1. Executive Summary

Roads play an important role in our daily lives. Accidents and their consequences are inevitable, but they can be minimised by making our roads safer by implementing appropriate safety measures. Safer Roads UK, a pressure group dedicated to social responsibility, is studying historical accident data from 2017 to observe how UK roads may be made safer.
The purpose of this coursework is to investigate the variations of accidents across UK's geography, identify the factors that contribute to accidents and their severity with suitable graphical and statistical evidence. It also aims at providing necessary recommendations to improve the roads, techniques to be adopted to make UK roads safer.

The data set majorly contains the geographical coordinates of each accident, date and time of occurrence, accident severity, casualties, and other attributes defining the spot of the incident.

2. Data Cleaning

Dataset has missing values and errors. Hence, it's essential to clean the data before the commencement of analysis. Initially, the encoded data was recoded in R replacing the missing values with NA characters followed by converting the attributes to appropriate data types. Literature has cited several imputation methods for handling missing values, most common being- dropping, replacing with mode, employing neural networks, etc (Margot Peeters et al., 2015). Imputation methods can inject bias into the data hence need to be chosen wisely. Percentage of NA's and unknown values were analysed for each attribute for choosing the best imputation method as shown in the table below. If the total percentage of NA and unknown values was less than 2%, then they were dropped during analysis. For percentages < 8, they were replaced with mode values. Attributes with NA% >40 were not considered for analysis.

Table 1: Missing data analysis

Categorical variable	No. of missing	Unknown	% of missing	% of unknown	Total %
Road_Type	0	2552	0.00%	1.96%	1.96%
Junction_Detail	609	0	0.47%	0.00%	0.47%
Junction_Control	56296	0	43.31%	0.00%	43.31%
Road_Class_2nd	54412	0	41.86%	0.00%	41.86%
Pedestrian_Crossing_Human_Control	2574	0	1.98%	0.00%	1.98%
Pedestrian_Crossing_Physical_Facilities	2765	0	2.13%	0.00%	2.13%
Road_Surface_Conditions	1937	0	1.49%	0.00%	1.49%
Special_Conditions_at_Site	2206	0	1.70%	0.00%	1.70%
Carriageway_Hazards	2073	0	1.59%	0.00%	1.59%
Weather_Conditions	1	7354	0.00%	5.66%	5.66%

3. Results and Discussions

Question 1

The preliminary data analysis entailed the study of the cleaned dataset and its continuous and categorical variables, to look for patterns, variations, and association with other variables. The geographical attributes are used to plot a GIS map and create a 'big picture' of the problem.

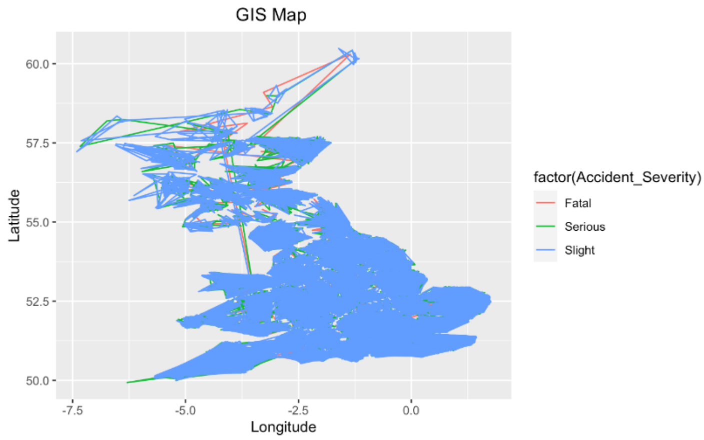

Figure1: GIS mapping of accidents in UK

The major metrics to categorize the accidents are the number of casualties and the number of vehicles. However, their distributions are highly skewed to the right proving most of the accidents are of slight severity. Hence, we use the statistic 'casualties per accident' to analyse various scenarios. The above statistic is normally distributed with a mean of 1.315 and a standard deviation of 0.76. The skewness of -1.75 indicates slight left skewness and a kurtosis greater than 3 (5.94) indicates the leptokurtic condition. The z-value of skewness (skew/standard error) falls within the range -7 to +7 (-5.11) for n>300 implies that the skew is not too much to consider the variable normally distributed.

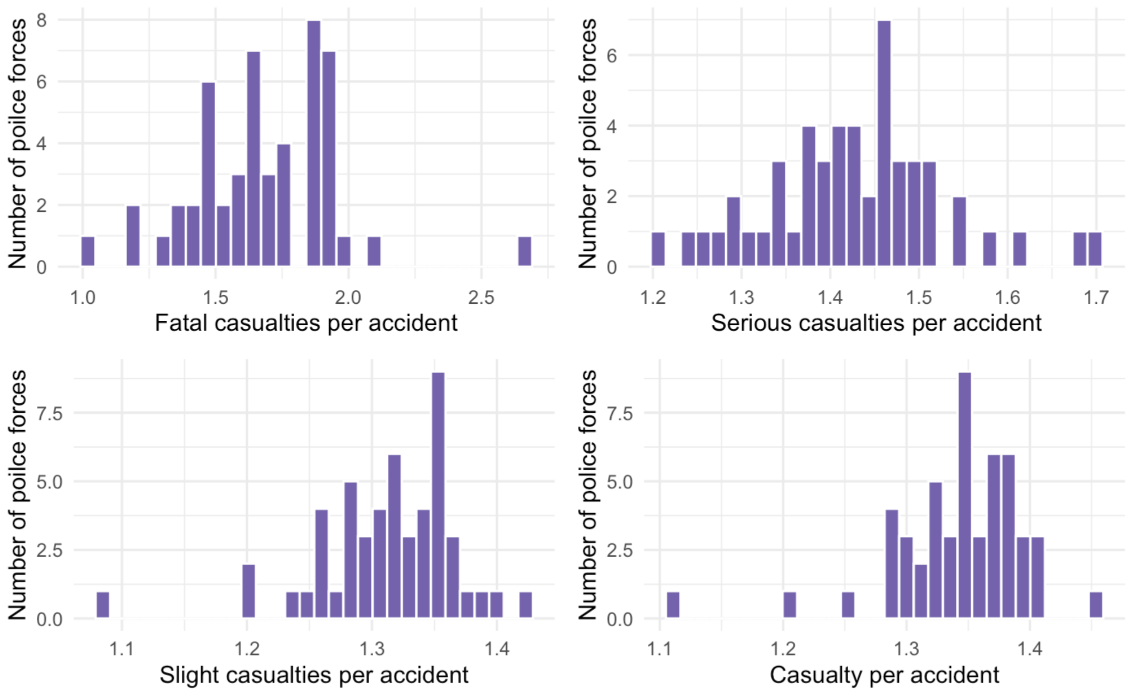

Figure 2: Casualty rate histograms

To analyse the variation of accident severity across the police forces, a chi-squared test was conducted assuming the below hypothesis:

H0: Accident severities are the same across the police forces
HA: Accident severities are different across the police forces

The results yield a p-value < 0.05 (2.2E-16) proving that there is a very small probability that the null hypothesis is true. Therefore, be rejected. The standardized residuals (in the appendix) depict that even though the "Metropolitan Police" force has the highest number of accidents, fewer fatal and serious accidents have occurred than expected (std. res<1.96) (Harris, Jenine K., 2021)

To statistically prove significant variations across the police forces, analysis of variance was done assuming the sample means are equal. The hypothesis formulations are as follows:

H0: The sample means are equal
HA: The sample means are not equal

Table 2: ANOVA summary

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Police_Force	50	227261752	4545235	1.559	0.0306
Accident_Severity	2	118951621	59475811	20.399	3.72E-08
Residuals	100	291561061	2915611		

The p-value 0.0306 and 3.72E-08 are < 0.05 and are sufficient to reject the null hypothesis inferring that these groups likely came from a population with different means making the variations statistically significant.

Question 2

It was found that 97,779 accidents have occurred on weekdays and 32,203 accidents have occurred on weekends. The number of accidents represent each observation in the dataset, it wouldn't be the right metric to perform the hypothesis testing as it's not possible to compute the mean and standard deviation. Hence, various statistics have been made use of to analyse the situation.

Initially, we consider the proportion of accidents occurring on weekdays and weekends instead of counting.

Let P1 be the proportion of accidents occurring on weekdays = 0.7522
Let P2 be the proportion of accidents occurring on weekends = 0.2477

Then, our null and alternate hypothesis would be,

H0: P2-P1=0
HA: P2-P1≠0

$$Z = \frac{Observed\ value - Expected\ value}{Standrad\ Error}$$

$$Z = \frac{(P2-P1)-0}{\sqrt{(SE(P1))^2+(SE(P2))^2}}$$

Substituting proportion values and using SE(P) = $\sqrt{\frac{P(1-P)}{n}}$ where n is the number of observations, we get Z= 297. Since the Z-value is very high p-value→0 <<0.05. Hence, rejecting the null hypothesis.

Similarly, we consider two other statistics - Number of vehicles and number of casualties and computed t-test in R. If X1 and X2 are the numbers of vehicles/ number of casualties occurring on weekdays and weekends respectively. The null, alternate hypothesis, and p-values are as follows:

Table 3: Hypothesis and p-value summary

	Number of vehicles	Number of casualties
Null hypothesis	X1-X2=0	X1-X2=0
Alternate hypothesis	X1-X2! =0	X1-X2! =0
p-value	8.35E-16	2.20E-16

The null hypothesis is rejected as p-value < 0.05 implying that the difference in the number of accidents occurring on weekdays and weekends is statistically significant.

Question 3

It was observed that fatal accidents occur majorly during the day. 986 accidents with 1625 casualties have occurred in the day and 690 accidents with 1198 casualties in the night. The lighting condition characteristic was used to distinguish between day and night. Even though more fatal accidents occur in the day, casualties per accident are higher at the night (1.73). This can be due to various other contributing factors such as higher cases of alcohol consumption, weather conditions, etc. Maximum fatal casualties happen inroads with a speed limit of 60 miles/hour during day and night with casualty per accident of 1.89 and 1.96 respectively

To check if the difference is statistically significant, a t-test was performed assuming:

H0: Mean differences of casualties are the same in day and night
HA: Mean differences of casualties are different in day and night

The null hypothesis is rejected as p-value <0.05 implying that mean differences of casualties are different in day and night and are statistically significant

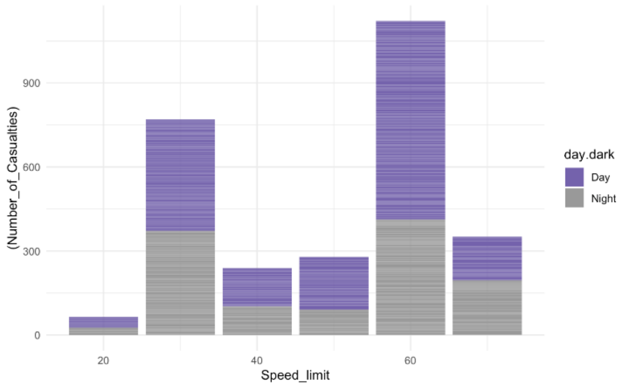

Figure 3: Stacked bar graph representing number of casualties across various speed limits and daytime

Question 4

Uni-variate analysis of various road types was performed to establish whether certain road types are dangerous, comparing the number of casualties across various factors. Single carriageways were found to be the most dangerous, with a highest number of accidents (122424 casualties) and casualty per accident of 1.3. It could be due of the lack of a divider that separates the traffic in opposite directions. However, dual carriageways have the highest casualty rate of 1.44. ANOVA results yielded a p-value <0.05 implying the mean casualties which are different across road types is statistically significant and not occurring by chance.

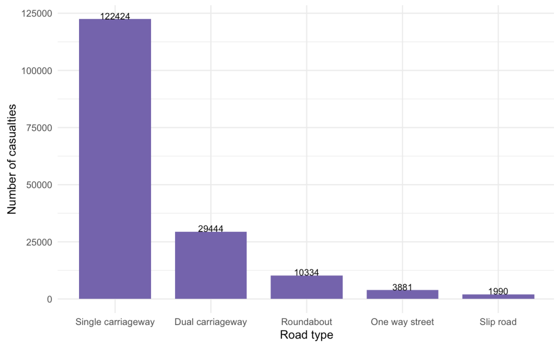

Figure: 4: Casualties across road types

A large number of accidents do occur in ‘A’ class and unclassified roads with a casualty rate of 1.35 and 1.24 respectively. Owing to broad motorways, the vehicle per accident is high as 2.28. It is observed that higher speed limit roads have higher casualty rates which are evident from the below figure.

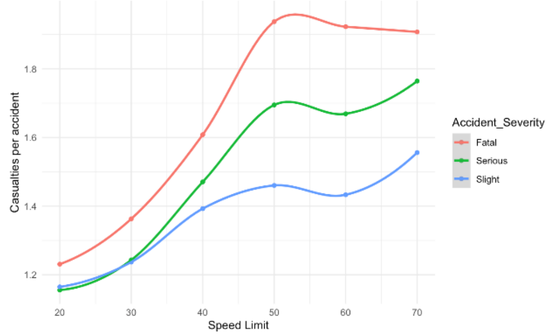

Figure 5: Variation of speed limit with casualty rate

Question 5

Some of the explanatory factors contributing to a number of accidents include certain junctions. Around 60% of the accidents occur at the junctions with staggered, crossroads, and roundabouts being major contributors. Slip roads have the highest casualty rate of 1.43. 80% of the accidents occur where there are no pedestrian crossing facilities within 50m. On the contrary, lesser accidents have occurred in traffic junctions and zebra crossings. High winds and rains are major causes of accidents. However, the casualty rates are almost the same across all weather conditions. The lack of dividers in single carriage roads can be another factor as over 73% of the accidents occur here. Even though over 70% of the accidents occur in dry road conditions, wet roads and floods over 3cm have high casualty rates of 1.35 and 1.45 respectively. ANOVA was conducted for each of the above contributing factors above whose p-values were <0.05 implying the difference of mean casualties across various factors was statistically significant and did not occur by chance.

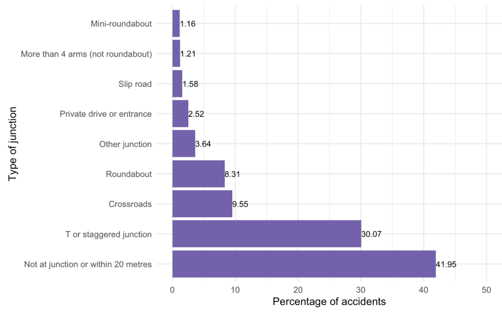

Figure 6: Analysis of accidents at junctions

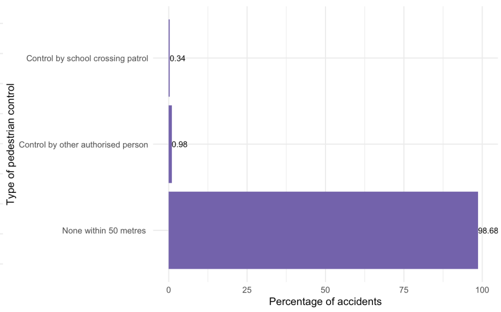

Figure 7: Pedestrian crossing controls

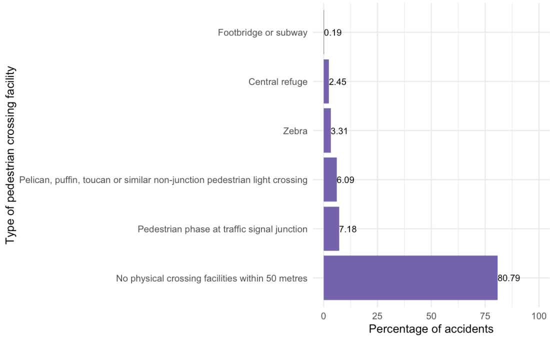

Figure 8: Pedestrian crossing facilities.

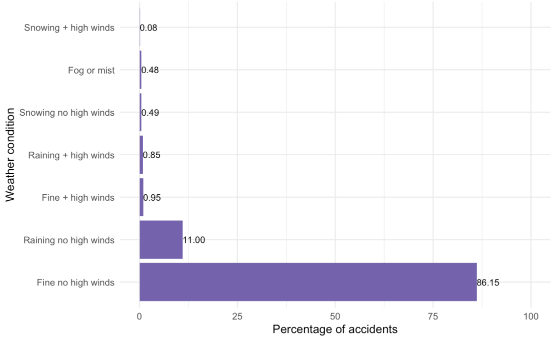

Figure 9: Accidents across weather conditions

Bi-variate analysis of road surface-weather, pedestrian crossing facilities-road type, and pedestrian control-road types was conducted to analyse the association through the chi-squared test. In all three cases, the p-value was <0.05 indicating their association is statistically significant. Residual analysis proved that the number of accidents is more than expected when roads are dry and no high winds. Also, they occur more than expected in single carriage ways without crossing facilities, subways in a slip road, traffic junctions in dual carriageways, and one-way streets.

4. Recommendations

Various contributing factors for road accidents in the UK are junction controls, pedestrian crossing facilities, weather conditions, etc. Junctions being a major contributor of accidents (60%) it's essential to install junction controls such as traffic signals at least in the staggered, crossroads and roundabouts. However, it has been observed that the remaining 40% of the accidents that don't occur at junctions are due to a lack of pedestrian crossing controls or lack of pedestrian crossing facilities. Hence, pedestrian crossing controls such as signals, authorized persons need to be deployed. As 58% of the accidents in places without pedestrian crossing control within 50m have occurred at the junctions themselves, it is essential to prioritise junction controls.
Histogram of accidents across the day depicts that a maximum number of accidents occur during the evening rush hour of 4 pm to 5 pm followed by morning peak hours around 8 am inroads with speed limit 30miles/hour with single carriageways being the highest. This implies that it's not speed rather the rush/volume of traffic that causes accidents. Also, it is evident from the below figure that the volume of traffic is higher on unclassified roads. Hence, it is recommended to divert the traffic to broader roads during rush hours to prevent accidents due to congestion.

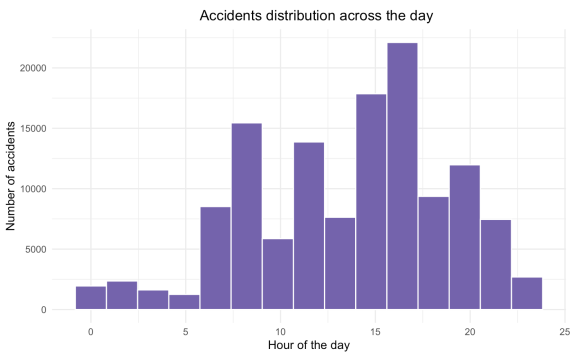

Figure 10: Histogram of accidents across the day

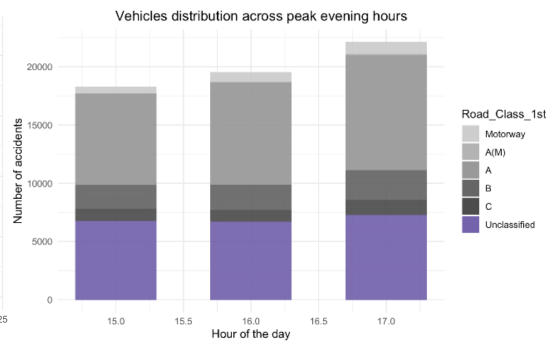

Figure 11: Vehicles involved in accident during peak hours

Although the casualty rate is almost the same in all weather conditions, more than 70% of accidents occur in dry road conditions due to drivers letting their guard down. Therefore, it is recommended to construct more placards for cautious driving on the roadside and strengthen the safety education for drivers regularly. The frequency of accidents on single carriage roads may be due to the lack of dividers. It is recommended to build a little vegetation separation belt in the centre of the road to separate the two side lanes. Although there are more fatal accidents during the day, the average number of casualties (1.73) is higher at night. This may be due to people's restricted vision when driving at night. It is recommended to improve road lighting conditions, increase streetlights, reflective strips, and other facilities.

5. References

[1] Margot Peeters, Mariëlle Zondervan-Zwijnenburg, Gerko Vink & Rens van de Schoot (2015), How to handle missing data: A comparison of different approaches, European Journal of Developmental Psychology, 12:4, 377-394

[2] Harris, Jenine K.(2021), Statistics with R : solving problems using real-world data. Sage, London, UK

6. Appendix

Code repository: https://github.com/kbkn11597/FBAMS/blob/master/Code/FBAMS_R_CODE.pdf

1. Below table explains the fact that even though maximum accidents occur in the metropolitan area, the fatal accidents are fewer than expected (with a std. res. Value <-1.96)

```
=====================================================================
##                                    Accident.cleaned$Accident_Severity
## Accident.cleaned$Police_Force      Fatal   Serious    Slight    Total
## ------------------------------------------------------------------------
## Metropolitan Police                  129      3566     23052    26747
##                                    344.9    4636.9   21765.2
##                                    0.005     0.133     0.862    0.206
##                                  -11.625   -15.727     8.722
## ------------------------------------------------------------------------
```

2. Skewness and kurtosis values of casualties per accidents across accident severity

```
skew.fatal

## skew (g1)        se         z         p
## 0.4198722 0.3429972 1.2241272 0.2209043

skew.serious

## skew (g1)        se         z         p
## 0.3010081 0.3429972 0.8775819 0.3801707

skew.slight

##     skew (g1)            se             z             p
## -1.438620e+00  3.429972e-01 -4.194262e+00  2.737613e-05

kurtosis.fatal

## Excess Kur (g2)              se                z                p
##    2.3385613722    0.6859943406     3.4090097161     0.0006519916

kurtosis.serious

## Excess Kur (g2)              se                z                p
##       0.6420749       0.6859943        0.9359770        0.3492850

kurtosis.slight

## Excess Kur (g2)              se                z                p
##    4.339628e+00    6.859943e-01     6.326041e+00     2.515306e-10
```

3. Test of significance for accidents occurring in day and night

```
t.test(formula=Accident.cleaned$Number_of_Casualties~Accident.cleaned$day.dark)

##  Welch Two Sample t-test
##
## data:  Accident.cleaned$Number_of_Casualties by Accident.cleaned$day.dark
## t = -7.8674, df = 67093, p-value = 3.674e-15
## alternative hypothesis: true difference in means between group Day and group
Night is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.04669795 -0.02807086
## sample estimates:
##   mean in group Day mean in group Night
##            1.304786            1.342171
```

4. Standard residual table for road surface condition and weather

```
================================================================================
##                 Accident.cleaned$Road_Surface_Conditions
## Accdn.$W_C          Dry   Wet or dmp      Snow   Frst or ic   Fld o 3. d
Total
## ------------------------------------------------------------------------------
-
## Fn n hgh w      91370        16633        81         1979           12
110075
##              79628.5      27993.9     367.9       1993.5         91.1
##                0.830        0.151     0.001        0.018        0.000
0.860
##               41.609      -67.902   -14.959       -0.326       -8.289
## ------------------------------------------------------------------------------
-
## Rnng n h w        259        13855        14           83           63
14274
##              10325.8       3630.1      47.7        258.5         11.8
##                0.018        0.971     0.001        0.006        0.004
0.111
##              -99.067      169.706    -4.881      -10.916       14.890
## ------------------------------------------------------------------------------
-
## Snwn n h w         17          238       273          109            0
637
##                460.8        162.0       2.1         11.5          0.5
##                0.027        0.374     0.429        0.171        0.000
0.005
##              -20.674        5.971   185.632       28.695       -0.726
## ------------------------------------------------------------------------------
-
## Fn + hgh w        875          337         5           18            0
1235
##                893.4        314.1       4.1         22.4          1.0
##                0.709        0.273     0.004        0.015        0.000
0.010
##               -0.616        1.293     0.429       -0.923       -1.011
## ------------------------------------------------------------------------------
-
## Rnng + h w         16         1042         2           12           30
1102
##                797.2        280.3       3.7         20.0          0.9
##                0.015        0.946     0.002        0.011        0.027
0.009
##              -27.668       45.502    -0.877       -1.781       30.454
## ------------------------------------------------------------------------------
-
## Snwn + h w          3           26        50           20            0
99
##                 71.6         25.2       0.3          1.8          0.1
##                0.030        0.263     0.505        0.202        0.000
0.001
##               -8.108        0.164    86.343       13.597       -0.286
```

```
## ---------------------------------------------------------------------------
-
## Fog or mst          88         433         3         98          1
623
##                  450.7       158.4       2.1       11.3        0.5
##                  0.141       0.695     0.005      0.157      0.002
0.005
##                -17.084      21.813     0.636     25.816      0.674
## ---------------------------------------------------------------------------
-
## Total            92628       32564       428       2319        106
128045
## ===========================================================================
```

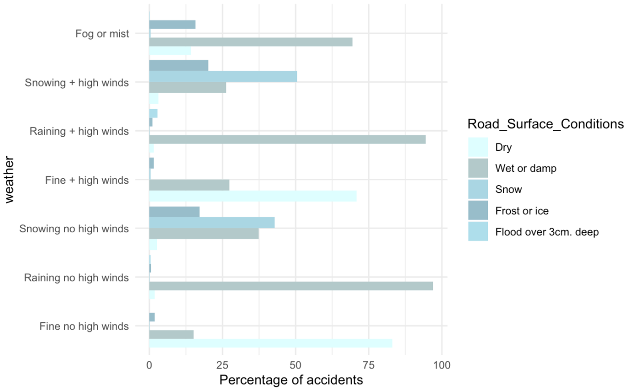

Figure: Comparison of weather and road surface conditions

5. Standard residual table for road types and pedestrian crossing facilities

```
================================================================================
##              Accident.cleaned$Road_Type
## A.$P_C_P_    Roundabot    On wy str    Dl crrgwy    Sngl crrg    Slip road
Total
## ---------------------------------------------------------------------------
-
## N p c f w         6667         2284        15750        76653         1292
102646
##                 6780.0       2727.5      16384.1      75565.6       1188.9
##                  0.065        0.022        0.153        0.747        0.013
0.806
##                 -1.372       -8.491       -4.954        3.956        2.989
## ---------------------------------------------------------------------------
-
## Zebra              343          254          305         3290           31
4223
##                  278.9        112.2        674.1       3108.9         48.9
##                  0.081        0.060        0.072        0.779        0.007
0.033
##                  3.836       13.385      -14.215        3.248       -2.561
## ---------------------------------------------------------------------------
-
## P, p, t o          433          379         1573         5449           43
7877
##                  520.3        209.3       1257.3       5798.9         91.2
##                  0.055        0.048        0.200        0.692        0.005
0.062
##                 -3.827       11.730        8.903       -4.594       -5.050
## ---------------------------------------------------------------------------
```

```
-
## P p a t s          308         399        2342        6141          82
9272
##                  612.4       246.4      1480.0      6825.8       107.4
##                  0.033       0.043       0.253       0.662       0.009
0.073
##                -12.302       9.724      22.408      -8.289      -2.451
## ----------------------------------------------------------------------------
-
## Ftbrd o s           73           5          81          67          12
238
##                   15.7         6.3        38.0       175.2         2.8
##                  0.307       0.021       0.340       0.282       0.050
0.002
##                 14.447      -0.526       6.978      -8.175       5.567
## ----------------------------------------------------------------------------
-
## Cntrl rfg          593          65         289        2211          16
3174
##                  209.6        84.3       506.6      2336.6        36.8
##                  0.187       0.020       0.091       0.697       0.005
0.025
##                 26.476      -2.106      -9.669      -2.599      -3.425
## ----------------------------------------------------------------------------
-
## Total             8417        3386       20340       93811        1476
127430
## =========================================================================
```

7. Standard residual table for road surface condition and road types

```
===============================================================================
##                Accident.cleaned$Road_Type
## Ac.$R_S_C     Roundabot   On wy str   Dl crrgwy   Sngl crrg   Slip road
Total
## ----------------------------------------------------------------------------
-
## Dry               6050        2596       14520       67129        1063
91358
##                 6034.4      2408.0     14610.5     67250.7      1054.4
##                  0.066       0.028       0.159       0.735       0.012
0.723
##                  0.201       3.832      -0.749      -0.469       0.265
## ----------------------------------------------------------------------------
-
## Wt or dmp         2189         705        5295       23673         374
32236
##                 2129.3       849.7      5155.4     23729.7       372.0
##                  0.068       0.022       0.164       0.734       0.012
0.255
##                  1.295      -4.963       1.945      -0.368       0.101
## ----------------------------------------------------------------------------
-
## Snow                15           4          81         317           7
424
##                   28.0        11.2        67.8       312.1         4.9
##                  0.035       0.009       0.191       0.748       0.017
0.003
##                 -2.458      -2.146       1.602       0.276       0.952
## ----------------------------------------------------------------------------
```

```
-
## Frst or i              94          27         279        1877          15
2292
##                     151.4        60.4       366.5      1687.2        26.5
##                     0.041       0.012       0.122       0.819       0.007
0.018
##                    -4.664      -4.299      -4.573       4.621      -2.227
## -------------------------------------------------------------------------
-
## Fl o 3. d               2           0          42          61           0
105
##                       6.9         2.8        16.8        77.3         1.2
##                     0.019       0.000       0.400       0.581       0.000
0.001
##                    -1.874      -1.664       6.152      -1.853      -1.101
## -------------------------------------------------------------------------
-
## Total                8350        3332       20217       93057        1459
126415
## ===================================================================
```

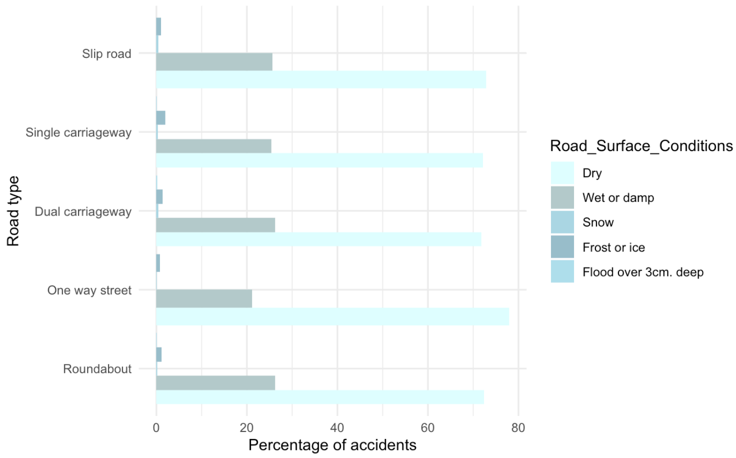

Figure: Comparison of road surface conditions and road types

8. ANOVA test for junction detail

```
oneway.test(formula = Number_of_Casualties~Junction_Detail,data=Accident.c
leaned,var.equal = TRUE)

##
##  One-way analysis of means
##
## data:  Number_of_Casualties and Junction_Detail
## F = 56.535, num df = 8, denom df = 129364, p-value < 2.2e-16
```

9. ANOVA test for pedestrian human control

```
oneway.test(formula = Number_of_Casualties~Pedestrian_Crossing_Human_Contr
ol,data=Accident.cleaned,var.equal = TRUE)

##
##  One-way analysis of means
##
## data:  Number_of_Casualties and Pedestrian_Crossing_Human_Control
## F = 8.7893, num df = 2, denom df = 127405, p-value = 0.0001525
```