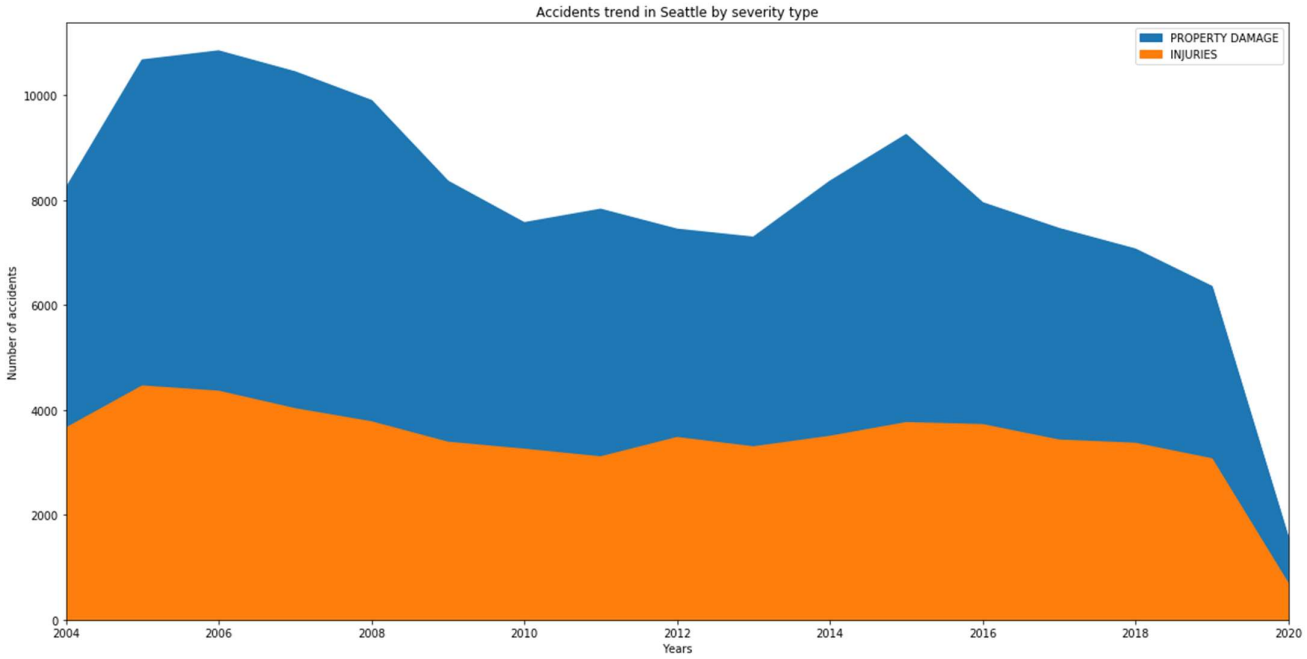




It is clear from the maps that the majority of accidents in Seattle happen around downtown and major highways.

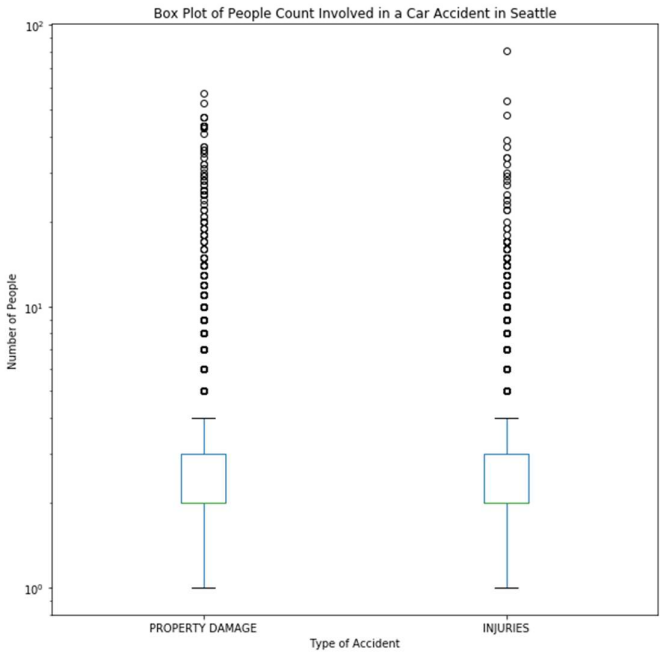
Next, let us have a brief look at accidents trends over the past 15 years. For that purpose, I will create a new data frame containing SEVERITYCODE and INCDATE (date of the incident),

group the dataset based on years, and count the number of occurrences of each accident by severity within each year. Below is the plot I was able to generate after processing the data:



It appears that overall, the trend of property damage incidents is decreasing. The trend of incidents that ended up with injury is decreasing too, but not as sharp. The peak value for both accident types has been recorded in 2005.

Now let us have a quick look at the statistical summary of the number of people involved in each accident type, to determine whether there is any significant statistical difference of those values for our future classification algorithm. To achieve this, I have created a separate data frame which contains SEVERITYCODE and PERSONCOUNT. Statistical summary and the boxplot are displayed below side-by-side:

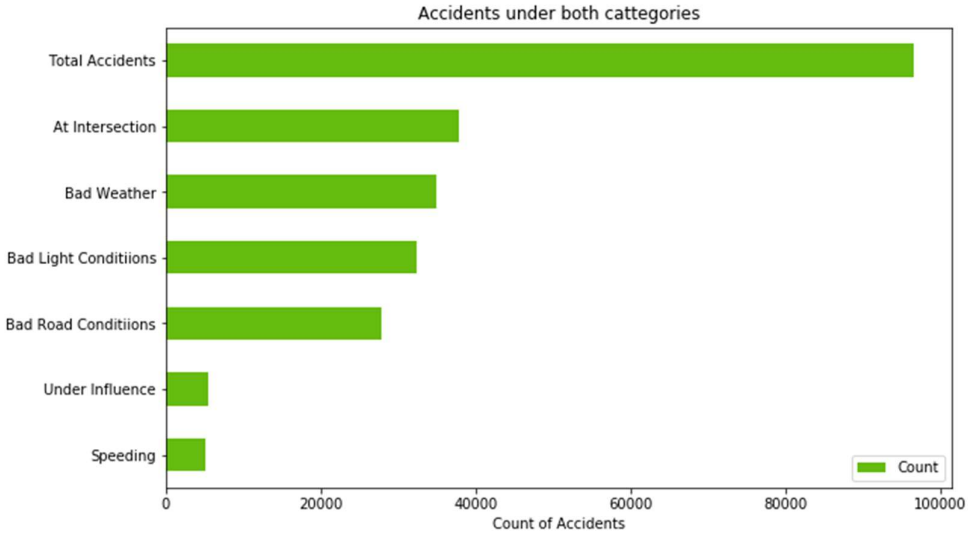


	PROPERTY DAMAGE	INJURIES
count	136,485.00	58,188.00
mean	2.33	2.71
std	1.23	1.54
min	0.00	0.00
25%	2.00	2.00
50%	2.00	2.00
75%	3.00	3.00
max	57.00	81.00

It seems like there is no significant difference between the count of people involved in both types of accidents, and therefore we should not include it in the classification algorithm.

Let us start investigating the core data by looking at the visual representation of accident features, to have an idea about their frequency of occurring.

The figure below is a bar chart of features count in our processed data set:

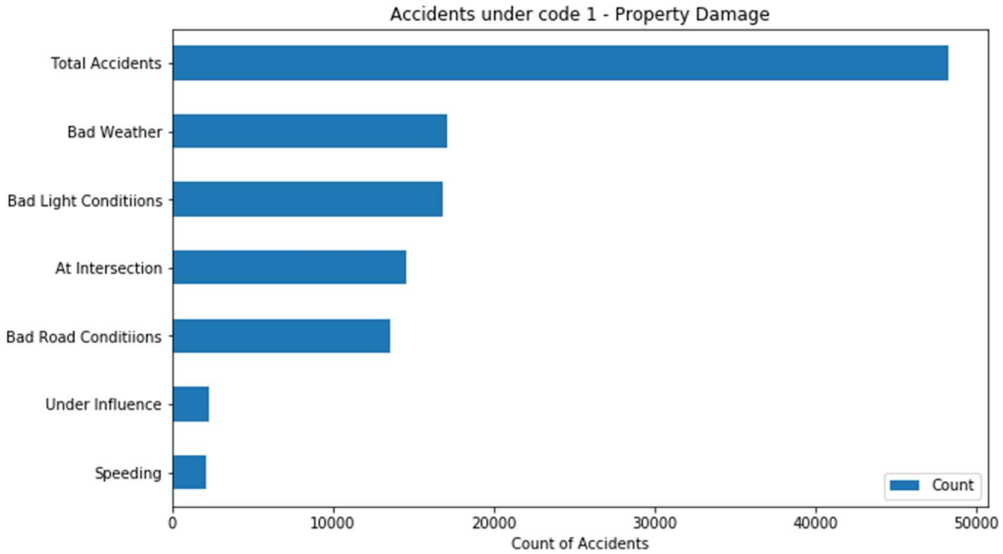


	Count
Total Accidents	96,662
At Intersection	37,897
Bad Weather	34,887
Bad Light Conditiions	32,295
Bad Road Conditiions	27,799
Under Influence	5,426
Speeding	5,171

We can see that most of the accidents overall happen at intersections and during bad weather conditions, such as rain, snow, etc. Speeding and alcohol/drugs influence does not occur very often.

Next, we will split our dataset into two based on the severity code (category). This will help us to uncover if there is any significant correlation between severity and features.

The below figure is a bar chart of features count for accidents under code 1 (property damage).

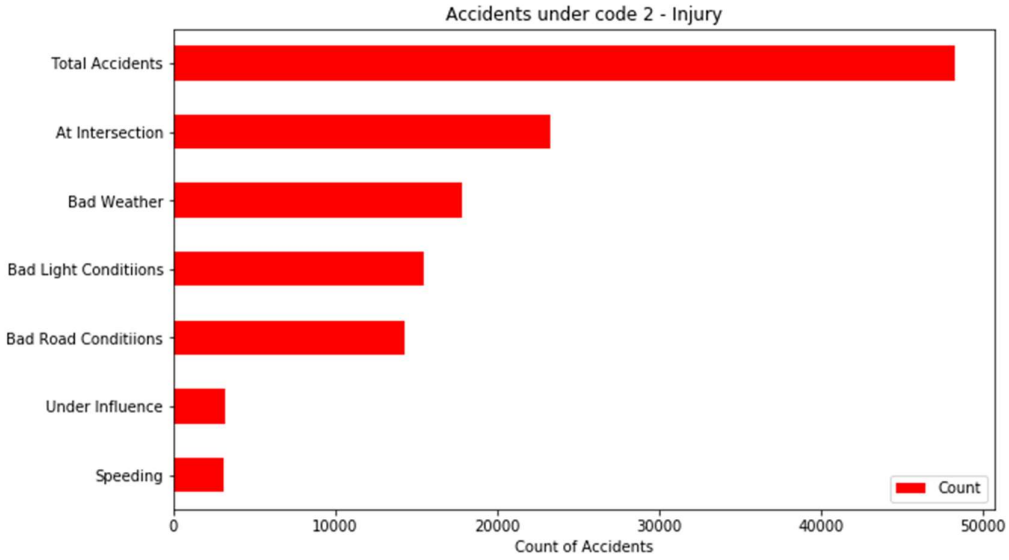


	Count
Total Accidents	48,331
Bad Weather	17,063
Bad Light Conditiions	16,846
At Intersection	14,578
Bad Road Conditiions	13,527
Under Influence	2,283
Speeding	2,064

Here we can see that accidents that caused property damage mainly have bad weather (17, 063 cases or 35%) and bad light conditions attributes (16,846 cases or 34%).

Over-speeding and under influence conditions are met quite rarely.

The next figure represents the same details for accidents under code 2, which caused the injury.



	Count
Total Accidents	48,331
At Intersection	23,319
Bad Weather	17,824
Bad Light Conditiions	15,449
Bad Road Conditiions	14,272
Under Influence	3,143
Speeding	3,107

Here we can see that the main attribute has changed. It appears that accidents that caused the injury, happen at an intersection more frequently than the same of property damage (23,319 cases against 14,578).

The number of drivers who were under influence of drugs or were disobeying speed limits is significantly higher as well (+50%), even though the absolute values remain low.

Bad weather, light, and road conditions do not have any visible difference.

The intermediate conclusion that can be made is that Intersections, driving under alcohol/drugs, and over-speeding are more prompt to cause injuries during the car accident.

Let us try to use some Machine Learning algorithms and see if they confirm our intuition.

Since our data is categorical, we are going to use classification algorithms and investigate which features have more significant coefficients during the classification process.

For this task, I decided to use Logistic Regression from sklearn library. You may refer to the notebook for a detailed description of the code.

A quick summary of model parameters:

- Train/test split: 80/20%
- Random state: 4
- Solver: sag

- Parameter “C”: 0.05

The results are available in the next section

4. RESULTS

Our model did not score a lot in terms of statistical metrics. You may refer to the screenshot on the side. However, the predictability of the model was not my main intention. More than that, I was using this tool to determine the significance of each selected attribute, so we can derive a generalized conclusion regarding the cause of accidents with injury.

	METRICS	SCORE
0	Jaccard	0.60
1	F1	0.60
2	LogLoss	0.66

As I suspected earlier, ADDRTYPE (Intersection or not), Influence of Drugs/Alcohol, and Speeding are the main factors taken into consideration by our model, while classifying the accident, and therefore have higher coefficients.

Weather, Road, and Light condition are not of major importance.

	FEATURE	COEFFICIENT
0	ADDRTYPE	0.38
1	UNDERINFL	0.11
2	SPEEDING	0.11
3	WEATHER	0.02
4	ROADCOND	0.01
5	LIGHTCOND	-0.09

5. DISCUSSION

Within this small research, we were able to look at causes of the injuries that occurred during car crashes under a different angle. Common sense tells us that such factors as bad weather or road conditions should be the most dangerous, however, it turns out that such factors as incidents happened at intersection, alcohol, and speeding outweigh everything else and are the biggest risk factors.

Drivers should be especially careful at intersections, obey speed limits, and never drive drunk or high to avoid severe consequences.

It is the responsibility of the government to introduce high fines and effective prosecution for over-speeding and drunk driving, however, a lot here depends on the culture and responsibility of drivers themselves.

Something that depends on drivers less is the metropolitan transport infrastructure. Cities should try to reduce the number of dangerous intersections and consider overhead bridges, underground tunnels, or multi-layer highways, wherever possible. Offloading traffic from city-centers will help a lot too.

It seems though that city of Seattle is doing a good job to reduce the number of accidents, which has been discovered by plotting the trend earlier. But it is still not enough and far from perfection.

6. CONCLUSION

In this research, we were able to find the most dangerous factors causing injuries during a car accident. To achieve this, we used Seattle Traffic Collisions Dataset, processed it according to our needs, reviewed it with the help of graphics, and applied Logistic Regression to determine the factors which weigh the most.

Intersections, drunk driving, and over-speeding contribute the most to injuries, and as the result, to a terrible human lives toll which roads and highways claim every year.

We hope that fellow readers, equipped with this knowledge, will do their best to stay safe while driving.

Our safety depends on our decisions and attitude. Let us prove the ability to be responsible for ourselves as well as for those who we love.