ASSESSMENT 2 - REPORT

CMP5353 - Introduction to Data Science A S1 2020/2021

Calderdale Accident Data Frame 2020/2021

Name - Jaspreet singh

StudentID - 19150299

Calderdale Accident Casualty Info
2020/2021

Contents

Introduction

The pile of data that this report is analysing and transforming is information regarding road traffic collisions (RTCs) in the Calderdale area nearby to the city of Leeds. This data was produced by Calderdale Council and made available at 'https://dataworks.calderdale.gov.uk/dataset/calderdale-accident-data-'.

In this report can be found the better organized code visualization samples extracted from the main raw file containing all R code that is included in the file named "19150299_assessment2.R" and the the CSV files of the cleaned data and the regression predictions included as "regression.csv" and "clean_accidents.csv" respectively.

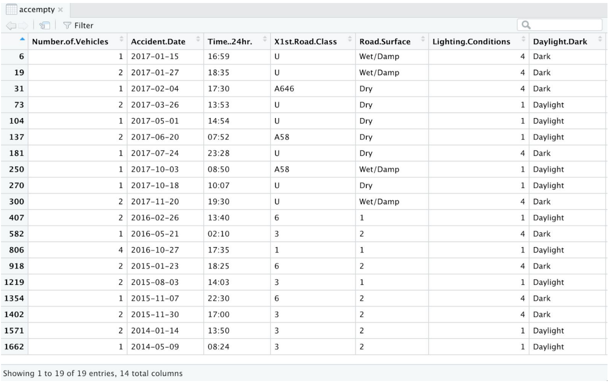
Data Wrangling

Examining the columns in the data

This dataset contains 14 columns, some of them, such as “Accident Date” and “Time (24hr)” tell us when exactly the accident took place, while other columns such as “Weather Conditions”, “Lighting Conditions” and “Road Surface” help us understand the situation in which the accident occured and then columns such as “Sex of Casualty” and “Age of Casualty” help us to understand the type of the people involved in it. Last but not least the columns “Casualty Severity”, “Casualty Class” and “Number of Vehicles” give us an insight on how big the accident was.

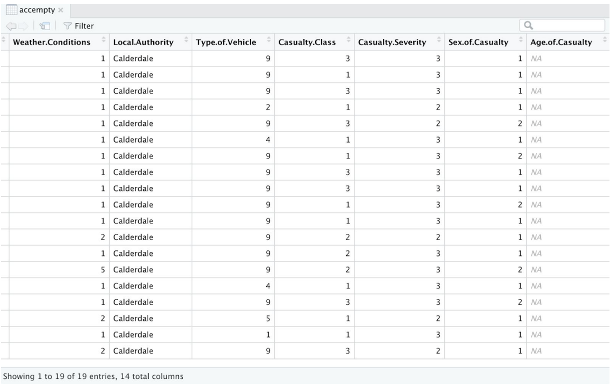
Missing data

In this dataset there are in total 19 missing values all found in the “Age of Casualty” column (1)



	Number.of.Vehicles	Accident.Date	Time..24hr.	X1st.Road.Class	Road.Surface	Lighting.Conditions	Daylight.Dark
6	1	2017-01-15	16:59	U	Wet/Damp	4	Dark
19	2	2017-01-27	18:35	U	Wet/Damp	4	Dark
31	1	2017-02-04	17:30	A646	Dry	4	Dark
73	2	2017-03-26	13:53	U	Dry	1	Daylight
104	1	2017-05-01	14:54	U	Dry	1	Daylight
137	2	2017-06-20	07:52	A58	Dry	1	Daylight
181	1	2017-07-24	23:28	U	Dry	4	Dark
250	1	2017-10-03	08:50	A58	Wet/Damp	1	Daylight
270	1	2017-10-18	10:07	U	Dry	1	Daylight
300	2	2017-11-20	19:30	U	Wet/Damp	4	Dark
407	2	2016-02-26	13:40	6	1	1	Daylight
582	1	2016-05-21	02:10	3	2	4	Dark
806	4	2016-10-27	17:35	1	1	1	Daylight
918	2	2015-01-23	18:25	6	2	4	Dark
1219	2	2015-08-03	14:03	3	1	1	Daylight
1354	1	2015-11-07	22:30	6	2	4	Dark
1402	2	2015-11-30	17:00	3	2	4	Dark
1571	2	2014-01-14	13:50	3	2	1	Daylight
1662	1	2014-05-09	08:24	3	2	1	Daylight

Showing 1 to 19 of 19 entries, 14 total columns



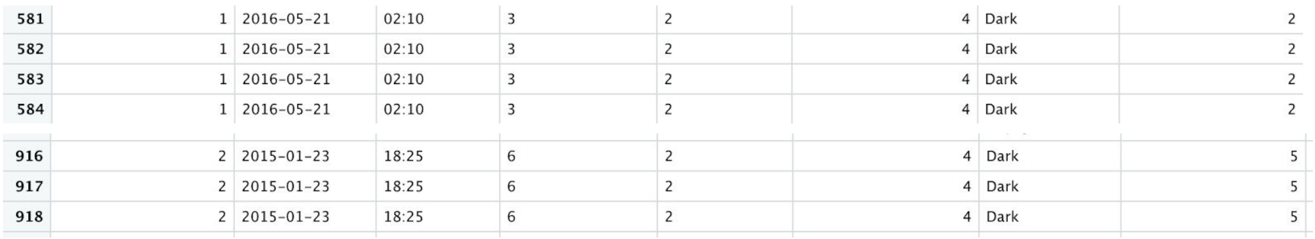
Weather.Conditions	Local.Authority	Type.of.Vehicle	Casualty.Class	Casualty.Severity	Sex.of.Casualty	Age.of.Casualty
1	Calderdale	9	3	3	1	NA
1	Calderdale	9	1	3	1	NA
1	Calderdale	9	3	3	1	NA
1	Calderdale	2	1	2	1	NA
1	Calderdale	9	3	2	2	NA
1	Calderdale	4	1	3	1	NA
1	Calderdale	9	1	3	2	NA
1	Calderdale	9	3	3	1	NA
1	Calderdale	9	3	3	1	NA
1	Calderdale	9	1	3	2	NA
1	Calderdale	9	1	3	1	NA
2	Calderdale	9	2	2	1	NA
1	Calderdale	9	2	3	1	NA
5	Calderdale	9	2	3	2	NA
1	Calderdale	4	1	3	1	NA
1	Calderdale	9	3	3	2	NA
2	Calderdale	5	1	2	1	NA
1	Calderdale	1	1	3	1	NA
2	Calderdale	9	3	2	1	NA

Showing 1 to 19 of 19 entries, 14 total columns

All the 19 rows where the value for “Age of Casualty” is missing. (1)

Looking at the missing values, one Missing Not At Random(MNAR) reason could be that the drivers involved were young people and didn’t want to disclose their age for fear of embarrassment at the potential stereotype that new young drivers are irresponsible at the wheel.

Looking at the rows near the end of the list we can see that there cases where more vehicles were involved in the accident and that could suggests that this was a major road traffic collision so it may be possible that despite the “Slight” injuries recorded some drivers may have left from the fear of a higher insurance cost before their age being registered. This would be an instance of data being Missing At Random(MAR). (2)

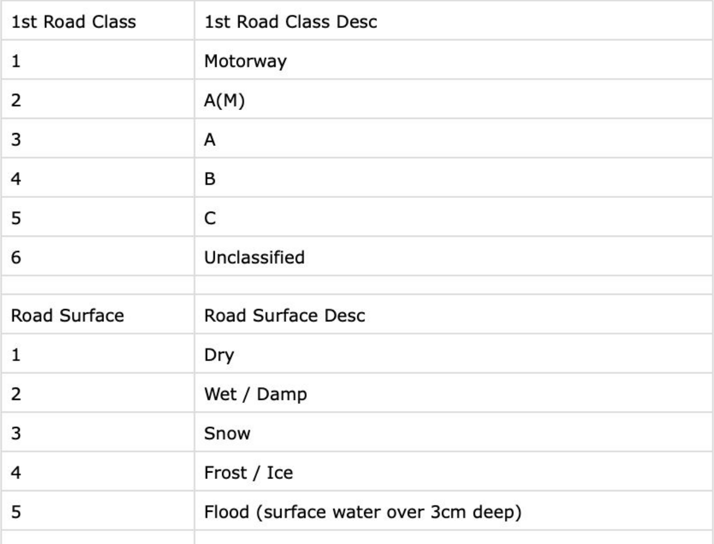


581	1	2016-05-21	02:10	3	2	4	Dark	2
582	1	2016-05-21	02:10	3	2	4	Dark	2
583	1	2016-05-21	02:10	3	2	4	Dark	2
584	1	2016-05-21	02:10	3	2	4	Dark	2
916	2	2015-01-23	18:25	6	2	4	Dark	5
917	2	2015-01-23	18:25	6	2	4	Dark	5
918	2	2015-01-23	18:25	6	2	4	Dark	5

Examples of possible MAR cases. (2)

Anomalies & Inconsistencies

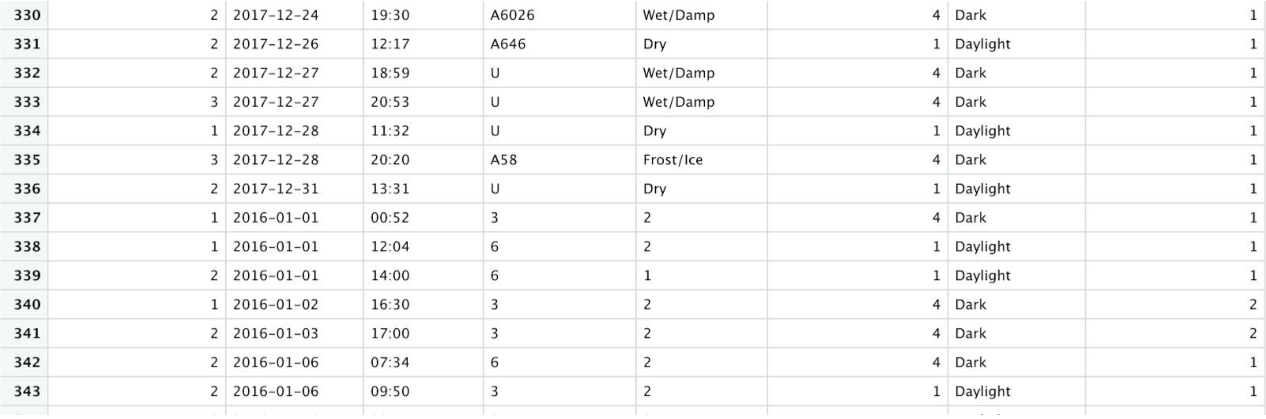
The “Guidance.csv” provided by Calderdale Council has the goal to help us translate the number presented in the CSV file to the actual real world case scenario.This system is used for the columns “Road Class” and “Road Surfaces”, the first telling us which type or road and the second how was the state of the surface of the road when the collision happened. (3)



1st Road Class	1st Road Class Desc
1	Motorway
2	A(M)
3	A
4	B
5	C
6	Unclassified
Road Surface	Road Surface Desc
1	Dry
2	Wet / Damp
3	Snow
4	Frost / Ice
5	Flood (surface water over 3cm deep)

Numbers and real world scenario correlation in the Guidance.csv. (3)

However, looking at the different values present in these columns we can see that not every record matches the numeric category structure represented in the “Guidance.csv”. (4)

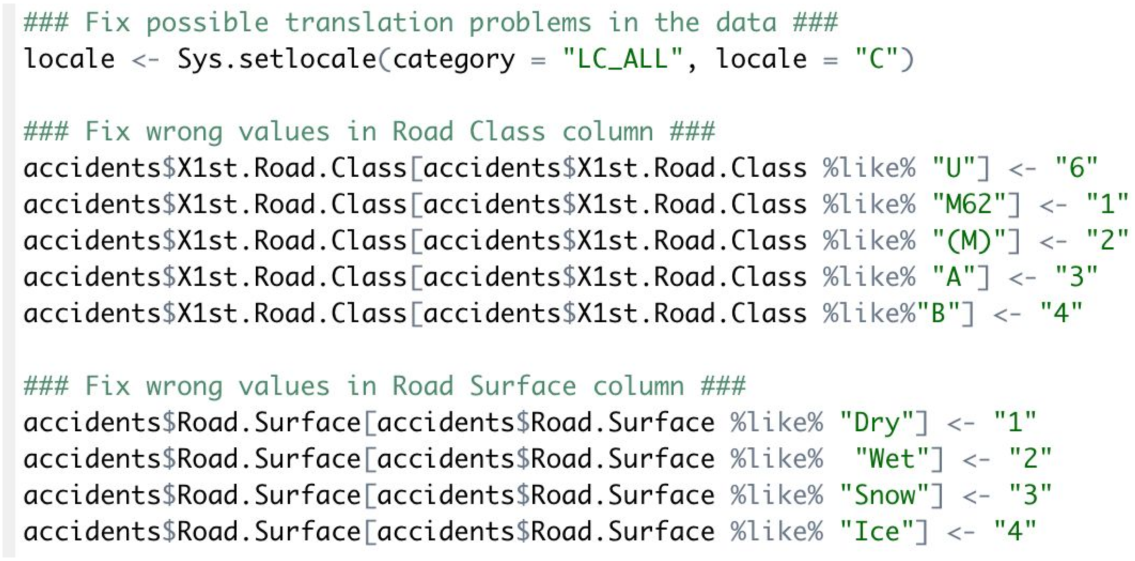


330	2	2017-12-24	19:30	A6026	Wet/Damp	4	Dark	1
331	2	2017-12-26	12:17	A646	Dry	1	Daylight	1
332	2	2017-12-27	18:59	U	Wet/Damp	4	Dark	1
333	3	2017-12-27	20:53	U	Wet/Damp	4	Dark	1
334	1	2017-12-28	11:32	U	Dry	1	Daylight	1
335	3	2017-12-28	20:20	A58	Frost/Ice	4	Dark	1
336	2	2017-12-31	13:31	U	Dry	1	Daylight	1
337	1	2016-01-01	00:52	3	2	4	Dark	1
338	1	2016-01-01	12:04	6	2	1	Daylight	1
339	2	2016-01-01	14:00	6	1	1	Daylight	1
340	1	2016-01-02	16:30	3	2	4	Dark	2
341	2	2016-01-03	17:00	3	2	4	Dark	2
342	2	2016-01-06	07:34	6	2	4	Dark	1
343	2	2016-01-06	09:50	3	2	1	Daylight	1

“Road Class” and “Road Surfaces” columns with different types of value recorded. (4)

Looking at the data the first 336 entries out of the total 2069 are inserted incorrectly and have not followed the guidelines set out in the “Guidance.csv” file.

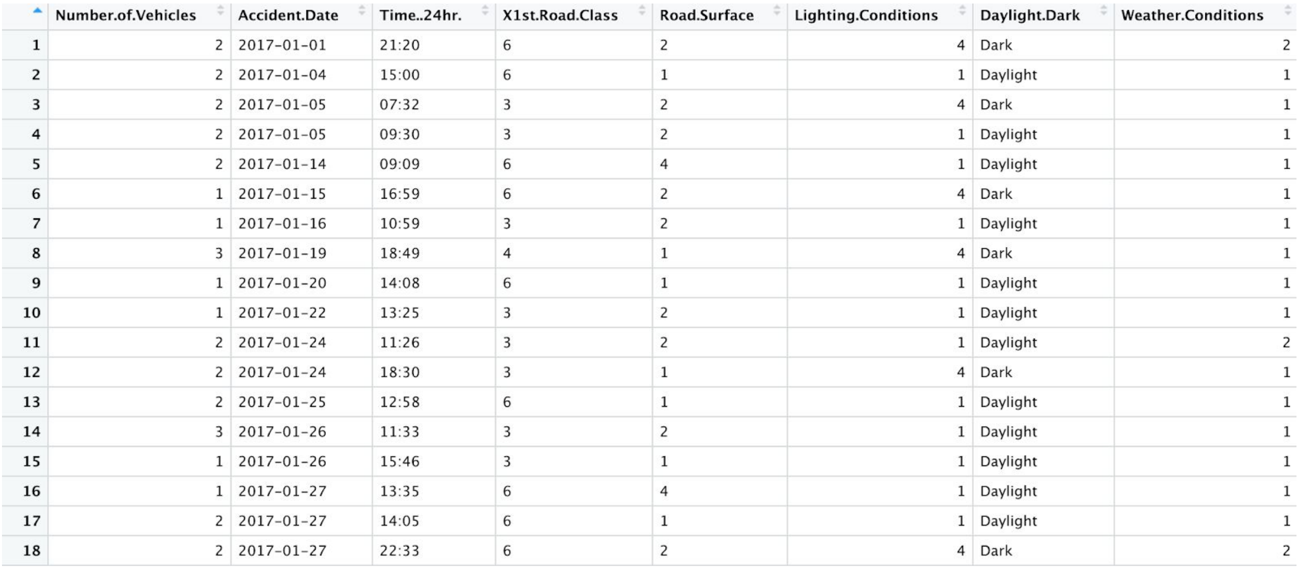
After we found the various anomalies we need to fix them, fortunately all the registered data seems to be still accurate to real life conditions and they can be correctly replaced with the correct numeric values. (5)



```
### Fix possible translation problems in the data ###
locale <- Sys.setlocale(category = "LC_ALL", locale = "C")

### Fix wrong values in Road Class column ###
accidents$X1st.Road.Class[accidents$X1st.Road.Class %like% "U"] <- "6"
accidents$X1st.Road.Class[accidents$X1st.Road.Class %like% "M62"] <- "1"
accidents$X1st.Road.Class[accidents$X1st.Road.Class %like% "(M)"] <- "2"
accidents$X1st.Road.Class[accidents$X1st.Road.Class %like% "A"] <- "3"
accidents$X1st.Road.Class[accidents$X1st.Road.Class %like%"B"] <- "4"

### Fix wrong values in Road Surface column ###
accidents$Road.Surface[accidents$Road.Surface %like% "Dry"] <- "1"
accidents$Road.Surface[accidents$Road.Surface %like%  "Wet"] <- "2"
accidents$Road.Surface[accidents$Road.Surface %like% "Snow"] <- "3"
accidents$Road.Surface[accidents$Road.Surface %like% "Ice"] <- "4"
```

Code to make “Road Class” and “Road Surfaces” columns comply with the “Guidance.csv” file. (5)

Here we have the dataset finally corrected and that completely follows the given guidelines file. (6)

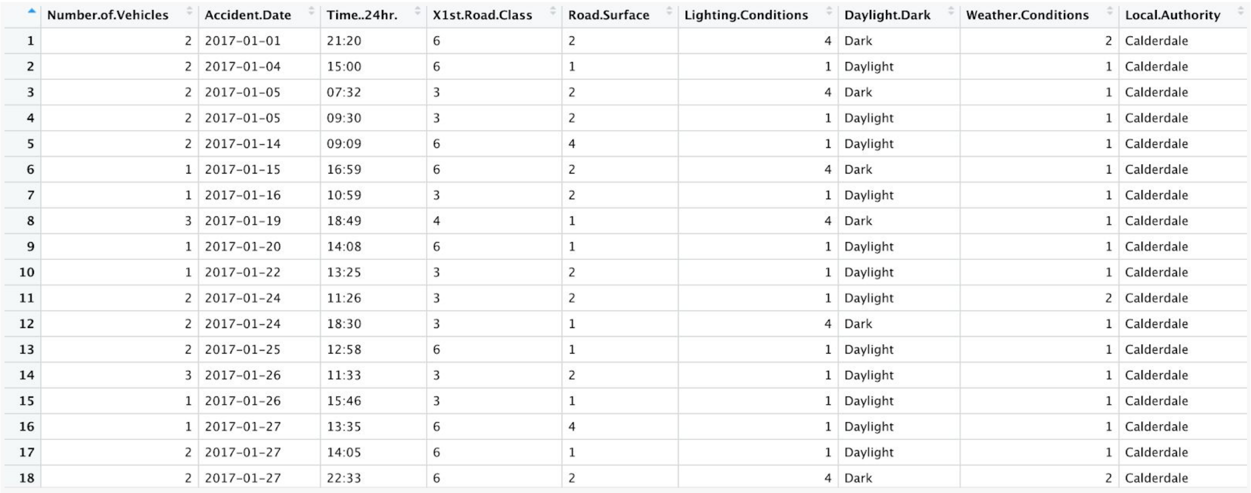


	Number.of.Vehicles	Accident.Date	Time..24hr.	X1st.Road.Class	Road.Surface	Lighting.Conditions	Daylight.Dark	Weather.Conditions
1	2	2017-01-01	21:20	6	2	4	Dark	2
2	2	2017-01-04	15:00	6	1	1	Daylight	1
3	2	2017-01-05	07:32	3	2	4	Dark	1
4	2	2017-01-05	09:30	3	2	1	Daylight	1
5	2	2017-01-14	09:09	6	4	1	Daylight	1
6	1	2017-01-15	16:59	6	2	4	Dark	1
7	1	2017-01-16	10:59	3	2	1	Daylight	1
8	3	2017-01-19	18:49	4	1	4	Dark	1
9	1	2017-01-20	14:08	6	1	1	Daylight	1
10	1	2017-01-22	13:25	3	2	1	Daylight	1
11	2	2017-01-24	11:26	3	2	1	Daylight	2
12	2	2017-01-24	18:30	3	1	4	Dark	1
13	2	2017-01-25	12:58	6	1	1	Daylight	1
14	3	2017-01-26	11:33	3	2	1	Daylight	1
15	1	2017-01-26	15:46	3	1	1	Daylight	1
16	1	2017-01-27	13:35	6	4	1	Daylight	1
17	2	2017-01-27	14:05	6	1	1	Daylight	1
18	2	2017-01-27	22:33	6	2	4	Dark	2

Dataset with the corrected “Road Class” and “Road Surfaces” columns. (6)

Removing unnecessary columns

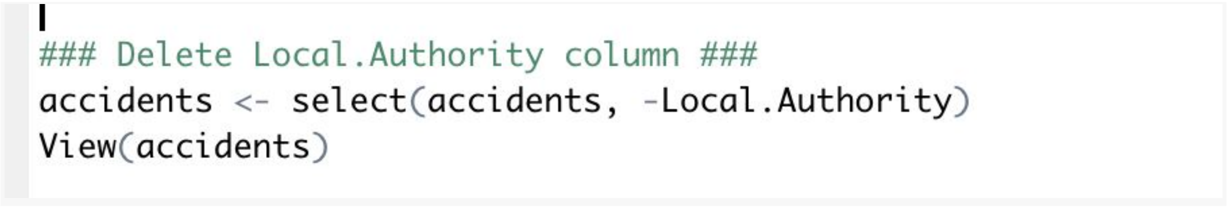
Looking at the various columns inside the “accidents.csv” file it’s easy to spot one that isn’t actually necessary to keep as all it’s value in every row doesn't change and it’s always the same. This column is the “Local Authority” one, that has to tell us which local authority in the UK is in charge of tracking the events registered in the file. Since we are only looking at accidents that happened in the area where only this authority operates, every event that occurred is bound to have Calderdale Council as the value for its local authority. (7)



	Number.of.Vehicles	Accident.Date	Time..24hr.	X1st.Road.Class	Road.Surface	Lighting.Conditions	Daylight.Dark	Weather.Conditions	Local.Authority
1	2	2017-01-01	21:20	6	2	4	Dark	2	Calderdale
2	2	2017-01-04	15:00	6	1	1	Daylight	1	Calderdale
3	2	2017-01-05	07:32	3	2	4	Dark	1	Calderdale
4	2	2017-01-05	09:30	3	2	1	Daylight	1	Calderdale
5	2	2017-01-14	09:09	6	4	1	Daylight	1	Calderdale
6	1	2017-01-15	16:59	6	2	4	Dark	1	Calderdale
7	1	2017-01-16	10:59	3	2	1	Daylight	1	Calderdale
8	3	2017-01-19	18:49	4	1	4	Dark	1	Calderdale
9	1	2017-01-20	14:08	6	1	1	Daylight	1	Calderdale
10	1	2017-01-22	13:25	3	2	1	Daylight	1	Calderdale
11	2	2017-01-24	11:26	3	2	1	Daylight	2	Calderdale
12	2	2017-01-24	18:30	3	1	4	Dark	1	Calderdale
13	2	2017-01-25	12:58	6	1	1	Daylight	1	Calderdale
14	3	2017-01-26	11:33	3	2	1	Daylight	1	Calderdale
15	1	2017-01-26	15:46	3	1	1	Daylight	1	Calderdale
16	1	2017-01-27	13:35	6	4	1	Daylight	1	Calderdale
17	2	2017-01-27	14:05	6	1	1	Daylight	1	Calderdale
18	2	2017-01-27	22:33	6	2	4	Dark	2	Calderdale

The “Local Authority” remains always the same as “Calderdale”. (7)

Here the raw code to delete the unnecessary column and a successive check of the action executed. (8)

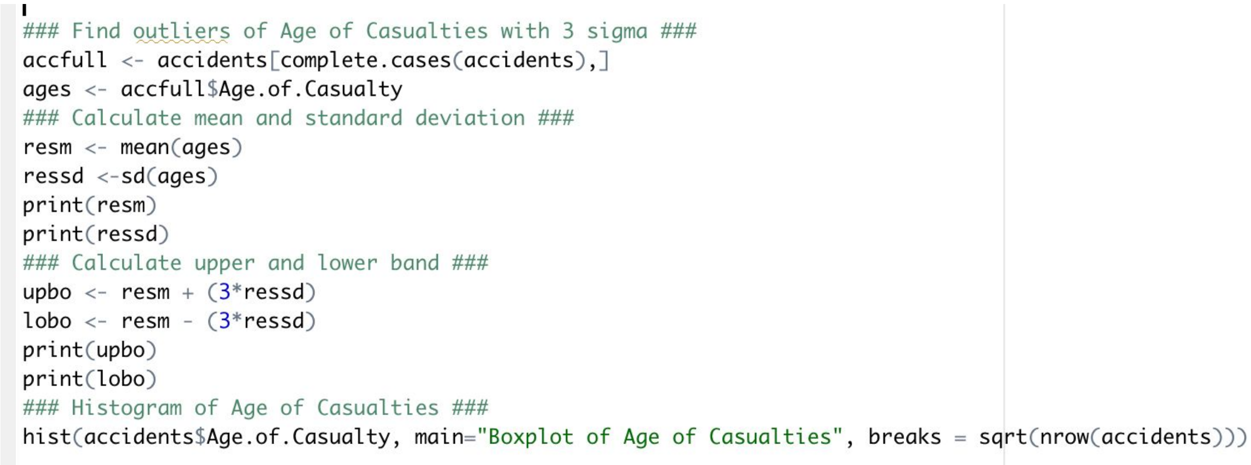


```
### Delete Local.Authority column ###
accidents <- select(accidents, -Local.Authority)
View(accidents)
```

The raw code to delete the “Local Authority” column. (8)

Identifying outliers in the ‘Age of Casualty’ column

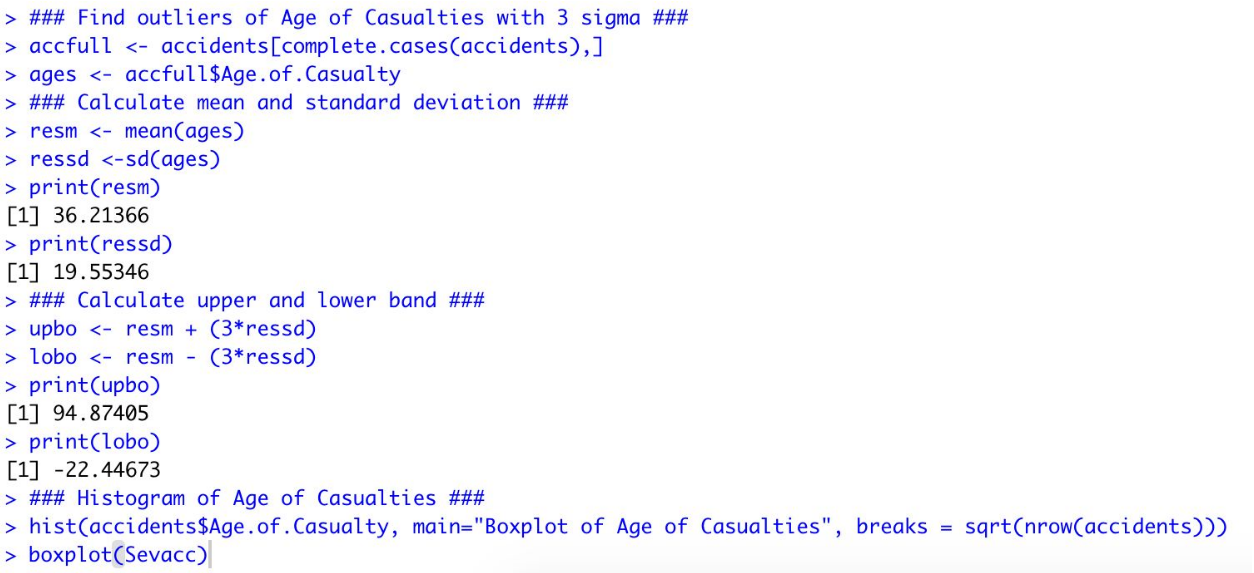
There are three main methods in order to identify outliers in a set of data. We're gonna start with using the 3 Sigma rule. (9)



```
### Find outliers of Age of Casualties with 3 sigma ###
accfull <- accidents[complete.cases(accidents),]
ages <- accfull$Age.of.Casualty
### Calculate mean and standard deviation ###
resm <- mean(ages)
ressd <-sd(ages)
print(resm)
print(ressd)
### Calculate upper and lower band ###
upbo <- resm + (3*ressd)
lobo <- resm - (3*ressd)
print(upbo)
print(lobo)
### Histogram of Age of Casualties ###
hist(accidents$Age.of.Casualty, main="Boxplot of Age of Casualties", breaks = sqrt(nrow(accidents)))
```

Code for the outliers analysis using the 3 Sigma rule. (9)

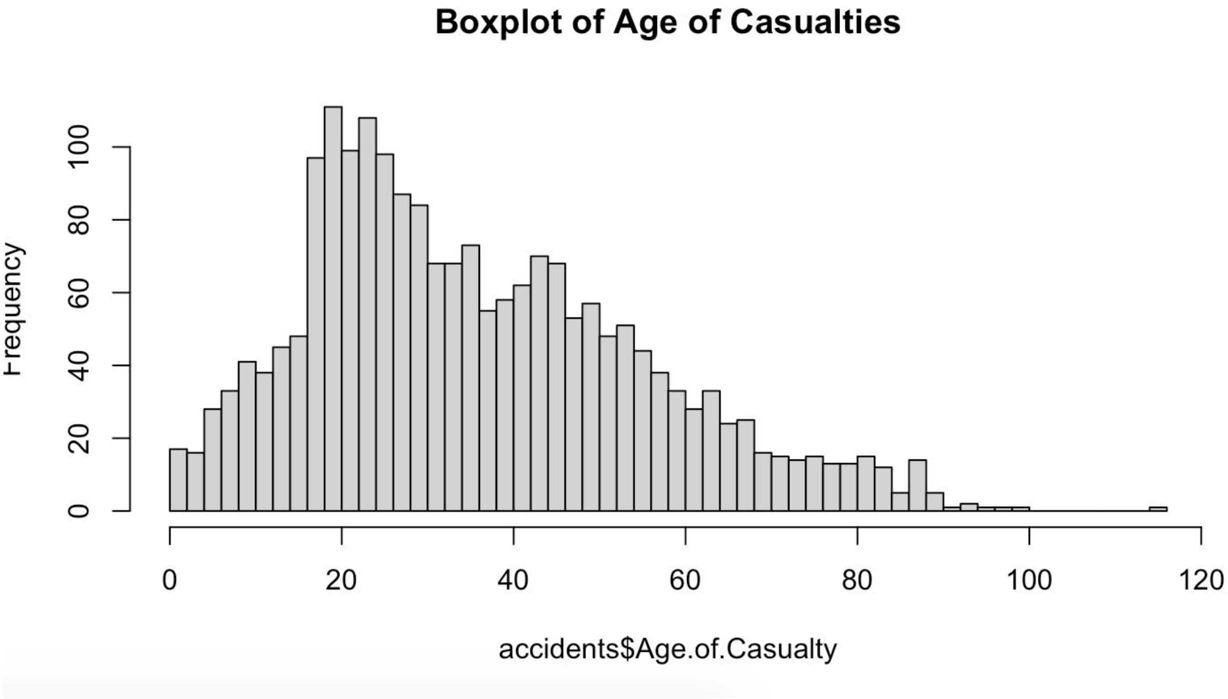
Here the output where we can see there are only 4 outliers. (10)



```
> ### Find outliers of Age of Casualties with 3 sigma ###
> accfull <- accidents[complete.cases(accidents),]
> ages <- accfull$Age.of.Casualty
> ### Calculate mean and standard deviation ###
> resm <- mean(ages)
> ressd <-sd(ages)
> print(resm)
[1] 36.21366
> print(ressd)
[1] 19.55346
> ### Calculate upper and lower band ###
> upbo <- resm + (3*ressd)
> lobo <- resm - (3*ressd)
> print(upbo)
[1] 94.87405
> print(lobo)
[1] -22.44673
> ### Histogram of Age of Casualties ###
> hist(accidents$Age.of.Casualty, main="Boxplot of Age of Casualties", breaks = sqrt(nrow(accidents)))
> boxplot(Sevacc)
```

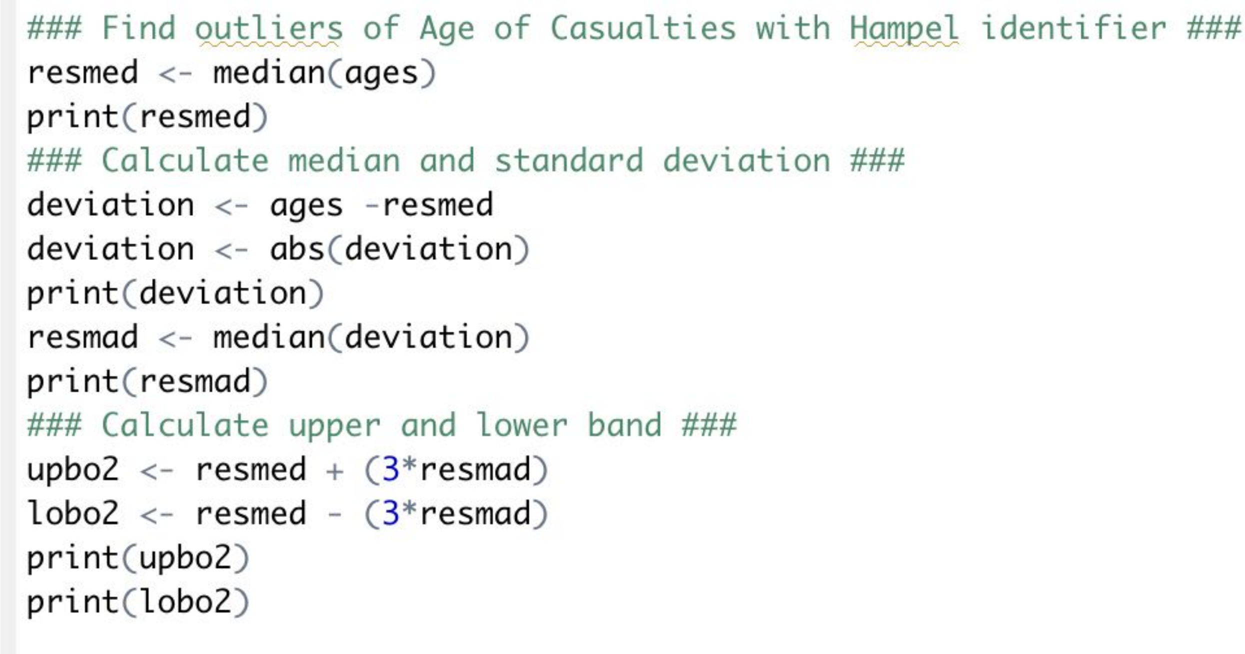
Code output for the outliers analysis using the 3 Sigma rule. (10)

Here a better visual representation of the data. (11)



Visual output for the outliers analysis using the 3 Sigma rule. (11)

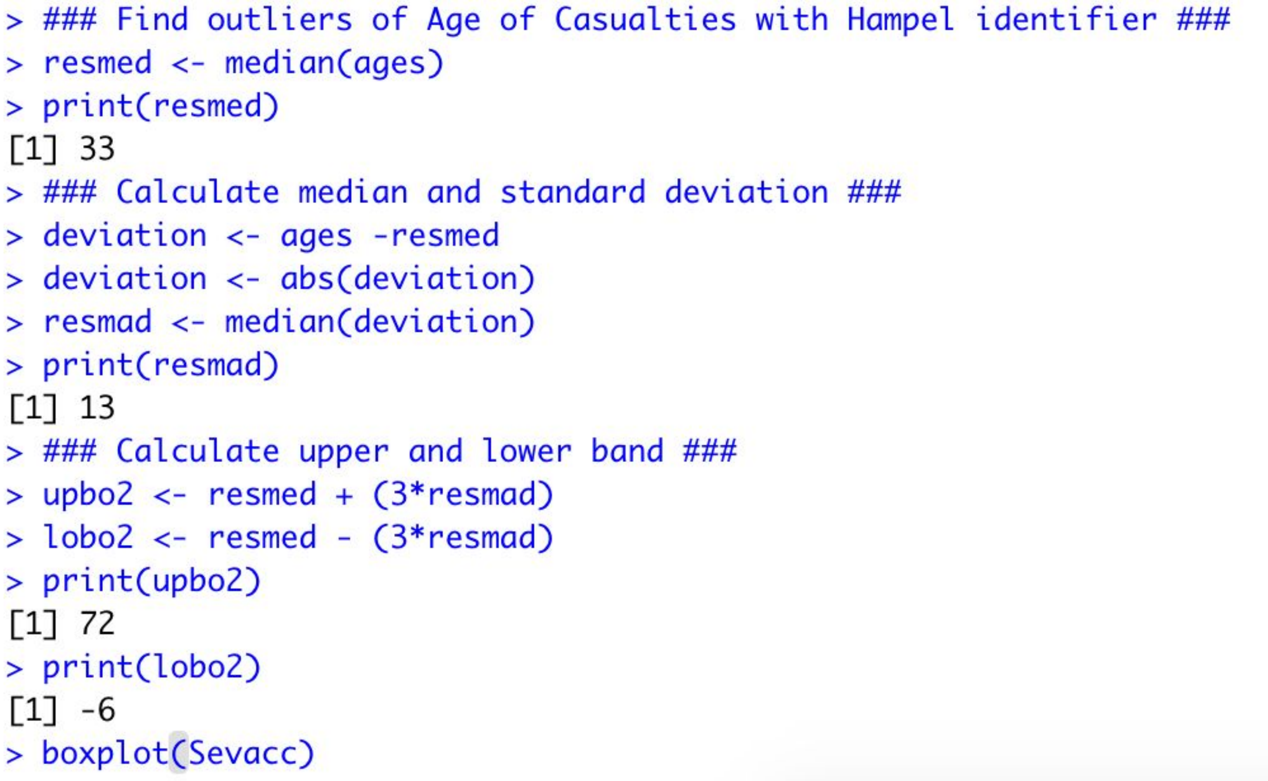
Another method to identify outliers is the Hampel identifier. (12)



```
### Find outliers of Age of Casualties with Hampel identifier ###
resmed <- median(ages)
print(resmed)
### Calculate median and standard deviation ###
deviation <- ages -resmed
deviation <- abs(deviation)
print(deviation)
resmad <- median(deviation)
print(resmad)
### Calculate upper and lower band ###
upbo2 <- resmed + (3*resmad)
lobo2 <- resmed - (3*resmad)
print(upbo2)
print(lobo2)
```

Code for the outliers analysis using the Hampel identifier. (12)

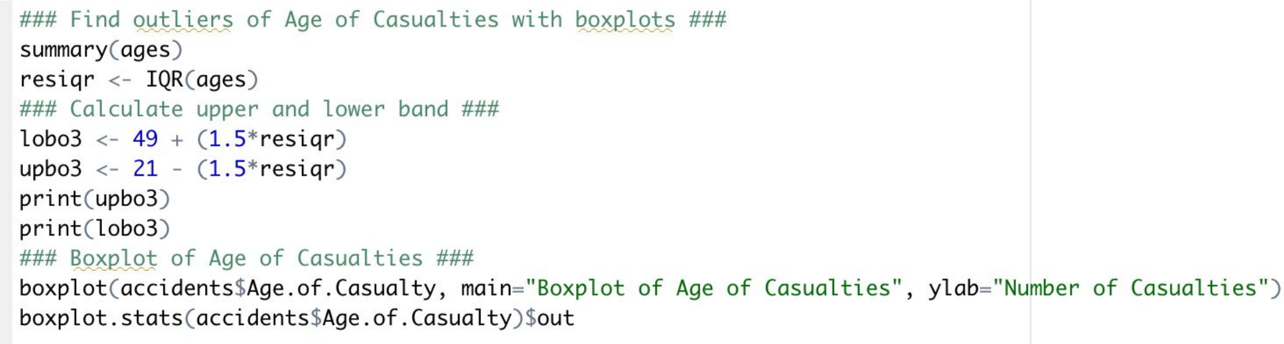
Here the output where we using an additional “sum(upbo2<(ages))” we can see there are 113 outliers. (13)



```
> ### Find outliers of Age of Casualties with Hampel identifier ###
> resmed <- median(ages)
> print(resmed)
[1] 33
> ### Calculate median and standard deviation ###
> deviation <- ages -resmed
> deviation <- abs(deviation)
> resmad <- median(deviation)
> print(resmad)
[1] 13
> ### Calculate upper and lower band ###
> upbo2 <- resmed + (3*resmad)
> lobo2 <- resmed - (3*resmad)
> print(upbo2)
[1] 72
> print(lobo2)
[1] -6
> boxplot(Sevacc)
```

Code output for the outliers analysis using the Hampel identifier. (13)

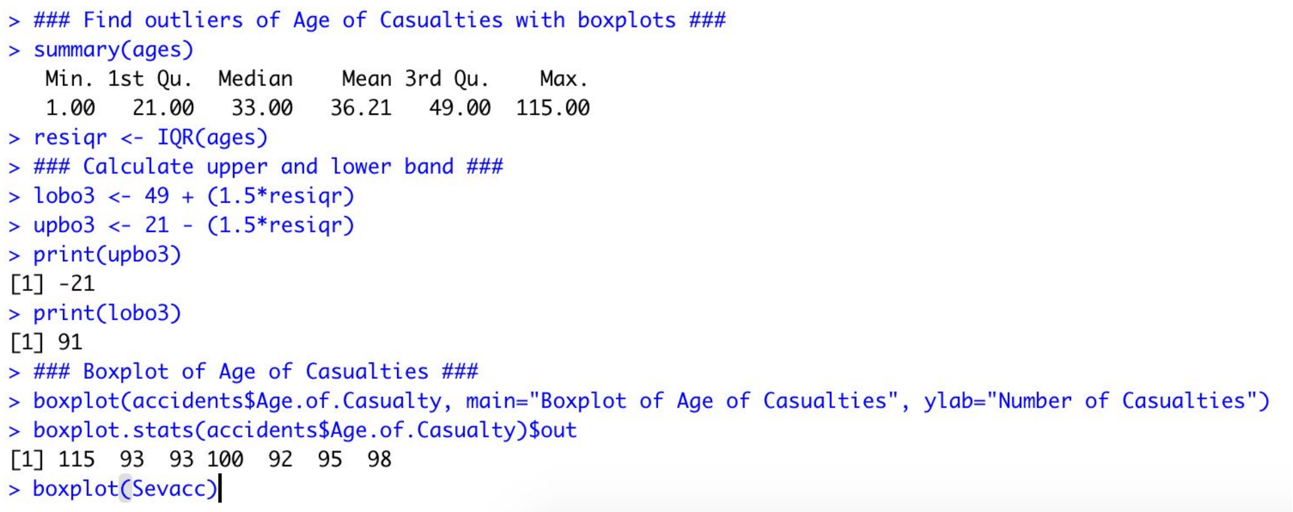
And the final method we can use here to identify outliers is using the boxplot and it’s code is as follows. (14)



```
### Find outliers of Age of Casualties with boxplots ###
summary(ages)
resiqr <- IQR(ages)
### Calculate upper and lower band ###
lobo3 <- 49 + (1.5*resiqr)
upbo3 <- 21 - (1.5*resiqr)
print(upbo3)
print(lobo3)
### Boxplot of Age of Casualties ###
boxplot(accidents$Age.of.Casualty, main="Boxplot of Age of Casualties", ylab="Number of Casualties")
boxplot.stats(accidents$Age.of.Casualty)$out
```

Code for the outliers analysis using the boxplots.. (14)

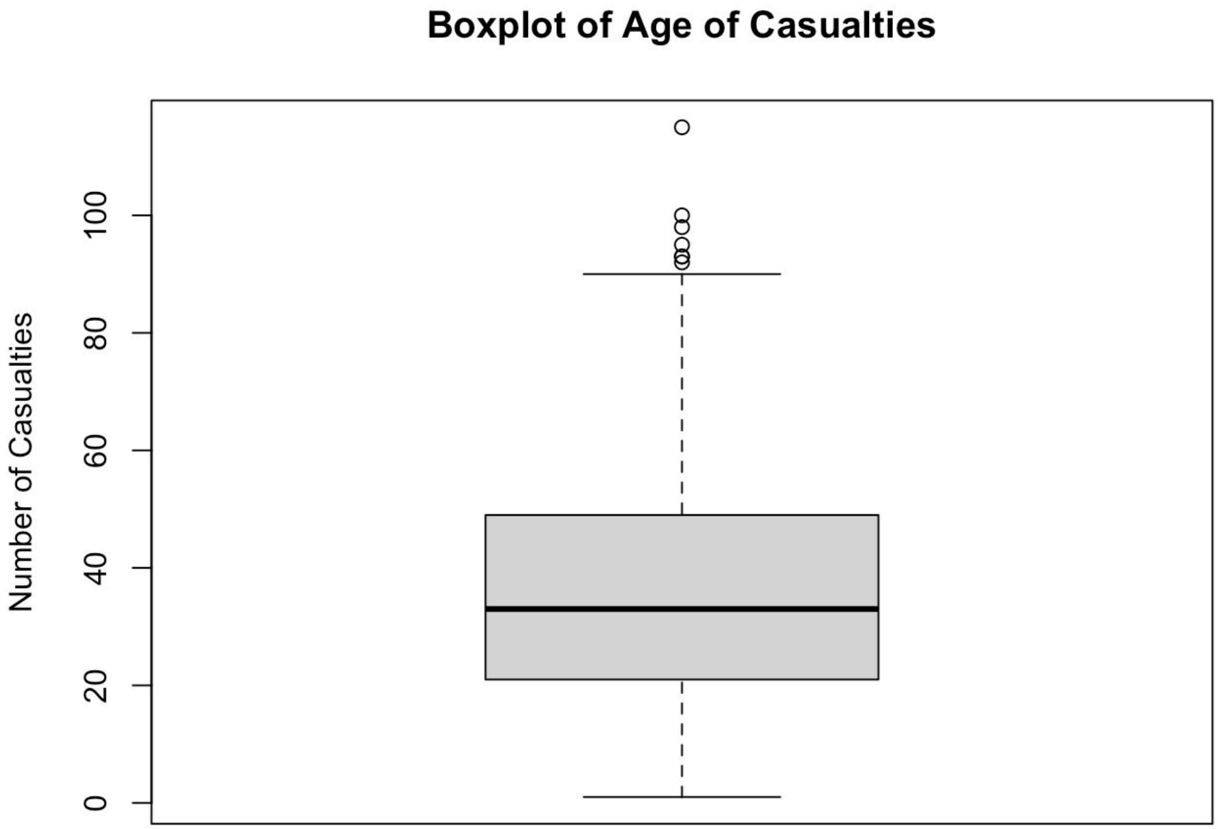
Here the output where we can see there are only 7 outliers. (15)



```
> ### Find outliers of Age of Casualties with boxplots ###
> summary(ages)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   21.00   33.00   36.21   49.00  115.00 
> resiqr <- IQR(ages)
> ### Calculate upper and lower band ###
> lobo3 <- 49 + (1.5*resiqr)
> upbo3 <- 21 - (1.5*resiqr)
> print(upbo3)
[1] -21
> print(lobo3)
[1] 91
> ### Boxplot of Age of Casualties ###
> boxplot(accidents$Age.of.Casualty, main="Boxplot of Age of Casualties", ylab="Number of Casualties")
> boxplot.stats(accidents$Age.of.Casualty)$out
[1] 115  93  93 100  92  95  98
> boxplot(Sevacc)
```

Code output for the outliers analysis using the Boxplots. (15)

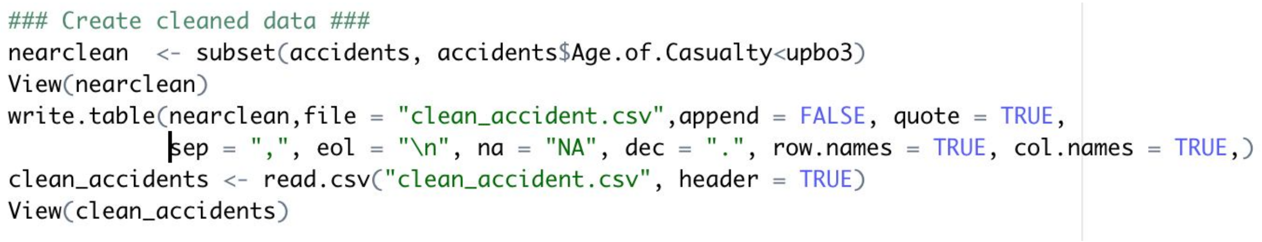
Here a better visual representation of the data using the boxplot. (16)



Visual output for the outliers analysis using the Boxplot. (16)

Since the Hampel Identifier method has a result that would take out a big quantity of data (5,5%) it won't be considered and the boxplot method will be chosen for better identifying outliers, the 3 sigma method has less possibility to reflex the real world scenario drivers where 90+ people drive daily.

After removing the outlier values, the dataset is now clean and ready for us to analyse its data. (17)



```
### Create cleaned data ###
nearclean  <- subset(accidents, accidents$Age.of.Casualty<upbo3)
View(nearclean)
write.table(nearclean,file = "clean_accident.csv",append = FALSE, quote = TRUE,
            sep = ",", eol = "\n", na = "NA", dec = ".", row.names = TRUE, col.names = TRUE,)
clean_accidents <- read.csv("clean_accident.csv", header = TRUE)
View(clean_accidents)
```

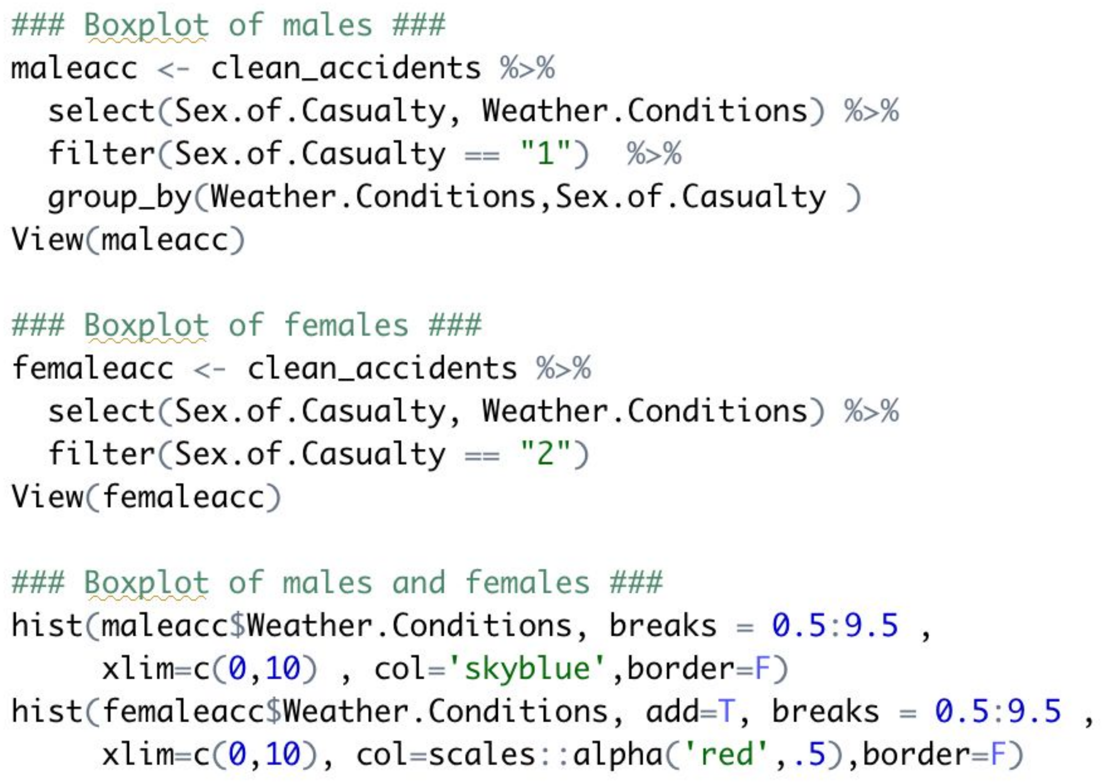
Code to create and store clean data. (17)

Data Exploration

Weather conditions and their effects on drivers of different genders

1 - "Is there any weather condition where male drivers/riders have more accidents than female drivers?".

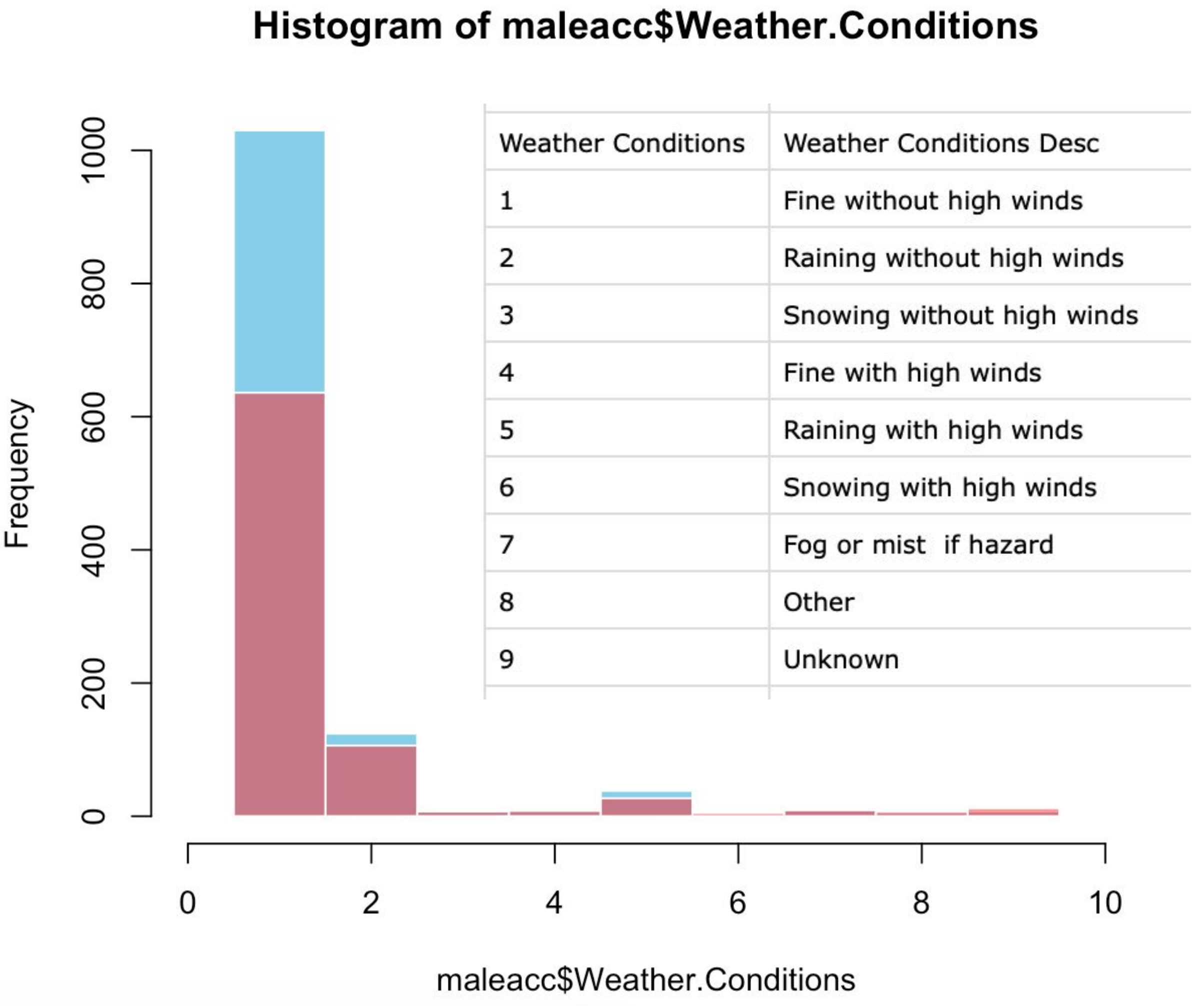
Here the code to obtain a bar graph with both males and females on each other in order to get who have more accidents with certain conditions. (18)



```
### Boxplot of males ###
maleacc <- clean_accidents %>%
  select(Sex.of.Casualty, Weather.Conditions) %>%
  filter(Sex.of.Casualty == "1")  %>%
  group_by(Weather.Conditions,Sex.of.Casualty )
View(maleacc)

### Boxplot of females ###
femaleacc <- clean_accidents %>%
  select(Sex.of.Casualty, Weather.Conditions) %>%
  filter(Sex.of.Casualty == "2")
View(femaleacc)

### Boxplot of males and females ###
hist(maleacc$Weather.Conditions, breaks = 0.5:9.5 ,
     xlim=c(0,10) , col='skyblue',border=F)
hist(femaleacc$Weather.Conditions, add=T, breaks = 0.5:9.5 ,
     xlim=c(0,10), col=scales::alpha('red',.5),border=F)
```

Code to analyse which gender does more accidents. (18)

Here a better visualization of the data with a reminder of the conversion in terms of a real life weather condition scenario. (19)



Weather Conditions	Weather Conditions Desc
1	Fine without high winds
2	Raining without high winds
3	Snowing without high winds
4	Fine with high winds
5	Raining with high winds
6	Snowing with high winds
7	Fog or mist if hazard
8	Other
9	Unknown

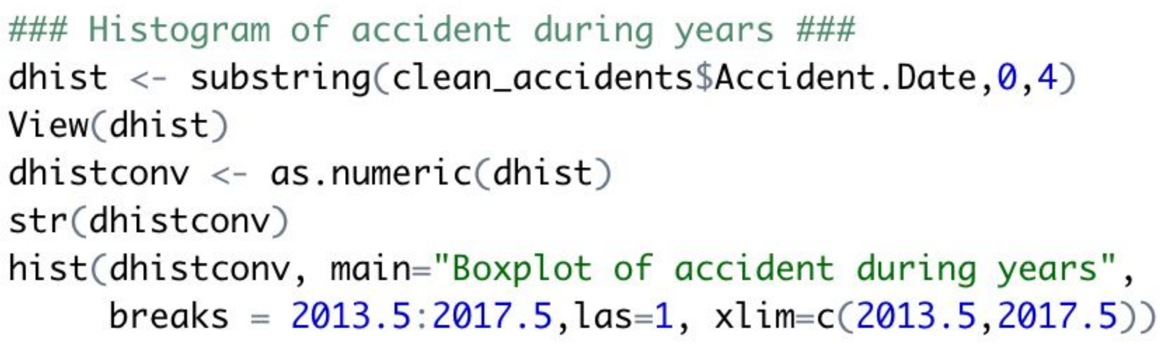
Histogram of male and female accidents (BLUE = MALE , RED = FEMALE). (19)

From the graph we can say that males and females are usually nearly identical as being involved in accidents, however the males make more accidents than females when the weather condition is fine without high winds. This can be caused by the fact that males are more inclined to go faster than the limits and possibly causing more events.

2 - Casualty Numbers on a year by year basis

Unfortunately the dataset does not procure us the number of casualties involved in each accident, so we are gonna have to assume that there is always just one casualty per accident.

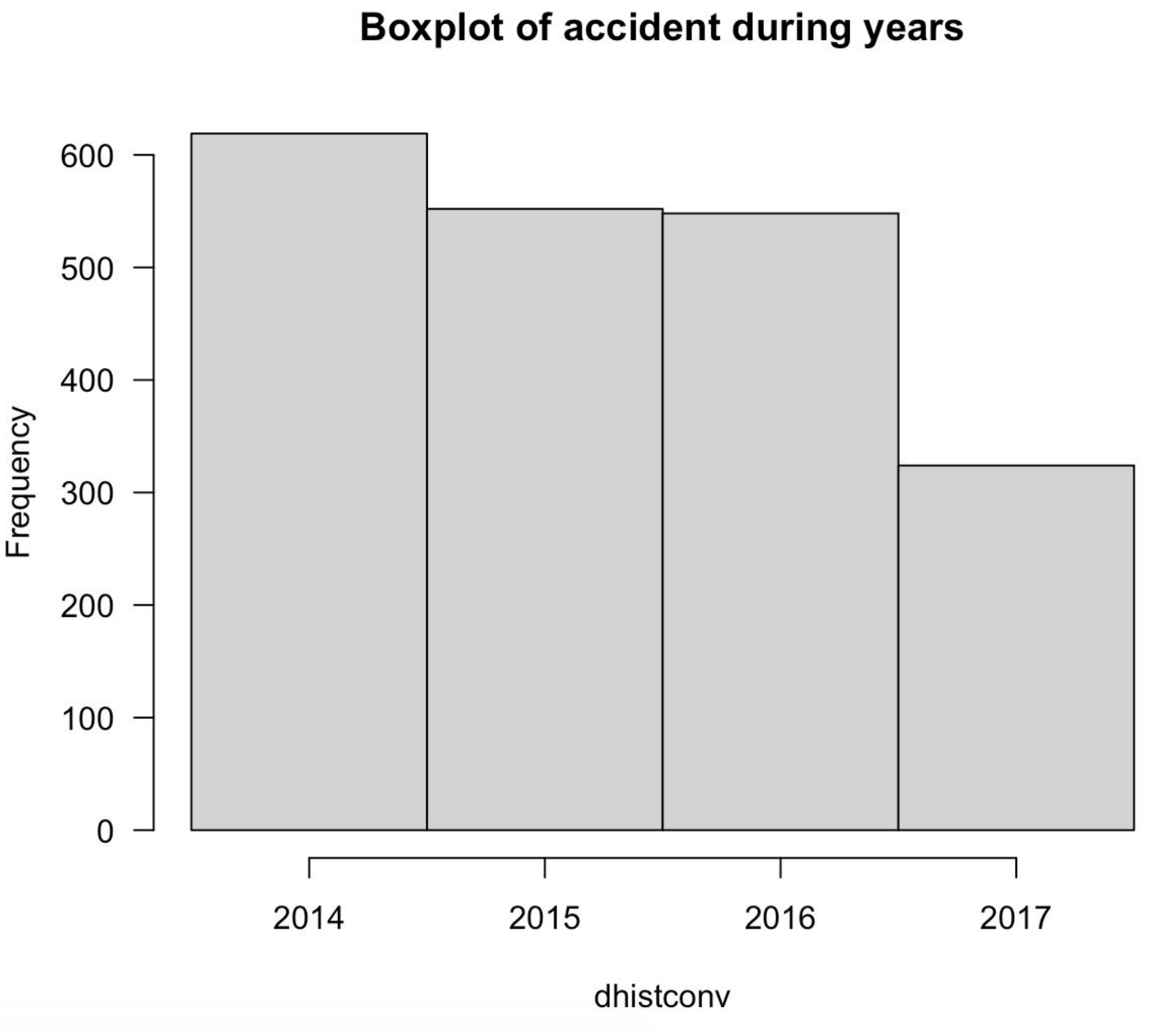
Another problem is that the data is not grouped by year, to do so we are gonna create another column that we're gonna fill with the first 4 characters of the date, later on we'll make it a numeric value and f then we'll be able to use the data to create the histogram that is gonna tell us how the casualties changed during this timeframe. (20)



```
### Histogram of accident during years ###
dhist <- substring(clean_accidents$Accident.Date,0,4)
View(dhist)
dhistconv <- as.numeric(dhist)
str(dhistconv)
hist(dhistconv, main="Boxplot of accident during years",
     breaks = 2013.5:2017.5,las=1, xlim=c(2013.5,2017.5))
```

Code used to manipulate dates in order to create an histogram.. (20)

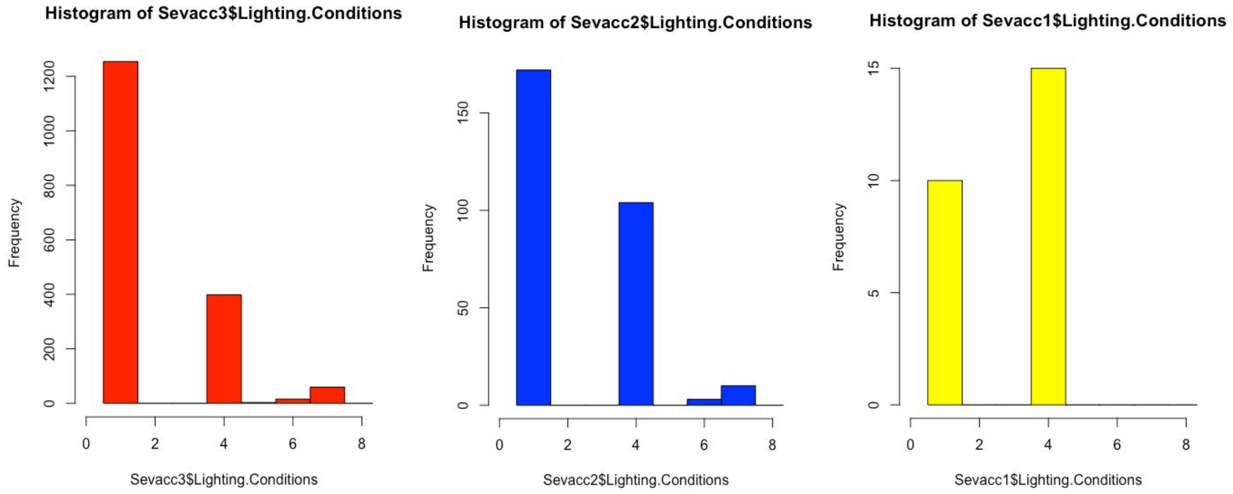
From the chart we can see a steadily decreasing number of accidents occurring each year. Since 2014 the number of accidents decreased for a ~47.5% till 2017 that counts just 325 cases compared to the 620 of 2014. (21)



Code used to manipulate dates in order to create an histogram. (21)

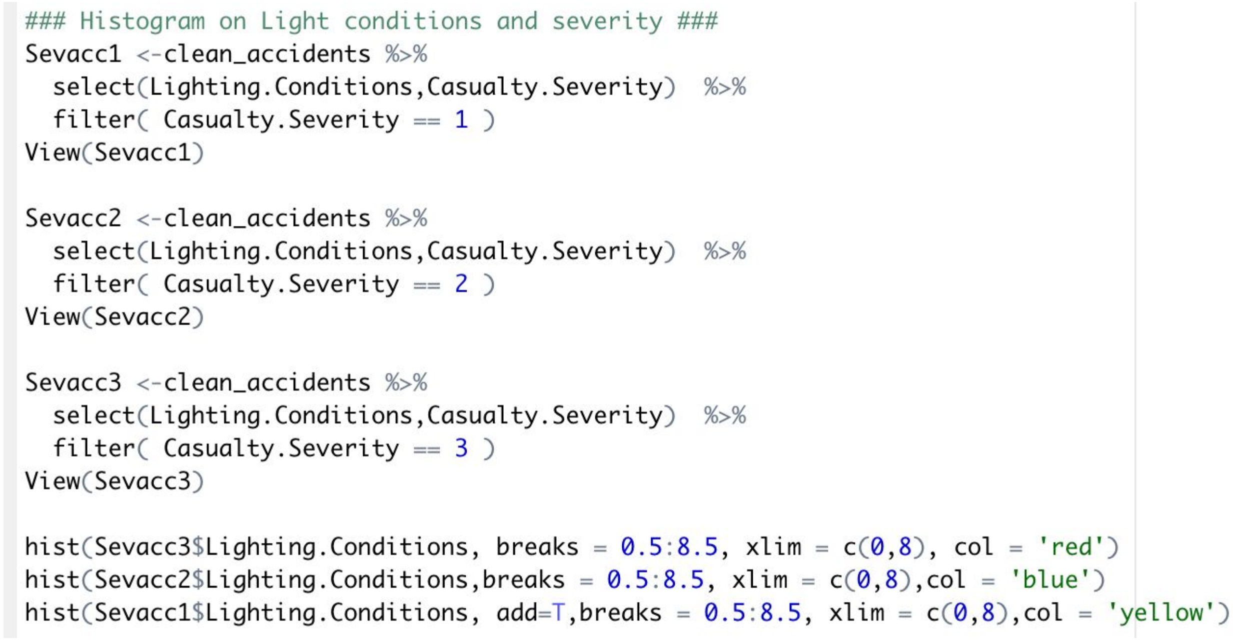
3 - Exploring the relationship between light condition and the severity of the casualty

I decided to use some histograms to explore the relationship between these two elements. (22)



Histograms for each severity (RED = 3 , BLUE = 2 , YELLOW = 1) . (22)

In the histograms we can see that as severity increases the number of its frequency decreases as it's easily noticeable from the various frequencies axis. We can also notice that severe cases happen only with conditions that have lights on (1-4) and none as people tend to be more aware when it's harder to see the road, this would also explain why most of the cases happen during the best condition to drive (1). As follows there is the raw code to reach the result. (23)

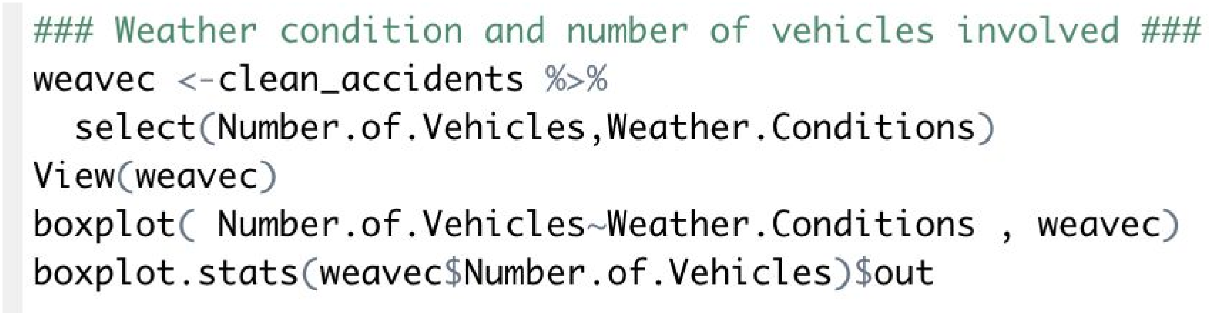


```
### Histogram on Light conditions and severity ###
Sevacc1 <-clean_accidents %>%
  select(Lighting.Conditions,Casualty.Severity)  %>%
  filter( Casualty.Severity == 1 )
View(Sevacc1)

Sevacc2 <-clean_accidents %>%
  select(Lighting.Conditions,Casualty.Severity)  %>%
  filter( Casualty.Severity == 2 )
View(Sevacc2)

Sevacc3 <-clean_accidents %>%
  select(Lighting.Conditions,Casualty.Severity)  %>%
  filter( Casualty.Severity == 3 )
View(Sevacc3)

hist(Sevacc3$Lighting.Conditions, breaks = 0.5:8.5, xlim = c(0,8), col = 'red')
hist(Sevacc2$Lighting.Conditions,breaks = 0.5:8.5, xlim = c(0,8),col = 'blue')
hist(Sevacc1$Lighting.Conditions, add=T,breaks = 0.5:8.5, xlim = c(0,8),col = 'yellow')
```

Raw code to create and print the histograms for each severity. (23)

4 - Exploring the relationship between Weather condition and number of vehicles involved in an accident

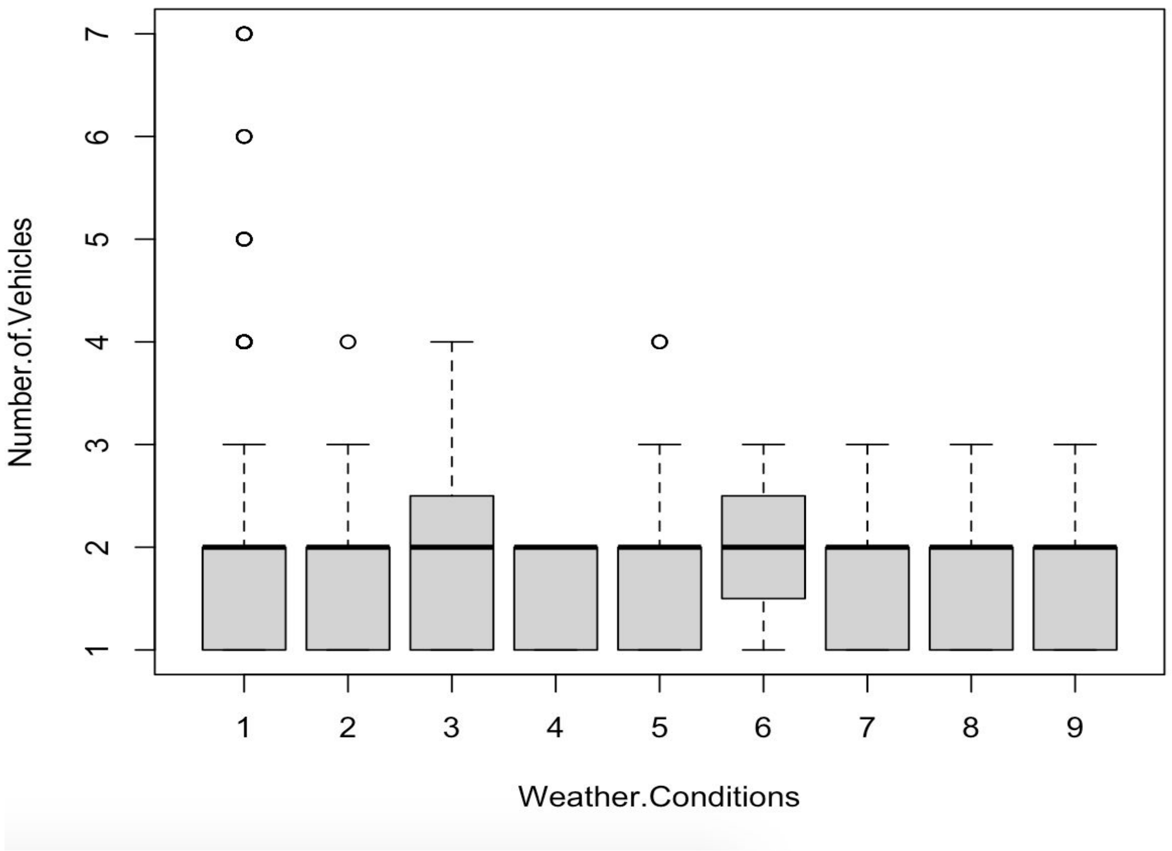
To explore this relationship I’ve decided to use a boxplot in order to find how many vehicles are usually involved based on the weather. (24)



```
### Weather condition and number of vehicles involved ###
weavec <-clean_accidents %>%
  select(Number.of.Vehicles,Weather.Conditions)
View(weavec)
boxplot( Number.of.Vehicles~Weather.Conditions , weavec)
boxplot.stats(weavec$Number.of.Vehicles)$out
```

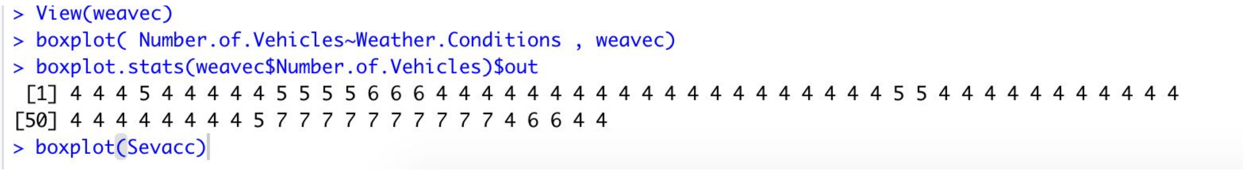
Code used to create a boxplot for the correlation between weather and vehicles. (24)

Looking at the boxplot we can see that usually every event is more likely to involve at least 2 vehicles or less, for snowing conditions (2-6) it's more possible to reach a higher number of cars or bikes involved. Finally a fine weather condition (1) it’s the only case where the involved vehicles could reach 4+. (25)



Boxplot for the correlation between weather and vehicles. (25)

We should also notice that the most of the outliers are from the first weather condition (1) and the most frequent value of vehicles involved are 4. (26)



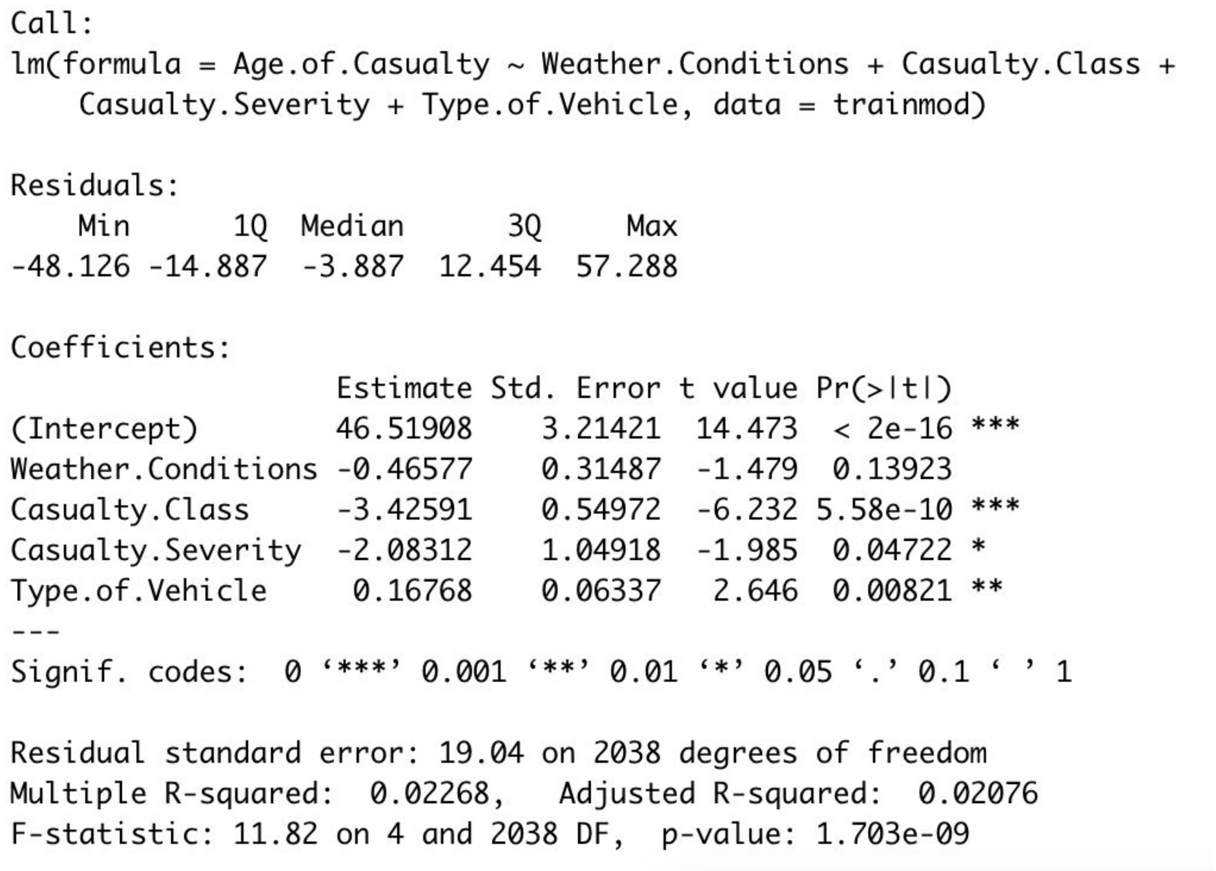
```
> View(weavec)
> boxplot( Number.of.Vehicles~Weather.Conditions , weavec)
> boxplot.stats(weavec$Number.of.Vehicles)$out
 [1] 4 4 4 5 4 4 4 4 4 5 5 5 5 6 6 6 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 4 4 4 4 4 4 4 4 4 4 4 4
[50] 4 4 4 4 4 4 4 4 5 7 7 7 7 7 7 7 7 7 7 4 6 6 4 4
> boxplot(Sevacc)
```

Raw code for the correlation between weather and vehicles and it’s outliers. (26)

Regression

Training and imputing new values using linear regression

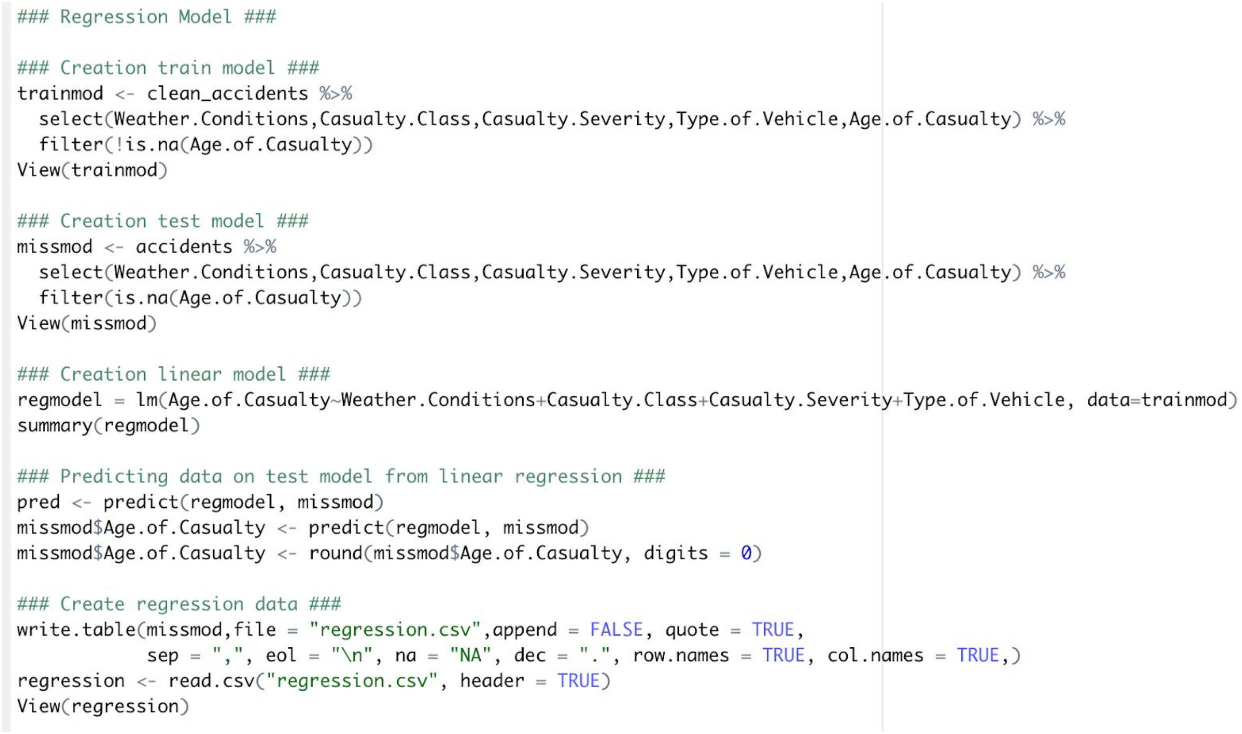
Linear regression helps visualize the relationship between variables, so it can be used to predict values. After splitting our data into a Training set and a Test set, I created the linear model based on the four columns of our choice and printed the summary to have a better understanding of the model. (27) (28)



```
Call:
lm(formula = Age.of.Casualty ~ Weather.Conditions + Casualty.Class + 
    Casualty.Severity + Type.of.Vehicle, data = trainmod)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.126 -14.887  -3.887  12.454  57.288 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        46.51908    3.21421  14.473  < 2e-16 ***
Weather.Conditions -0.46577    0.31487  -1.479  0.13923    
Casualty.Class     -3.42591    0.54972  -6.232 5.58e-10 ***
Casualty.Severity  -2.08312    1.04918  -1.985  0.04722 *  
Type.of.Vehicle     0.16768    0.06337   2.646  0.00821 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.04 on 2038 degrees of freedom
Multiple R-squared:  0.02268,	Adjusted R-squared:  0.02076 
F-statistic: 11.82 on 4 and 2038 DF,  p-value: 1.703e-09
```

Summary of the Linear Model. (27)

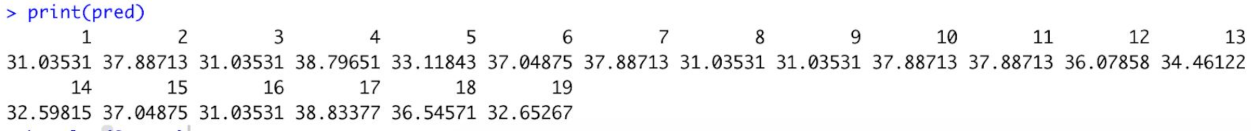
We can see that some elements aren’t that significant and makes our linear model not the best one to predict possible values in the dataframe as it’s R*2 values are quite low.



```
### Regression Model ###

### Creation train model ###
trainmod <- clean_accidents %>%
  select(Weather.Conditions,Casualty.Class,Casualty.Severity,Type.of.Vehicle,Age.of.Casualty) %>%
  filter(!is.na(Age.of.Casualty))
View(trainmod)

### Creation test model ###
missmod <- accidents %>%
  select(Weather.Conditions,Casualty.Class,Casualty.Severity,Type.of.Vehicle,Age.of.Casualty) %>%
  filter(is.na(Age.of.Casualty))
View(missmod)

### Creation linear model ###
regmodel = lm(Age.of.Casualty~Weather.Conditions+Casualty.Class+Casualty.Severity+Type.of.Vehicle, data=trainmod)
summary(regmodel)

### Predicting data on test model from linear regression ###
pred <- predict(regmodel, missmod)
missmod$Age.of.Casualty <- predict(regmodel, missmod)
missmod$Age.of.Casualty <- round(missmod$Age.of.Casualty, digits = 0)

### Create regression data ###
write.table(missmod,file = "regression.csv",append = FALSE, quote = TRUE,
            sep = ",", eol = "\n", na = "NA", dec = ".", row.names = TRUE, col.names = TRUE,)
regression <- read.csv("regression.csv", header = TRUE)
View(regression)
```

Raw code for all the regression model. (28)

Once the data has been predicted we can see that the values stay near the average of the values in the age of the casualty column and aren’t rounded as they should. After inserting the value we create the regression file and we can notice that the data predicted aren’t perfectly fit for the real world scenario as there could be someone with a different age behind the wheel. (29)



```
> print(pred)
       1        2        3        4        5        6        7        8        9       10       11       12       13 
31.03531 37.88713 31.03531 38.79651 33.11843 37.04875 37.88713 31.03531 31.03531 37.88713 37.88713 36.07858 34.46122 
      14       15       16       17       18       19 
32.59815 37.04875 31.03531 38.83377 36.54571 32.65267 
```

Predicted data from the trained linear model. (29)

Conclusion

The conclusions we were able to draw from this data set are that overall, crashes in the Calderdale area declined over the years and the majority of accidents results in only slight injuries with all types of casualties happening mostly in normal conditions. In snowy and high winds weather conditions we can see the highest average number of vehicles involved in the crash and finally, women are less likely to crash than men when driving.