SI618 Project 2 Mingyu Li
Analysis of car accident causes and measures to prevent

Motivation

In this project, I am going to look into the factors that are related to car accidents, and what steps could we take to reduce their number and casualties. Car accidents are causing a lot of deaths, while driving is really a convenient way for commuting or going on a trip. This is a topic that may be important to most people, that is, what should we care about, and how can we prevent these tragedies from happening?

To be more specific, I've picked these three research questions:

1. How is the number of accidents related to time?
    a. Does it increase or decrease with years?
    b. Which month, or which day of a week and which hours within a day are most likely to have car accidents?
2. Do driver characteristics and driving conditions influence accidents?
    a. How are driver characteristics: driver gender and driver age related to accident number as well as severity? Is the bias true that women are fatal drivers?
    b. How does light condition, vehicle type and age influence accident number and severity?
3. There are multiple ways to ensure road safety, like adding road signs, limiting the speed, etc. However, unexpected things always happen on the road, making the road condition more complicated.
    a. Do those actions that try to prevent traffic accidents: speed limit, junction control and pedestrian crossing facilities have an effect?
    b. How are the unexpected events: special conditions at site, road hazards and pedestrian movement affect the number of accidents?

Data source

The dataset I've used for this project is from Kaggle: [UK Accidents 10 years history with many variables | Kaggle]. It contains 3 csv files that record the detailed road accident information in the UK from 2005 to 2014. There are about 1,640,000 rows in this dataset, and it contains a lot of objective aspects that are related to each accident. Most of them are categorical, and the references are also provided in a separate csv file.
Since there are so many columns and not all of them will be used, I will provide the sample data in the method part.

Methods

Question 1: How is the number of accidents related to time?

For this question, I'm only using accidents0514.csv, which shows the accurate time for each accident. The null values in time and date are dropped and the data are re-indexed.

Before visualizing the data, I parsed the Date string to add two new columns: year and month. Another column representing the hour of a day is added by parsing the time column. Then I added these three new integer type columns to the original dataframe.

Samples from the cleaned dataset is:

	Accident_Index	Date	Day_of_Week	Time	Year	Month	Hour
0	200501BS00001	04/01/2005	3	17:42	2005	1	17
1	200501BS00002	05/01/2005	4	17:36	2005	1	17
2	200501BS00003	06/01/2005	5	00:15	2005	1	0
3	200501BS00004	07/01/2005	6	10:35	2005	1	10
4	200501BS00005	10/01/2005	2	21:13	2005	1	21

Manipulation of this dataset is relatively easy. When it returns type error by parsing the data, I realized that there are some null values not reflected by Kaggle and used dropna to deal with them.

Question 2: Do driver characteristics and driving conditions influence accidents?

The factors used in this question include: vehicle type, age of vehicle, light conditions, driver gender and driver age. The variables to evaluate are total number of accidents as well as accident severity.

First, combine the accidents with vehicles (vehicles0514.csv) dataframe on same accident index. Then exclude the meaningless data by filtering out the categorical values that are equal to -1 (missing), or 'others'. Before deciding which visualization to use, get an overview distribution of accidents under each category by df.value_counts(). If the values vary too much so that it's hard to view the distribution through histogram, get the count of accidents for each category, or separate into subcategory groups.

Samples for the combined dataset is:

	Accident_Index	Light_Conditions	Sex_of_Driver	Age_Band_of_Driver	Vehicle_Type	Age_of_Vehicle	Accident_Severity
1	200501BS00002	4	1	7	11	3	3
2	200501BS00003	4	1	6	11	5	3
3	200501BS00003	4	1	9	9	6	3
4	200501BS00004	1	2	8	9	4	3
5	200501BS00005	7	1	8	3	10	3

For the light conditions, since the total number of accidents of each condition differs too much, I made the dataframe separated by creating independent datasets for each of them. Since there is no age band for vehicles, I used pd.qcut() to separate the column values into 5 equally sized

subsets. And for the vehicle type, I assigned them into several subcategories: motorcycles, cars, buses, so that the results make more sense. The table that relates category id to its meaning will be listed in the analysis section.
The difficulty in this part is that I need to get the accident distribution table for each column to decide which kind of visualization to use. Some of them have extremely large/small values, so I need to divide them into subcategories and rearrange the data. The reference table is long, hence it took me a lot of time to look up which values represent missing/meaningless data.

Question 3: Do the actions that try to prevent traffic accidents work? How do those unexpected events affect the accidents?
Factors that are considered to be actions preventing accidents are: junction control, speed limit, and pedestrian cross physical facilities. The unexpected events include: pedestrian movement, carriageway hazards, and special conditions at site. Merge the accidents with casual dataframe (casualty0514.csv) on Accident_Index and only keep the columns we need. Samples for the merged dataframe are:

	Accident_Index	Speed_limit	Junction_Control	ped_cross	Special_Conditions_at_Site	Carriageway_Hazards	Pedestrian_Movement
0	200501BS00001	30	-1	1	0	0	1
1	200501BS00002	30	2	5	0	0	0
2	200501BS00003	30	-1	0	0	0	0
3	200501BS00004	30	-1	0	0	0	2
4	200501BS00005	30	-1	0	0	0	0

When performing the analysis for a certain factor, filter the missing values with -1 in that column (can be seen in the sample data), or the values that represent null. However, I keep a complete dataframe that holds all the rows, so that other columns won't be affected when filtering out the missing values in one column.
The tough problems for this question are similar to question 2, that is how to deal with the extreme values, and how to separate them as subcategories. And since there are very few numeric values, it's hard to decide which visualization methods to use, therefore I spent a lot of time on manipulating the data.

Analysis and result

Question 1: How is the number of accidents related to time?
First of all, count the number of accidents for each year, month, hour and day of week separately. A seaborn count plot would be enough to observe the distributions. Figure 1 shows how they are distributed by time.

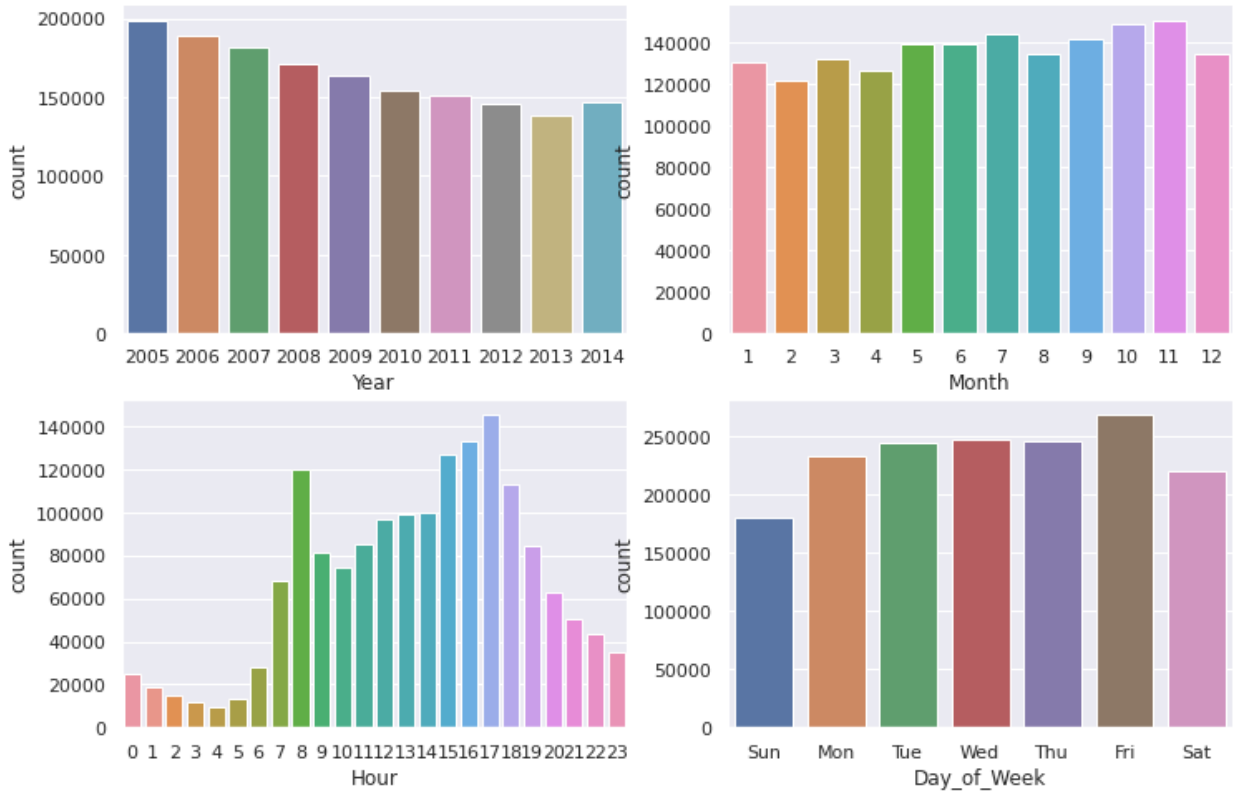

Figure 1. The distribution of accidents for each year/month/hour/day of week

From this plot, we can see that the total number of accidents are decreasing as the year grows, although there is a slight rebound in 2013-2014. The reason may be from many aspects. For example, more difficult assessment for becoming a new driver, more complete road safety facilities, or more road restriction rules, etc.

From the month subplot, there are the most accidents in July, October and November, while there are the fewest accidents in February, April and December. It's harder to analyze why this happens, but I think part of the reason depends on the climatic characteristics in the UK, and also festival effects.

There is a significant local maximum at 8am, and the accident values are quite high from 3pm to 5pm. 8am is the morning rush hour, and it's so early that drivers may be too tired to concentrate. 5pm is the evening rush hour, and it is when the most traffic accidents occur. For the peaks from 3pm to 4pm, maybe it's the time that people are most likely to go out when it is not a workday.

For the last plot, clearly there are the most accidents on Friday, and less accidents on weekends. Friday is the day with the most horrible traffic jams. People are rushing home or planning to go on a short trip, so it is not a surprising result.

I also generated a cross table to observe at which time on a certain day has the most accidents, and then created a seaborn heatmap based on it. This could combine day and hour to get some more precise observations.

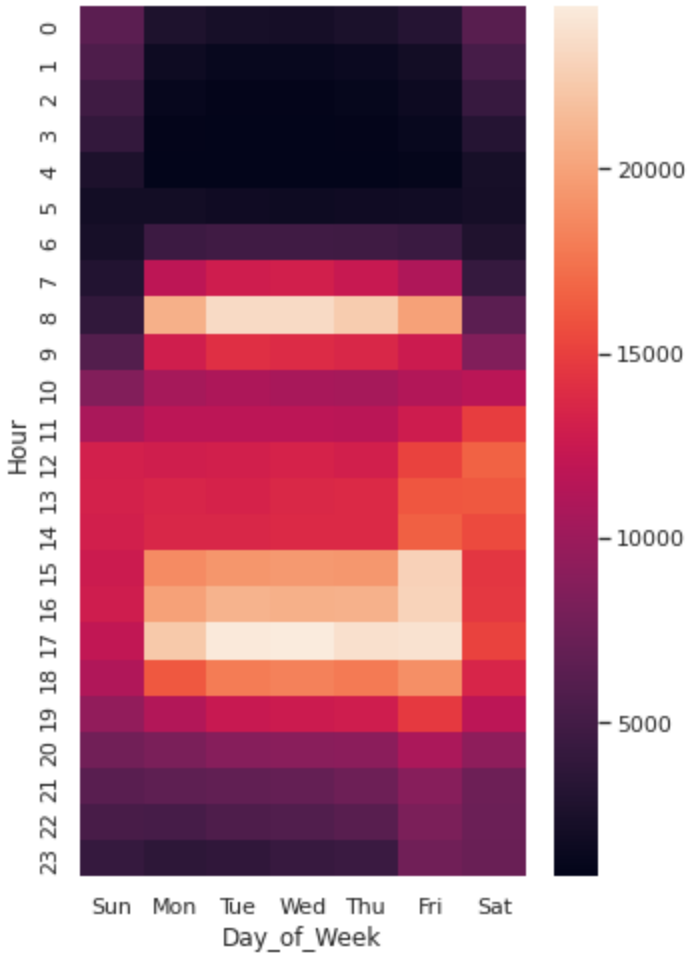

Figure 2. Heatmap of day of week vs. hour

From this map we can get some new conclusions. First, although Friday has the most accidents, the highest density is on Tuesday and Wednesday, at 8am and 5pm. Another interesting conclusion is that weekends have more accidents at midnight. Those who stay up late on weekends are likely to drive carelessly and cause traffic accidents.
In this part, I think all the results concluded from the visualizations are reasonable. They clearly state how the number of accidents varies with time, and also show some interesting results of when traffic accidents are most likely to happen.

Question 2: Do driver characteristics and driving conditions influence accidents?
We will first look into driver's sex. 1 representing male, and 2 representing female. I only keep the records that the Driver_sex column matches these two values. Use a mosaic plot to compare the distribution of accident severity between male and female drivers so that the result would be clearer.

Figure 3. Mosaic plot for male and female drivers' accident severity

Meaning of the severity labels:

code	label
1	Fatal
2	Serious
3	Slight

From this plot, it's clear that male drivers are more likely to create severe accidents. As the severity goes up, the percentage of male drivers keeps increasing. Hence, it is not true that women are more likely to cause fatal accidents.

To visualize the impact of driver's age, since there are 8 possible age ranges in the dataset, and the label increases as the age grows, it would be possible to use a line plot or a count plot. First, I visualized how many accidents of each severity the drivers under different age ranges have caused.

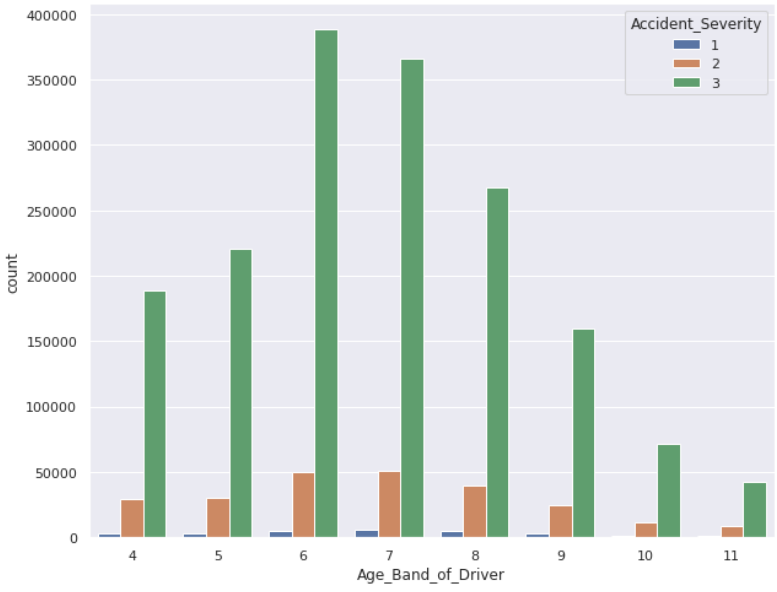

Figure 4. Distribution of driver ages in accidents by severity

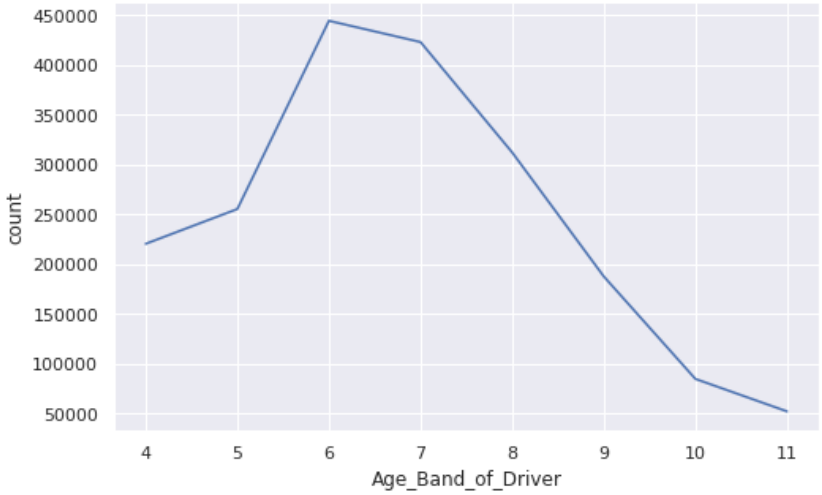

Figure 5. Total accident distribution by age

The representation of the age code is:

code	label
1	0 - 5
2	6 - 10
3	11 - 15
4	16 - 20
5	21 - 25
6	26 - 35
7	36 - 45
8	46 - 55
9	56 - 65
10	66 - 75
11	Over 75
-1	Data missing or out of range

Obviously, those within 26-35 have caused the most accidents, then 36-45. The distributions of each severity are similar. This should be caused by the large number of drivers under this age range, and in age 26-35 there may be more new drivers.

Then we will go into the driving condition, first by considering light conditions on the road. By getting the value counts for light conditions, I noticed that the accident number for different light conditions varies too much, so a hued histogram may not be a good choice. Therefore, I choose the pie chart from plotly to observe how accident severity is distributed under different light conditions. From the count table, most of the accidents happen at daylight because the traffic flow is huge.

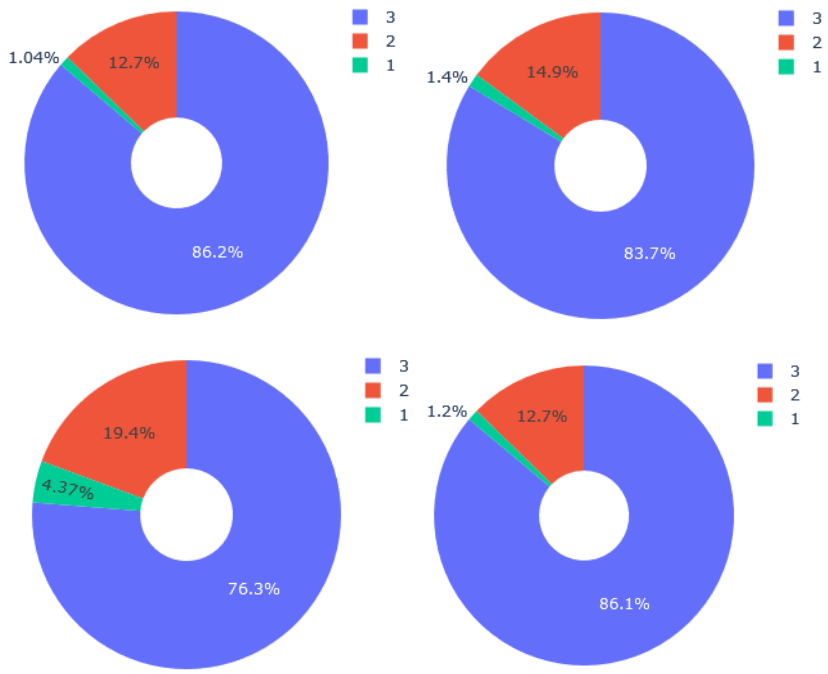

Figure 6. Pie chart of accident severity distribution (From top to bottom, left to right: daylight, darkness-lights lit, darkness-lights unlit, darkness-no lighting, darkness)

Comparing these 4 pie charts, we can see that the 1, 2, 4 condition have similar distributions, while under the lights unlit condition, the severe accidents have much higher rate (serious and fatal). So lighting is still very important for road safety, and drivers should be extremely careful when the lights are broken.

Then consider the vehicle type. These are the representation of vehicle type code that I'll include in this part:

code	label
1	Pedal cycle
2	Motorcycle 50cc and under
3	Motorcycle 125cc and under
4	Motorcycle over 125cc and up to 500cc
5	Motorcycle over 500cc
8	Taxi/Private hire car
9	Car
10	Minibus (8 - 16 passenger seats)
11	Bus or coach (17 or more pass seats)

The subsets could be divided as: 1-5 (cycle/motorcycle), 8-9 (cars), 10-11 (buses), and visualize them separately.

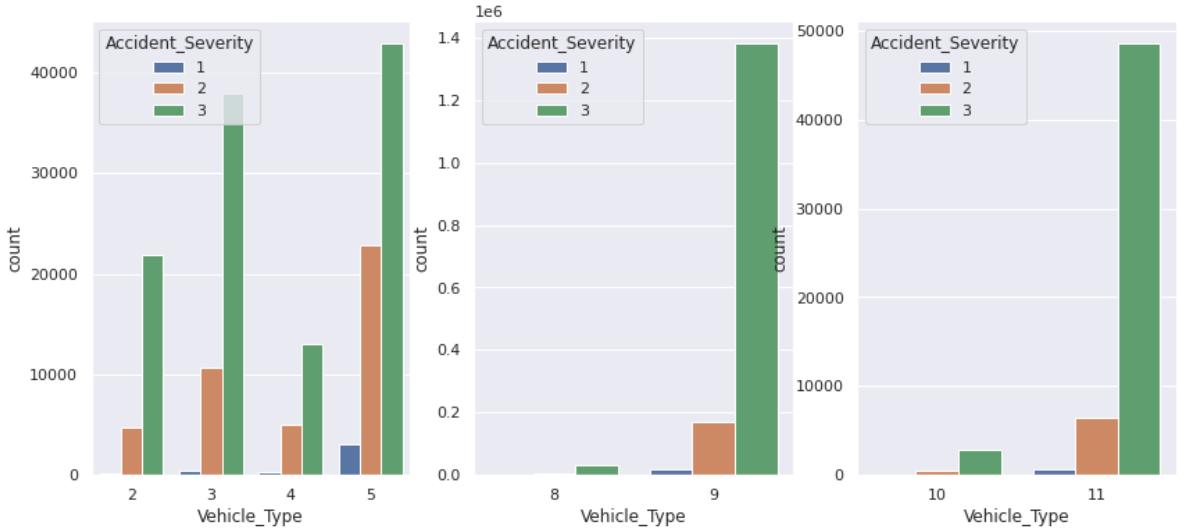

Figure 7. Accident severity distribution across different vehicle types

In the first subset (cycles/motorcycles), Motorcycles over 500cc have caused the most accidents, and also with highest severe accident rates. Drivers of motorcycles with over 500cc may drive quite fast and race with cars, so it's not surprising that they cause more accidents. However, the second largest one is motorcycles with 125cc under. I'm not sure why it has a much larger number than label 4, which is above 125cc.

The majority of accidents are caused by cars (label 9) excluding taxi/private hire cars. The fatality rate of cars is not so high compared with motorcycles. And from the last subplot in Figure 7, large buses are causing more accidents (17 passengers and above), and the fatality rate is similar to cars. Therefore, there is no clear evidence that large buses are more likely to cause deaths than cars. Comparing through all the subplots, the over 500cc motorcycles have the most severe safety problem.

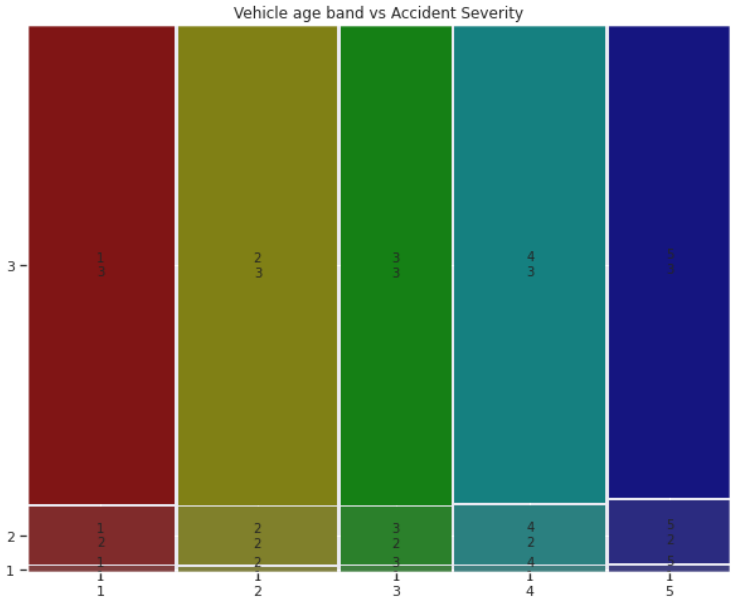

Figure 8. Distribution of vehicle age band vs accidents

The last factor is vehicle age. I divided the vehicle age into 5 equally sized groups, and label them as 1-5 from smallest to largest values. For this part, I only use the rows with vehicle type equal to label 9, that is cars, because this range has a dominant size of data, and restricting the vehicle type would show a more convincing result. Then use a mosaic plot to show the severity distribution.
Overall, the distributions look similar. The label 5, that is the oldest vehicles, have a larger fatality rate. Hence, we may conclude that older vehicles are really more dangerous.

In this part, most of the conclusions are within expectation. However, due to the unbalanced data in vehicle type, it's hard to get further from the current conclusion. Maybe there are better ways to deal with this column, and the method in my project is not a suitable way.

Question 3: Do the actions that try to prevent traffic accidents work? How do those unexpected events affect the accidents?
First, we use a count plot to look at the speed limit.

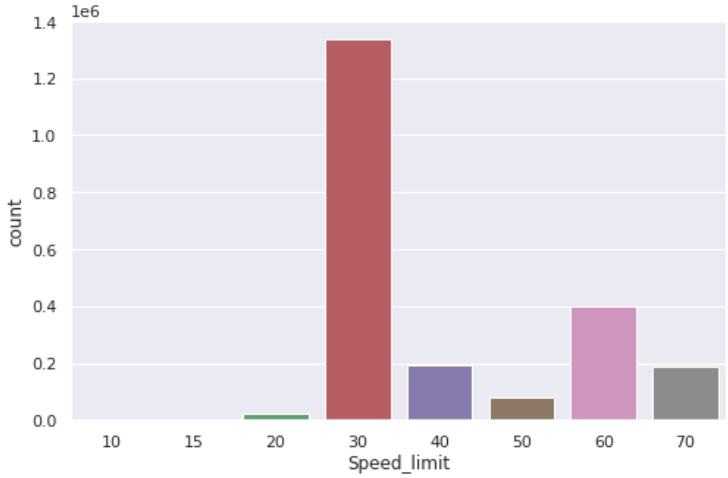

Figure 9. Speed limit vs. accident numbers

From Figure 9, we can see a quite strange distribution. I cannot find any pattern in this plot, so I didn't go on looking into the distribution of severity. I think it may only be related to the number of roads that have a certain speed limit. No matter what the speed limit is, people will always exceed it and get fined.
Then we look into junction controls and pedestrian crossing facilities. First use countplot to see how these two factors make an effect on accident numbers independently.

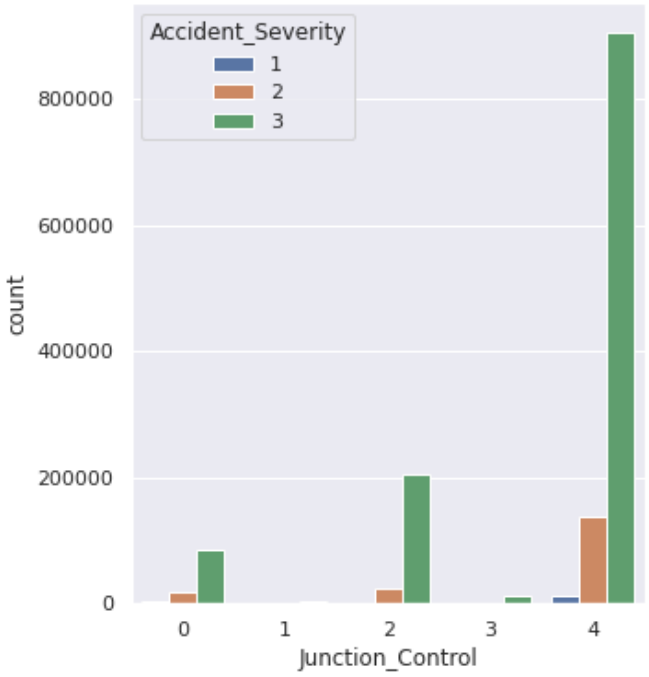

Figure 10. Distribution of accident number for different junction controls

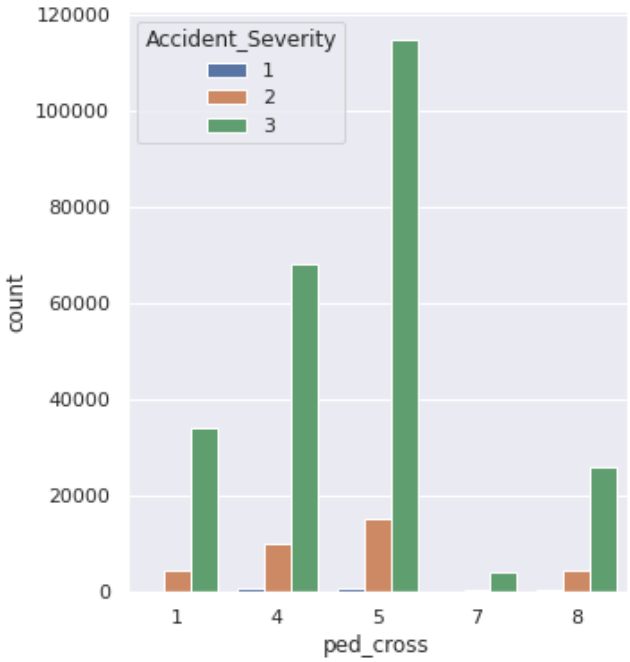

Figure 11. Distribution of accident number for different pedestrian crossing facilities

The reference table for junction control labels is:

code	label
0	Not at junction or within 20 metres
1	Authorised person
2	Auto traffic signal
3	Stop sign
4	Give way or uncontrolled
-1	Data missing or out of range

The reference table for pedestrian crossing is:

code	label
0	No physical crossing facilities within 50 metres
1	Zebra
4	Pelican, puffin, toucan or similar non-junction pedestrian light crossing
5	Pedestrian phase at traffic signal junction
7	Footbridge or subway
8	Central refuge
-1	Data missing or out of range

From Figure 10, at a give way or uncontrolled junction, a traffic accident is most likely to happen. The total number of accidents at such a junction is much higher than any other type. This corresponds to what I've anticipated, that a controlled junction either by a physical facility or by humans would be better than drivers just going through the intersection whenever they want. From Figure 11, most accidents happen at label 5 and 4, which are Pedestrian phase at traffic signal junction and Pelican, puffin, toucan or similar non-junction pedestrian light crossing. This also fits our common sense that a zebra crossing or footbridge/subway is helpful to reduce accidents.

After that, I also used a heatmap to see when these conditions are combined, how likely it is that an accident happens.

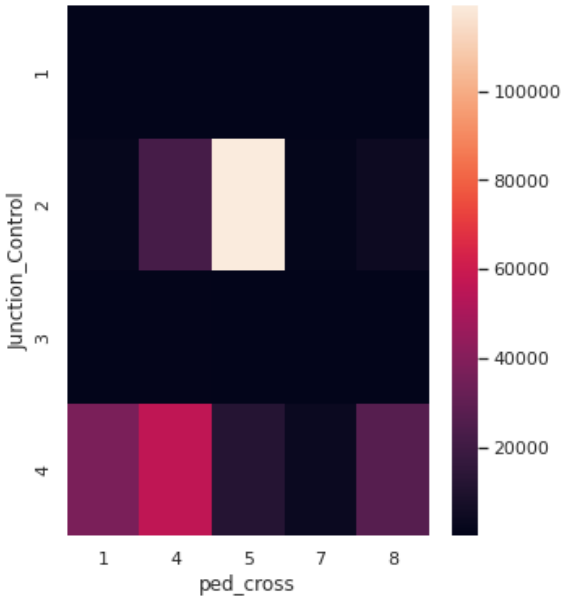

Figure 12. Heatmap for pedestrian crossing facilities and junction controls

Combining these results, we can find that the condition when most accidents happen is when there is an auto traffic signal with pedestrian phase. This one looks safe, but maybe drivers drive fast when they think it is safe, and pedestrians think it doesn't matter even if they cross on a red light, therefore an accident happens.

Finally, we are going to explore some unexpected factors on the road. The process is similar to the previous one. The reference tables are:
Pedestrian movement:

code	label
0	Not a Pedestrian
1	Crossing from driver's nearside
2	Crossing from nearside - masked by parked or stationary vehicle
3	Crossing from driver's offside
4	Crossing from offside - masked by parked or stationary vehicle
5	In carriageway, stationary - not crossing (standing or playing)
6	In carriageway, stationary - not crossing (standing or playing) - masked by parked or stationary vehicle
7	Walking along in carriageway, facing traffic
8	Walking along in carriageway, back to traffic
9	Unknown or other
-1	Data missing or out of range

Special conditions at site:

code	label
0	None
1	Auto traffic signal - out
2	Auto signal part defective
3	Road sign or marking defective or obscured
4	Roadworks
5	Road surface defective
6	Oil or diesel
7	Mud
-1	Data missing or out of range

Carriageway hazards:

code	label
0	None
1	Vehicle load on road
2	Other object on road
3	Previous accident
4	Dog on road
5	Other animal on road
6	Pedestrian in carriageway - not injured
7	Any animal in carriageway (except ridden horse)
-1	Data missing or out of range

Visualizing the accident numbers with these factors, and excluding the non pedestrian values. The count plots are:

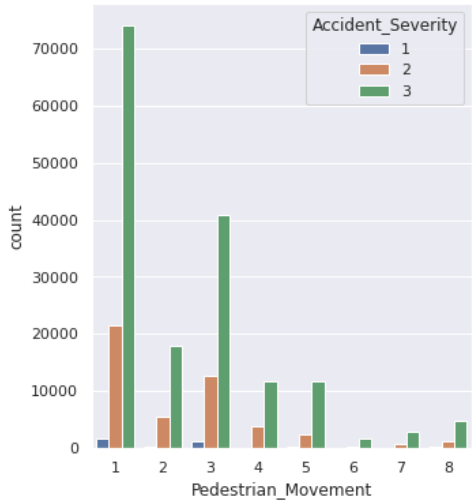

Figure 13. Distribution of pedestrian movement with accidents

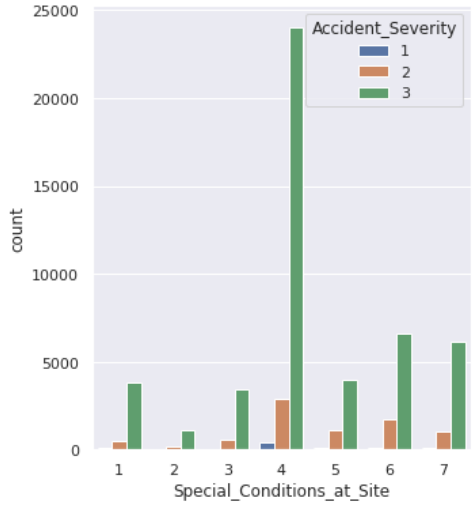

Figure 14. Distribution of pedestrian movement with special condition at site

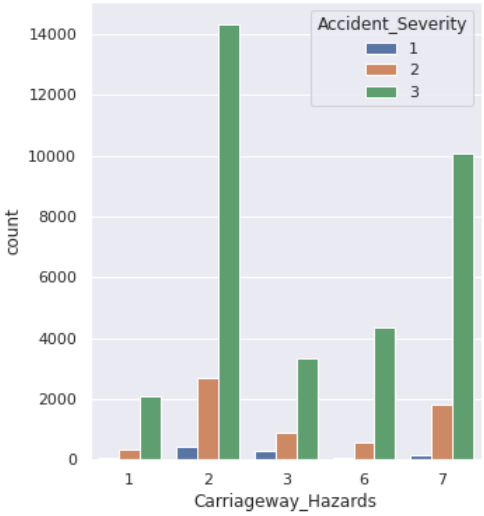

Figure 15. Distribution of pedestrian movement with carriageway hazards

From Figure 13, when the pedestrian is crossing from the driver's nearside, the accident is most likely to happen. Although people may think all the other pedestrian movements are more dangerous than just crossing from the nearside without a block, this is actually where most accidents come from. From Figure 14, the accidents are mostly caused by roadworks/out of traffic signals. A junction out of a traffic signal is just like an uncontrolled one, which aligns with our previous conclusions. And here we can observe that road work is very dangerous for drivers. The distribution for 3 levels of severity for these 2 figures are similar.

From Figure 15, objects other than vehicles on the road are most dangerous for drivers, and the second one is the animals on the road. I also made a mosaic plot based on the data for Figure 15:

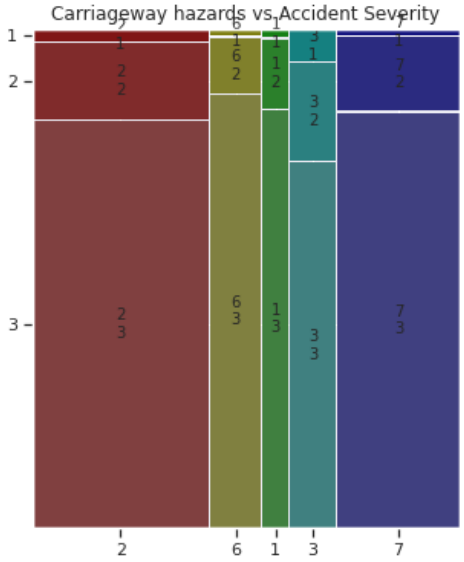

Figure 16. Mosaic plot for carriageway hazards and accident severity

From Figure 16, we can draw a new conclusion that when there are previous accidents on this road, the severity is more likely to be severe.

For this question, most of the answers to the proposed questions are in accordance with people's common sense. However, there are several problems. First, the explanation to some columns is not very clear so it's hard to make an analysis. Second, there should be some better methods to visualize the categorical data, or get an insight of how they are interrelated to each other. Overall, there is still a lot that could be improved.