Analysis of factors contributing to the crashes occurred between 2011 and 2020.

Kusum Mohan Warad

Department of Engineering, San Jose State University

ISE 201: Math Foundations for Decision and Data Sciences

Dr. Shilpa Gupta

December 1, 2022

**Table of Contents**

[Exploratory Data Analysis 4](#_Toc120814756)

[Hypothesis Formulation 18](#_Toc120814757)

[Regression Analysis 22](#_Toc120814758)

[Logistic regression analysis model 1 29](#_Toc120814759)

[Logistic regression analysis model 2 34](#_Toc120814760)

[Conclusion 39](#_Toc120814761)

[References 39](#_Toc120814762)

#### 

#### **Introduction**

This dataset contains statistics from crashes that happened between 2011 and 2020. The weather, road conditions, and time of day at which the incident happened are all addressed. This information may be used to analyze the severity of the accident and to provide methods for preventing future mishaps.

I want to analyse the following from the dataset: a) How many pedestrians were involved and what was the severity of their injuries? b) The presence of street lights and traffic controls at the time of the incident c) What were the weather conditions like, and what can be done to improve the safety of both pedestrians and car drivers? d) How many drivers are involved in a crash as a result of a violation? e) From 2011 through 2020, how many incidents occurred each year?

#### **Source**

This data was obtained from the **City of San Jose’s** website and was compiled from government records for public domain information.

#### **Link**

<https://data.sanjoseca.gov/dataset/crashes-data/resource/c19a01f2-33e1-4c66-9498-85d489f90da4>

#### **Context**

There are **56044** observations in this dataset.

This is an **observational** study. The observations incorporate crash level information such as weather and road conditions, as well as the hour of day. The involved party (the vehicle), principal collision factor, fatality rate, severe, moderate, and minor injuries per crash are all provided. The units are mostly numeric indicates the total number of injuries, hour of day, whether a pedestrian was involved, condition of road (in terms of wet, dry, snowy,etc), weather (clear,rain, cloudy,etc)

There are **25 variables** that contain relevant information required for the above analysis.

I will be studying the CrashFactId, Minor, Moderate, Severe and Fatal injuries, CrashDateTime, Roadway Surface (wet or dry), Roadway Condition (presence of any obstructions, flooded, repair zone, etc), Lighting, Primary Collision Factor, Traffic Control, CollisionType, Weather and Vehicle Involved With.

## Exploratory Data Analysis

getwd()

## [1] "/Users/kusumwarad/Downloads/201 project"

setwd("~/Downloads/201 project")  
dataset <- read.csv("crashdata.csv", header = TRUE, sep = ",")  
head(dataset)

## CrashFactId Name MinorInjuries ModerateInjuries SevereInjuries  
## 1 591079 CR-0000071607 0 0 0  
## 2 591080 CR-0000071780 0 0 0  
## 3 591081 CR-0000060418 0 0 0  
## 4 591082 CR-0000060410 0 1 0  
## 5 591083 CR-0000060514 2 0 0  
## 6 591084 CR-0000072455 0 0 0  
## FatalInjuries TcrNumber CityDamageFlag ShortFormFlag Distance  
## 1 0 18-073-0962 TRUE FALSE 228  
## 2 0 18-060-0123 TRUE FALSE 148  
## 3 0 16-033-0204 FALSE FALSE 1583  
## 4 0 16-041-0882 FALSE FALSE 295  
## 5 0 16-063-0761 FALSE FALSE 0  
## 6 0 18-126-0676 TRUE FALSE 54  
## CrashDateTime PedestrianAction RoadwaySurface  
## 1 3/14/2018 23:17 No Pedestrians Involved Wet  
## 2 3/1/2018 7:30 No Pedestrians Involved Wet  
## 3 2/2/2016 9:02 No Pedestrians Involved Wet  
## 4 2/10/2016 20:33 No Pedestrians Involved Dry  
## 5 3/3/2016 19:04 Crossing - Not In Crosswalk Dry  
## 6 5/6/2018 18:58 No Pedestrians Involved Dry  
## RoadwayCondition Lighting PrimaryCollisionFactor  
## 1 No Unusual Conditions Dark - Street Light Violation Driver 1  
## 2 No Unusual Conditions Daylight Violation Driver 1  
## 3 No Unusual Conditions Daylight Violation Driver 1  
## 4 No Unusual Conditions Dark - Street Light Violation Driver 1  
## 5 No Unusual Conditions Dark - Street Light Violation Driver 1  
## 6 No Unusual Conditions Daylight Violation Driver 1  
## TrafficControl Weather CollisionType ProximityToIntersection  
## 1 No Controls Present/Factor Clear Hit Object Non-Related  
## 2 No Controls Present/Factor Rain Hit Object Non-Related  
## 3 No Controls Present/Factor Rain Overturned Non-Related  
## 4 No Controls Present/Factor Clear Head On Non-Related  
## 5 No Controls Present/Factor Cloudy Vehicle/Pedestrian Intersection  
## 6 No Controls Present/Factor Clear Hit Object Non-Related  
## VehicleInvolvedWith PedestrianDirectionFrom PedestrianDirectionTo  
## 1 Fixed Object Not Applicable Not Applicable  
## 2 Fixed Object Not Applicable Not Applicable  
## 3 Fixed Object Not Applicable Not Applicable  
## 4 Fixed Object Not Applicable Not Applicable  
## 5 Pedestrian South North  
## 6 Fixed Object Not Applicable Not Applicable  
## DirectionFromIntersection Comment  
## 1 East Of V1 HIT CEMENT GUARD RAIL  
## 2 West Of v1 hit city pole  
## 3 South Of NULL  
## 4 East Of NULL  
## 5 At NULL  
## 6 West Of V1 HIT NO PARKING POLE

#### **The columns required for the analysis**

dataset\_n <- subset(dataset, select = -c(Name,TcrNumber,Distance,CityDamageFlag,ShortFormFlag, PedestrianAction, ProximityToIntersection,PedestrianDirectionFrom,PedestrianDirectionTo,DirectionFromIntersection,Comment))  
head(dataset\_n)

## CrashFactId MinorInjuries ModerateInjuries SevereInjuries FatalInjuries  
## 1 591079 0 0 0 0  
## 2 591080 0 0 0 0  
## 3 591081 0 0 0 0  
## 4 591082 0 1 0 0  
## 5 591083 2 0 0 0  
## 6 591084 0 0 0 0  
## CrashDateTime RoadwaySurface RoadwayCondition Lighting  
## 1 3/14/2018 23:17 Wet No Unusual Conditions Dark - Street Light  
## 2 3/1/2018 7:30 Wet No Unusual Conditions Daylight  
## 3 2/2/2016 9:02 Wet No Unusual Conditions Daylight  
## 4 2/10/2016 20:33 Dry No Unusual Conditions Dark - Street Light  
## 5 3/3/2016 19:04 Dry No Unusual Conditions Dark - Street Light  
## 6 5/6/2018 18:58 Dry No Unusual Conditions Daylight  
## PrimaryCollisionFactor TrafficControl Weather CollisionType  
## 1 Violation Driver 1 No Controls Present/Factor Clear Hit Object  
## 2 Violation Driver 1 No Controls Present/Factor Rain Hit Object  
## 3 Violation Driver 1 No Controls Present/Factor Rain Overturned  
## 4 Violation Driver 1 No Controls Present/Factor Clear Head On  
## 5 Violation Driver 1 No Controls Present/Factor Cloudy Vehicle/Pedestrian  
## 6 Violation Driver 1 No Controls Present/Factor Clear Hit Object  
## VehicleInvolvedWith  
## 1 Fixed Object  
## 2 Fixed Object  
## 3 Fixed Object  
## 4 Fixed Object  
## 5 Pedestrian  
## 6 Fixed Object

In the above code chunk, I have removed the columns that are not required for my analysis, therefore making the dataset compact.

#### **Data quality**

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(dplyr)  
glimpse(dataset\_n)

## Rows: 56,044  
## Columns: 14  
## $ CrashFactId <int> 591079, 591080, 591081, 591082, 591083, 591084,…  
## $ MinorInjuries <int> 0, 0, 0, 0, 2, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0,…  
## $ ModerateInjuries <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,…  
## $ SevereInjuries <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…  
## $ FatalInjuries <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…  
## $ CrashDateTime <chr> "3/14/2018 23:17", "3/1/2018 7:30", "2/2/2016 9…  
## $ RoadwaySurface <chr> "Wet", "Wet", "Wet", "Dry", "Dry", "Dry", "Dry"…  
## $ RoadwayCondition <chr> "No Unusual Conditions", "No Unusual Conditions…  
## $ Lighting <chr> "Dark - Street Light", "Daylight", "Daylight", …  
## $ PrimaryCollisionFactor <chr> "Violation Driver 1", "Violation Driver 1", "Vi…  
## $ TrafficControl <chr> "No Controls Present/Factor", "No Controls Pres…  
## $ Weather <chr> "Clear", "Rain", "Rain", "Clear", "Cloudy", "Cl…  
## $ CollisionType <chr> "Hit Object", "Hit Object", "Overturned", "Head…  
## $ VehicleInvolvedWith <chr> "Fixed Object", "Fixed Object", "Fixed Object",…

z1 <- dataset\_n  
sum(is.na(z1))

## [1] 0

colSums(is.na(z1))

## CrashFactId MinorInjuries ModerateInjuries   
## 0 0 0   
## SevereInjuries FatalInjuries CrashDateTime   
## 0 0 0   
## RoadwaySurface RoadwayCondition Lighting   
## 0 0 0   
## PrimaryCollisionFactor TrafficControl Weather   
## 0 0 0   
## CollisionType VehicleInvolvedWith   
## 0 0

sum(duplicated(z1))

## [1] 0

is.ts(z1)

## [1] FALSE

The modified dataset has 0 missing values, 0 duplicated rows, and is not a time series one.

#### **Transforming variables**

z1$Lighting = factor(z1$Lighting, level = c("Daylight", "Dark - Street Light", "Dark - No Street Light","Dusk - Dawn", "Dark - Street Light Not Functioning", "Unknown"),  
 labels = c(1,2,3,4,5,6))  
head(z1)

## CrashFactId MinorInjuries ModerateInjuries SevereInjuries FatalInjuries  
## 1 591079 0 0 0 0  
## 2 591080 0 0 0 0  
## 3 591081 0 0 0 0  
## 4 591082 0 1 0 0  
## 5 591083 2 0 0 0  
## 6 591084 0 0 0 0  
## CrashDateTime RoadwaySurface RoadwayCondition Lighting  
## 1 3/14/2018 23:17 Wet No Unusual Conditions 2  
## 2 3/1/2018 7:30 Wet No Unusual Conditions 1  
## 3 2/2/2016 9:02 Wet No Unusual Conditions 1  
## 4 2/10/2016 20:33 Dry No Unusual Conditions 2  
## 5 3/3/2016 19:04 Dry No Unusual Conditions 2  
## 6 5/6/2018 18:58 Dry No Unusual Conditions 1  
## PrimaryCollisionFactor TrafficControl Weather CollisionType  
## 1 Violation Driver 1 No Controls Present/Factor Clear Hit Object  
## 2 Violation Driver 1 No Controls Present/Factor Rain Hit Object  
## 3 Violation Driver 1 No Controls Present/Factor Rain Overturned  
## 4 Violation Driver 1 No Controls Present/Factor Clear Head On  
## 5 Violation Driver 1 No Controls Present/Factor Cloudy Vehicle/Pedestrian  
## 6 Violation Driver 1 No Controls Present/Factor Clear Hit Object  
## VehicleInvolvedWith  
## 1 Fixed Object  
## 2 Fixed Object  
## 3 Fixed Object  
## 4 Fixed Object  
## 5 Pedestrian  
## 6 Fixed Object

z1$CrashDateTime <- format(as.Date(z1$CrashDateTime, format="%m/%d/%Y %H:%M"),"%Y")  
head(z1)

## CrashFactId MinorInjuries ModerateInjuries SevereInjuries FatalInjuries  
## 1 591079 0 0 0 0  
## 2 591080 0 0 0 0  
## 3 591081 0 0 0 0  
## 4 591082 0 1 0 0  
## 5 591083 2 0 0 0  
## 6 591084 0 0 0 0  
## CrashDateTime RoadwaySurface RoadwayCondition Lighting  
## 1 2018 Wet No Unusual Conditions 2  
## 2 2018 Wet No Unusual Conditions 1  
## 3 2016 Wet No Unusual Conditions 1  
## 4 2016 Dry No Unusual Conditions 2  
## 5 2016 Dry No Unusual Conditions 2  
## 6 2018 Dry No Unusual Conditions 1  
## PrimaryCollisionFactor TrafficControl Weather CollisionType  
## 1 Violation Driver 1 No Controls Present/Factor Clear Hit Object  
## 2 Violation Driver 1 No Controls Present/Factor Rain Hit Object  
## 3 Violation Driver 1 No Controls Present/Factor Rain Overturned  
## 4 Violation Driver 1 No Controls Present/Factor Clear Head On  
## 5 Violation Driver 1 No Controls Present/Factor Cloudy Vehicle/Pedestrian  
## 6 Violation Driver 1 No Controls Present/Factor Clear Hit Object  
## VehicleInvolvedWith  
## 1 Fixed Object  
## 2 Fixed Object  
## 3 Fixed Object  
## 4 Fixed Object  
## 5 Pedestrian  
## 6 Fixed Object

All the values of the Lighting variable have been converted to factor type from character because it makes analysis easier, where:

1. Daylight Dark - Street Light
2. Dark - Street Light
3. Dark - No Street Light
4. Dusk - Dawn
5. Dark - Street Light Not Functioning
6. Unknown

z1$TrafficControl = factor(z1$TrafficControl, level = c("No Controls Present/Factor", "Controls Functioning", "Controls Not Functioning","Controls Obscured", "Unknown"),  
 labels = c(1,2,3,4,5))  
head(z1)

## CrashFactId MinorInjuries ModerateInjuries SevereInjuries FatalInjuries  
## 1 591079 0 0 0 0  
## 2 591080 0 0 0 0  
## 3 591081 0 0 0 0  
## 4 591082 0 1 0 0  
## 5 591083 2 0 0 0  
## 6 591084 0 0 0 0  
## CrashDateTime RoadwaySurface RoadwayCondition Lighting  
## 1 2018 Wet No Unusual Conditions 2  
## 2 2018 Wet No Unusual Conditions 1  
## 3 2016 Wet No Unusual Conditions 1  
## 4 2016 Dry No Unusual Conditions 2  
## 5 2016 Dry No Unusual Conditions 2  
## 6 2018 Dry No Unusual Conditions 1  
## PrimaryCollisionFactor TrafficControl Weather CollisionType  
## 1 Violation Driver 1 1 Clear Hit Object  
## 2 Violation Driver 1 1 Rain Hit Object  
## 3 Violation Driver 1 1 Rain Overturned  
## 4 Violation Driver 1 1 Clear Head On  
## 5 Violation Driver 1 1 Cloudy Vehicle/Pedestrian  
## 6 Violation Driver 1 1 Clear Hit Object  
## VehicleInvolvedWith  
## 1 Fixed Object  
## 2 Fixed Object  
## 3 Fixed Object  
## 4 Fixed Object  
## 5 Pedestrian  
## 6 Fixed Object

All the values of the Traffic Control variable have been converted to factor type from character because it makes analysis easier, where:

1. No Controls Present/Factor
2. Controls Functioning
3. Controls Not Functioning
4. Controls Obscured
5. Unknown

dataset\_n$Lighting <- as.factor(dataset\_n$Lighting)  
summary(dataset\_n$Lighting)

## Dark - No Street Light Dark - Street Light   
## 744 18839   
## Dark - Street Light Not Functioning Daylight   
## 171 33280   
## Dusk - Dawn Unknown   
## 2392 618

z1$FatalInjuries <- as.factor(z1$FatalInjuries)  
summary(z1$FatalInjuries)

## 0 1 2 3   
## 55600 427 16 1

z1$PrimaryCollisionFactor <- as.factor(z1$PrimaryCollisionFactor)  
summary(z1$PrimaryCollisionFactor)

## Bike At Fault Fell Asleep Other Improper Driving   
## 1194 27 189   
## Other Than Driver Parked/Rolling Pedestrian At Fault   
## 228 15 873   
## Unknown Violation Driver 1 Violation Driver 2   
## 14835 38336 347

z1$VehicleInvolvedWith <- as.factor(z1$VehicleInvolvedWith)  
summary(z1$VehicleInvolvedWith)

## Animal Bike   
## 60 3089   
## Fixed Object Ice Cream Truck   
## 6065 4   
## Light Rail Vehicle Motor Vehicle On Other Roadway   
## 66 198   
## Motorcycle Non-Collision   
## 683 246   
## Other Object Other Vehicle   
## 178 30993   
## Parked Vehicle Pedestrian   
## 11457 2696   
## Scooter Motorized Scooter Non-Motorized   
## 75 15   
## Skateboard Train   
## 90 12   
## Unknown Wheelchair   
## 47 70

z1$CollisionType <- as.factor(z1$CollisionType )  
summary(z1$CollisionType)

## Broadside Head On Hit Object Other   
## 10555 4004 3447 13710   
## Overturned Rear End Sideswipe Vehicle/Bike   
## 151 13888 7545 1238   
## Vehicle/Pedestrian   
## 1506

z1$TrafficControl <- as.factor(z1$TrafficControl )  
summary(z1$TrafficControl)

## 1 2 3 4 5   
## 26354 28060 259 32 1339

z1$RoadwayCondition <- as.factor(z1$RoadwayCondition)  
summary(z1$RoadwayCondition)

## Construction - Repair Zone Flooded   
## 358 19   
## Holes Deep Rut Loose Material On Roadway   
## 83 57   
## No Unusual Conditions Obstruction On Roadway   
## 53235 81   
## Other Reduced Roadway Width   
## 205 82   
## Unknown   
## 1924

Five conclusions can be drawn from the above:

1. We can infer that there were around 22,146 crashes at night and approximately 33,280 during the day.
2. The majority of the incidents took place because of drivers violating traffic rules.
3. Most crashes have taken place due to crashing into the rear end of another vehicle.
4. A total of 26,354 accidents were caused due to the absence of traffic controls over the years.
5. 358 crashes have occurred due to construction-repair zones on roads.

#### **Summary statistics**

r\_mode <- function(x) {  
 which.max(tabulate(x))  
}  
dataset\_n %>%  
 summarise(average = mean(SevereInjuries),  
 median = median(SevereInjuries),  
 mode = r\_mode(SevereInjuries),  
 stddev = sd(SevereInjuries),  
 variance = var(SevereInjuries),  
 range = max(SevereInjuries) - min(SevereInjuries))

## average median mode stddev variance range  
## 1 0.03136821 0 1 0.1895158 0.03591626 5

Based on the output we can observe that the minimum number of severe injuries seems to be 1 and maximum is 5 in a day.

r\_mode <- function(x) {  
 which.max(tabulate(x))  
}  
dataset\_n %>%  
 summarise(average = mean(MinorInjuries),  
 median = median(MinorInjuries),  
 mode = r\_mode(MinorInjuries),  
 stddev = sd(MinorInjuries),  
 variance = var(MinorInjuries),  
 range = max(MinorInjuries) - min(MinorInjuries))

## average median mode stddev variance range  
## 1 0.3893548 0 1 0.6965736 0.4852147 8

The maximum number of people that suffered minor injuries is 8 in a day.

r\_mode <- function(x) {  
 which.max(tabulate(x))  
}  
dataset\_n %>%  
 summarise(average = mean(ModerateInjuries),  
 median = median(ModerateInjuries),  
 mode = r\_mode(ModerateInjuries),  
 stddev = sd(ModerateInjuries),  
 variance = var(ModerateInjuries),  
 range = max(ModerateInjuries) - min(ModerateInjuries))

## average median mode stddev variance range  
## 1 0.142745 0 1 0.395495 0.1564163 10

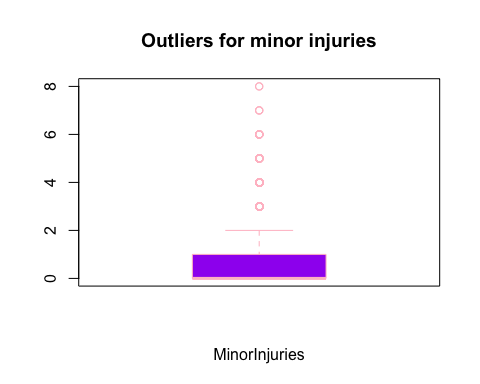
Maximum moderate injuries occurred is 10 per day.

r\_mode <- function(x) {  
 which.max(tabulate(x))  
}  
dataset\_n %>%  
 summarise(average = mean(FatalInjuries),  
 median = median(FatalInjuries),  
 mode = r\_mode(FatalInjuries),  
 stddev = sd(FatalInjuries),  
 variance = var(FatalInjuries),  
 range = max(FatalInjuries) - min(FatalInjuries))

## average median mode stddev variance range  
## 1 0.008243523 0 1 0.09409444 0.008853764 3

Highest number of fatal injuries in a day is seen to be 3.

boxplot(z1$MinorInjuries,  
 main= "Outliers for minor injuries",  
 xlab= "MinorInjuries", col="purple", border="pink")

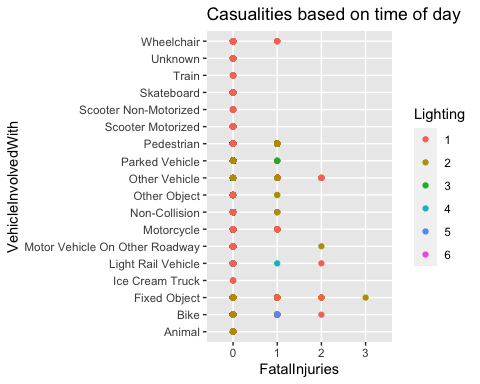


quantile(z1$MinorInjuries, probs=c(0,0.25,0.5,0.75,1))

## 0% 25% 50% 75% 100%   
## 0 0 0 1 8

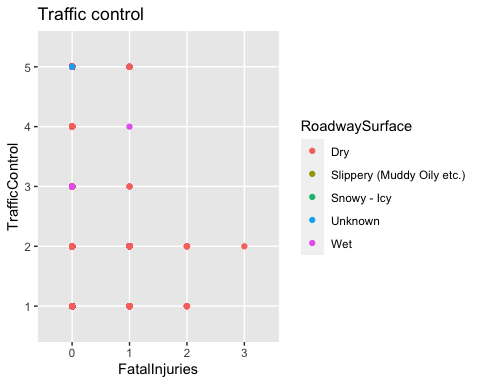
The outliers for Minor Injuries range from 3 to 8. More than 75% of the people were injured because of the accidents.

ggplot(data = z1) +  
 ggtitle("Casualities based on time of day") +  
 geom\_point(mapping = aes(x = FatalInjuries, y = VehicleInvolvedWith, color = Lighting))



It can be noted that most casualties have occurred during the day (daylight- Lighting) and also because of crashing into a bike, fixed object, motorcycle, other vehicle, and/or wheelchair. As it gets darker, we can notice the number of crashes or injuries occurring is decreasing.

ggplot(data = z1) +  
 ggtitle("Traffic control") +  
 geom\_point(mapping = aes(x = FatalInjuries, y = TrafficControl, color = RoadwaySurface))

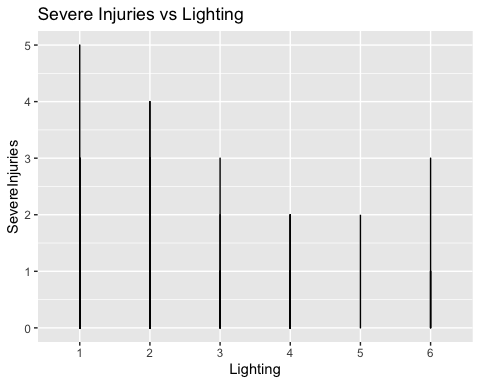


All the values of the Traffic Control variable have been converted to factor type from character because it makes analysis easier, where:

1. No Controls Present/Factor
2. Controls Functioning
3. Controls Not Functioning
4. Controls Obscured
5. Unknown

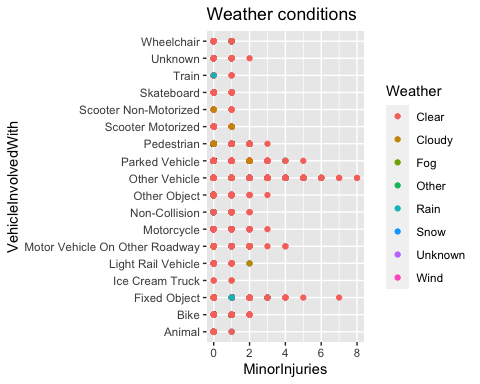
From the plot above, we can see that people involved in the crash have suffered greatly because of a lack of traffic controls and the roadway surface being dry, which should be fixed to prevent further mishaps. Also, even if the traffic controls are functioning, it can be seen that there are casualties occurring. This could be because either driver violated traffic rules.

ggplot(data = z1) +  
 ggtitle("Severe Injuries vs Lighting") +  
 geom\_line(mapping = aes(x = Lighting, y = SevereInjuries))

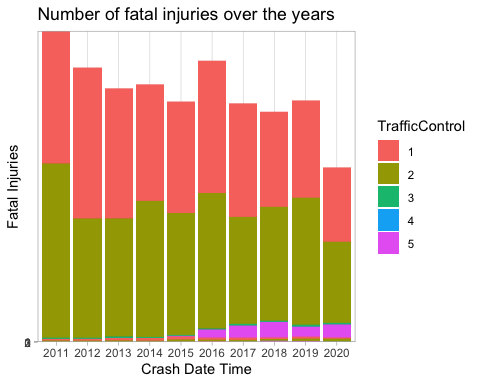


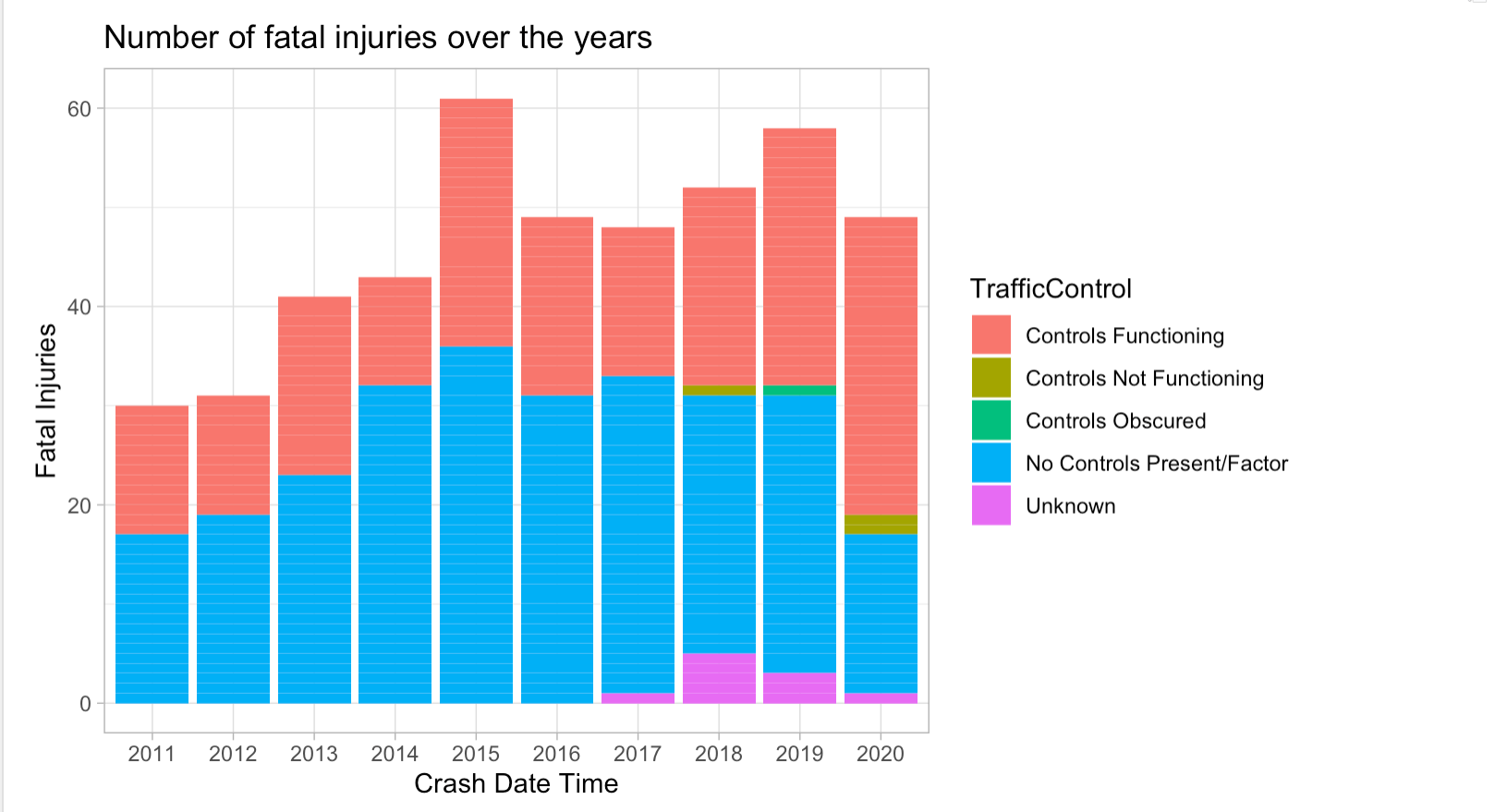
Most of the people suffered severe injuries during the night with streetlights functioning as compared to the day (4 injuries in a day), and regardless of the time of day, the accidents occurred because (could be) of violations of traffic rules.

ggplot(data = z1) +  
 ggtitle("Weather conditions") +  
 geom\_point(mapping = aes(x = MinorInjuries, y = VehicleInvolvedWith , color = Weather))

 According to the plot above, a few of pedestrians (approximately three each day) were involved in or injured as a result of the accident. Most of the accidents involve another vehicle on a clear day, lesser number of incidents have taken place due to unusual weather conditions like rain or cloudy.

z1%>%  
ggplot(aes(x=CrashDateTime,y=FatalInjuries,fill=TrafficControl))+geom\_bar(stat ="identity")+  
 theme\_light()+  
 labs(x="Crash Date Time",y="Fatal Injuries",title = "Number of fatal injuries over the years")





From the above barplot, we can infer that in 2015, there were a maximum of about 60 fatalities, and the trend seems to decrease for the next three years, but in 2019, there was a sudden increase in the number again, this time around 55. Also, we can observe that most deaths have occurred because of the absence of traffic controls.

#### **Questions for the next stage**

1. If all conditions like weather, traffic controls, and lighting were ideal, would the number of accidents be less?
2. Find out how many crashes have occurred over the years; is there a decreasing or an increasing trend?
3. Is it possible that traffic controls work better during the day than they do at night?

## Hypothesis Formulation

The following are the two kinds of statistical hypothesis testing:

1. **Null Hypothesis**: Hypothesis testing is carried out in order to determine the validity of a claim or assumption that is made about the larger population. Null Hypothesis is the claim that is associated with the trial. H0 represents the null hypothesis.
2. **Alternate Hypothesis**: the hypothesis is valid if the null hypothesis is rejected. The evidence presented in the trial consists primarily of data and the statistical computations that go with it. Denoted by H1 or Ha.

Hypothesis testing uses a p-value to assess the strength of the evidence or to indicate what the data reveals about the population. The p-value ranges between 0 and 1. It could be interpreted as: ~ It is usually reasonable to reject the null hypothesis with a small p-value (< 0.05). ~ When the p-value is large (> 0.05), there is little evidence to reject the null hypothesis.

#### A) Is it possible that traffic controls work better during the day than they do at night?

We can use the **chi-square test of independence** (as it indicates whether two variables are likely to be related or not) to check the aforementioned condition. Assuming the null hypothesis is H0 and the alternate hypothesis is H1.

two\_way=table(z1$TrafficControl,z1$Lighting)  
two\_way

##   
## 1 2 3 4 5 6  
## 1 14417 9857 522 1156 98 304  
## 2 17876 8651 196 1137 57 143  
## 3 180 45 6 14 13 1  
## 4 22 9 1 0 0 0  
## 5 785 277 19 85 3 170

prop.table(two\_way)

##   
## 1 2 3 4 5  
## 1 2.572443e-01 1.758797e-01 9.314110e-03 2.062665e-02 1.748626e-03  
## 2 3.189637e-01 1.543609e-01 3.497252e-03 2.028763e-02 1.017058e-03  
## 3 3.211762e-03 8.029405e-04 1.070587e-04 2.498037e-04 2.319606e-04  
## 4 3.925487e-04 1.605881e-04 1.784312e-05 0.000000e+00 0.000000e+00  
## 5 1.400685e-02 4.942545e-03 3.390193e-04 1.516665e-03 5.352937e-05  
##   
## 6  
## 1 5.424309e-03  
## 2 2.551567e-03  
## 3 1.784312e-05  
## 4 0.000000e+00  
## 5 3.033331e-03

prop.table(two\_way, 1)

##   
## 1 2 3 4 5 6  
## 1 0.547051681 0.374022919 0.019807240 0.043864309 0.003718601 0.011535251  
## 2 0.637063435 0.308303635 0.006985032 0.040520314 0.002031361 0.005096222  
## 3 0.694980695 0.173745174 0.023166023 0.054054054 0.050193050 0.003861004  
## 4 0.687500000 0.281250000 0.031250000 0.000000000 0.000000000 0.000000000  
## 5 0.586258402 0.206870799 0.014189694 0.063480209 0.002240478 0.126960418

prop.table(two\_way, 2)

##   
## 1 2 3 4 5  
## 1 0.4332031250 0.5232231010 0.7016129032 0.4832775920 0.5730994152  
## 2 0.5371394231 0.4592069643 0.2634408602 0.4753344482 0.3333333333  
## 3 0.0054086538 0.0023886618 0.0080645161 0.0058528428 0.0760233918  
## 4 0.0006610577 0.0004777324 0.0013440860 0.0000000000 0.0000000000  
## 5 0.0235877404 0.0147035405 0.0255376344 0.0355351171 0.0175438596  
##   
## 6  
## 1 0.4919093851  
## 2 0.2313915858  
## 3 0.0016181230  
## 4 0.0000000000  
## 5 0.2750809061

We can see that the probability of traffic controls functioning during daylight is higher than other lighting conditions (at 53.7%). The chi-square test of independence can be used to determine whether this discovery is statistically significant.

**H0** - The 2 variables Traffic Control and Lighting are independent of each other.

**H1** - The two variables are related to each other.

z <- table(z1$TrafficControl,z1$Lighting)  
head(z)

##   
## 1 2 3 4 5 6  
## 1 14417 9857 522 1156 98 304  
## 2 17876 8651 196 1137 57 143  
## 3 180 45 6 14 13 1  
## 4 22 9 1 0 0 0  
## 5 785 277 19 85 3 170

chisq.test(z1$TrafficControl,z1$Lighting)

## Warning in chisq.test(z1$TrafficControl, z1$Lighting): Chi-squared approximation  
## may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: z1$TrafficControl and z1$Lighting  
## X-squared = 2571.8, df = 20, p-value < 2.2e-16

We may reject the null hypothesis and infer that the alternate hypothesis is accurate because the p value is substantially lower than 0.05. So, it is possible that there’s a higher presence of traffic controls in the day

#### B) To test if the mean severe injury index is higher than moderate injury index

Since there’s a higher number of moderate and severe injuries reported, we can test if their means are equal. For this, we’ll be using a T-test. A T-test is an inferential statistic that is used to assess whether or not there is a significant difference in the means of two populations and how they are correlated.

µ1 = mean severe injuries index

µ2 = mean moderate injuries index

**H0** - µ1 - µ2 = 0

**H1** - µ1 - µ2 > 0

#### **Test Statistic**

Difference in sample mean of Severe injuries and Moderate injuries

#### **Reference distribution**

We can perform a test to verify if one is greater than the other. While taking a large number of samples into consideration, we assume that the populations are normally distributed and use the t test to conclude the tests. As the data assumes a normal distribution, the level of significance is assumed to be 5%.

We can use the pooled t-test to evaluate the proposition indicated in Hypothesis B

t.test(z1$SevereInjuries, z1$ModerateInjuries, alternative = "greater", paired = FALSE, var.equal = TRUE, conf.level = 0.95)

##   
## Two Sample t-test  
##   
## data: z1$SevereInjuries and z1$ModerateInjuries  
## t = -60.122, df = 112086, p-value = 1  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -0.1144239 Inf  
## sample estimates:  
## mean of x mean of y   
## 0.03136821 0.14274499

Since the p-value is 1 we need not reject the null hypothesis. Therefore, the difference between the mean of sample 1 (Severe Injuries) and sample 2 (Moderate Injuries) are equal to 0 (also the mean can be equal to each other).

In conclusion, the chi-square test of independence conducted determines whether or not two variables are connected. In hypothesis formulation A, the computation reveals that the traffic controls could be functioning more frequently during the day than at night.

Hypothesis Formulation B tells us that the mean of both samples (Severe Injuries and Moderate Injuries) are equal to each other, which means that the number of injuries of each of the two samples is centered around the same value.

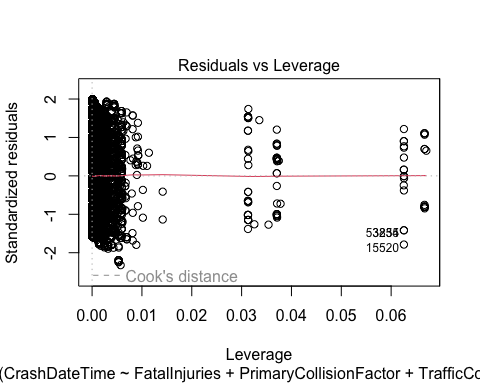
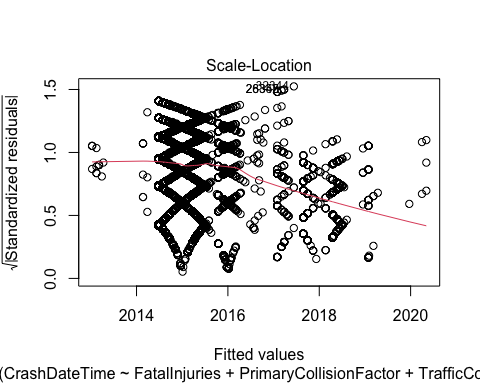
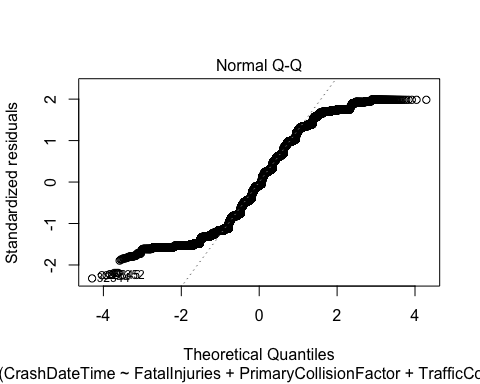
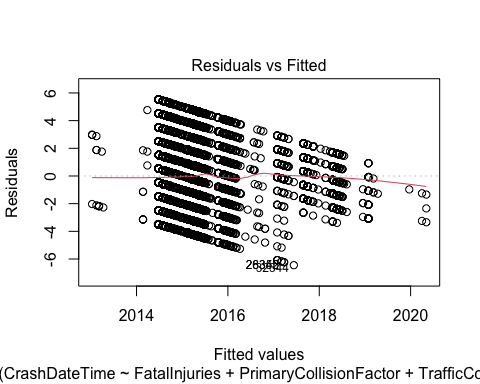
## Regression Analysis

output = lm(CrashDateTime ~ FatalInjuries + PrimaryCollisionFactor + TrafficControl + Lighting, data = z1)  
summary(output)

##   
## Call:  
## lm(formula = CrashDateTime ~ FatalInjuries + PrimaryCollisionFactor +   
## TrafficControl + Lighting, data = z1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.4438 -2.3698 -0.2192 2.3659 5.5165   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 2014.97125 0.08198 24579.755  
## FatalInjuries1 0.79774 0.13612 5.861  
## FatalInjuries2 1.44365 0.69535 2.076  
## FatalInjuries3 -0.36980 2.78093 -0.133  
## PrimaryCollisionFactorFell Asleep -0.25413 0.54129 -0.469  
## PrimaryCollisionFactorOther Improper Driving 2.10883 0.21777 9.684  
## PrimaryCollisionFactorOther Than Driver -0.16165 0.20112 -0.804  
## PrimaryCollisionFactorParked/Rolling -1.94402 0.72253 -2.691  
## PrimaryCollisionFactorPedestrian At Fault 0.24947 0.12445 2.004  
## PrimaryCollisionFactorUnknown -0.48774 0.08377 -5.822  
## PrimaryCollisionFactorViolation Driver 1 0.15298 0.08185 1.869  
## PrimaryCollisionFactorViolation Driver 2 -0.32700 0.16962 -1.928  
## TrafficControl2 0.15061 0.02401 6.272  
## TrafficControl3 0.85399 0.17403 4.907  
## TrafficControl4 -0.34155 0.49195 -0.694  
## TrafficControl5 3.16844 0.07936 39.927  
## Lighting2 0.09496 0.02556 3.715  
## Lighting3 0.66454 0.10336 6.430  
## Lighting4 0.21307 0.05891 3.617  
## Lighting5 0.20239 0.21373 0.947  
## Lighting6 0.78376 0.11487 6.823  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## FatalInjuries1 4.63e-09 \*\*\*  
## FatalInjuries2 0.037886 \*   
## FatalInjuries3 0.894211   
## PrimaryCollisionFactorFell Asleep 0.638725   
## PrimaryCollisionFactorOther Improper Driving < 2e-16 \*\*\*  
## PrimaryCollisionFactorOther Than Driver 0.421540   
## PrimaryCollisionFactorParked/Rolling 0.007135 \*\*   
## PrimaryCollisionFactorPedestrian At Fault 0.045023 \*   
## PrimaryCollisionFactorUnknown 5.83e-09 \*\*\*  
## PrimaryCollisionFactorViolation Driver 1 0.061626 .   
## PrimaryCollisionFactorViolation Driver 2 0.053891 .   
## TrafficControl2 3.59e-10 \*\*\*  
## TrafficControl3 9.26e-07 \*\*\*  
## TrafficControl4 0.487505   
## TrafficControl5 < 2e-16 \*\*\*  
## Lighting2 0.000203 \*\*\*  
## Lighting3 1.29e-10 \*\*\*  
## Lighting4 0.000299 \*\*\*  
## Lighting5 0.343670   
## Lighting6 9.01e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.781 on 56023 degrees of freedom  
## Multiple R-squared: 0.04089, Adjusted R-squared: 0.04055   
## F-statistic: 119.4 on 20 and 56023 DF, p-value: < 2.2e-16

plot(output)

## Warning: not plotting observations with leverage one:  
## 15237



It can be inferred that the year 2015 had the highest number of fatalities compared to other years. The most common occurrences are fatal injuries to one or two people per day. Three of the major reasons (Primary Collision Factors) for a mishap are: other drivers’ improper driving, pedestrians at fault, and vehicles parked. The other Primary Collision Factors do not contribute as much to the deaths of pedestrians or drivers. The most significant factors that worsen the situation and cause numerous casualties are “dusk-dawn, dark, street lights on, and dark, no street light” (lighting conditions), “traffic controls not functioning,” and “traffic controls functional.” The aforementioned variables connect in a statistically significant way to the dependent variable Crash Date Time, which implies that during that specific year (2015), those were the major factors that contributed to the casualties. We can disregard the other factors because their p-values are greater than 0.05.

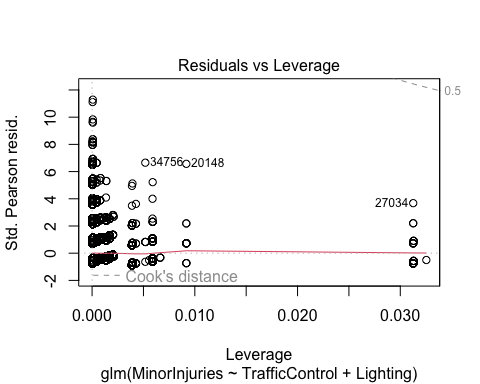
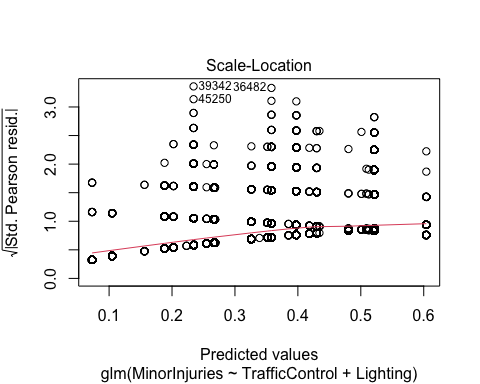
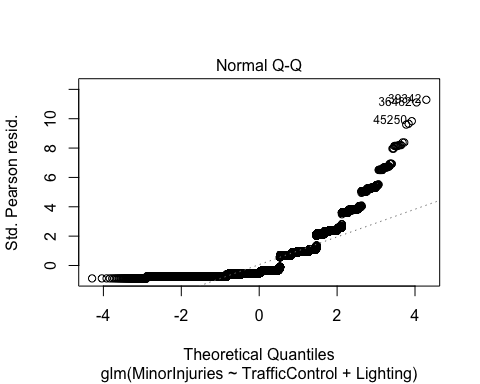
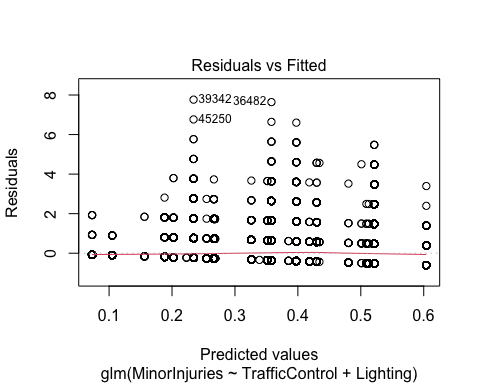
The scatter plot’s center part has the largest density of data points, which indicates that 2015 is a critical year. The linearity is violated, and there seems to be a quadratic relationship. In the normal QQ plot we notice that the residuals are somewhat normally distributed. The scale location plot suggests some non-linearity here, but what we see is that the spread of magnitudes seems to be lowest in the fitted values close to 2014, highest in the fitted values around 2015–2016, and medium around 2017–2020. This suggests heteroskedasticity.

## Logistic regression analysis model 1

model <- glm(MinorInjuries ~ TrafficControl + Lighting, data = z1)  
summary(model)

##   
## Call:  
## glm(formula = MinorInjuries ~ TrafficControl + Lighting, data = z1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.6041 -0.3975 -0.3580 0.4787 7.7658   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.357987 0.005022 71.280 < 2e-16 \*\*\*  
## TrafficControl2 0.163326 0.005933 27.530 < 2e-16 \*\*\*  
## TrafficControl3 0.246114 0.043042 5.718 1.08e-08 \*\*\*  
## TrafficControl4 0.150880 0.121688 1.240 0.2150   
## TrafficControl5 -0.031909 0.019565 -1.631 0.1029   
## Lighting2 -0.123767 0.006296 -19.659 < 2e-16 \*\*\*  
## Lighting3 -0.169860 0.025552 -6.648 3.01e-11 \*\*\*  
## Lighting4 -0.091447 0.014568 -6.277 3.48e-10 \*\*\*  
## Lighting5 -0.103094 0.052839 -1.951 0.0511 .   
## Lighting6 -0.253096 0.028369 -8.922 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.4732485)  
##   
## Null deviance: 27193 on 56043 degrees of freedom  
## Residual deviance: 26518 on 56034 degrees of freedom  
## AIC: 117130  
##   
## Number of Fisher Scoring iterations: 2

plot(model)



This analysis concludes that the majority of Minor Injuries occurred as a result of an accident that occurred at night when street lights were operational, and when they were not, some of them occurred around dusk or dawn. The insignificant factors here are obscured traffic controls and lighting conditions (night and street lights not working).

The deviance residuals are somewhat symmetric. In the residuals vs fitted plot the linearity seems to hold reasonably well. As we move to the right on the x axis we notice that the residuals seem to be increasing. Furthermore, we can see that the residuals aren’t normally distributed. The points above the dotted red line in the residuals vs leverage plot are extremely influential.

1-pchisq(27193-26518,56043-56034)

## [1] 0

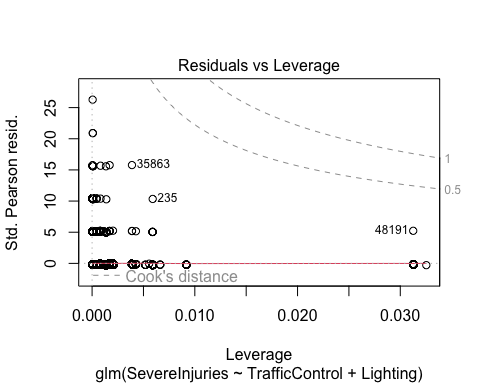
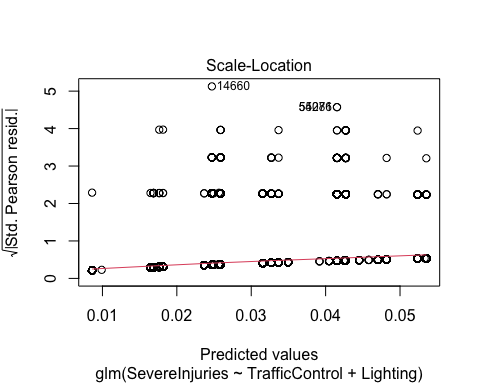
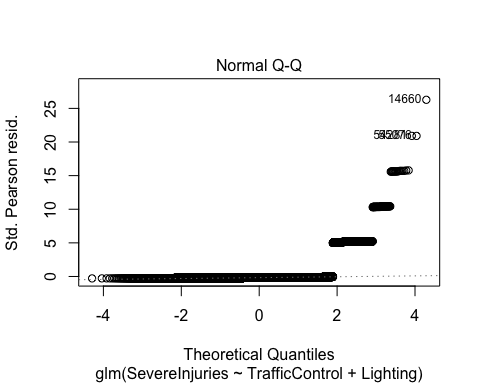
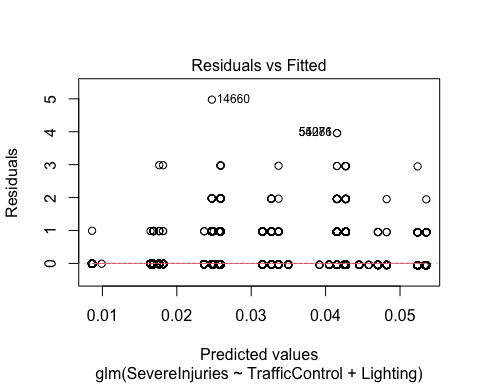
Since the chi-squared statistics for the null model and the model with two independent variables are exactly 0, the **null hypothesis** that the deviance of the model with the constant term and the deviance of the model with two independent variables added to it are the same, **doesn’t hold good**. The deviations aren’t the same; the latter are much lower, statistically significantly lower.

## Logistic regression analysis model 2

model1 <- glm(SevereInjuries ~ TrafficControl + Lighting, data = z1)  
summary(model1)

##   
## Call:  
## glm(formula = SevereInjuries ~ TrafficControl + Lighting, data = z1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.0535 -0.0415 -0.0259 -0.0247 4.9753   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.024711 0.001382 17.877 < 2e-16 \*\*\*  
## TrafficControl2 0.001171 0.001633 0.717 0.4734   
## TrafficControl3 -0.006565 0.011846 -0.554 0.5795   
## TrafficControl4 0.000945 0.033492 0.028 0.9775   
## TrafficControl5 -0.007854 0.005385 -1.459 0.1447   
## Lighting2 0.016816 0.001733 9.705 < 2e-16 \*\*\*  
## Lighting3 0.027652 0.007033 3.932 8.44e-05 \*\*\*  
## Lighting4 0.006822 0.004010 1.701 0.0889 .   
## Lighting5 0.022319 0.014543 1.535 0.1249   
## Lighting6 -0.008248 0.007808 -1.056 0.2908   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.03584941)  
##   
## Null deviance: 2012.9 on 56043 degrees of freedom  
## Residual deviance: 2008.8 on 56034 degrees of freedom  
## AIC: -27480  
##   
## Number of Fisher Scoring iterations: 2

plot(model1)



This analysis concludes that the majority of severe injuries occurred as a result of an accident that occurred at night when the street lights were not functioning; some of them occurred around dusk or dawn. The rest of the factors seem to be insignificant and do not contribute to the occurrence of severe mishaps.

In conclusion, the linearity appears to hold quite well in the residuals vs. fitted plot. The residuals appear to be growing as we move to the right on the x axis. We can also see that the residuals are not normally distributed. The residuals vs. leverage plot points above the dashed red line are particularly influential.

1-pchisq(2012.9-2008.8, 56043-56034)

## [1] 0.9047082

Because the chi-squared statistics for the null model and the model with two independent variables are close to 1, the **null hypothesis** that the deviance of the model with the constant term and the deviance of the model with two independent variables added to it are the same, **holds good**. The deviations are the same.

Also, the Akaike Information Criterion (AIC) value for the second regression model is much lower than the first, so the former indicates a better-fit model.

## Conclusion

From the above analyses, it can be concluded that 2015 was one of the years where the majority of the crashes resulted in fatalities, injuring many pedestrians and drivers. Most of the crashes have taken place during the day when traffic controls were functioning properly, some in the night when the controls were both functioning and not functioning, and a few during dusk. After the year 2015, we could observe a decreasing trend in the number of crashes and casualties.

The most important factors that contributed to the mishaps were: a) Over a ten-year period, it is estimated that the absence of traffic controls caused around 26,000 accidents. b) Violation of traffic rules by the primary drivers. c) Vehicles ramming into pedestrians, other vehicles, fixed objects and parked vehicles.

The hypothesis formulations: a) provided evidence using chi-square test of independence that the likelihood of traffic controls operating during the day is greater (about 54%) than under other lighting situations. b) indicated that the mean of both severe and moderate injuries are equal to each other and also are centered around the same value.

Regression analysis aided in understanding the major factors contributing to crashes over time, as well as determining which model best fits the data based on AIC values. As seen, the regression model for severe injuries fits the dataset better than minor injuries, it indicates the number of incidents, the factors contributing and the significant factors to cause a crash.

## References

1. Alice, M. (2020, July 5). How to perform a Logistic Regression in R. R-bloggers. <https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/>
2. n.d.). 11.3 - Chi-Square Test of Independence. <https://online.stat.psu.edu/stat200/book/export/html/230>
3. Chouldechova, P. (n.d.). Lecture 8 - Hypothesis testing in R. <https://www.andrew.cmu.edu/user/achoulde/94842/lectures/lecture07/lecture07-94842.html>
4. GeeksforGeeks. (2020, June 22). Hypothesis Testing in R Programming. <https://www.geeksforgeeks.org/hypothesis-testing-in-r-programming/>
5. Grolemund, H. W. A. G. (n.d.). 7 Exploratory Data Analysis | R for Data Science. <https://r4ds.had.co.nz/exploratory-data-analysis.html>
6. Linear Regression with R -. (2016, January 1). Visual Studio Magazine. <https://visualstudiomagazine.com/articles/2016/01/01/linear-regression-with-r.aspx>
7. Logit Regression | R Data Analysis Examples. (n.d.). <https://stats.oarc.ucla.edu/r/dae/logit-regression/>
8. (2022, September 28). Understanding Deviance Residuals | University of Virginia Library Research Data Services + Sciences. <https://data.library.virginia.edu/understanding-deviance-residuals/>