Analysis of “Road Safety Data - Accidents 2019”

Nowres Al-Rubaie

Introduction

The dataset that I have collected is the "Road Safety Data - Accidents 2019" dataset, available at: https://data.gov.uk/dataset/road-accidents-safety-data

The dataset contains 117,536 records, with each record representing a single accident. It includes information about the location of the accident, the types of vehicles involved, the severity of the accident, the date and time and the number of casualties.

The dataset is reliable as it is from the .gov website and has been collected, recorded and published by the authorities.

A different dataset (named Road-Safety-Open-Dataset-Data-Guide) on the same webpage (refer to the top of this page for the link) had a key defining what each number represents in each column. Thus, I implemented this key onto the dataset.

Pre-processing Steps Discussion

I have used Microsoft Excel to pre-process the data. The following pre-processing steps were taken:

1. Irrelevant columns (noisy data) were removed. This is known as data cleaning. E.g: year_of_accident, local_authority_district and accident_index were removed as such infso is not necessary in investigating the hypothesis.
2. Data integration. A different dataset (named Road-Safety-Open-Dataset-Data-Guide) on the same webpage (refer to the top of this page for the link) had a key defining what each number represents in each column. Thus, I implemented this key onto the dataset.
3. Data transformation: Attribute selection – only important attributes from the keys were selected. Information such as “data missing” and “unknown” were not considered in the selection.
4. The "Weather_Conditions" column was recoded to 4 categories: "Fine", "Rain", “Snow” and “Fog” (removing noisy data) (“Windy” data not considered as important).
5. The "Road_surface_conditions" column was recoded to 5 categories: "Dry", "Wet or damp", “snow”, “frost or ice” and “flood over 3cm deep”.
6. The “Light_Conditions” column was recorded to 4 categories: “Daylight”, “Darkness – lights lit”, “Darkness – lights unlit” and “Darkness – no lighting”.
7. The “junction_detail” column was recorded to 4 categories: “Roundabout”, “T or staggered junction”, “Slip road” and “Crossroads”.

Raw Dataset Screenshot

	A	B	C	D	E	F	G	H	I	J	K
1	accident_index	accident_year	accident_reference	location_easting_osgr	location_northing_osgr	longitude	latitude	police_force	accident_severity	number_of_vehicles	number_of_casualties
2	2.02E+12	2019	10128300	528218	180407	-0.153842	51.508057	1	3	2	3
3	2.02E+12	2019	10152270	530219	172463	-0.127949	51.436208	1	3	2	1
4	2.02E+12	2019	10155191	530222	182543	-0.124193	51.526795	1	3	2	1
5	2.02E+12	2019	10155192	525531	184605	-0.191044	51.546387	1	2	1	1

	L	M	N	O	P	Q	R	S	T	U	V
1	date	day_of_week	time	local_authority_district	local_authority_ons_district	local_authority_highway	first_road_class	first_road_number	road_type	speed_limit	junction_detail
2	18/02/2019	2	17:50	1	E09000033	E09000033	3	4202	1	30	1
3	15/01/2019	3	21:45	9	E09000022	E09000022	3	23	2	30	0
4	01/01/2019	3	01:50	2	E09000007	E09000007	4	504	6	30	3
5	01/01/2019	3	01:20	2	E09000007	E09000007	4	510	6	20	3

	W	X	Y	Z	AA	AB	AC
1	junction_control	second_road_class	second_road_number	pedestrian_crossing_human_control	pedestrian_crossing_physical_facilities	light_conditions	weather_conditions
2	2	3	4202	0	5	1	1
3	-1	-1	-1	9	9	4	1
4	4	6	0	0	0	4	1
5	4	4	510	0	0	4	1

	AD	AE	AF	AG	AH	AI	AJ
1	road_surface_conditions	special_conditions_at_site	carriageway_hazards	urban_or_rural_area	did_police_officer_attend_scene_of_accident	trunk_road_flag	lsoa_of_accident_location
2	1	0	0	1	3	2	E01004762
3	1	0	0	1	3	2	E01003117
4	1	0	0	1	1	2	E01000943
5	1	0	0	1	1	2	E01000973

urban_or_rural_area	1	Urban
urban_or_rural_area	2	Rural
urban_or_rural_area	3	Unallocated
urban_or_rural_area	-1	Data missing or out of range

junction_detail	0	Not at junction or within 20 metres
junction_detail	1	Roundabout
junction_detail	2	Mini-roundabout
junction_detail	3	T or staggered junction
junction_detail	5	Slip road
junction_detail	6	Crossroads
junction_detail	7	More than 4 arms (not roundabout)
junction_detail	8	Private drive or entrance
junction_detail	9	Other junction
junction_detail	99	unknown (self reported)
junction_detail	-1	Data missing or out of range

light_conditions	1	Daylight
light_conditions	4	Darkness - lights lit
light_conditions	5	Darkness - lights unlit
light_conditions	6	Darkness - no lighting
light_conditions	7	Darkness - lighting unknown
light_conditions	-1	Data missing or out of range

road_surface_conditions	1	Dry
road_surface_conditions	2	Wet or damp
road_surface_conditions	3	Snow
road_surface_conditions	4	Frost or ice
road_surface_conditions	5	Flood over 3cm. deep
road_surface_conditions	6	Oil or diesel
road_surface_conditions	7	Mud
road_surface_conditions	-1	Data missing or out of range
road_surface_conditions	9	unknown (self reported)

accident_severity	1	Fatal
accident_severity	2	Serious
accident_severity	3	Slight

weather_conditions	1	Fine no high winds
weather_conditions	2	Raining no high winds
weather_conditions	3	Snowing no high winds
weather_conditions	4	Fine + high winds
weather_conditions	5	Raining + high winds
weather_conditions	6	Snowing + high winds
weather_conditions	7	Fog or mist
weather_conditions	8	Other
weather_conditions	9	Unknown
weather_conditions	-1	Data missing or out of range

Pre-processed Data

	A	B	C	D	E	F	G	H
1	time	speed_limit	junction_detail	light_conditions	weather_conditions	road_surface_conditions	urban_or_rural_area	accident_severity
2	17:50	30	1	1	1	1	1	3
3	01:50	30	2	2	1	1	1	3
4	01:20	20	2	2	1	1	1	2
5	00:40	30	4	2	1	1	1	3
6	01:35	30	4	2	1	1	1	3
				1	1	1	1	3

light_conditions	1	Daylight
light_conditions	2	Darkness - lights lit
light_conditions	3	Darkness - lights unlit
light_conditions	4	Darkness - no lighting
weather_conditions	1	Fine no high winds
weather_conditions	2	Raining no high winds
weather_conditions	3	Snowing no high winds
weather_conditions	4	Fog or mist
road_surface_conditions	1	Dry
road_surface_conditions	2	Wet or damp
road_surface_conditions	3	Snow
road_surface_conditions	4	Frost or ice
road_surface_conditions	5	Flood over 3cm. deep

accident_severity	1	Fatal
accident_severity	2	Serious
accident_severity	3	Slight
junction_detail	1	Roundabout
junction_detail	2	T or staggered junction
junction_detail	3	Slip road
junction_detail	4	Crossroads

urban_or_rural_area	1	Urban
urban_or_rural_area	2	Rural

Hypothesis Discussion:

Based on the data, I have built the following hypothesis:

Analysing the data will help to identify the factors that contribute to road accidents in the UK, which will allow the government to implement measures to reduce the number of accidents and casualties on the roads. Specifically, I will explore the following problems:

1. "More accidents occurred when the speed limit is higher and during the day time". Are there any trends in time and speed limit when accidents occurred? (E.g: Are there more accidents when it is day time and the speed limit is 70?)
2. "More accidents occurred at night time when it was raining". Are there any trends between time and raining weather in regards to accidents? (E.g: More accidents when it's raining at night?)
3. "More accidents occur during daylight than darkness time with no lighting, because at night there are less cars on the road"

These problems are crucial for the UK government because road accidents are a significant public health issue, with over 1,700 deaths and 25,000 serious injuries reported in 2019. By identifying the causes of road accidents and finding areas of high risk, the government can implement measures to decrease the number of accidents and casualties on the roads.

Technique that will be used for proposed solution

For the proposed solution, I shall be using the K-Means Clustering technique. K-Means Clustering is an unsupervised learning algorithm that is used to group similar data points together. Thus, this technique is particularly helpful for finding patterns in data and detecting structures within the data. My dataset contains multiple attributes and I desire to group similar data points together based on these attributes. This will help to discover the areas and factors where accidents are more sever and frequent. As a result, necessary steps can be taken to reduce the number and severity of accidents based on the findings.

K-Means Clustering is suitable for my dataset because it can handle numerical data and is appropriate for identifying patterns in the data. Indeed, the information in the dataset is represented in a numerical form, with the key defining what each number represents. Additionally, K-Means Clustering is a simple and efficient algorithm, making it suitable for handling large datasets.

An advantage of K-Means Clustering is that it is easy to implement and interpret the results. The algorithm works by randomly selecting K number of centroids and then assigning each data point to the nearest centroid. After this, the algorithm updates the centroids based on the mean of the data points assigned to each centroid. The process continues until the centroids no longer change, and the data points are grouped together based on their proximity to each centroid.

K-Means Clustering is also beneficial for identifying outliers in the data. Outliers are data points that do not fit well within any of the clusters. This could help to identify attributes that may not have correlation and effect with road accidents.

I shall be comparing the data columns listed in points 1 to 3 (located under my hypothesis) with the “accident_severity” column for the clustering. For example: I shall cluster and compare weather conditions, light conditions and geographical areas with “accident_severity”. This should help me to conclude which factors affect the severity of an accident.

In conclusion, K-Means Clustering is an appropriate technique for my proposed solution. It is efficient, easy to implement, and can handle large datasets. Further, it can group similar data points together based on the attributes of the data. Finally, it is also useful for identifying outliers in the data, allowing me to identify attributes that may not be related to road accidents.

Section 1 of hypothesis

1. “More accidents occurred when the speed limit is higher and during the day time”. Are there any trends in time and speed limit when accidents occurred? (E.g: Are there more accidents when it is day time and the speed limit is 70?)

	time	speed_limit
0	1750	30
1	150	30
2	120	20
3	40	30
4	135	30
...	...	...
50473	2220	30
50474	1450	30
50475	1055	60
50476	1530	60
50477	1410	30

50478 rows × 2 columns

Step 1: Displaying “time” and “speed_limit” column

```
array([[1750,   30],
       [ 150,   30],
       [ 120,   20],
       ...,
       [1055,   60],
       [1530,   60],
       [1410,   30]])
```

Step 2: Placing columns in an array

	time	speed_limit	cluster
0	1750	30	1
1	150	30	0
2	120	20	0
3	40	30	0
4	135	30	0
...	...	...	...
50473	2220	30	1
50474	1450	30	2
50475	1055	60	0
50476	1530	60	2
50477	1410	30	2

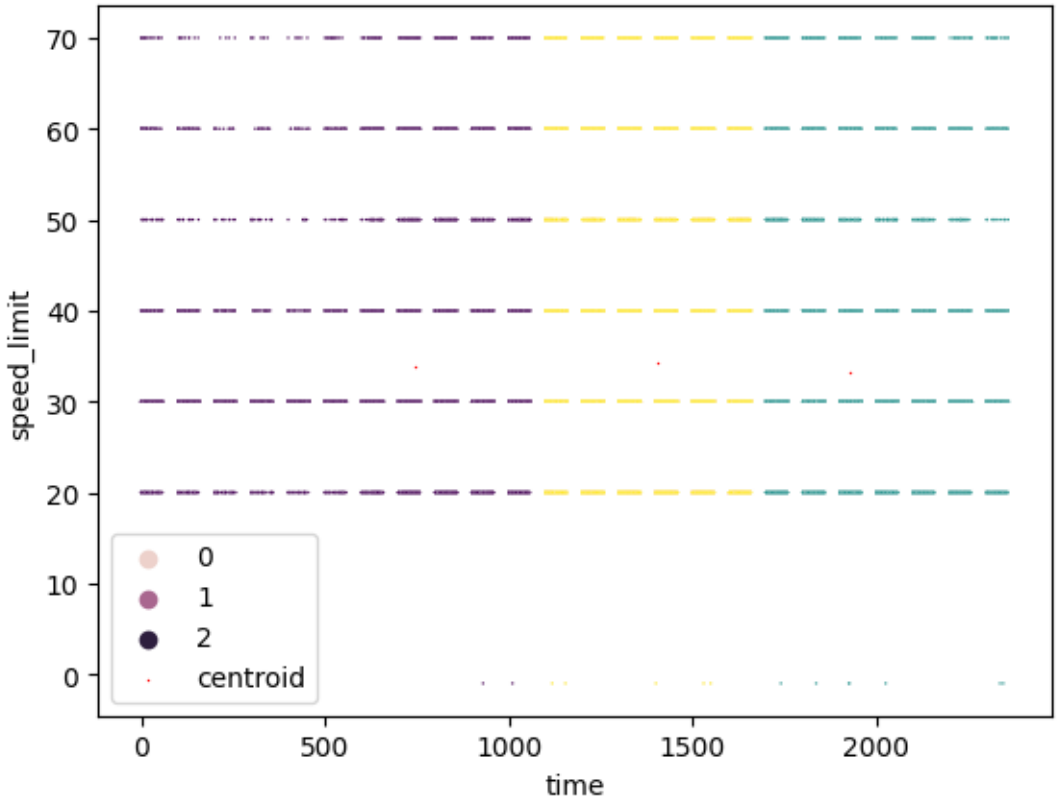
50478 rows × 3 columns

Step 3: Each value assigned a cluster (0,1,2) (k = 3)

	0	1
0	743.363829	33.924491
1	1927.690039	33.177698
2	1405.669758	34.199242

Step 4: Centroids for each cluster calculated

The four diagrams above display the general steps taken to investigate hypothesis 1. After importing the csv file, I displayed only the “time” and “speed_limit” column. I then placed it into an array, calculated the k value as 3. After, each column was assigned a cluster and the centroids for clusters 0, 1 and 2 were calculated. The final diagram is produced below:



Section 2 of hypothesis

2. "More accidents occurred at night time when it was raining".
   Are there any trends between time and raining weather in regards to accidents?
   (E.g: More accidents when it's raining at night?)

	time	weather_conditions
0	1750	1
1	150	1
2	120	1
3	40	1
4	135	1
...	...	...
50473	2220	1
50474	1450	1
50475	1055	1
50476	1530	1
50477	1410	1

50478 rows × 2 columns

Step 1: Displaying "time" and "weather_conditions" column

```
array([[1750,    1],
       [ 150,    1],
       [ 120,    1],
       ...,
       [1055,    1],
       [1530,    1],
       [1410,    1]])
```

Step 2: Placing columns in an array

	time	weather_conditions	cluster
0	1750	1	0
1	150	1	1
2	120	1	1
3	40	1	1
4	135	1	1
...	...	...	...
50473	2220	1	0
50474	1450	1	2
50475	1055	1	1
50476	1530	1	2
50477	1410	1	2

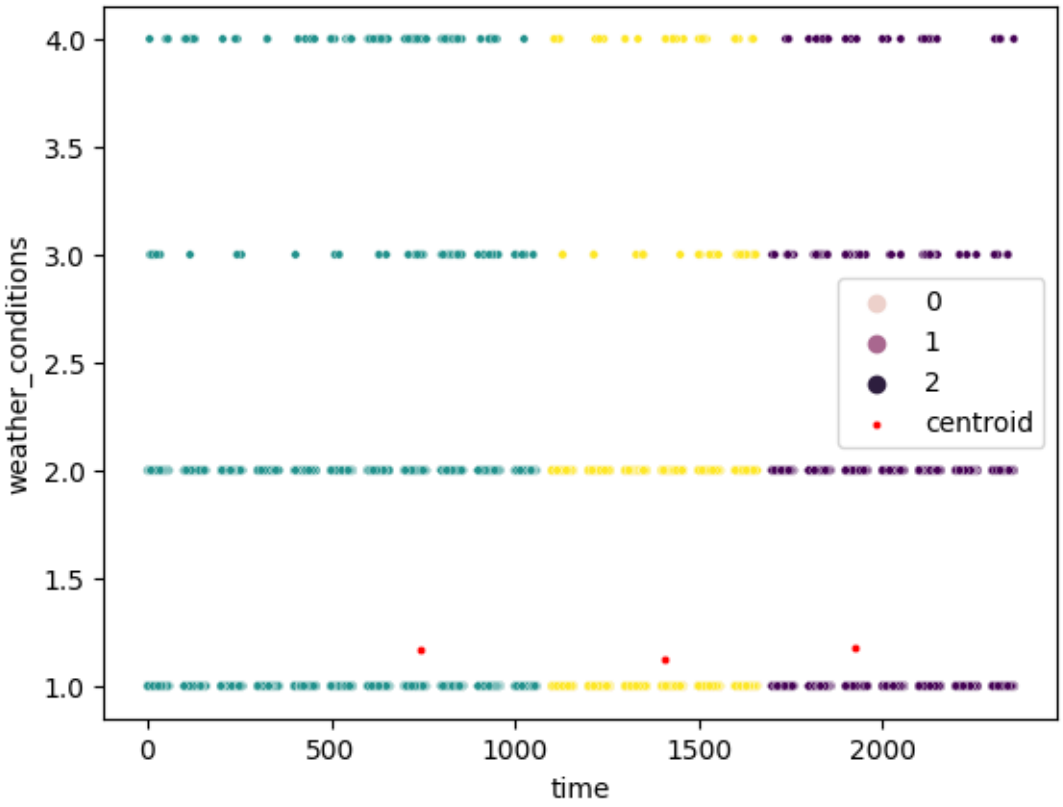
50478 rows × 3 columns

Step 3: Each value assigned a cluster (0,1,2) (k = 3)

	0	1
0	1927.690039	1.180494
1	743.363829	1.166997
2	1405.669758	1.123623

Step 4: Centroids for each cluster calculated

The four diagrams above display the general steps taken to investigate hypothesis 2. After importing the csv file, I displayed only the "time" and "weather_conditions" column. I then placed it into an array, calculated the k value as 3. After, each column was assigned a cluster and the centroids for clusters 0, 1 and 2 were calculated. The final diagram is produced below:



Section 3 of hypothesis

3. “More accidents occur during daylight than darkness time with no lighting, because at night there are less cars on the road”

	time	light_conditions
0	1750	1
1	150	2
2	120	2
3	40	2
4	135	2
...	...	...
50473	2220	2
50474	1450	1
50475	1055	1
50476	1530	1
50477	1410	1

Step 1: Displaying “time” and “light_conditions” column

```
array([[1750,    1],
       [ 150,    2],
       [ 120,    2],
       ...,
       [1055,    1],
       [1530,    1],
       [1410,    1]])
```

Step 2: Placing columns in an array

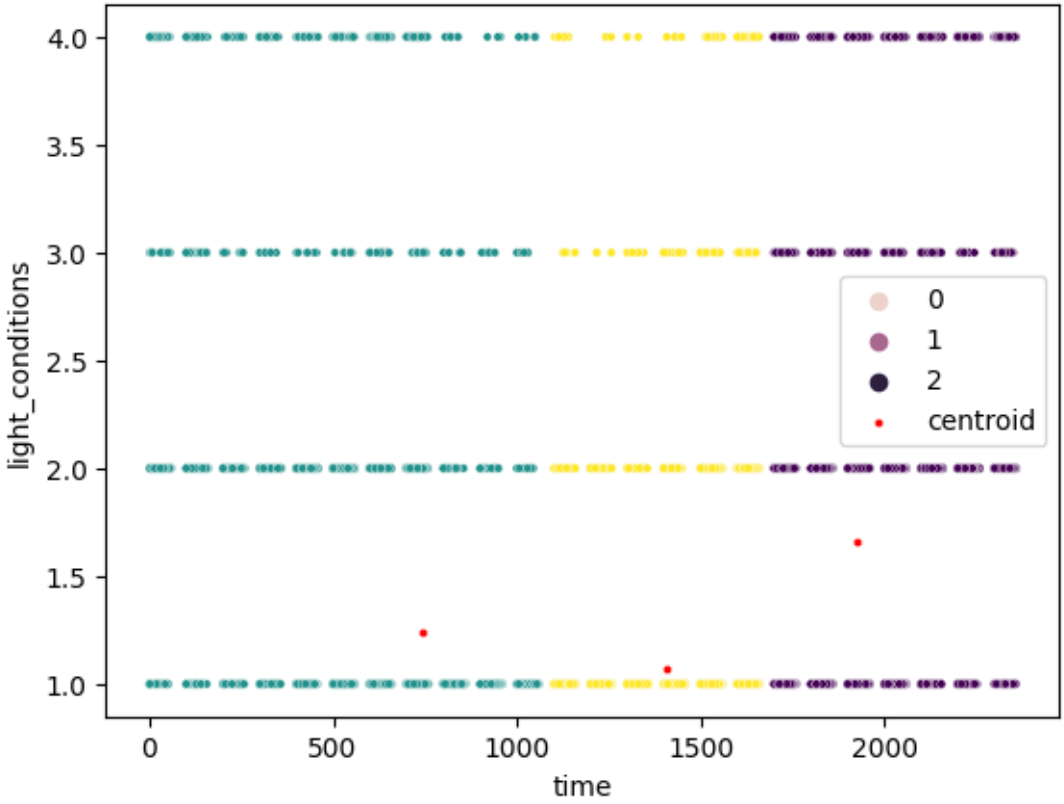
	time	light_conditions	cluster
0	1750	1	0
1	150	2	1
2	120	2	1
3	40	2	1
4	135	2	1
...	...	...	...
50473	2220	2	0
50474	1450	1	2
50475	1055	1	1
50476	1530	1	2
50477	1410	1	2

Step 3: Each value assigned a cluster (0,1,2) (k = 3)

	0	1
0	1927.690039	1.659054
1	743.363829	1.242788
2	1405.669758	1.067370

Step 4: Centroids for each cluster calculated

The four diagrams above display the general steps taken to investigate hypothesis 2. After importing the csv file, I displayed only the “time” and “light_conditions” column. I then placed it into an array, calculated the k value as 3. After, each column was assigned a cluster and the centroids for clusters 0, 1 and 2 were calculated. The final diagram is produced below:



Arguments

Section 1 of hypothesis:

More accidents occur when the speed limit is 20mph compared to the speed limits above 20mph (30,40,50,60,70mph). Therefore, the statement that more accidents occur when the speed limit is higher is incorrect.

Further, there is a significant drop of accidents occurring during the early hours of the morning (between midnight and 5am) on roads with speed limits of 50,60 and 70mph. However, on roads with speed limits of 40mph or less, there is not a huge drop of accidents compared to the day time.

To conclude, more accidents occur when the speed limit is 20mph, this is most likely because roads that are of 20mphs are more used compared to motorways (especially between midnight and 5am). Also, 20mph roads have more complexity when driving such as roundabouts, turns or crossings, etc.

Moreover, accidents occurred at night-time roughly the same amount as during the day (apart from high speed roads of 50, 60 and 70mph).

As a result, a scheme could be introduced to teach drivers how to drive more safely on 20mphs roads. Ensuring that they are aware that there are more cars on a 20mph road and that they should maintain a good distance from other cars. Also, drivers could be taught how to drive more safely at night (especially on roads that are 20mph), ensuring that they take a stop and rest if they feel tired, looking out at zebra crossings, turns and having sharper vision. Moreover, drivers should never drive under the influence of drugs or alcohol.

Section 2 of hypothesis:

More accidents occurred when the weather was fine and raining compared to snow and fog. Both fine and raining conditions had a similar amount of accidents throughout the day and the night. Therefore, the hypothesis is partially incorrect, as both rain and fine conditions showed a consistent and equal amount of accidents occurring throughout the whole day and night. However, it is partially correct as less accidents occurred at night in snow and fog conditions compared to rain.

There were less accidents when it was raining and foggy throughout the day and night. This is most likely because these two weather conditions do not occur that often in the UK.

As a result, a scheme could be introduced to teach drivers how to drive more safely during wet and fine weather conditions. In wet conditions, this could be by teaching drivers to be more alert when the road is slippery and wet, and to drive more slowly. In dry conditions, this could be by teaching drivers not to get distracted by their mobile phones and other devices when driving. Further, drivers should be aware of the other drivers on the road and anticipate their actions to avoid accidents.

Section 3 of hypothesis:

In general, the hypothesis is correct. More accidents occurred throughout the “daylight” compared to “darkness – no lighting”, “darkness – lights unlit” and “darkness – lights unlit”. However, many accidents still occurred in the “darkness” conditions.

It may be the case that when it is dark, drivers automatically drive more safely as their vision is more limited in the dark. Meanwhile, during the daylight, drivers may be less alert and cautious and driving, which ends up in more accidents occurring.

As a result, a scheme can be implemented to ensure that drivers drive throughout the daylight in a similar way to darkness time. I.e: Checking mirrors and blind spots: Drivers should regularly check their mirrors and blind spots to be aware of their surroundings and avoid collisions. Also, drivers should use turn signals to indicate their intentions when turning, changing lanes, or merging.