UNIVERSITY OF Southampton

Group Coursework

MANG6513- Foundation of Business Analytics and Management Sciences

Group No.9

Student ID:
31302815
32519001
31552552
32246013

Word count: 2104

Table of Contents

1. Introduction *3*

2. Problem description *3*

3. Analytical method *4*

4. Analysis *5*

4.1 Question 1 .... **5**

4.2 Question 2 .... **8**

4.3 Question 3 .... **10**

4.4 Question 4 .... **12**

4.5 Question 5 .... **14**

4.6 Question 6 .... **17**

5. Conclusion *20*

6. Recommendations *21*

7. References *22*

8. Appendices *23*

1. Introduction

In modern society, with the increasing number of accidents, factors causing accidents need to be analyzed and acted upon. A study shows that occurrence of accidents will be affected by road, traffic, environment and other related factors (Angelica & Olga, 2016). Correspondingly, a pressure group called 'Safer Roads UK' also highlighted the importance of road safety.

If the number of traffic accidents is to be reduced, road safety needs to be improved. In addition, R studio is a very useful software for analyzing such problems based on past recorded data. This report will use R studio to analyze data related to accidents. Therefore, this report will draw some conclusions based on the analysis of road accident data in 2017 and make recommendations based on the analysis results to make UK roads safer in 2019.

2. Problem description

Road Accident data of 2017 from UK has been provided for analysis. Data has 29 variables and 129982 observations.

Aim is to analyse this data using statistics with the help of R language and write a report to provide insights to make UK roads safer. The 'To Be' system is to make UK roads safer so that number of accidents can be reduced significantly. The desired system needs to be created based on statistical analysis of provided data which can explain factors affecting increasing number of accidents.

3. Analytical method

There are six analysis methods used in this report:

1. Descriptive statistics: Refers to various activities that use tabulation and classification, graphics, and calculation of summary data to describe data characteristics (Wienclaw, 2021).
2. Student's T test: Student's T test uses T distribution theory to infer the probability of the difference, to compare whether the difference between two averages is significant (Ruth, 2019).
3. Chi-square test: Chi-square test is the degree of deviation between the actual observation value of the statistical sample and the theoretical inferred value (Turhan, 2020).
4. Analysis of Variance (ANOVA): It is used to test the significance of the difference between the means of three or more samples (Holt, 2020).
5. Linear regression: Linear regression is a regression analysis that uses the least square function of the linear regression equation to model the relationship between one or more independent variables and dependent variables (Tantawi, 2021).
6. Multinominal logistic regression model: It is a kind of generalized linear regression. GLM is an extension of the linear model, which establishes the relationship between the mathematical expectation of the response variable and the predictor variable of the linear combination through the link function (Annette & Adrian, 2018).

4. Analysis

4.1 Question 1

Whether there are significant variations in the number of accidents and in the accidents' severity across police forces?

There are significant variations in number of accidents and accidents' severity across police forces. To begin with, the bar chart (figure1) shows that the police force1 (Metropolitan Police) has the highest total number of accidents. Subsequently, table has been created using the "f table" method (figure2). The clear distinction as shown in the table, is that the first police force has the most accidents among the three severity levels which are fatal, serious, and slight followed by Two-way ANOVA with replication as shown in Figure3. Null hypothesis (H0) indicates that the number of accidents from the first to the 98th police force is the same, and alternate hypothesis(H1) illustrates at least one police force will have a different number of accidents. The result shows that P-value=2e-16 < α=0.05. As a result, Null hypothesis is rejected which implies that at least one police force will have a different number of accidents.

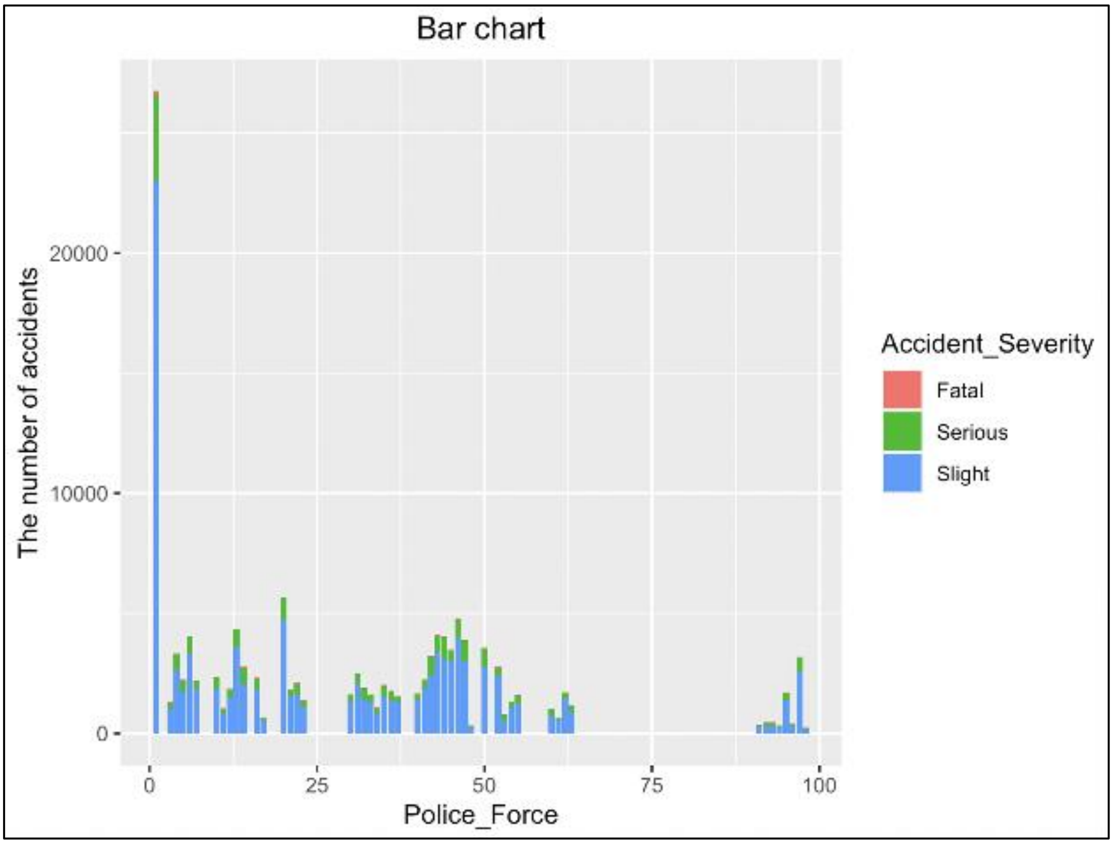

Figure (1): Number of Accidents against Police Forces

	Accident Severity				Accident Severity		
Police Force	Fatal	Serious	Slight	Police Force	Fatal	Serious	Slight
1	129	3566	23052	40	19	250	1390
3	27	250	1014	41	23	351	1868
4	40	641	2674	42	45	806	2384
5	26	500	1715	43	59	677	3393
6	50	687	3332	44	42	902	3128
7	21	327	1862	45	33	482	2987
10	31	460	1871	46	56	758	3974
11	19	204	825	47	50	849	3013
12	38	361	1457	48	2	56	255
13	39	742	3589	50	54	714	2800
14	42	692	2060	52	46	295	2433
16	28	475	1825	53	19	255	525
17	7	137	535	54	23	225	1072
20	54	885	4743	55	26	291	1279
21	31	217	1559	60	21	238	745
22	53	438	1630	61	19	94	516
23	34	271	1063	62	32	239	1460
30	34	297	1312	63	26	265	899
31	32	394	2066	91	17	63	272
32	45	454	1409	92	14	149	304
33	19	248	1356	93	22	120	317
34	40	207	821	94	5	72	239
35	42	377	1585	95	22	294	1376
36	28	351	1388	96	6	88	312
37	31	233	1309	97	44	544	2597
				98	11	43	182

Figure (2): Accident Severity against Police Forces

```
                             Df    Sum Sq Mean Sq F value Pr(>F)    
data1$Accident_Severity       2    161394   80697   128.7 <2e-16 ***
Residuals                129979  81495857     627                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure (3): Result of ANOVA Analysis

4.2 Question 2

Whether the difference in the number of accidents between weekends and weekdays is significant?

There is no significant variation in number of accidents in weekends and weekdays. The result of the analysis are presented below. Firstly, the number of accidents on weekdays (2-6) and weekends (1,7) have been calculated and divided into two groups. Thereafter, there means are compare in histograms. Therefore, the number of accidents on weekdays is higher than on weekends and weekends.

T-test is performed to test the hypothesis. H0 defines that there is no distinction between weekdays and weekends. There is a distinction between weekdays and weekends for H1. It can be deduced from the result that P-value=0.1506 >α =0.05, which means null hypothesis cannot be rejected. This states that there is no significant difference between the two.

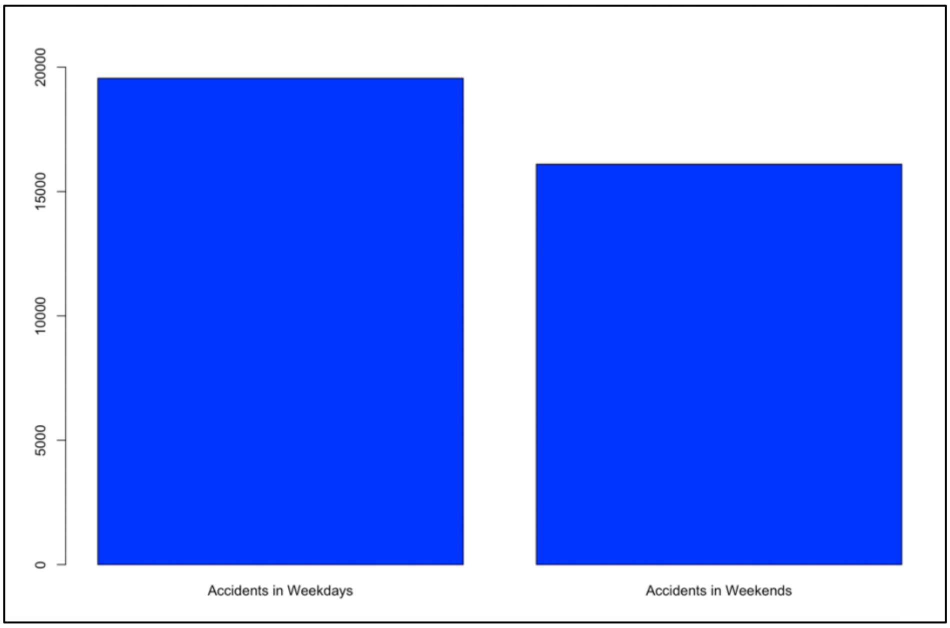

Figure (4): Difference in Number of Accidents in Weekdays and Weekend

```
> t.test(df_weekday_group$Accident_Group, df_weekend_group$Accident_Group)

	Welch Two Sample t-test

data:  df_weekday_group$Accident_Group and df_weekend_group$Accident_Group
t = 2.7796, df = 1.4735, p-value = 0.1506
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4213.607 11122.207
sample estimates:
mean of x mean of y 
  19555.8   16101.5 
```

Figure (5): T-test result

4.3 Question 3

Whether more fatal accidents happen at night?

According to data, 1676 accidents out of a total of 129,982 are fatal. Using the count and table functions, it can see that 1005 fatal accidents occur during the day and 671 occur at night. The chi-squared test is then used as hypothesis testing to see whether the output has the same distribution. The null hypothesis will be the fatal accidents happening at night and during the day are the same. The alternative hypothesis (H1) is that there is a variation in the accident between day and night. The result indicates a P-value of 3.393e-16 <0.05, meaning H0 is rejected. Consequently, it can concluded that total number of fatal accidents during day and night vary and more fatal accidents happen during day.

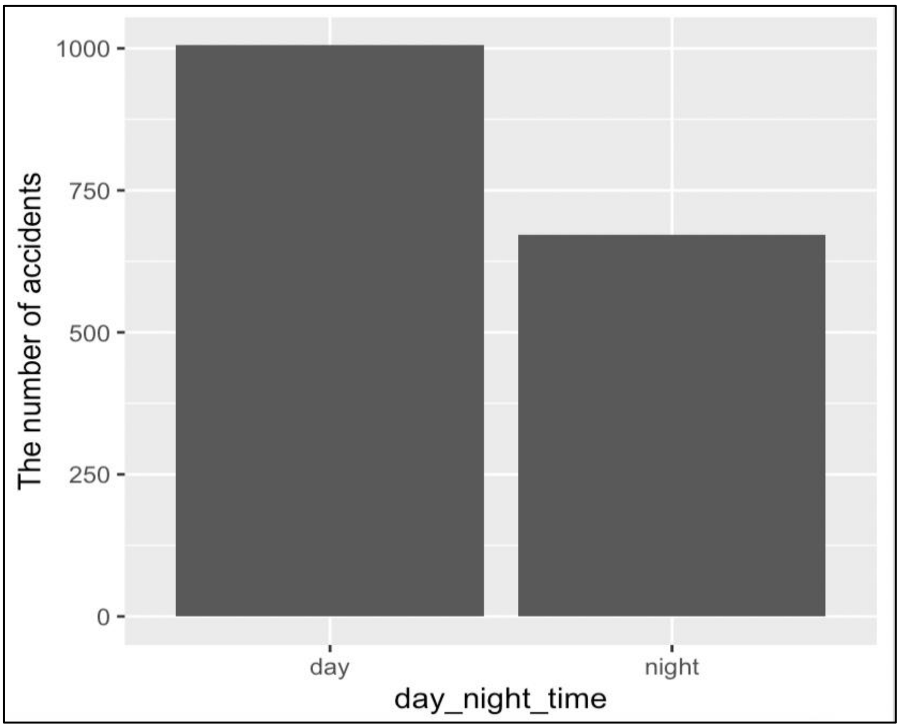

Figure (6): Difference in number of accidents in day and night time

	day	night
Fatal	1005	671

Figure (7): Total number of fatal accidents in day and night time

```
        Chi-squared test for given probabilities

data:  fatal_num
X-squared = 66.561, df = 1, p-value = 3.393e-16
```

Figure (8): Chi-square test result

4.4 Question 4

Whether certain road types are more dangerous?

Single carriageway is more dangerous of all road types. Detailed analysis is as follow: Figure (9) shows the sum of the three levels of accidents for each road type. Figure (10) shows the detailed fatal level, serious level, and slight level of accidents for each road type. From the plot (Figure9), it can be deduced that the total number of accidents on a single carriageway is the highest. From the table data (Figure10), it can be observed that the number of fatal accidents, serious accidents, and slight accidents on the single carriageway are the highest. Among them, the number of fatal accidents on single carriageway was 1,287, the number of serious accidents was 17,443, and the number of slight accidents was 75,081. By comparison with other road types, both the total number of accidents and the number of accidents of different severities on single carriageway are the highest, so single carriageway is most dangerous.

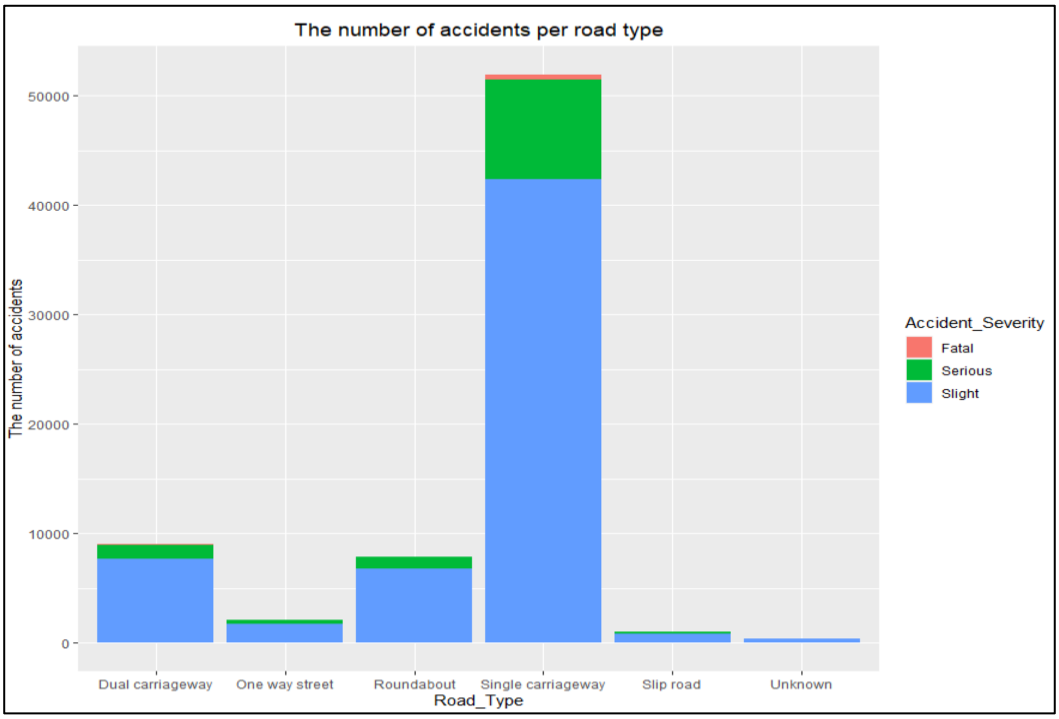

Figure (9): Number of accidents against road types

	Fatal	Serious	Slight
Dual carriageway	323	3034	16983
One way street	21	557	2808
Roundabout	28	1066	7323
Single carriageway	1287	17443	75081
Slip road	9	193	1274
Unknown	8	241	2303

Figure (10): Accident Severity against Road Types

4.5 Question 5

What explanatory factors contribute to the number of accidents?

Based on data provided, using human intelligence, correlation matrix (figure 14), and using the analytical method- linear regression it is found that Weather Conditions, Light Conditions, and Road Surface Conditions play a significant role compared to other variables in the number of accidents. These three factors were found to be significant basis p-value of linear regression where critical value was 0.1, kindly refer below figure -(11), (12), and (13). On further analysis of each of the significant variables it is deduced that among all types of weather conditions, a major number of accidents is recorded in 'Daylight' weather in contrast to other light conditions, refer figure (15). On examining Weather Conditions discretely, it has been observed that the highest number of accidents have been recorded in 'Fine no high winds' weather type, refer figure (16) whereas on investigating Road Surface Conditions it is found that the maximum number of accidents have taken place on 'Dry' road surfaces, refer figure (17).

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          55882       9234   6.051  0.00905 **
Light_Conditions     -9006       1832  -4.915  0.01613 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8436 on 3 degrees of freedom
Multiple R-squared:  0.8895,    Adjusted R-squared:  0.8527
F-statistic: 24.16 on 1 and 3 DF,  p-value: 0.01613
```

Figure (11): Linear Regression Result- Light Conditions

```
lm(formula = No_of_Accidents_grouped ~ Weather_Conditions, data = df_weather_group)

Residuals:
   Min      1Q Median      3Q     Max
-16186 -11579  -3778    6011   34658

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)           29080      12187   2.386   0.0484 *
Weather_Conditions    -4210       2166  -1.944   0.0930 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure (12): Linear Regression Result- Weather Conditions

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                 51495      14684   3.507   0.0393 *
Road_Surface_Conditions    -12347       4427  -2.789   0.0685 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14000 on 3 degrees of freedom
Multiple R-squared:  0.7216,    Adjusted R-squared:  0.6289
F-statistic: 7.778 on 1 and 3 DF,  p-value: 0.06849
```

Figure (13): Linear Regression Result- Road Surface Conditions

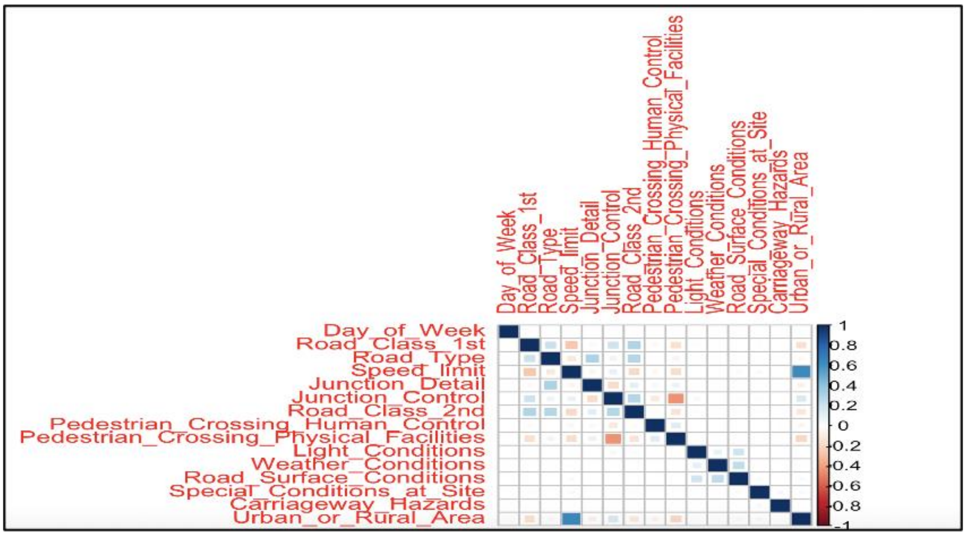

Figure (14): Correlation Matrix of Independent Variables

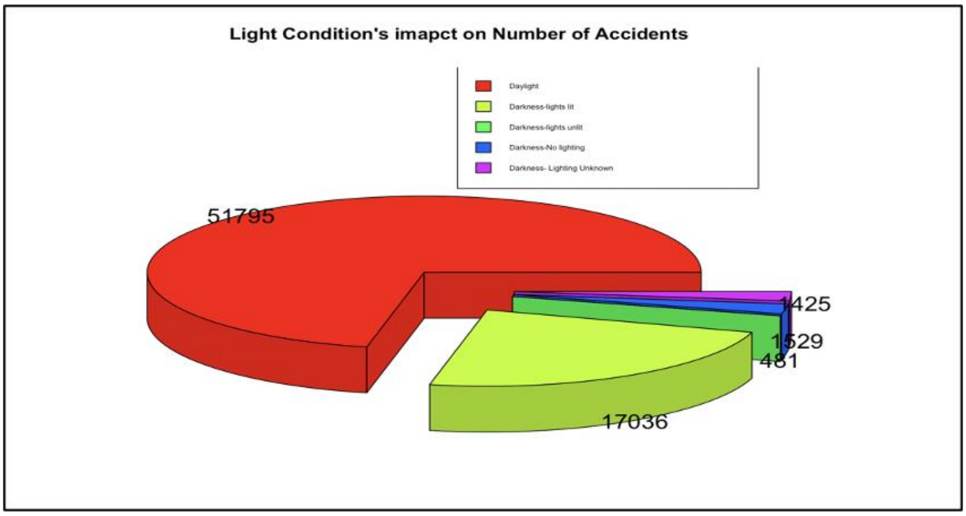

Figure (15): Light Condition’s Impact on Accidents

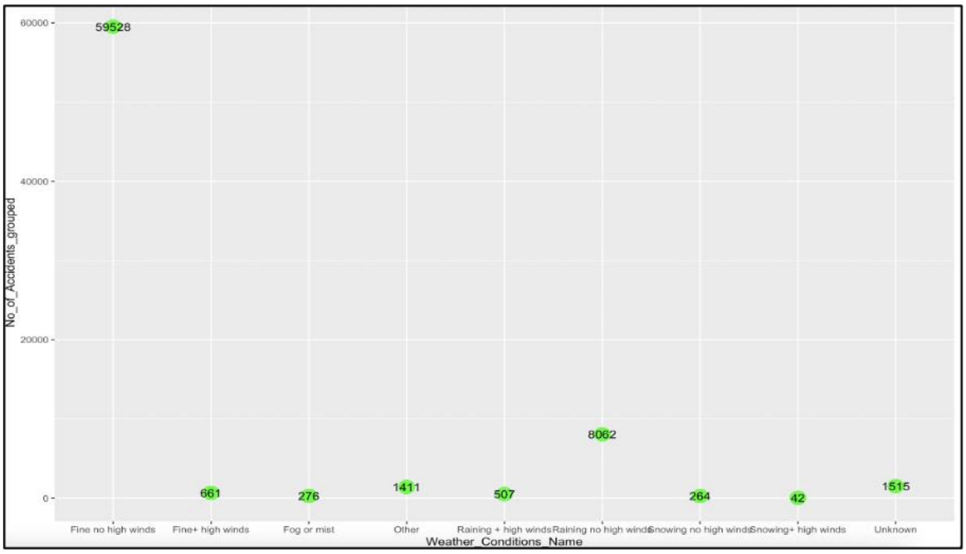

Figure (16): Weather Condition's Impact on Accidents

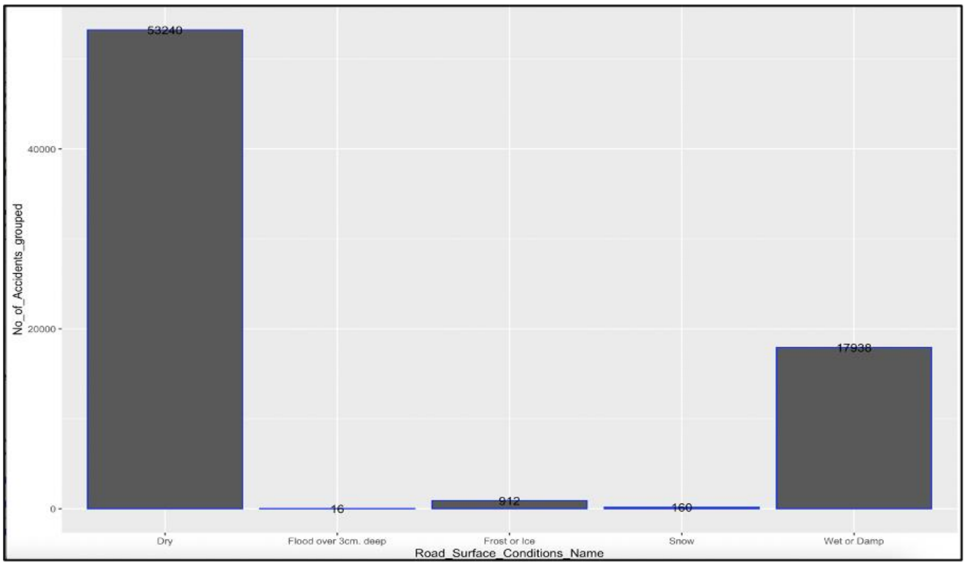

Figure (17): Road Surface Condition's Impact on Accidents

4.6 Question 6

Remember that Safer Roads UK values your creativity and good understanding of Business Analytics methods; hence, they want you to conduct any relevant and useful analyses from the given data that can provide important insights on how UK roads could be made safer in the future

To make UK roads safer and save people's lives is a long and cumbersome process. However, in immediate action plan, primarily the focus must be on how to reduce accident severity and action must be taken to improve the factors linked to severe accidents. Factors associated with 'fatal' accidents have been elicited from the data using analytical method- Multinomial Logistic Regression. Based on regression analysis result, it is found that factors having p-value<0.05 (refer Figure 18,19) and exponential coefficient with respect to fatal accidents for respective factors<0.8 (refer Figure 19,20) are leading to fatal accidents. It is observed that following factors and sub-factors have led to increase in 'fatal' accidents:

Factors	Sub-Factors
Road Class 1st	A
Road Type	Single Carriageway
Junction Detail	Mini Roundabout, T or staggered junction, Slip Road, Private drive, or entrance
Junction Control	Auto traffic signal, Stop Sign, Give way or uncontrolled
Pedestrian Cross Physical Facilities	Pelican, puffin, toucan or similar non-junction pedestrian light crossing, Pedestrian phase at traffic signal junction, Footbridge or subway, Central refuge
Light Conditions	Darkness-light lit, Darkness- no lighting
Weather Conditions	Fine + High wind
Road Surface Conditions	Frost or ice
Urban or Rural Area	Urban

```
> p_value_acc_severity
  (Intercept) Day_of_Week2 Day_of_Week3 Day_of_Week4 Day_of_Week5 Day_of_Week6 Day_of_Week7
2           0 1.939062e-04 0.0064616823 5.386192e-04 0.0400458844 1.453038e-05   0.08674718
3           0 6.467983e-07 0.0001515809 3.682881e-06 0.0006925884 9.880803e-08   0.01142962
  Road_Class_1st2 Road_Class_1st3 Road_Class_1st4 Road_Class_1st5 Road_Class_1st6 Road_Type2 Road_Type3
2               0     0.033819263       0.2719634       0.1829386       0.2824800  0.4017479  0.5406098
3               0     0.009351572       0.1211234       0.1844058       0.1576229  0.4824649  0.8997846
  Road_Type6 Road_Type7 Road_Type9  Speed_limit Junction_Detail2 Junction_Detail3 Junction_Detail5
2 0.02368726  0.3783039  0.6314978 6.661338e-16       0.01351630       0.92146805     4.345503e-05
3 0.01503727  0.2369113  0.9798456 0.000000e+00       0.01497222       0.07343321     7.614192e-06
  Junction_Detail6 Junction_Detail7 Junction_Detail8 Junction_Detail9 Junction_Control2
2        0.6210957        0.5093706       0.37648041        0.4224966                 0
3        0.1095073        0.5722102       0.02965661        0.8900127                 0
  Junction_Control3 Junction_Control4 Road_Class_2nd2 Road_Class_2nd3 Road_Class_2nd4 Road_Class_2nd5
2      2.340635e-04                 0       0.4950944       0.2022074       0.9295769       0.3800713
3      2.609325e-05                 0       0.6042317       0.2573008       0.7232378       0.5342427
  Road_Class_2nd6 Pedestrian_Crossing_Human_Control1 Pedestrian_Crossing_Human_Control2
2       0.6736767                          0.9373236                          0.3087419
3       0.2681825                          0.9518211                          0.2524081
  Pedestrian_Crossing_Physical_Facilities1 Pedestrian_Crossing_Physical_Facilities4
2                               0.03157378                              0.045116023
3                               0.04277763                              0.005230064
  Pedestrian_Crossing_Physical_Facilities5 Pedestrian_Crossing_Physical_Facilities7
2                              0.043911626                               0.06556286
3                              0.005550746                               0.04963032
  Pedestrian_Crossing_Physical_Facilities8 Light_Conditions4 Light_Conditions5 Light_Conditions6
2                               0.15502930      1.043522e-04         0.3641869       0.032993651
3                               0.02461145      2.031534e-08         0.2551687       0.001571413
  Light_Conditions7 Weather_Conditions2 Weather_Conditions3 Weather_Conditions4 Weather_Conditions5
2         0.7483093           0.6202355           0.5059726         0.018387189           0.5829451
3         0.6272710           0.9396837           0.5911030         0.007693549           0.7622044
  Weather_Conditions6 Weather_Conditions7 Weather_Conditions8 Weather_Conditions9
2        1.705805e-09           0.2764072           0.8959271           0.3406092
3        0.000000e+00           0.3446047           0.9293162           0.2224906
  Road_Surface_Conditions2 Road_Surface_Conditions3 Road_Surface_Conditions4 Road_Surface_Conditions5
2                0.7910213                        0               0.09470589             3.080851e-04
3                0.4954167                        0               0.20267209             1.881877e-06
```

Figure (18): p-value of Factors

```
  Road_Surface_Conditions2 Road_Surface_Conditions3 Road_Surface_Conditions4 Road_Surface_Conditions5
2                0.7910213                        0               0.09470589             3.080851e-04
3                0.4954167                        0               0.20267209             1.881877e-06
  Special_Conditions_at_Site1 Special_Conditions_at_Site2 Special_Conditions_at_Site3
2                1.320738e-03                           0                   0.4640264
3                8.784370e-08                           0                   0.4211324
  Special_Conditions_at_Site4 Special_Conditions_at_Site5 Special_Conditions_at_Site6
2                   0.5525653                   0.4146666                           0
3                   0.4771581                   0.7166002                           0
  Special_Conditions_at_Site7 Carriageway_Hazards1 Carriageway_Hazards2 Carriageway_Hazards3
2                           0            0.2699943            0.8569979                    0
3                           0            0.6192834            0.7990092                    0
  Carriageway_Hazards6 Carriageway_Hazards7 Urban_or_Rural_Area2 Urban_or_Rural_Area3
2            0.3049927                    0           0.06667981            0.2420597
3            0.3158690                    0           0.01280865            0.9084147
> exp(coefficients(acc_severity_model))
  (Intercept) Day_of_Week2 Day_of_Week3 Day_of_Week4 Day_of_Week5 Day_of_Week6 Day_of_Week7
2    3966.822     1.999652     1.571149     1.839804     1.391660     2.147835     1.321591
3   70618.294     2.489536     1.855254     2.232389     1.709563     2.525212     1.498213
  Road_Class_1st2 Road_Class_1st3 Road_Class_1st4 Road_Class_1st5 Road_Class_1st6 Road_Type2 Road_Type3
2        64.52032       0.4549319       0.6461157       0.5669846       0.6573256   1.796741  0.8245540
3        85.78106       0.3929750       0.5495876       0.5766351       0.5854124   1.629885  0.9616007
  Road_Type6 Road_Type7 Road_Type9 Speed_limit Junction_Detail2 Junction_Detail3 Junction_Detail5
2  0.5031473   1.584174  0.6555138   0.9574541        0.3531515        0.9751440        0.2883759
3  0.4825349   1.835547  1.0222091   0.9487404        0.3644529        0.6373732        0.2667030
  Junction_Detail6 Junction_Detail7 Junction_Detail8 Junction_Detail9 Junction_Control2
2        0.8724276         1.447880        0.7588788         1.314387        0.07907829
3        0.6471099         1.368056        0.5128736         1.047646        0.06301397
  Junction_Control3 Junction_Control4 Road_Class_2nd2 Road_Class_2nd3 Road_Class_2nd4 Road_Class_2nd5
2         0.1756799        0.12755485       0.6254849        1.466984       1.0289060        1.351353
3         0.1401476        0.08031021       0.7086359        1.391160       0.8949814        1.230979
  Road_Class_2nd6 Pedestrian_Crossing_Human_Control1 Pedestrian_Crossing_Human_Control2
2       0.8843928                          1.0737119                           1.915400
3       0.7312432                          0.9473003                           2.065422
  Pedestrian_Crossing_Physical_Facilities1 Pedestrian_Crossing_Physical_Facilities4
2                                 4.099530                                0.6698089
3                                 3.769044                                0.5762687
```

Figure (19): p-value and exponential values of factors

```
 Junction_Control3 Junction_Control4 Road_Class_2nd2 Road_Class_2nd3 Road_Class_2nd4 Road_Class_2nd5
2         0.1756799        0.12755485       0.6254849        1.466984       1.0289060        1.351353
3         0.1401476        0.08031021       0.7086359        1.391160       0.8949814        1.230979
 Road_Class_2nd6 Pedestrian_Crossing_Human_Control1 Pedestrian_Crossing_Human_Control2
2       0.8843928                          1.0737119                           1.915400
3       0.7312432                          0.9473003                           2.065422
 Pedestrian_Crossing_Physical_Facilities1 Pedestrian_Crossing_Physical_Facilities4
2                                4.099530                                0.6698089
3                                3.769044                                0.5762687
 Pedestrian_Crossing_Physical_Facilities5 Pedestrian_Crossing_Physical_Facilities7
2                               0.6595531                                0.3030428
3                               0.5688548                                0.2945506
 Pedestrian_Crossing_Physical_Facilities8 Light_Conditions4 Light_Conditions5 Light_Conditions6
2                               0.6911416         0.6340122         0.6377332         0.6599189
3                               0.5620720         0.5222605         0.5754395         0.5509523
 Light_Conditions7 Weather_Conditions2 Weather_Conditions3 Weather_Conditions4 Weather_Conditions5
2         0.8872088           0.9096147           0.5813717           0.4555007            1.393684
3         0.8362576           1.0143425           0.6516127           0.4217260            1.198395
 Weather_Conditions6 Weather_Conditions7 Weather_Conditions8 Weather_Conditions9
2             8.35255           0.5951619           0.9530435            1.575567
3            23.93528           0.6478714           1.0326576            1.781203
 Road_Surface_Conditions2 Road_Surface_Conditions3 Road_Surface_Conditions4 Road_Surface_Conditions5
2                 1.038387                 787.3884                0.5823954                 4.069996
3                 1.100175                1599.8149                0.6697593                 6.526724
 Special_Conditions_at_Site1 Special_Conditions_at_Site2 Special_Conditions_at_Site3
2                    1.453007                    14.20206                   0.5579434
3                    1.871061                    25.15546                   0.5346842
 Special_Conditions_at_Site4 Special_Conditions_at_Site5 Special_Conditions_at_Site6
2                    1.418409                   2.4869702                    379.6706
3                    1.511672                   0.6667987                    308.2359
 Special_Conditions_at_Site7 Carriageway_Hazards1 Carriageway_Hazards2 Carriageway_Hazards3
2                    146.7654            0.3908558            1.1069428             79.28246
3                    140.7467            0.6668644            0.8675683             31.50551
 Carriageway_Hazards6 Carriageway_Hazards7 Urban_or_Rural_Area2 Urban_or_Rural_Area3
2            0.4102184             96.80233            0.7554470            1.9722514
3            0.4268710            154.93518            0.6870204            0.9349019
>
```

Figure (20): Exponential values of factors

Odd Ratio:

```
 Junction_Detail6 Junction_Detail7 Junction_Detail8 Junction_Detail9 Junction_Control2
2        0.8724276         1.447880        0.7588788         1.314387        0.07907829
3        0.6471099         1.368056        0.5128736         1.047646        0.06301397
 Junction_Control3 Junction_Control4 Road_Class_2nd2 Road_Class_2nd3 Road_Class_2nd4 Road_Class_2nd5
2         0.1756799        0.12755485       0.6254849        1.466984       1.0289060        1.351353
3         0.1401476        0.08031021       0.7086359        1.391160       0.8949814        1.230979
 Road_Class_2nd6 Pedestrian_Crossing_Human_Control1 Pedestrian_Crossing_Human_Control2
```

Figure (21): Odd Ratio Example

- In case of Junction Detail 8(highlighted section), for factor Junction Detail- Private Drive or Entrance the odds of fatal accident to serious accident is 0.7588788 i.e., the odds of serious accident is 1-0.7588788=0.241122 lesser than fatal accident for respective factors.
- In case of Junction Detail 9, for factor Junction Detail- Other Junction the odds of Fatal accident to Slight accident are 1.047646 i.e., the odds of fatal accident are 1.047646-1= 0.047646 more than of slight accident.

5. Conclusion

Following conclusions have been drawn based upon application of statistical tools and diagrams using R:

1. There are significant variations in the number of accidents and in the accidents' severity across police forces.

2. There is no significant difference in the number of accidents between weekends and weekdays.

3. Total number of fatal accidents that happen during day and night time varies. There are more fatal accidents happening during day time than night.

4. On comparing with other road types, total number of accidents and number of accidents with different severities on single carriageway is highest, therefore single carriageway is most dangerous road type.

5. Explanatory factors which contribute to 'number of accidents' are:

   a) Weather conditions- there are more accidents in 'Fine no high winds' condition than other weather conditions.

   b) Light conditions- there are more accidents in 'Day light' condition than other light conditions.

   c) Road surface conditions- maximum number of accidents have happened on 'Dry' road surfaces' than other road surface conditions.

6. Road junctions, pedestrian crossings, roads with dark lighting conditions and roads with frost and ice are the points where 'fatal accidents' are happening.

6. Recommendations

1. Police forces like Metropolitan Police, West Midlands, Kent and others where there are a greater number of accidents must get necessary instructions, training and counselling from police forces who are able to contain number of accidents well such as Dumfries and Galloway, and Fife.

2. Social Messages must be put across UK that a greater number of fatal accidents are happening during day time and people must be made aware and cautious about it.

3. As a greater number of accidents are happening on Single carriageway, increased vigilance, more checks, and signboards must be taken into consideration by concerned authorities.

4. Also, as a greater number of fatal accidents are happening at Junctions, and Pedestrian crossings. Therefore, while issuing driving license, drivers must be made aware of these, and necessary training must be provided to them. Word must be spread to pedestrians through media campaigns and other relevant means of communication.

7. References

Batrakova, A. and Gredasova, O. (2016) 'Influence of Road Conditions on Traffic Safety', Procedia Engineering, 134, pp. 196–204.
doi:10.1016/j.proeng.2016.01.060

Wienclaw, R. A. (2021) 'Descriptive Statistics (Business)', Salem Press Encyclopedia. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=89163648&site=eds-live (Accessed: 11 December 2021).

Ruth, M. (2019) 'Student's t-test', Salem Press Encyclopedia. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=87325028&site=eds-live (Accessed: 11 December 2021).

Turhan, N. S. (2020) 'Karl Pearson's Chi-Square Tests', Educational Research and Reviews, 16(9), pp. 575–580. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1267545&site=eds-live (Accessed: 11 December 2021).

Holt, M. P. . Ms. (2020) 'Analysis of variance (ANOVA)', Salem Press Encyclopedia of Science. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=113931283&site=eds-live (Accessed: 11 December 2021).

Tantawi, R., PhD (2021) 'Linear regression', Research Starter Topics. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=90558376&site=eds-live (Accessed: 11 December 2021).

Annette J. Dobson and Adrian G. Barnett (2018) An Introduction to Generalized Linear Models. Boca Raton: Chapman and Hall/CRC (Texts in Statistical Science). Available at: https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1867460&site=eds-live (Accessed: 11 December 2021).

8. Appendices

```
install.packages("ggplot2")
install.packages("tidyverse")
install.packages("sqldf")
install.packages("gsubfn")
install.packages("proto")
install.packages("RSQLite")
install.packages("caTools")
install.packages("xtable")
install.packages("DescTools")
install.packages("Hmisc")
install.packages("stringi")
install.packages("corrplot")
library(corrplot)
library(Hmisc)
library(stringi)
library(DescTools)
library(sqldf)
library(gsubfn)
library(proto)
library(RSQLite)
library(ggplot2)
library(tidyverse)
library(dplyr)
library(caTools)
library(readr)

#read the data
accidents=read.csv("/Users/vivian/Desktop/研究所/Foundation of Business Analytics and Management Science/assignment/Accidents.csv",header=T)

#1)
#get column(there is no NA for this two column)
data1=accidents[,c(5,6)]
#change the column name
data1$Accident_Severity <- ifelse(data1$Accident_Severity == '1', 'Fatal', data1$Accident_Severity)
```

```r
data1$Accident_Severity <- ifelse(data1$Accident_Severity == '2', 'Serious', data1$Accident_Severity)
data1$Accident_Severity <- ifelse(data1$Accident_Severity == '3', 'Slight', data1$Accident_Severity)
#plot
data1 <- data1 %>%
  mutate(Accident_Severity=as.factor(Accident_Severity))
plot1 <- ggplot(data=data1)+
  geom_bar(mapping=aes(x=Police_Force,  fill=Accident_Severity))
plot1 + scale_y_continuous("The number of accidents")+
  labs(title = bquote("Bar chart"))+
  theme(plot.title = element_text(hjust = 0.5))
#table
df1=ftable(data1, data1$Accident_Severity, data1$Police_Force)
#AnovaTest
anova1 = aov(data1$Police_Force ~ data1$Accident_Severity, data=data1)
summary(anova1)
#impoet table
write.ftable(df1,"/Users/vivian/Desktop/研究所/Foundation of Business Analytics and Management Science/assignment/ftable.csv",row.names=T,sep=",")

#2)
#get column
df_week <- subset(accidents,Day_of_Week>0, select=(Day_of_Week), drop=FALSE)
df_week
(df_weekday<-subset(df_week,Day_of_Week>1 & Day_of_Week<7,select=(Day_of_Week),drop=FALSE))
(df_weekday$Accident_count<-1)
df_weekday
(df_weekend<-subset(df_week,Day_of_Week==1 | Day_of_Week==7,select=(Day_of_Week),drop=FALSE))
(df_weekend$Accident_count<-1)
df_weekend
#calculating
(weekday_row_count <- nrow(df_weekday))
weekday_row_count_mean<-as.integer(weekday_row_count/5)
weekday_row_count_mean
#Number of rows is number of accidents
```

```r
weekend_row_count <- nrow(df_weekend)
weekend_row_count
weekend_row_count_mean= as.integer(weekend_row_count/2)
weekend_row_count_mean
#plot
vec<- append(weekday_row_count_mean,weekend_row_count_mean)
vec
M<- c('Accidents in Weekdays', 'Accidents in Weekends')
barplot(vec,names.arg =M,ylim = c(0,20000), col='blue')
df_weekday_group<- df_weekday%>%
  group_by(Day_of_Week)%>%
  summarise(Accident_Group=sum(Accident_count))
df_weekday_group
df_weekend_group<- df_weekend%>%
  group_by(Day_of_Week)%>%
  summarise(Accident_Group=sum(Accident_count))
df_weekend_group
#t-test
t.test(df_weekday_group$Accident_Group, df_weekend_group$Accident_Group)

#3)
#gets column
data6 <- accidents[,c(6,11)]
#change name of column Accident_Severity(fatal)
data6$Accident_Severity <- ifelse(data6$Accident_Severity == '1', 'Fatal',
data6$Accident_Severity)
#only take the fatal data
data7 <- filter(data6,Accident_Severity %in% c("Fatal"))
fatal_accidents_time <- data7[order(data7$Time, decreasing = F),]
#separate the time in to hr and min
hr_min = str_split_fixed(fatal_accidents_time$Time, ":", 2)
#merge the data with only hr(column)
new_fatal_accidents_time <- cbind(fatal_accidents_time, hr_min[,c(1)])
 #change the columns
 colnames(new_fatal_accidents_time) <- c("Accident_Severity","Time","Hour")
 #set the rules: day(6am-17pm),night(18pm-0am & 1am-5am)
 day_night_time <- ifelse(as.integer(new_fatal_accidents_time$Hour) >= 6 &
as.integer(new_fatal_accidents_time$Hour) <= 17, "day","night")
```

```r
#new data frame
new_fatal_accidents_time <- cbind(new_fatal_accidents_time, day_night_time)
new_fatal_accidents_time1 <- new_fatal_accidents_time[,c(1,4)]
#table
fatal_num <- table(new_fatal_accidents_time1)
#1=day 2=night
data_mod <- data_frame(fatal_num = as.vector(fatal_num), day_night= c(1,2))
#plot
fatal_num_polt <- ggplot(data= new_fatal_accidents_time1 )+geom_bar(mapping= aes(x= day_night_time))
fatal_num_polt + scale_y_continuous("The number of accidents")
#Chi-square test
chisq.test(fatal_num)
#import table
write.table(fatal_num,"/Users/vivian/Desktop/研究所/Foundation of Business Analytics and Management Science/assignment/fatal_num.csv",row.names=T,sep=",")


#4)
#select the column to use
data2=accidents[,c(6,15)]
#clear data
data2[data2==-1]<-NA
data2=na.omit(data2)
#convert data
data2$Accident_Severity <- ifelse(data2$Accident_Severity == '1', 'Fatal', data2$Accident_Severity)
data2$Accident_Severity <- ifelse(data2$Accident_Severity == '2', 'Serious', data2$Accident_Severity)
data2$Accident_Severity <- ifelse(data2$Accident_Severity == '3', 'Slight', data2$Accident_Severity)
data2$Road_Type <- ifelse(data2$Road_Type == '1', 'Roundabout', data2$Road_Type)
data2$Road_Type <- ifelse(data2$Road_Type == '2', 'One way street', data2$Road_Type)
data2$Road_Type <- ifelse(data2$Road_Type == '3', 'Dual carriageway', data2$Road_Type)
data2$Road_Type <- ifelse(data2$Road_Type == '6', 'Single carriageway', data2$Road_Type)
data2$Road_Type <- ifelse(data2$Road_Type == '7', 'Slip road', data2$Road_Type)
data2$Road_Type <- ifelse(data2$Road_Type == '9', 'Unknown', data2$Road_Type)
```

```r
data2$Road_Type <- ifelse(data2$Road_Type == '12', 'One way street/Slip road', data2$Road_Type)
#make a histogram
data2 <- data2 %>%
  mutate(Accident_Severity=as.factor(Accident_Severity))
plot2 <- ggplot(data=data2)+
  geom_bar(mapping=aes(x=Road_Type,  fill=Accident_Severity))
plot2 + scale_y_continuous("The number of accidents")+
  labs(title = bquote("The number of accidents per road type"))+
  theme(plot.title = element_text(hjust = 0.5))
#make a table
data2$accident_Severity<-recode(data2$Accident_Severity, "1"="Fatal","2"="Serious","3"="Slight")
data2$road_Type<-recode(data2$Road_Type, "1"="Roundabout","2"="One way street","3"="Dual carriageway","6"="Single carriageway","7"="Slip road","9"="Unknown")
data3=data2[,c(3,4)]
table(data3$road_Type,data3$accident_Severity)
#5)
#get the column
df_accident_factors<- subset(accidents, select=c(Day_of_Week, Road_Class_1st, Road_Type,Speed_limit, Junction_Detail, Junction_Control, Road_Class_2nd, 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))
df_accident_factors
#clean the data
df_accident_factors[df_accident_factors<0] <-NA
df_accident_factors2<-df_accident_factors
df_accident_factors_clean <-na.omit(df_accident_factors)
df_accident_factors_clean
#correlation_matrix
(accident_factors_matrix<- as.matrix(df_accident_factors_clean))
accident_correlation_matrix<-cor(df_accident_factors_clean)
round(accident_correlation_matrix,2)
```

```

corrplot(accident_correlation_matrix, method = "square" )

#each variable's grouping with total number of accidents and significant relation with number of Accidents#
df_accident_factors_clean$Accident_count<-1

df_road_type_group <- df_accident_factors_clean%>%
  group_by(Road_Type)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_road_type<- lm(No_of_Accidents_grouped ~ Road_Type, data=df_road_type_group)
summary(lm_road_type)

df_day_of_week_group<-df_accident_factors_clean%>%
  group_by(Day_of_Week)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_day_of_week<- lm(No_of_Accidents_grouped ~ Day_of_Week, data=df_day_of_week_group)
summary(lm_day_of_week)

df_road_surface_group<-df_accident_factors_clean%>%
  group_by(Road_Surface_Conditions)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_road_surface_conditions<- lm(No_of_Accidents_grouped ~ Road_Surface_Conditions, data=df_road_surface_group)
summary(lm_road_surface_conditions)

df_weather_group<-df_accident_factors_clean%>%
  group_by(Weather_Conditions)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_weather_conditions<- lm(No_of_Accidents_grouped ~ Weather_Conditions, data=df_weather_group)
summary(lm_weather_conditions)

df_urban_rural_group<-df_accident_factors_clean%>%
```

```r
  group_by(Urban_or_Rural_Area)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_urban_rural<- lm(No_of_Accidents_grouped ~ Urban_or_Rural_Area, data=df_urban_rural_group)
summary(lm_urban_rural)

df_pedestrian_physical_group<-df_accident_factors_clean%>%
  group_by(Pedestrian_Crossing_Physical_Facilities)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_pedestrian_physical_group<- lm(No_of_Accidents_grouped ~ Pedestrian_Crossing_Physical_Facilities, data=df_pedestrian_physical_group)
summary(lm_pedestrian_physical_group)

df_Road_Class_1st_group<-df_accident_factors_clean%>%
  group_by(Road_Class_1st)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_road_class_1st<- lm(No_of_Accidents_grouped ~ Road_Class_1st, data=df_Road_Class_1st_group)
summary(lm_road_class_1st)

df_speed_limit_group<-df_accident_factors_clean%>%
  group_by(Speed_limit)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_speed_limit<- lm(No_of_Accidents_grouped ~ Speed_limit, data=df_speed_limit_group)
summary(lm_speed_limit)

df_junction_detail_group<-df_accident_factors_clean%>%
  group_by(Junction_Detail)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_junction_detail<- lm(No_of_Accidents_grouped ~ Junction_Detail, data=df_junction_detail_group)
summary(lm_junction_detail)

```

```
df_junction_control_group<-df_accident_factors_clean%>%
  group_by(Junction_Control)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_junction_control<- lm(No_of_Accidents_grouped ~ Junction_Control,
data=df_junction_control_group)
summary(lm_junction_control)

df_road_class2_group<-df_accident_factors_clean%>%
  group_by(Road_Class_2nd)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_road_class2<- lm(No_of_Accidents_grouped ~ Road_Class_2nd,
data=df_road_class2_group)
summary(lm_road_class2)

df_ped_human_group<-df_accident_factors_clean%>%
  group_by(Pedestrian_Crossing_Human_Control)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_ped_human_group<- lm(No_of_Accidents_grouped ~ Pedestrian_Crossing_Human_Control,
data=df_ped_human_group)
summary(lm_ped_human_group)

df_special_condition_group<-df_accident_factors_clean%>%
  group_by(Special_Conditions_at_Site)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_special_conditions<- lm(No_of_Accidents_grouped ~ Special_Conditions_at_Site,
data=df_special_condition_group)
summary(lm_special_conditions)

df_light_conditions_group<-df_accident_factors_clean%>%
  group_by(Light_Conditions)%>%
  summarise('No_of_Accidents_grouped'=sum(Accident_count))

lm_light_conditions<- lm(No_of_Accidents_grouped ~ Light_Conditions,
data=df_light_conditions_group)
```

```
summary(lm_light_conditions)

#plotting_light_conditions#
df_light_conditions_group_graph<-df_light_conditions_group%>%
  mutate(Light_Conditions_Name=case_when(
    (Light_Conditions==1) ~ "Daylight",
    (Light_Conditions==4) ~ "Darkness-lights lit",
    (Light_Conditions==5) ~ "Darkness-lights unlit",
    (Light_Conditions==6) ~ "Darkness-No lighting",
    (Light_Conditions==7) ~ "Darkness- Lighting Unknown"
  ))

install.packages("plotrix")
library(plotrix)

slices<-(df_light_conditions_group_graph$No_of_Accidents_grouped)
label=df_light_conditions_group_graph$No_of_Accidents_grouped
pie3D(slices, labels=label,explode=0.2,radius=1,
      main= "Light Condition's imapct on Number of Accidents")
legend("topright", c("Daylight", "Darkness-lights lit","Darkness-lights unlit","Darkness-No lighting","Darkness- Lighting Unknown"),cex=0.5, fill=rainbow(length(df_light_conditions_group_graph$Light_Conditions_Name)))

ggplot(data = df_light_conditions_group_graph,aes(x=Light_Conditions_Name, y=No_of_Accidents_grouped))+
  geom_bar(stat="identity",fill="blue")+
  #geom_smooth(method = "loess")
  geom_text(aes(label=No_of_Accidents_grouped),position = position_stack(vjust=1.5))

#plotting_Weather_Conditions#
df_weather_group_graph<-df_weather_group%>%
  mutate(Weather_Conditions_Name=case_when(
    (Weather_Conditions==1) ~ "Fine no high winds",
    (Weather_Conditions==2) ~ "Raining no high winds",
    (Weather_Conditions==3) ~ "Snowing no high winds",
    (Weather_Conditions==4) ~ "Fine+ high winds",
    (Weather_Conditions==5) ~ "Raining + high winds",
    (Weather_Conditions==6) ~ "Snowing+ high winds",
```

```
    (Weather_Conditions==7) ~ "Fog or mist",
    (Weather_Conditions==8) ~ "Other",
    (Weather_Conditions==9) ~ "Unknown",
  ))


ggplot(data = df_weather_group_graph,aes(Weather_Conditions_Name,No_of_Accidents_grouped))+
  geom_point(shape=19, colour="Green", size=6)+
  geom_smooth(method = "loess")+
  geom_text(aes(label=No_of_Accidents_grouped))

ggplot(df_weather_group_graph, aes(x=Weather_Conditions_Name, y=No_of_Accidents_grouped))+
  geom_bar(stat='identity', colour='Blue')+
  geom_text(aes(label=No_of_Accidents_grouped),position = position_stack(vjust=1.0))



##plotting_road_surfaces#
df_road_surface_group_graph<-df_road_surface_group%>%
  mutate(Road_Surface_Conditions_Name=case_when(
    (Road_Surface_Conditions==1) ~ "Dry",
    (Road_Surface_Conditions==2) ~ "Wet or Damp",
    (Road_Surface_Conditions==3) ~ "Snow",
    (Road_Surface_Conditions==4) ~ "Frost or Ice",
    (Road_Surface_Conditions==5) ~ "Flood over 3cm. deep",
    (Road_Surface_Conditions==6) ~ "Oil or Diesel",
    (Road_Surface_Conditions==7) ~ "Mud"
  ))

ggplot(df_road_surface_group_graph, aes(x=Road_Surface_Conditions_Name, y=No_of_Accidents_grouped))+
  geom_bar(stat='identity', colour="blue")+
  geom_text(aes(label=No_of_Accidents_grouped),position = position_stack(vjust=1.0))


#6)
```

```
#get the column
df_accident_severity<- subset(accidents, select=c(Day_of_Week,
  Accident_Severity,Road_Class_1st, Road_Type,Speed_limit, Junction_Detail,
  Junction_Control, Road_Class_2nd,
  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))
df_accident_severity
#clean the data
df_accident_severity[df_accident_severity<0] <-NA
df_accident_severity_clean <-na.omit(df_accident_severity)
df_accident_severity_clean
#Making all categorical variables as factors to run multinomial logistic regression#
df_accident_severity_clean$Accident_Severity<-
  as.factor(df_accident_severity_clean$Accident_Severity)
df_accident_severity_clean
#convert variable into factor
df_accident_severity_clean$Road_Type<-as.factor(df_accident_severity_clean$Road_Type)
df_accident_severity_clean$Day_of_Week<-
  as.factor(df_accident_severity_clean$Day_of_Week)
df_accident_severity_clean$Road_Class_1st<-
  as.factor(df_accident_severity_clean$Road_Class_1st)
df_accident_severity_clean$Road_Class_2nd<-
  as.factor(df_accident_severity_clean$Road_Class_2nd)
df_accident_severity_clean$Junction_Control<-
  as.factor(df_accident_severity_clean$Junction_Control)
df_accident_severity_clean$Junction_Detail<-
  as.factor(df_accident_severity_clean$Junction_Detail)
df_accident_severity_clean$Pedestrian_Crossing_Human_Control<-
  as.factor(df_accident_severity_clean$Pedestrian_Crossing_Human_Control)
df_accident_severity_clean$Pedestrian_Crossing_Physical_Facilities<-
  as.factor(df_accident_severity_clean$Pedestrian_Crossing_Physical_Facilities)
df_accident_severity_clean$Light_Conditions<-
  as.factor(df_accident_severity_clean$Light_Conditions)
df_accident_severity_clean$Weather_Conditions<-
  as.factor(df_accident_severity_clean$Weather_Conditions)
df_accident_severity_clean$Special_Conditions_at_Site<-
  as.factor(df_accident_severity_clean$Special_Conditions_at_Site)
```

```r
df_accident_severity_clean$Road_Surface_Conditions<-as.factor(df_accident_severity_clean$Road_Surface_Conditions)
df_accident_severity_clean$Carriageway_Hazards<-as.factor(df_accident_severity_clean$Carriageway_Hazards)
df_accident_severity_clean$Urban_or_Rural_Area<-as.factor(df_accident_severity_clean$Urban_or_Rural_Area)
str(df_accident_severity_clean)
#multinomial logistic regression
library(nnet)
df_accident_severity_clean$Accident_Severity<-relevel(df_accident_severity_clean$Accident_Severity, ref = '1')
acc_severity_model<- multinom(Accident_Severity~.,data=df_accident_severity_clean)
summary(acc_severity_model)
#2-tailed z-test
z_test_acc_severity<-summary(acc_severity_model)$coefficients/summary(acc_severity_model)$standard.errors
z_test_acc_severity
p_value_acc_severity<- (1-pnorm(abs(z_test_acc_severity),0,1))*2
p_value_acc_severity
#to calculate exponential values of coefficients calculated from model. Exponential values will be used to find odd ratio of each variable with levels of accident severity
exp(coefficients(acc_severity_model))

```