Reducing the Number of High Fatality Accidents in the UK

Can Data Analysis Inform Interventions to Reduce High Fatality Road Accidents?

John King'athia Karuitha

Tuesday, March 14, 2023

Contents

Background

The World Health Organization (WHO) estimates that approximately 1.3 million people globally die in road accidents every year. Road accidents occur in every country. However, about 93% of the accidents happen in low- and middle-income countries, with 60% of the total vehicles in the world. The monetary loss occasioned by these road accidents amounts to about 3% of the GDP of each country. Importantly, road accidents are the leading cause of death for children and young adults between 5-29 years [1]. The social and economic costs associated with road accidents make it paramount to craft strategies to reduce the number of road accidents and, critically, the fatalities. Coincidentally, at the time of writing this article, the United Nations is leading the world in commemorating victims of road accidents on Sunday, November 21, 2021, dubbed the `World Day of Remembrance for Road Traffic Victims` [2].

In this article, I use data from the department of transport in the UK to garner insights into road accidents and associated fatalities. As a (fictional) employee of the road safety team in the department, my task is

[1] See the WHO summary on road accidents fatalities on https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] See https://www.un.org/en/observances/road-traffic-victims-day for further information about World remembrance day, which happens annually on the third Sunday of November.

to use data to develop insights that would help reduce the number and fatalities of major incidents. Major accidents have at least three casualties. The department seeks to uncover the characteristics of these major incidents to brainstorm interventions that could lower the number of deaths. To this end, the department has asked me to examine the data and answer the following questions.

1. What time of day and day of the week do most major incidents happen?
2. Are there any patterns in the time of day/ day of the week when major incidents occur?
3. What characteristics stand out in major incidents compared with other accidents?
4. On what areas would I recommend the planning team focus their brainstorming efforts to reduce major incidents?

NB: Please click on `code` to unhide the `R` code underlying the results presented.

Data

The reporting department has been collecting data on every reported accident. They've included this along with a lookup file for 2020's accidents. I start by reading in both data sets into `R`. [3]

```
## Read in the accidents data from the data folder.
accidents <- read_csv("data/accident-data.csv", na = "-1") %>%

  ## Convert date and time to date/time format
  mutate(date = dmy(date),

         time = hms(time)) %>%

  mutate(day_of_week = factor(day_of_week,

                              labels = c("Sun", "Mon", "Tue",

                                         "Wed", "Thur", "Fri", "Sat")))

## Structure of the accidents data.
#str(accidents)

## Read in the accidents data from the data folder.
lookup <- read_csv("data/road-safety-lookups.csv")

## Structure of the lookup data
#str(lookup)
```

The `accidents` table has 27 variables and 91199 observations. The `lookup` table has 5 variables, and 129 observations contains the metadata for the main data set, `accidents`.

I begin by examining the data, with reference to missing values and possible duplicates. Table 1 below variables with missing data points and the corresponding number of missing observations.

```
sapply(accidents, is.na) %>%

  colSums() %>%

  tibble(variable_name = names(accidents), missing = .) %>%

  arrange(desc(missing)) %>%
```

[3] *Published by the department for transport. , https://data.gov.uk/dataset/road-accidents-safety-data. Contains public sector information licensed under the Open Government Licence v3.0.*

```
  filter(missing > 0) %>%

  kbl(., booktabs = TRUE, caption = "Table 1: Missing Values") %>%

  kable_classic(full_width = FALSE,

                latex_options = "hold_position")
```

Table 1: Table 1: Missing Values

variable_name	missing
junction_control	38298
road_surface_conditions	316
special_conditions_at_site	218
carriageway_hazards	208
pedestrian_crossing_human_control	143
pedestrian_crossing_physical_facilities	135
longitude	14
latitude	14
speed_limit	12
second_road_number	7
junction_detail	2
light_conditions	1
weather_conditions	1

The data set is reasonably complete except for a glaring gaps in the junction control variable.

```
accidents %>%

        filter(duplicated(.))
```

```
## # A tibble: 0 x 27
## # ... with 27 variables: accident_index <chr>, accident_year <dbl>,
## #   accident_reference <chr>, longitude <dbl>, latitude <dbl>,
## #   accident_severity <dbl>, number_of_vehicles <dbl>,
## #   number_of_casualties <dbl>, date <date>, day_of_week <fct>, time <Period>,
## #   first_road_class <dbl>, first_road_number <dbl>, road_type <dbl>,
## #   speed_limit <dbl>, junction_detail <dbl>, junction_control <dbl>,
## #   second_road_class <dbl>, second_road_number <dbl>, ...
```

Again, the data has no duplicate rows that may indicate an accident captured more than once.

Data Exploration

We summarise the variables in Appendix 1. For these summary statistics, I have eliminated some variables that would yield no meaningful information when summarised, for instance, longitude and latitude. Table 2 below shows the correlation matrix. Of interest are variables with higher correlations with `number of casualties`, the target variable, while bearing in mind that correlation may not capture non-linear relationships. The two variables that show a high correlation with `number of casualties` are `number of vehicles` involved in an accident (0.197) and `speed limit` (0.153).

```
accidents %>%

        ## Seselect redunndant variables
        select(-accident_index, -accident_reference,

         -accident_year, -day_of_week,

         -date, -time,

         -longitude, -latitude, -junction_control) %>%

  relocate(number_of_casualties) %>%

        ## Remove NA rows
        na.omit() %>%

        ## Do correlation and a nice table.
        cor() %>%

        data.frame() %>%

        ## Make a nice table
        kbl(., booktabs = TRUE, caption = "Table 2: Correlation Matrix") %>%

        kable_classic(full_width = TRUE,

                      latex_options = "hold_position",

                      font_size = 10)
```

The Analysis

In this section, we examine each of the questions in the analysis. I start by mapping out the areas where major fatal accidents happen- the hotspots. In this case, I use the density (2D) plot and the hexagonal heat map of 2D bin counts (geom_hex)[4]. The visualisation in Figure 1 shows that accidents happen all over the UK. However, there are hotspots where major accidents (with at least three casualties) are concentrated, leading us to our first insight. Figure 1 shows these accidents hotspots.

```
accidents %>%

        ## Filter out major accidents
        filter(number_of_casualties >= 3) %>%

        ## Plot locations of major accidents
        ggplot(mapping = aes(x = longitude, y = latitude)) +

         ## Add a hex geom
        geom_hex(alpha = 0.5) +

        ## Add a gem dendity 2D
```

[4]See https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjpv8Kkp8f0AhXHDGMBHXKKCPsQFnoECAIQAQ&url=https%3A%2F%2Fggplot2.tidyverse.org%2Freference%2Fgeom_hex.html&usg=AOvVaw1k9D7ncWuUhdgbBeCkxFWQ for details on hexagonal heat maps.

Table 2: Table 2: Correlation Matrix

	number_of_casualties	accident_severity	number_of_vehicles	first_road_class	first_road_number	road_type	speed_limit	junction_detail	second_road_class	second_road_number	pedestrian_crossing_human_control	pedestrian_crossing_physical_facilities	light_conditions	weather_conditions	road_surface_conditions	special_conditions_at_site	carriageway_hazards	urban_or_rural_area
number_of_casualties	1.0000000	-0.0721619	0.1969233	-0.0714877	0.0153349	-0.0107753	0.1527840	-0.0345430	0.0099575	0.0039677	-0.0460355	-0.0528777	0.0286223	-0.0148285	0.0007400	-0.0258306	-0.0274437	0.1179114
accident_severity	-0.0721619	1.0000000	0.0765720	0.0055638	-0.0127589	[illegible]	-0.1179176	0.0586559	-0.0502471	0.0154857	0.0708365	0.0596644	-0.0410348	0.0314524	0.0220876	0.0306033	0.0379883	-0.1142710
number_of_vehicles	0.1969233	0.0765720	1.0000000	-0.1003126	[illegible]	[illegible]	0.0361490	0.0068868	-0.0258223	0.0111854	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
first_road_class	-0.0714877	0.0055638	-0.1003126	1.0000000	-0.2636545	0.2800455	-0.3050582	0.0022734	0.1607087	-0.0682004	[illegible]	-0.1268533	-0.0406602	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
first_road_number	0.0153349	-0.0127589	[illegible]	-0.2636545	1.0000000	-0.0062438	-0.0206682	-0.0250309	-0.0206893	0.1205800	-0.0381798	0.0296077	0.0003184	-0.0055519	-0.0960280	-0.0209807	-0.0815060	0.0145706
road_type	-0.0107753	[illegible]	[illegible]	0.2800455	-0.0062438	1.0000000	-0.1667396	0.1141511	0.2361954	-0.1085774	0.1010956	0.0085348	-0.0115504	0.0387163	[illegible]	0.0750697	0.0973123	-0.0354984
speed_limit	0.1527840	-0.1179176	0.0361490	-0.3050582	-0.0206682	-0.1667396	1.0000000	-0.1425347	0.0316116	-0.0262743	-0.1438059	-0.2448763	0.0757321	-0.0129455	0.0306953	-0.0411696	-0.0407497	0.6712973
junction_detail	-0.0345430	0.0586559	0.0068868	0.0022734	-0.0250309	0.1141511	-0.1425347	1.0000000	-0.0636442	0.0152162	0.3544493	0.2683532	[illegible]	0.0871323	[illegible]	0.2370999	0.2327765	-0.1336158
second_road_class	0.0099575	-0.0502471	-0.0258223	0.1607087	-0.0206893	0.2361954	0.0316116	-0.0636442	1.0000000	-0.4933114	-0.0678294	-0.1829637	0.0042322	-0.0127640	[illegible]	-0.0260785	-0.0093906	0.0406732
second_road_number	0.0039677	0.0154857	0.0111854	-0.0682004	0.1205800	-0.1085774	-0.0262743	0.0152162	-0.4933114	1.0000000	-0.0060260	0.0903749	-0.0015055	-0.0005098	-0.0072195	-0.0084799	-0.0164678	-0.0236122
pedestrian_crossing_human_control	-0.0460355	0.0708365	[illegible]	[illegible]	-0.0381798	0.1010956	-0.1438059	0.3544493	-0.0678294	-0.0060260	1.0000000	0.4717488	[illegible]	[illegible]	[illegible]	[illegible]	0.3895807	-0.1348956
pedestrian_crossing_physical_facilities	-0.0528777	0.0596644	[illegible]	-0.1268533	0.0296077	0.0085348	-0.2448763	0.2683532	-0.1829637	0.0903749	0.4717488	1.0000000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-0.2708051
light_conditions	0.0286223	-0.0410348	[illegible]	-0.0406602	0.0003184	-0.0115504	0.0757321	[illegible]	0.0042322	-0.0015055	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]	[illegible]	[illegible]	0.0485156
weather_conditions	-0.0148285	0.0314524	[illegible]	[illegible]	-0.0055519	0.0387163	-0.0129455	0.0871323	-0.0127640	-0.0005098	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]	[illegible]	-0.0146518
road_surface_conditions	0.0007400	0.0220876	[illegible]	[illegible]	-0.0960280	[illegible]	0.0306953	[illegible]	[illegible]	-0.0072195	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]	[illegible]	0.0136347
special_conditions_at_site	-0.0258306	0.0306033	[illegible]	[illegible]	-0.0209807	0.0750697	-0.0411696	0.2370999	-0.0260785	-0.0084799	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	[illegible]	-0.0354214
carriageway_hazards	-0.0274437	0.0379883	[illegible]	[illegible]	-0.0815060	0.0973123	-0.0407497	0.2327765	-0.0093906	-0.0164678	0.3895807	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.0000000	-0.0404989
urban_or_rural_area	0.1179114	-0.1142710	[illegible]	[illegible]	0.0145706	-0.0354984	0.6712973	-0.1336158	0.0406732	-0.0236122	-0.1348956	-0.2708051	0.0485156	-0.0146518	0.0136347	-0.0354214	-0.0404989	1.0000000

```
      geom_density_2d() +

      ## Add labels and titles
      labs(x = "Longitude", y = "Latitude",

    title = "FATAL ACCIDENT HOT SPOTS",

    subtitle = "Locations of Fatal Accidents in the UK",

    caption = "John Karuitha, 2021 Using R and ggplot2") +

theme(legend.title = element_blank()) +

scale_fill_gradient(low = "lightgray", high = "red")
```

Insight 1: Major accidents happen all over the UK. However, the graph shows two primary regions with an unusually high concentration of accidents and fatalities. One of these regions is in the south-east (presumably

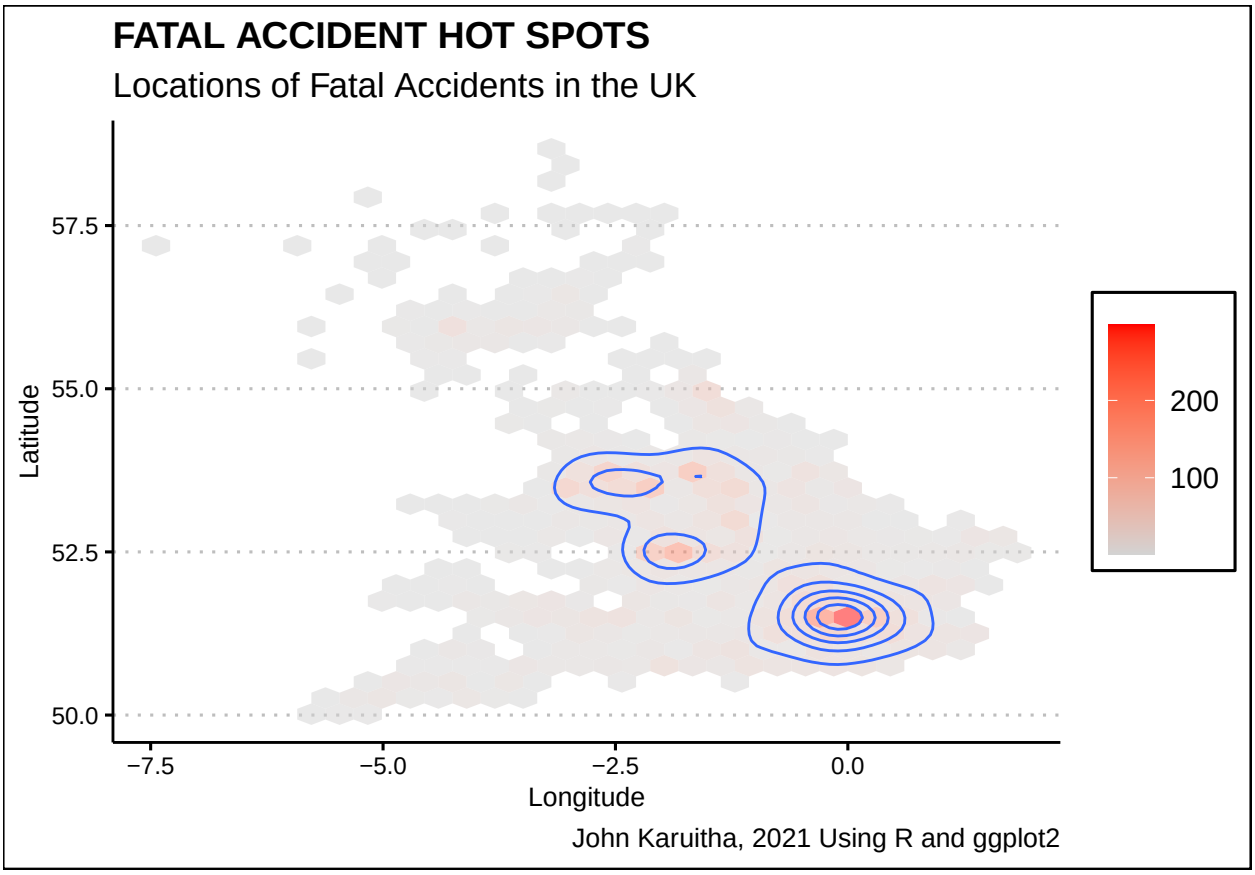


Figure 1: Figure 1: Major Accidents Hotspots in the UK

around London) while the other region spans Central UK. The accident hot spot in the central UK has three key sub-areas where accidents concentrate. There are additional pockets of accident hotspots, one towards the North East and another to the North(Scotland).

We also examine the distribution of the accident fatalities, given that this is the target variable. Figure 2 is a histogram of the distribution of accident fatalities.

```
accidents %>%

        ## Filter out the major accidents
        filter(number_of_casualties <= 10) %>%

        ## Do a histogram of casualties
        ggplot(mapping = aes(x = as.integer(number_of_casualties))) +

  geom_histogram(binwidth = 1, col = "black") +

        scale_fill_okabe_ito(name = "Cylinders", alpha = .9) +

        scale_color_okabe_ito(name = "Cylinders") +

        labs(x = "Number of Casualies",

             y = "Count", title = "Distribution of Accident Casualties")
```

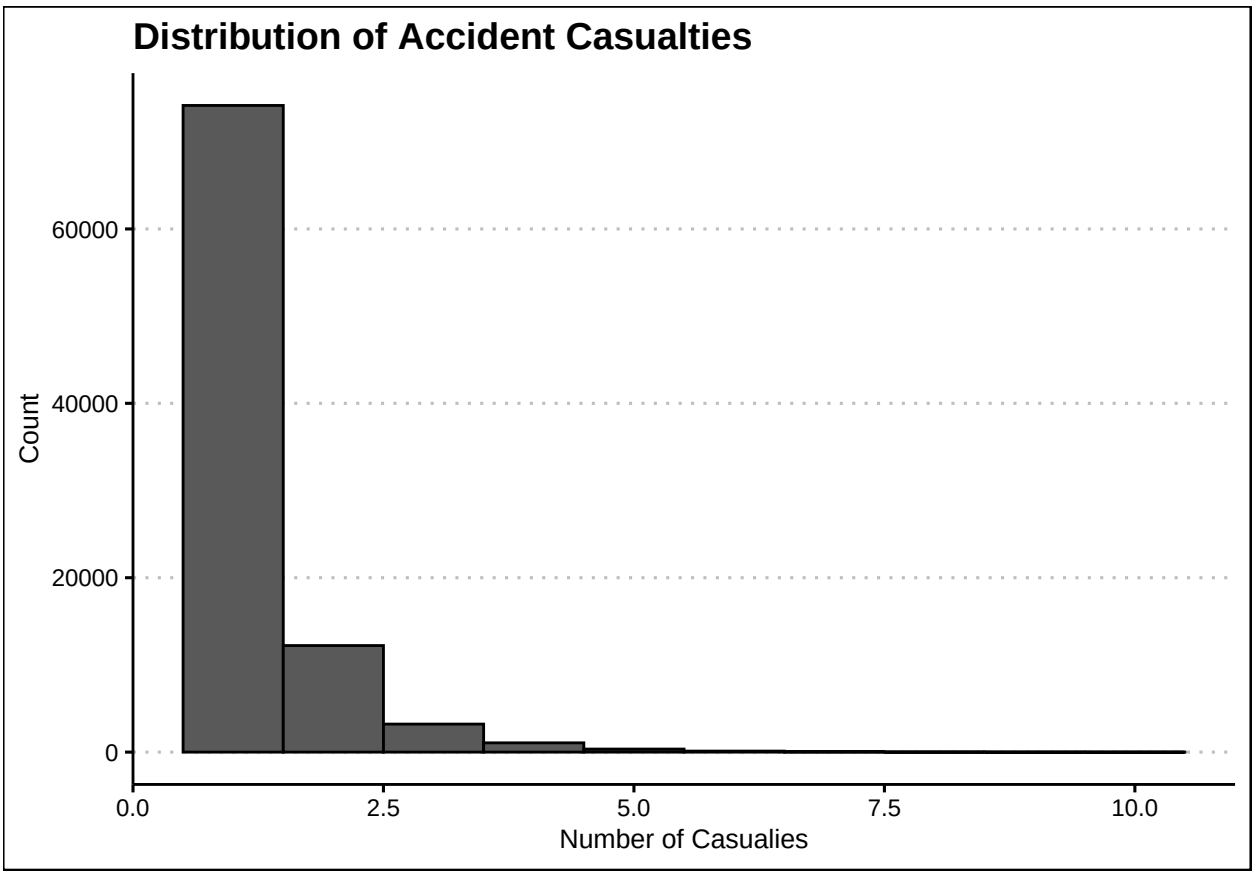


Figure 2: Figure 2: Distribution of Accident Casualties in the UK

Insight 2: The graph clearly shows that most accidents in the UK are minor, with at most two casualties. However, these minor accidents account for 94% of deaths. Only about 6% of deaths arise from major accidents. Table 3 below shows the number and proportions of major and minor accidents deaths.

```
accidents %>%

  mutate(major = case_when(

    number_of_casualties <= 3 ~ "Minor",

    TRUE ~ "Major"
  )) %>%

  group_by(major) %>%

  summarise(casualties = sum(number_of_casualties)) %>%

  mutate(prop = glue("{round(casualties / sum(casualties) * 100, 2)}%")) %>%

  kbl(., booktabs = TRUE,

  caption = "Table 3: Proportion of Deaths from Major and Minor Accidents") %>%

        kable_classic(full_width = FALSE,
                                                                                latex_options = "h
```

Table 3: Table 3: Proportion of Deaths from Major and Minor Accidents

major	casualties	prop
Major	7339	6.35%
Minor	108245	93.65%

Next, I examine the weekdays and times that major accidents happen most frequently.

What time of day and day of the week do most major incidents happen?

Suppose there is a pattern in the weekdays or hours where major accidents happen. In that case, this could prove a significant entry point for interventions to minimise accident casualties. This section examines the prevalence of accidents and casualties by day of the week and time of day.

What day of the week do most fatal accidents happen?

This section only examines the major accidents with at least three casualties. Figure 3 below shows the number of accidents and associated deaths for each day of the week. Weekends have markedly higher numbers of accidents and casualties. However, weekdays do have substantial numbers of accidents and fatalities as well. It is also notable but not surprising that the number of accidents correlates with the number of accidents. To this end, I do scatter plots of the number of accidents/casualties against the weekday.

```
(
### GRAPH 1: Number of Accidents.
accidents %>%

  ## Only major accidents with casualties >= 3.
  filter(number_of_casualties >= 3) %>%
```

```
  ## We want to see which day of week has more casualties.
  group_by(day_of_week) %>%

  ## Summary of total deaths by day of the week.
  summarise(accidents = n()) %>%

  ## plot casualties versus week day, colored by week day.
  ggplot(mapping = aes(x = day_of_week, y = accidents,

                       col = factor(day_of_week))) +

  ## Scatter plot.
  geom_point(shape = 1, size = 4, stroke = 4, show.legend = FALSE) +

  ## Axis labels, title and caption
  labs(x = "Day of Week", y = "Number of Accidents",

       title = "Number of Major Accidents by Week Day",

       subtitle = "Weekends Have the Highest Number of Major Accidents",

       caption = "Developed by John Karuitha, 2021 Using R and ggplot2") +

################################################################################
### GRAPH 2: Number of Casualties.

accidents %>%

  ## Only major accidents with casualties >= 3.
  filter(number_of_casualties >= 3) %>%

  ## We want to see which day of week has more casualties.
  group_by(day_of_week) %>%

  ## Summary of total deaths by day of the week.
  summarise(casualties = sum(number_of_casualties)) %>%

  ## plot casualties versus week day, colored by week day.
  ggplot(mapping = aes(x = day_of_week, y = casualties,

                       col = factor(day_of_week))) +

  ## Scatter plot.
  geom_point(shape = 1, size = 4, stroke = 4, show.legend = FALSE) +

  ## Axis labels, title and caption
  labs(x = "Day of Week", y = "Casualties",

       title = "Casualties of Major Accidents by Week Day",

       subtitle = "Weekends Have the Highest Casualties from Major Accidents",

       caption = "Developed by John Karuitha, 2021 Using R and ggplot2"))
```

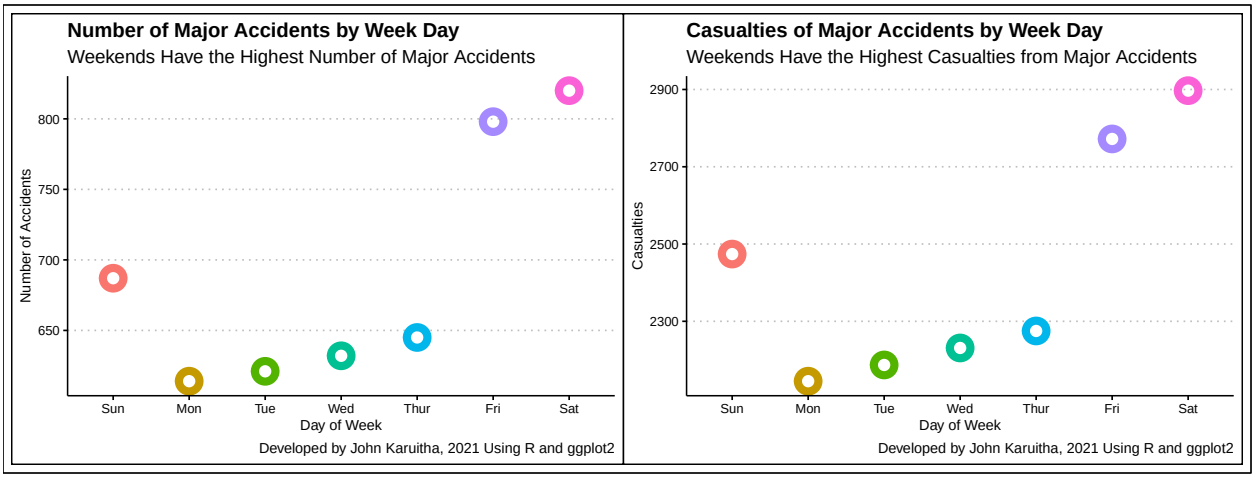


Figure 3: Figure 3: Number of Major Accidents by Week Day

Insight 3: Weekends have the highest number of major accidents resulting in the most casualties. Typically, accidents and deaths are lowest on Mondays. Accidents and fatalities rise gradually during the week and then spike on Friday, peaking on Saturday. The numbers then reduce on Sunday (to levels only lower than Friday and Saturday) and slump back to their minimum on Monday. Still, weekdays have a significant number of fatal road accidents and fatalities, as the Table 4 shows.

```
accidents %>%

  ## Only major accidents with casualties >= 3.
  filter(number_of_casualties >= 3) %>%

  ## We want to see which day of week has more casualties.
  group_by(day_of_week) %>%

  ## Summary of total deaths by day of the week.
  summarise(all_accidents = n(),

            percent_accidents = all_accidents/nrow(accidents) * 100,

            fatalities = sum(number_of_casualties),

            percent_fatalities = fatalities/sum(accidents %>%

                                                  select(number_of_casualties)) * 100) %>%

  ## Make a nice table
  kbl(., booktabs = TRUE,

      caption = "Table 4: Major Accidents and Fatalities by Week Day") %>%

  kable_classic(latex_option = "hold_position", full_width = FALSE) %>%

  add_footnote(c("The weekday count starts on Sunday", "Note: We only include major accidents with more
```

Table 4: Table 4: Major Accidents and Fatalities by Week Day

day_of_week	all_accidents	percent_accidents	fatalities	percent_fatalities
Sun	687	0.7532977	2474	2.140435
Mon	614	0.6732530	2145	1.855793
Tue	621	0.6809285	2187	1.892130
Wed	632	0.6929901	2231	1.930198
Thur	645	0.7072446	2275	1.968266
Fri	798	0.8750096	2772	2.398256
Sat	820	0.8991327	2897	2.506402

[a] The weekday count starts on Sunday
[b] Note: We only include major accidents with more than 3 fatalities

What time of day do most fatal accidents happen?

This section examines the number of major accident casualties by the time of day. I visualize this data using an area plot. Figure () below shows the number of deaths from accidents for each hour, which leads us to another insight.

Insight 4: There are two peak times of the day when most accidents happen. The first period starts at around 0500 hours, peaking at 0800 hours followed by a sudden drop. In the second and most deadly phase, the number of casualties starts rising at 1000 hours, reaching the maximum at 1700 hours, after which there is a sharp decline.

However, lumping data together may mask heterogeneity in accidents casualties patterns between the days of the week. In the next section, I break the data into the constituent days of the week.

```
accidents %>%

        ## Get only the major accidents
        filter(number_of_casualties >= 3) %>%

        ## Get the hour of day from the time column
        group_by(hour = hour(time)) %>%

        ## Get total casualties by hour of day
        summarise(death_by_hour = sum(number_of_casualties)) %>%

        ## Plot deaths against hour
        ggplot(mapping = aes(x = hour, y = death_by_hour)) +

  geom_area(col = "purple", fill = "lightgray") +

  labs(x = "Hour", y = "Casualties",

       title = "Number of Accident Casualties by Hour")
```

Are there any patterns in the time of day/ day of the week when major incidents occur?

This section examines unique patterns in the time of the day or day of the week when major accidents occur. Again, if there are regularities, then the ministry could have interventions at these days/ times to reduce incidences of fatal accidents.

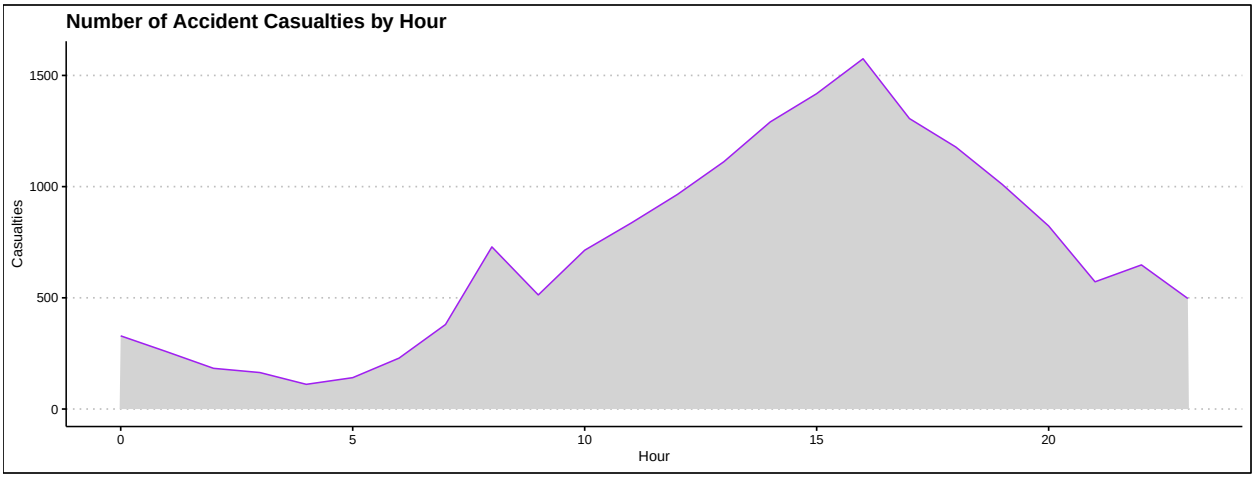


Figure 4: Figure 4: Number of Accident Casualties by Hour

Breaking down accident casualties by the time of day and day of the week.

As noted earlier, lumping together figures could mask heterogeneity in the patterns of accidents by hour of day or day of the week. In this section, I disaggregate this data revealing patterns to each day of the week.

Insight 5: Figure 5 below shows a breakdown of the casualties by the hour for each day of the week. Working days (which I define as Monday to Friday) have a similar trend that follows **Insight 1**.

However, Saturdays and Sundays have a markedly different pattern.

```
accidents %>%

        ## Get only the major accidents
        filter(number_of_casualties >= 3) %>%

        ## Group by time of day and day of week
        group_by(hour = hour(time), day_of_week) %>%

        ## Get number of deaths by hour
        summarise(death_by_hour = sum(number_of_casualties)) %>%

        ## Plot deaths against hour
        ggplot(mapping = aes(x = hour, y = death_by_hour, col = factor(hour))) +

  geom_point(show.legend = FALSE) +

  facet_wrap(~day_of_week, ncol = 3) +

  scale_color_viridis_d() +

  labs(x = "Hour", y = "Casualties",

       title = "Number of Accident Casualties by Day and Hour")
```

Insight 6: Unlike other days of the week, accidents on Sundays have one peak period that begins around midday and peaks at 1500 hours, followed by a rapid drop.

Insight 7: Like Sunday, Saturday has one peak in the number of accident casualties. However, the peak

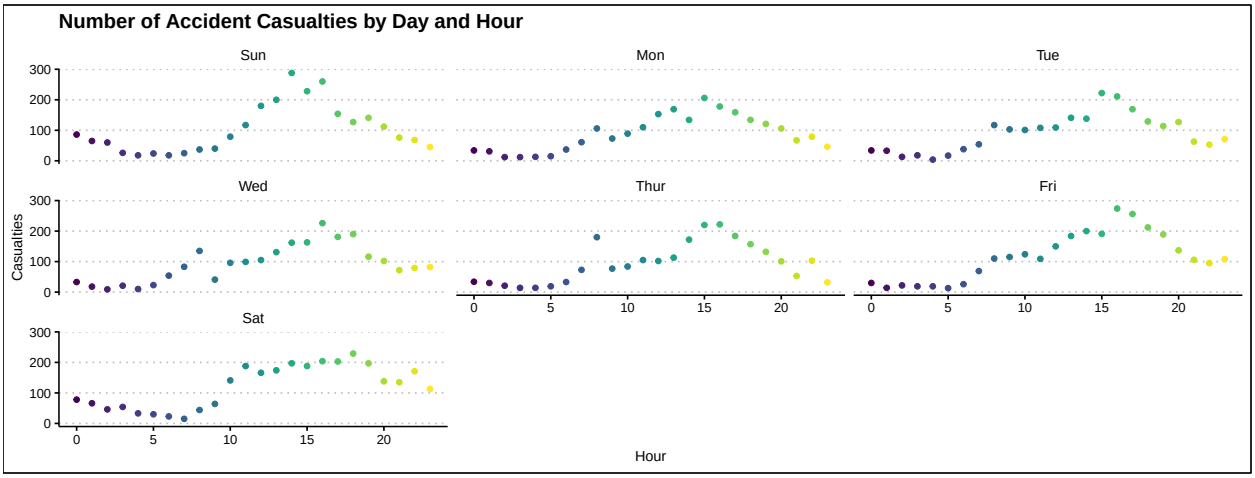


Figure 5: Figure 5: Number of Accident Casualties by Day and Hour

accident period is prolonged, lasting from around 11.00 hours to 1800 hours, after which there is a gradual decline.

Where do accidents happen at weekends vs weekdays?

Major Accident Casualties versus Speed Limits Speed is one major cause of road traffic accidents (Ashraf et al. 2019; Rolison et al. 2018). This section examines the speed limits that correspond with high casualties for major accidents. Figure 6 shows a plot of major accident casualties (y-axis) versus speed limits (x-axis), faceted by day of the week.

```
accidents %>%

        ## Get the major accidents
        filter(number_of_casualties >= 3) %>%

        ## Group data by speed limit and day of week
        group_by(speed_limit, day_of_week) %>%

        ## Summarise the total casualties
        summarise(casualties = sum(number_of_casualties)) %>%

        ## Get the sum of casualties
        mutate(perc_casualties = casualties / sum(casualties) * 100) %>%

        ## Plot speed limit versus proportion of casualties
        ggplot(mapping = aes(x = speed_limit, y = perc_casualties)) +

        geom_line() + facet_wrap(~ day_of_week) +

        labs(x = "Speed Limit", y = "Percent Casualties",

       title = "Speed Limits versus Proportion of major Accident Casualties")
```

Insight 8: At 21% and 20%, respectively, Fridays and Saturdays account for the highest proportion of major accident casualties at the speed limit of 20 mph. Friday also has a second peak at a speed limit of 50 mph. This observation means that the accidents could result from failure to observe road limit signs. Sunday has most accidents where speed limits are between 30 and 60 mph. Other days of the week have notable peaks at

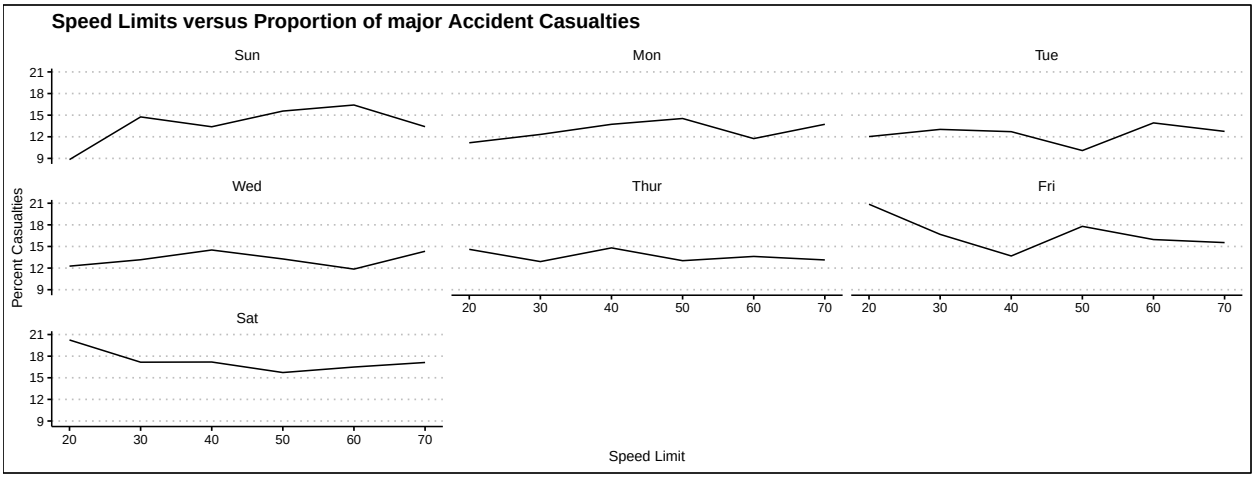


Figure 6: Figure 6: Speed Limits versus Proportion of major Accident Casualties

around 40 mph (Jacobs and Sayer 1983).

I also examine the patterns of major accidents in road stretches with speed limits. In this case, I plot the number of casualties versus days of week faceted by speed limits.

Insight 9: Overall, the road stretches with a speed limit of 30 mph have the highest incident of accidents and accident casualties across all days of the week. Figure () below shows this trend.

```
accidents %>%

        ## Create a column of major and minor accidents
        mutate(major = case_when(

                number_of_casualties >= 3 ~ "Major",

                TRUE ~ "Minor"
        )) %>%

        ## Group by day of week and speed limit
        group_by(day_of_week, speed_limit) %>%

        ## Summarise total casualties
        summarise(total = sum(number_of_casualties)) %>%

        ## Plot day of week versus casualties
        ggplot(mapping = aes(x = day_of_week, y = total,

                             col = day_of_week, fill = day_of_week)) +

        geom_col(show.legend = FALSE) +

        facet_wrap(~ speed_limit) +

        scale_fill_viridis_d() +

        labs(y = "Number of Casualties",
```

```
        title = "Accident Casualties by Weekday and Speed Limit")
```

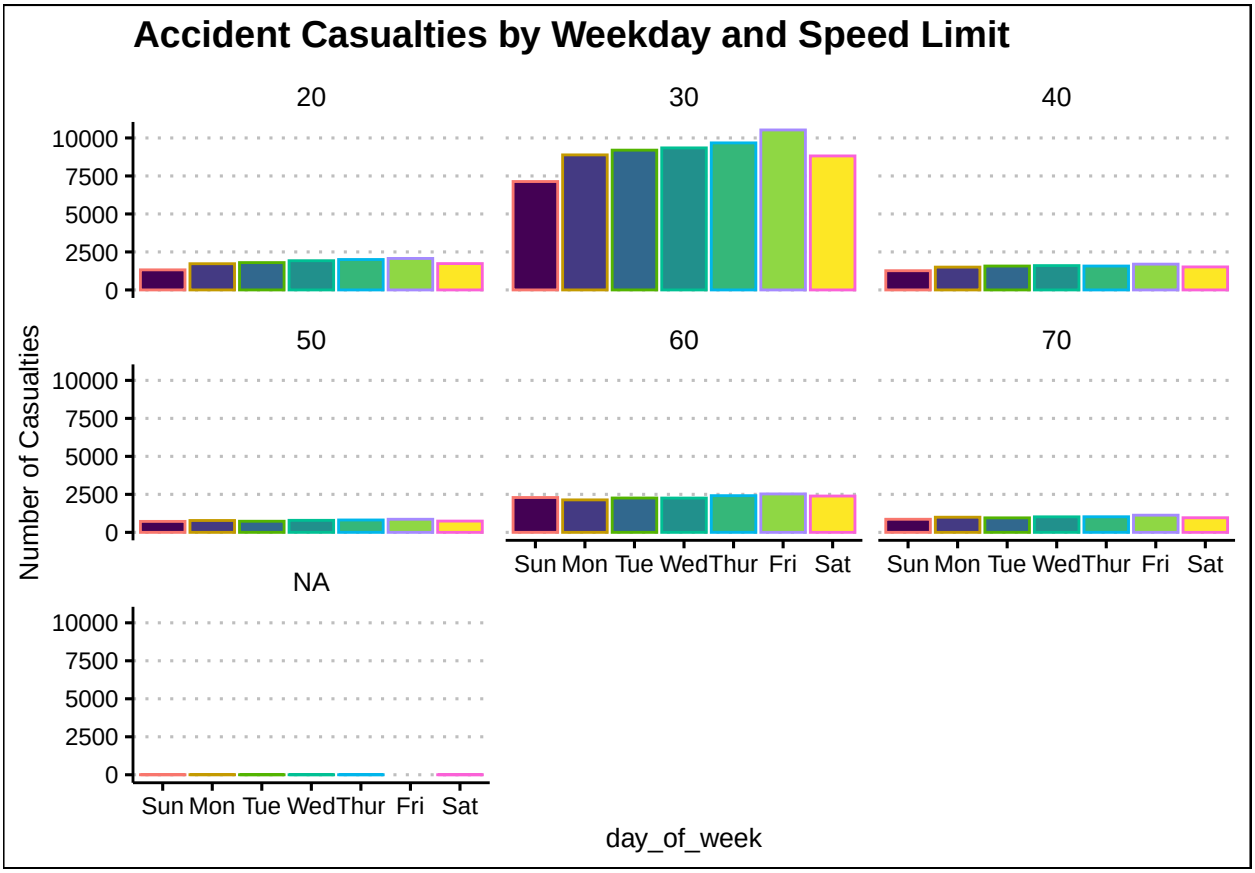


Figure 7: Figure 7: Accident Casualties by Weekday and Speed Limit

Major Accident Casualties and Weather Conditions Another probable driver of accidents is the weather. I tabulate weather conditions against the number of major accident casualties. The results are in Table 5.

Insight 9: Table 5 below shows that about 77% of major accident casualties happen when the weather is fine (that is, fine with no high winds). In comparison, 14% occurs when raining with no high winds.

```
accidents %>%

        ## Get only the major accidents
        filter(number_of_casualties >= 3) %>%

        ## Group by weather conditions
        group_by(weather_conditions) %>%

        ## Get the total number of casualties
        summarise(casualties = sum(number_of_casualties)) %>%

        ## Compute the proportion of casualties
        mutate(perc_casualties = casualties / sum(casualties) * 100) %>%

        ## Sort by casualties
```

Table 5: Table 5: Weather Cinditions and Accident Casualties

weather_conditions	casualties	perc_casualties
1	13093	77.1038219
2	2382	14.0274424
5	417	2.4556858
8	391	2.3025735
4	307	1.8079030
9	202	1.1895648
7	130	0.7655615
3	47	0.2767799
6	12	0.0706672

```
        arrange(desc(casualties)) %>%

        ## Make a nice table
        kbl(., booktabs = TRUE,

            caption = "Table 5: Weather Cinditions and Accident Casualties") %>%

        kable_classic(full_width = TRUE)
```

We next break down this weather data into days of the week.

Insight 10: Most accidents on any day of the week happen when the weather is fine with no high winds (Figure 8). The second riskiest weather pattern is raining with no high winds. It would follow that people are most likely to speed when the weather is clear, hence the high number of casualties.

```
accidents %>%

        ## Get the major accidents
        filter(number_of_casualties >= 3) %>%

        ## Group by weather conditions and day of the week
        group_by(weather_conditions, day_of_week) %>%

        ## Summarise total casualties
        summarise(casualties = sum(number_of_casualties)) %>%

        ## Create new variable for proportion of casualties
        mutate(perc_casualties = casualties / sum(casualties) * 100) %>%

        ## Plot casualties versus day of week
        ggplot(mapping = aes(x = day_of_week, y = casualties,

                       fill = day_of_week)) +

        geom_col(show.legend = FALSE) +

        facet_wrap(~ weather_conditions) +

        labs(x = "Day of the Week", y = "Casualties",
```

```
        title = "Casualties by Day of Week and Weather Conditions")
```

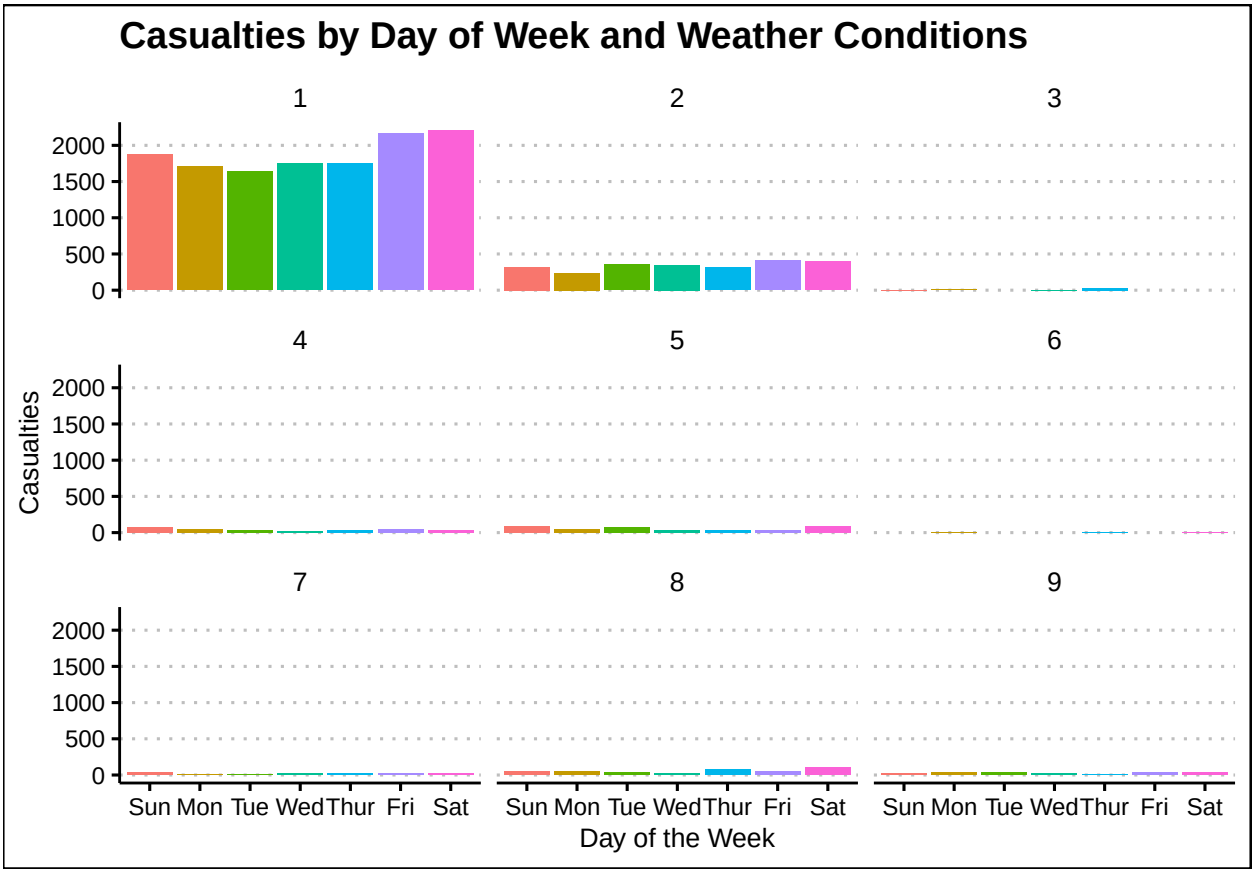


Figure 8: Figure 8: Casualties by Day of Week and Weather Conditions

What characteristics stand out in major incidents compared with other accidents?

First, most casualties arise from other accidents instead of major accidents, as noted earlier in **Insight 1**. Figure 9 visualizes this observation.

```
accidents %>%

        ## Create a new column major
        mutate(major = case_when(

                number_of_casualties >= 3 ~ "Major",

                TRUE ~ "Minor"
        )) %>%

        ## Group the data by major
        group_by(major) %>%

        ## Get the total casualties
        summarise(total_casualties = sum(number_of_casualties)) %>%
```

```
        ## Plot casualties against major
        ggplot(mapping = aes(x = major, y = total_casualties)) +

        geom_col(col = "skyblue", fill = "darkblue") +

        labs(x = "Accident Type- Major or Minor",

             y = "Number of Casualties",

             title = "Accident Type versus Number of Casualties")
```

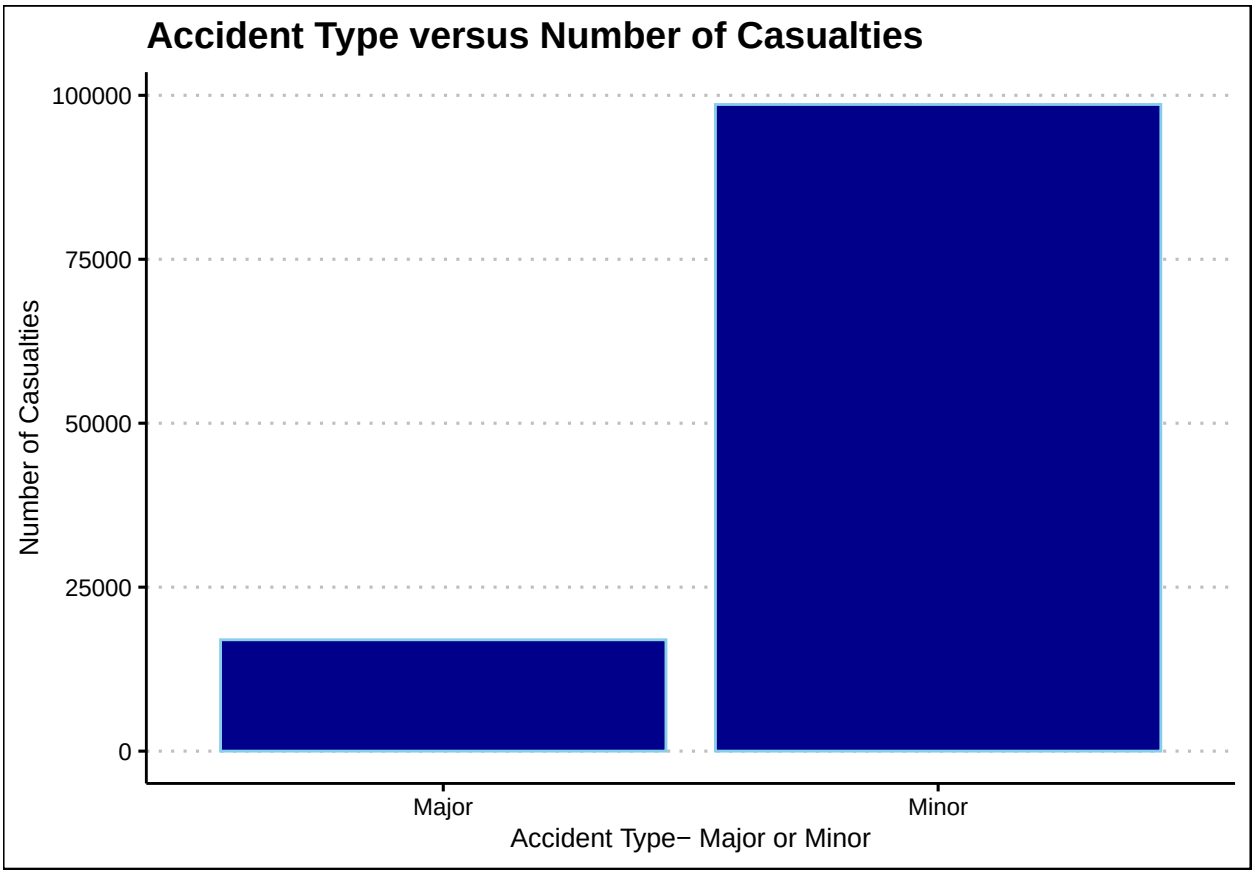


Figure 9: Figure 9: Accident Type versus Number of Casualties

Insight 11: In comparison to major accidents, minor accidents are more dispersed. Figure 10 shows that major accidents concentrate in three major hotspots, whereas minor accidents concentrate in more places. Note that the figure's colour `red` shows the accident concentration hotspots.

```
accidents %>%

        ## Create a new column for major and minor accidents
        mutate(major = case_when(

                number_of_casualties >= 3 ~ "Major",

                TRUE ~ "Minor"
        )) %>%
```

```
        ## Group by major/ minor accident
        group_by(major) %>%

        ## Map the locations of accidents
        ggplot(mapping = aes(x = longitude, y = latitude)) +

  geom_hex(alpha = 0.5) +

  geom_density_2d() +

  labs(x = "Longitude", y = "Latitude",

       title = "MAJOR/MINOR ACCIDENT HOT SPOTS",

       subtitle = "Locations of Major/ Minor Accidents in the UK",

       caption = "John Karuitha, 2021 Using R and ggplot2") +

  theme(legend.title = element_blank()) +

  scale_fill_gradient(low = "lightgray", high = "red") +

        ## Facet accidents by major/minor category
        facet_wrap(~ major)
```

Insight 12: Furthermore, major accidents fatalities appear evenly distributed across rural and urban settings. In contrast, over 2/3 of minor accidents happen in urban areas.

```
## The data
accidents %>%

        ## Create a major/minor columns
        mutate(major = case_when(

                number_of_casualties >= 3 ~ "Major",

                TRUE ~ "Minor"
        )) %>%

        ## Group by major and urban/rural area
        group_by(major, urban_or_rural_area) %>%

        ## Get total of casualties
        summarise(total = sum(number_of_casualties)) %>%

        ## Make a wide dataset
        pivot_wider(names_from = "major", values_from = "total") %>%

        ## Make a nice table
        kbl(., booktabs = TRUE,

  caption = "Table 6: Casualties in Rural/Urban Settings for Major and Minor Accidents") %>%

        kable_classic(full_width = FALSE,
```

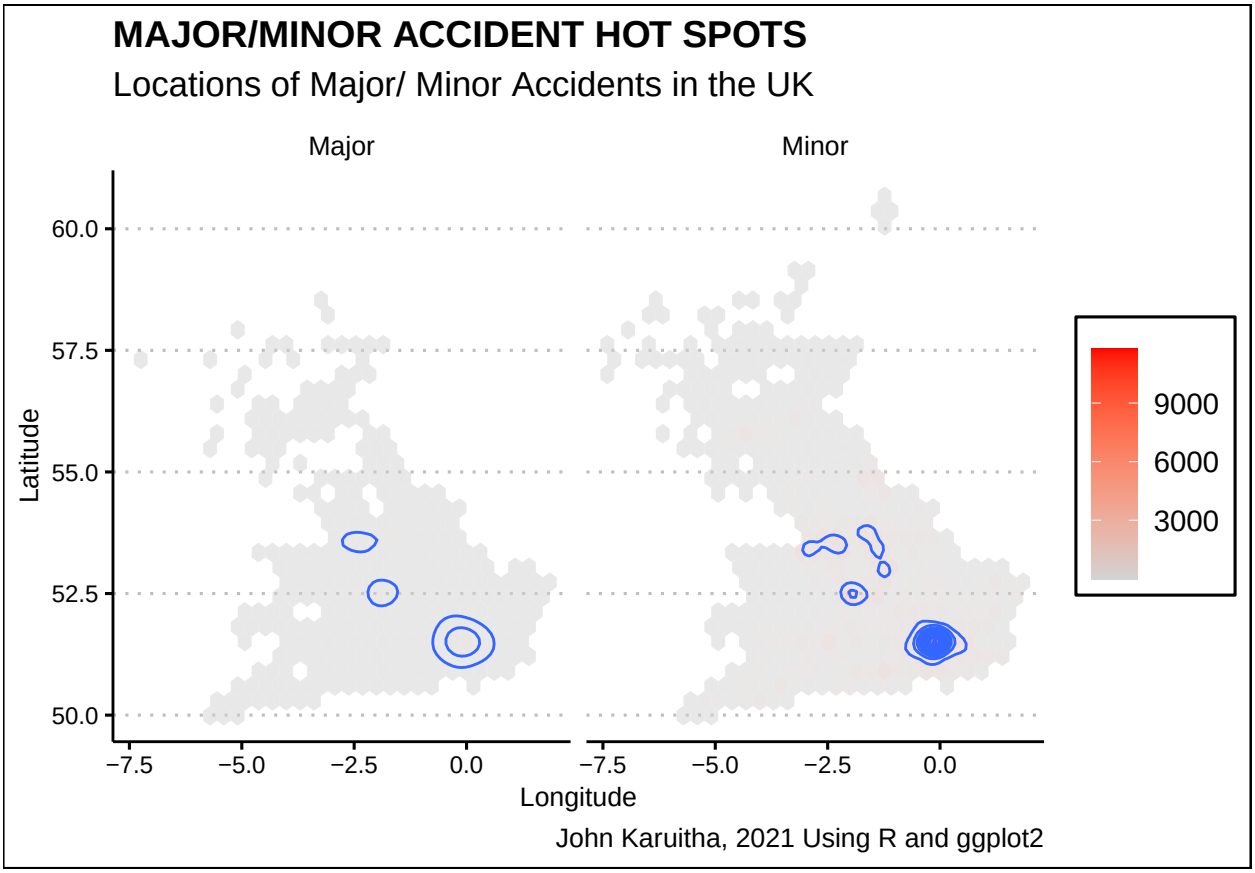


Figure 10: Figure 10: Major/Minor Accidents Hotspots

```
                                                                    latex_options = "h
```

Table 6: Table 6: Casualties in Rural/Urban Settings for Major and Minor Accidents

urban_or_rural_area	Major	Minor
1	8680	66142
2	8301	32444
3	-	17

Next, I examine the number of casualties by speed limit for major and minor accidents. The bar chart below shows the distribution showing a significant difference between major and minor accidents. Compared to other accidents, a significantly lower proportion of casualties happen at a speed limit of 30 mph. In contrast, more accident casualties occur at the 60 mph speed limit for major accidents than minor accidents.

```
accidents %>%

        ## Create a major/minor accidents column
        mutate(major = case_when(

                number_of_casualties >= 3 ~ "Major",

                TRUE ~ "Minor"
        )) %>%

        ## Group by major and speed limit
        group_by(major, speed_limit) %>%

        ## Summarise total casualties
        summarise(casualties = sum(number_of_casualties)) %>%

        ungroup() %>%

        group_by(major) %>%

        ## Create a new column for proportion of casualties
        mutate(prop = casualties/ sum(casualties)) %>%

        ## Plot prop against speed limit
        ggplot(mapping = aes(x = speed_limit, y = prop, fill = speed_limit)) +

        geom_col() +

        scale_fill_viridis_b() +

        facet_wrap(~ major) +

        ## Add labels and titles
        labs(x = "Speed Limit", y = "Proportion of Casualties",

             title = "Speed Limits versus Proportion of Accident Casualties")
```

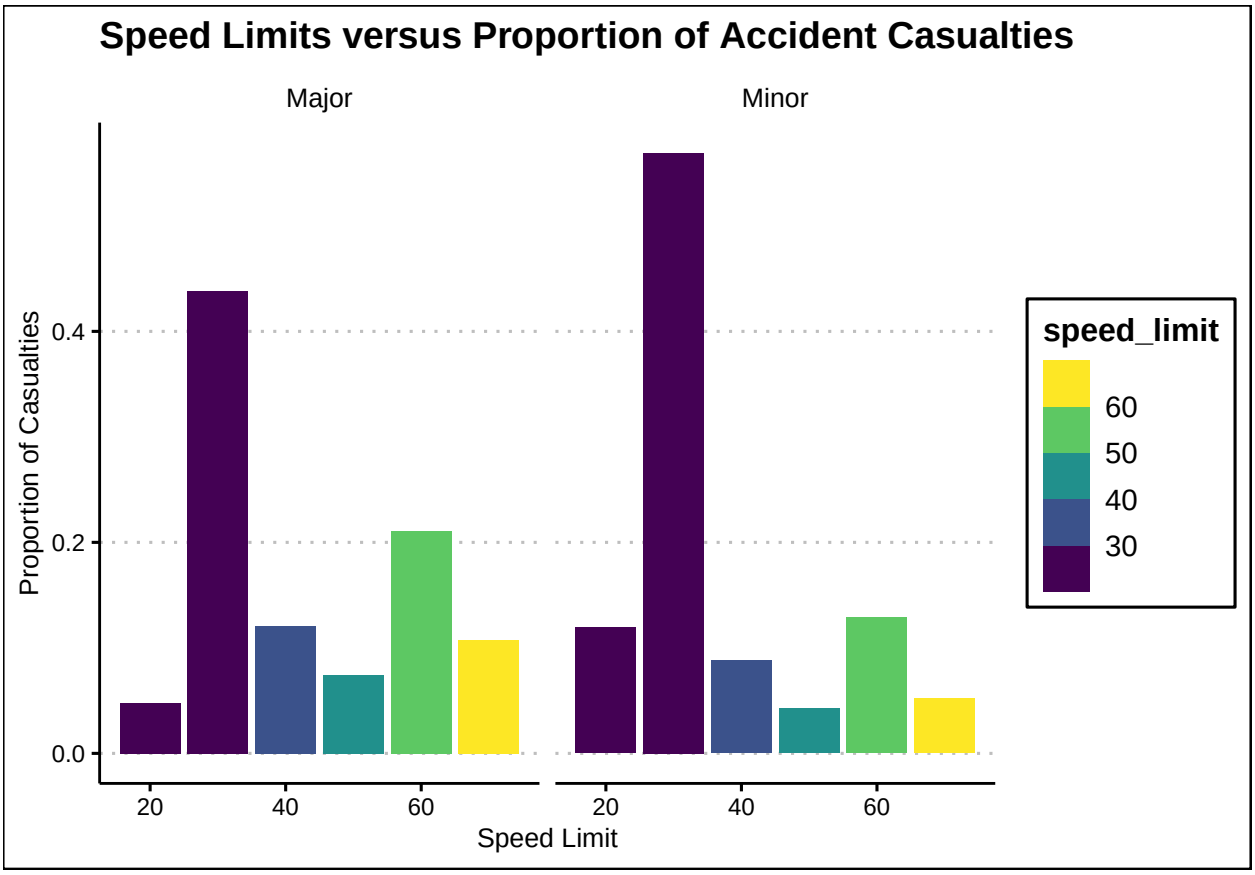


Figure 11: Figure 11: Speed Limits versus Proportion of Accident Casualties

On what areas would I recommend the planning team focus their brainstorming efforts to reduce major incidents?

One of the major takeaways from the analysis is that most road accidents happen when the weather is clear (77%), with another 14% happening when raining but with no high winds. 91% of major accidents occur during these two weather patterns to restate the point. The first recommendation is for the ministry to pay attention when the weather is clear and when raining with no high winds. However, we cannot take the weather pattern in isolation. After considering the weather, the ministry should also jointly focus on the recommendations below.

First, while accidents are spread all over the UK, there is a need to concentrate on areas with a high concentration of significant casualties. Figure 1 shows these two hotspots, one around London and another in the central UK. The hotspot in the central UK has a further three sub-hotspots that the ministry should pay particular attention to. After separating major and minor accidents, the areas that the ministry should concentrate on to reduce fatalities from major accidents stand out better. Figure 10, panel A, shows the two regions. Again we should not take this recommendation in isolation but with the other suggestions.

Another significant regularity in the data is the high incidence of accidents at stretches with 30mph and 60mph speed limits. This recommendation draws from insight 8 and insight 9. For some reason, the speed limit of 30 mph is especially notorious for both major and minor accidents, followed by the 60 mph speed limit. Thus, the ministry should put mechanisms to enforce adherence to these speed limits.

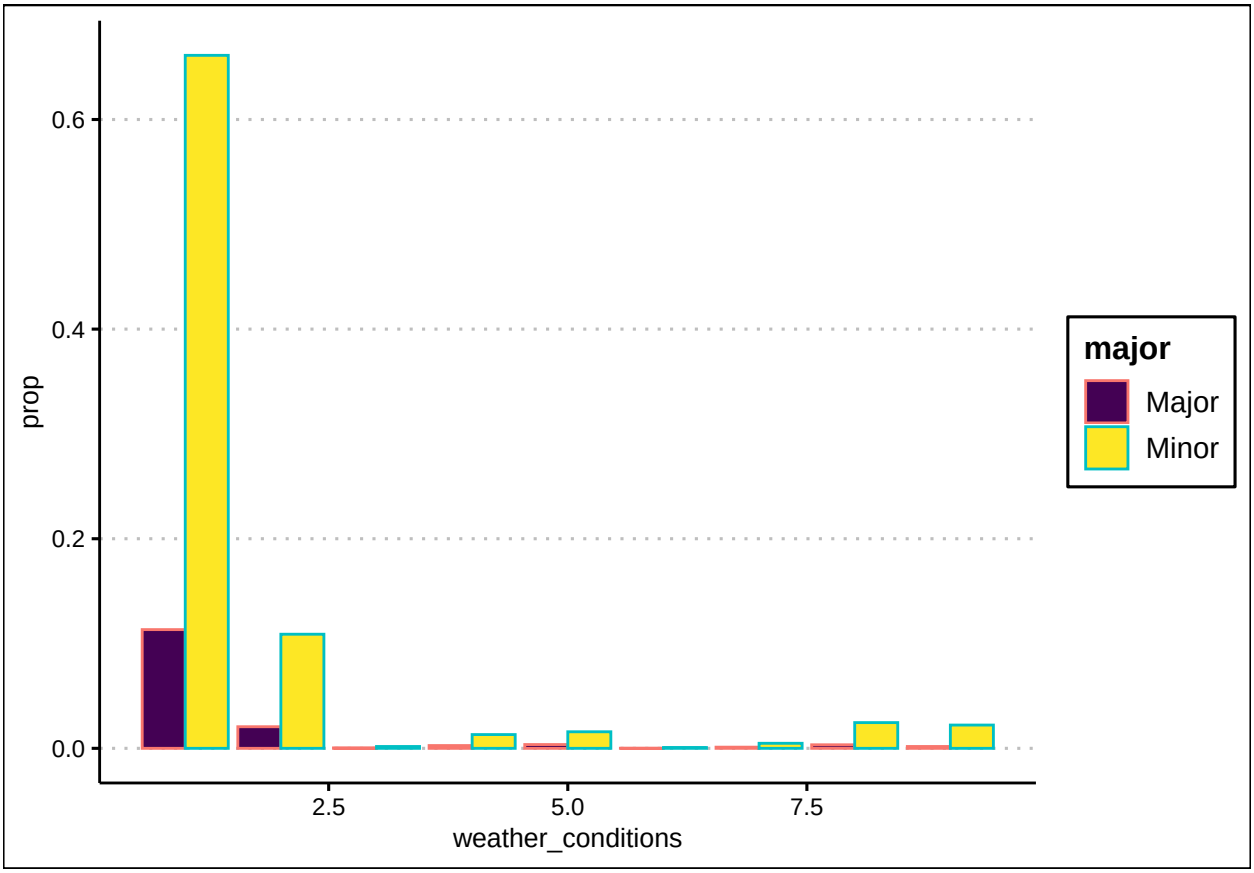
The day of the week effect is very prominent and a major entry point to minimise major accidents. As noted earlier in insight 3 and 5, major accidents spike on Fridays, peaking on Saturday and declining on Sundays to levels below Friday and Saturday. However, over weekends, there is a significant pattern of accidents across an hour of the day, as we discuss next.

Time of day is also a significant entry point for interventions to reduce fatalities from major accidents in line with insight 4 and insight 5. The times when major accidents and accident fatalities occur vary by day of the week. For instance, on Saturdays, the worst day in terms of major accidents fatalities, we observe an extended peak between 1100 hours and 1800 hours when the ministry should effect the interventions. We would follow a similar interpretation for other days of the week.

As noted earlier, it would be a mistake to take any one of these recommendations in isolation. Each recommendation should serve to reinforce the others. Take Saturday, a day notorious for deadly road accidents, as an example. The ministry should focus efforts on the two regions in the UK (Figure) with an unusually high concentration of major accidents. In these regions, the ministry should focus on the period between 1100 hours and 1800 hours at road stretches with a speed limit of 30 mph and 60 mph, respectively, by installing speed cameras. The ministry should intensify these interventions when the weather is clear, for instance, by having traffic corps on sight.

```
accidents %>%

        mutate(major = case_when(

                number_of_casualties >= 3 ~ "Major",

                TRUE ~ "Minor"
        )) %>%

        group_by(major, weather_conditions) %>%

        summarise(total = sum(number_of_casualties)) %>%

        ungroup() %>%

        mutate(prop = total / sum(total)) %>%
```

```
        ggplot(aes(x = weather_conditions, y = prop,
                   fill = major, col = major)) +
        geom_col(position = "dodge") +
        scale_fill_viridis_d()
```



Conclusion

In this analysis, I examined the road accidents data from the UK. The study of the data provided some valuable insights, the major ones being;

- Accident casualties peak during weekends, Starting from Friday and falling on Sunday.
- Accidents casualties vary by time of day.
- Major accidents and accident casualties mainly happen when the weather is fine.
- Major accidents and accident casualties mainly occur in road stretches with speed limits of 30 mph and 60 mph.

The recommendations to reduce significant accidents casualties should hence draw from these insights. None of the proposals would work very well in isolation. What is needed is a package of interventions that would help lower the tide of casualties from significant accidents.

References

Ashraf, Imran, Soojung Hur, Muhammad Shafiq, and Yongwan Park. 2019. “Catastrophic Factors Involved in Road Accidents: Underlying Causes and Descriptive Analysis.” PLoS One 14 (10): e0223473. Jacobs, GD, and I Sayer. 1983. “Road Accidents in Developing Countries.” Accident Analysis & Prevention 15 (5): 337–53. Rolison, Jonathan J, Shirley Regev, Salissou Moutari, and Aidan Feeney. 2018. “What Are the Factors That Contribute to Road Accidents? An Assessment of Law Enforcement Views, Ordinary Drivers’ Opinions, and Road Accident Records.” Accident Analysis & Prevention 115: 11–24.

Appendix

Appendix 1: Summary Statistics

```
accidents %>%

        ## Select the relevant variables
        select(-accident_index, -accident_reference,

         -accident_year, -day_of_week,

         -date, -time,

         -longitude, -latitude) %>%

        ## Get summary statistics
        skimr::skim_without_charts() %>%

        ## Remove redundant summary statistics
        select(-skim_type, -n_missing) %>%

        ## Make a nice table
        kbl(., booktabs = TRUE, caption = "Table 7: Summary Statistics") %>%

        kable_classic(full_width = TRUE,

                latex_options = "hold_position")
```

Ashraf, Imran, Soojung Hur, Muhammad Shafiq, and Yongwan Park. 2019. “Catastrophic Factors Involved in Road Accidents: Underlying Causes and Descriptive Analysis.” *PLoS One* 14 (10): e0223473.

Jacobs, GD, and I Sayer. 1983. “Road Accidents in Developing Countries.” *Accident Analysis & Prevention* 15 (5): 337–53.

Rolison, Jonathan J, Shirley Regev, Salissou Moutari, and Aidan Feeney. 2018. “What Are the Factors That Contribute to Road Accidents? An Assessment of Law Enforcement Views, Ordinary Drivers’ Opinions, and Road Accident Records.” *Accident Analysis & Prevention* 115: 11–24.

Table 7: Table 7: Summary Statistics

skim_variable	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100
accident_severity	1.0000000	2.7682321	0.4566821	1	3	3	3	3
number_of_vehicles	1.0000000	1.8352723	0.6772721	1	1	2	2	13
number_of_casualties	1.0000000	1.2673823	0.6814733	1	1	1	1	41
first_road_class	1.0000000	4.2203204	1.4434755	1	3	4	6	6
first_road_number	1.0000000	790.6660709	1580.8177432	0	0	34	538	9174
road_type	1.0000000	5.2560006	1.6848780	1	6	6	6	9
speed_limit	0.9998684	36.2750173	13.8903666	20	30	30	40	70
junction_detail	0.9999781	3.9349869	12.6128945	0	0	2	3	99
junction_control	0.9980612	3.7193248	1.2284939	1	4	4	4	9
second_road_class	1.0000000	5.5517714	1.0151126	1	6	6	6	6
second_road_number	0.9999232	220.2489692	913.7258441	0	0	0	0	9174
pedestrian_crossing_human_control	0.9984320	0.3551002	1.6986016	0	0	0	0	9
pedestrian_crossing_physical_facilities	0.9985197	1.1885487	2.4462881	0	0	0	0	9
light_conditions	0.9999890	2.0653413	1.7476701	1	1	1	4	7
weather_conditions	0.9999890	1.7020768	1.8457741	1	1	1	1	9
road_surface_conditions	0.9965356	1.3998988	0.9164743	1	1	1	2	9
special_conditions_at_site	0.9976096	0.2477221	1.3187224	0	0	0	0	9
carriageway_hazards	0.9977113	0.1832929	1.1497169	0	0	0	0	9
urban_or_rural_area	1.0000000	1.3232053	0.4680310	1	1	1	2	3