Assignment 1 - EISE, Data Visualization and Exploration - Erik Skipper

Assignment

Utilizing the visualization methods we learned in class to develop hypotheses about what actions the FRA could take to reduce the severity of rail accidents, the code (and associated graphs) in this `.RMD` file demonstrates the following:

1. 1 (a) 6 severity metrics the FRA should consider in evaluating their safety regulations
2. 1 (b) 2 metrics to describe accident severity
3. 2 (a) the current accident situation in terms of both frequency of accidents and severity of accidents from 2001-2019.
4. 2 (b) the contributors to accident severity in the extreme accidents data using the multivariate visualization techniques discussed in class
5. 3 (a) 2 well-formed, actionable hypotheses for each severity metric
6. BONUS

Setup

```
traindir <- "~/Documents/UVA/APMA 6430/Data/"
sourcedir <-"~/Documents/UVA/APMA 6430/R"
setwd(sourcedir)
source("AccidentInput.R")
source("SPM_Panel.R")
library(ggplot2)
library(lattice)
acts <- file.inputl(traindir)
totacts <- combine.data(acts)
```

Rather than sourcing the `.R` files and libraries within the code block they are associated with, I decided to perform this action at the beginning.

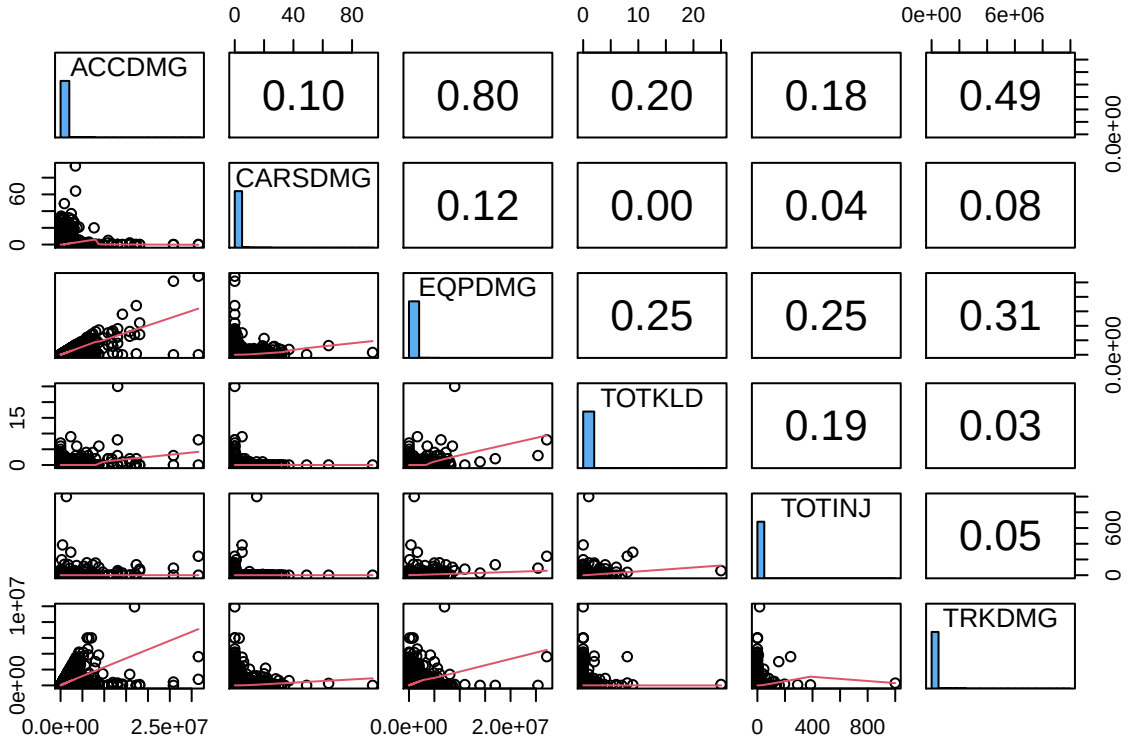
1. (a)

The 6 severity metrics the FRA should consider in evaluating their safety regulations are as follows:

- ACCDMG
- CARSDMG
- EQPDMG

- TOTKLD
- TOTINJ
- TRKDMG

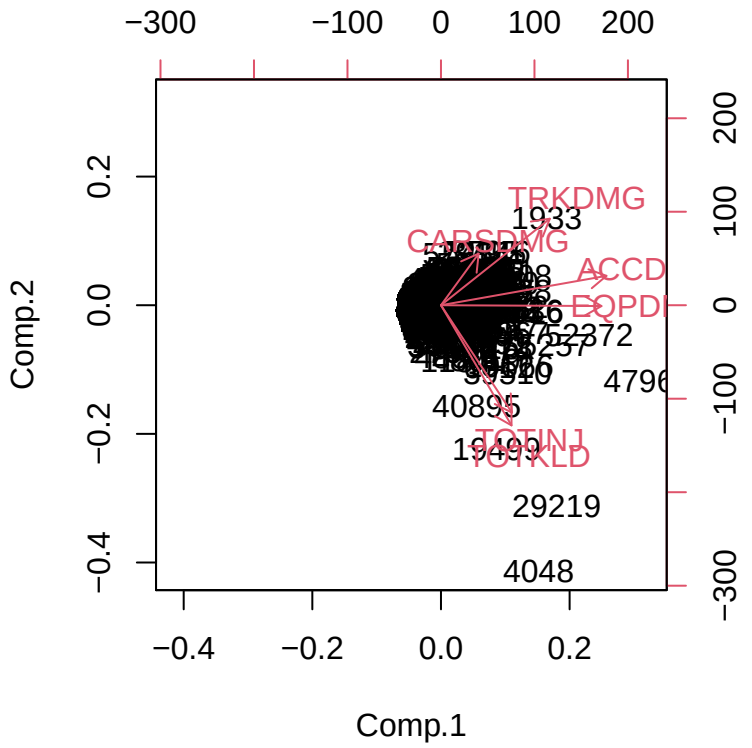
```
setwd(sourcedir)
# source("SPM_Panel.R")
uva.pairs(totacts[,c("ACCDMG", "CARSDMG", "EQPDMG", "TOTKLD", "TOTINJ", "TRKDMG")])
```



The 6 severity metrics were selected based on the code above, which demonstrates the relationship between each of the 6 severity metrics.

1. (b)

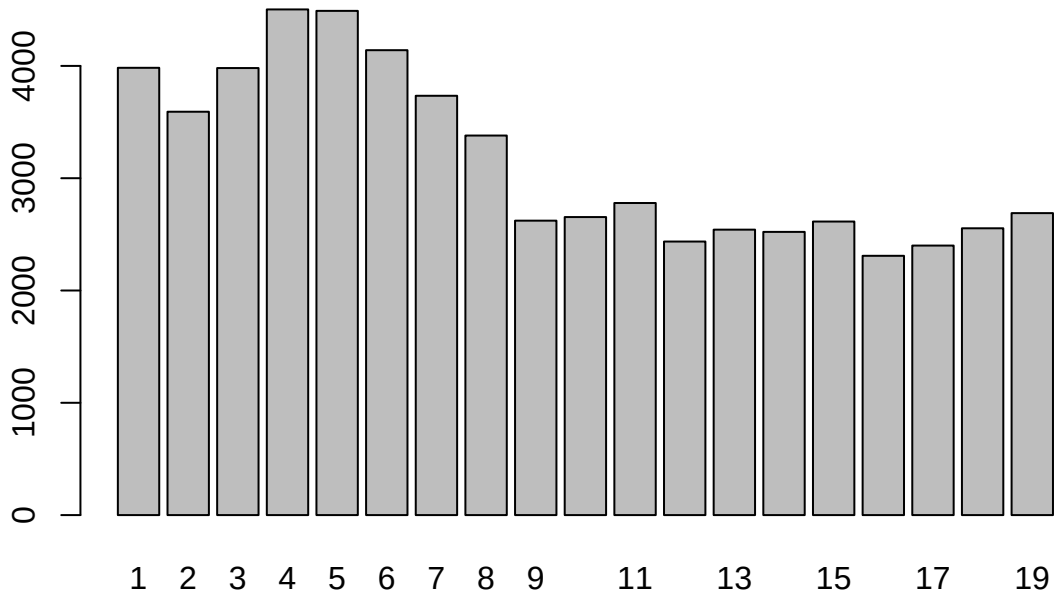
```
totactsnd <- totacts
totactsnd.pca <- princomp(totactsnd[,c("ACCDMG", "CARSDMG", "EQPDMG", "TOTKLD", "TOTINJ", "TRKDMG")], c
biplot(totactsnd.pca)
```



The biplot from the code above demonstrates the relationship between each of the 6 severity metrics using PCA. The length of each vector indicates the importance of the severity metric that it represents, and the direction in which the vector is pointed relative to the other vectors indicates the correlation of the severity metric with the other severity metrics. Based on the graph, we can see that ACCDMG, EQPDMG, TOTKLD, and TOTINJ appear to be the longest, meaning they have a high degree of importance. We can also see that ACCDMG and EQPDMG appear close to each other, meaning there is a high degree of correlation between them; similarly, TOTKLD and TOTINJ appear close to each other, meaning there is a high degree of correlation between them as well. Since ACCDMG and EQPDMG, and TOTINJ and TOTKLD, are so close to each other, we can choose two metrics by down-selecting from them both: ACCDMG and TOTKLD. ACCDMG and TOTKLD are roughly orthogonal/perpendicular to each other, meaning there is a degree of variability between the 2 severity metrics.

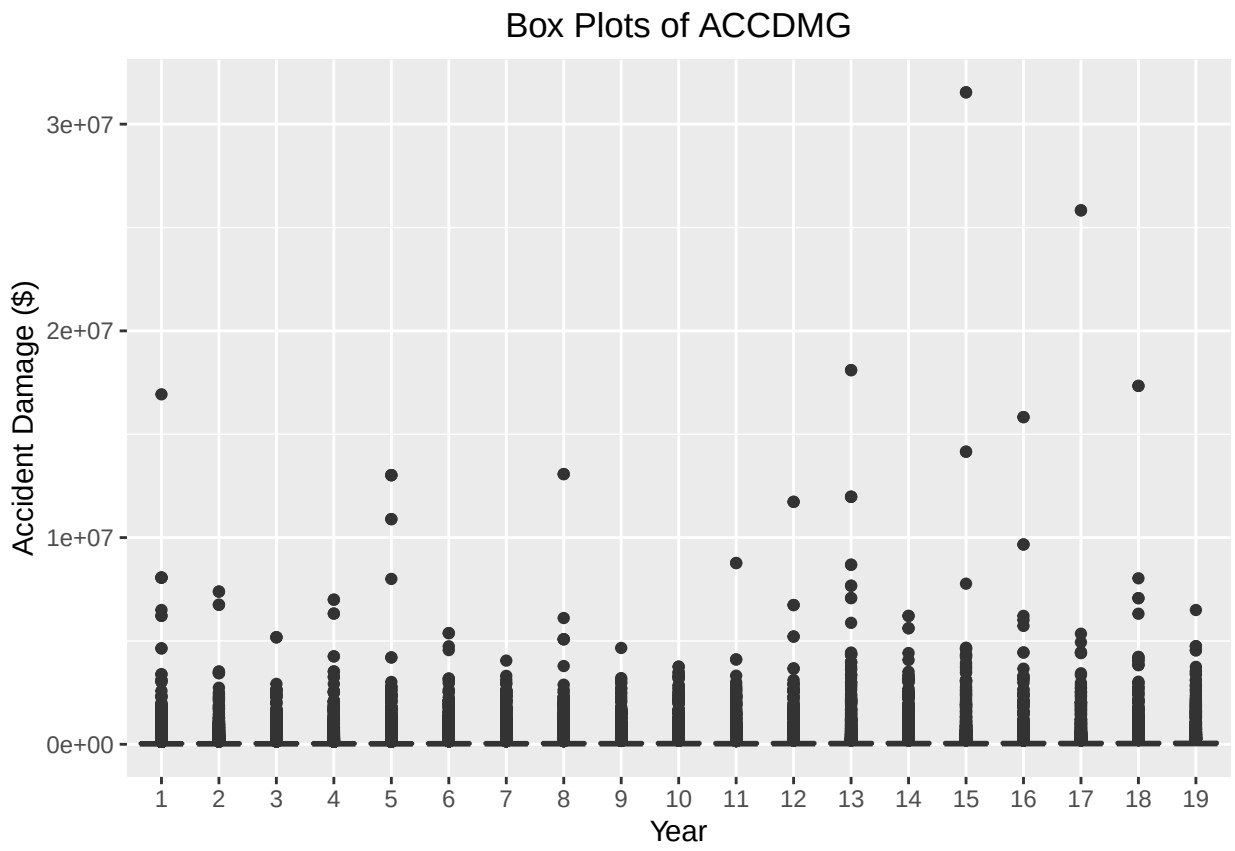
2. (a)

```
barplot(table(totacts$YEAR))
```



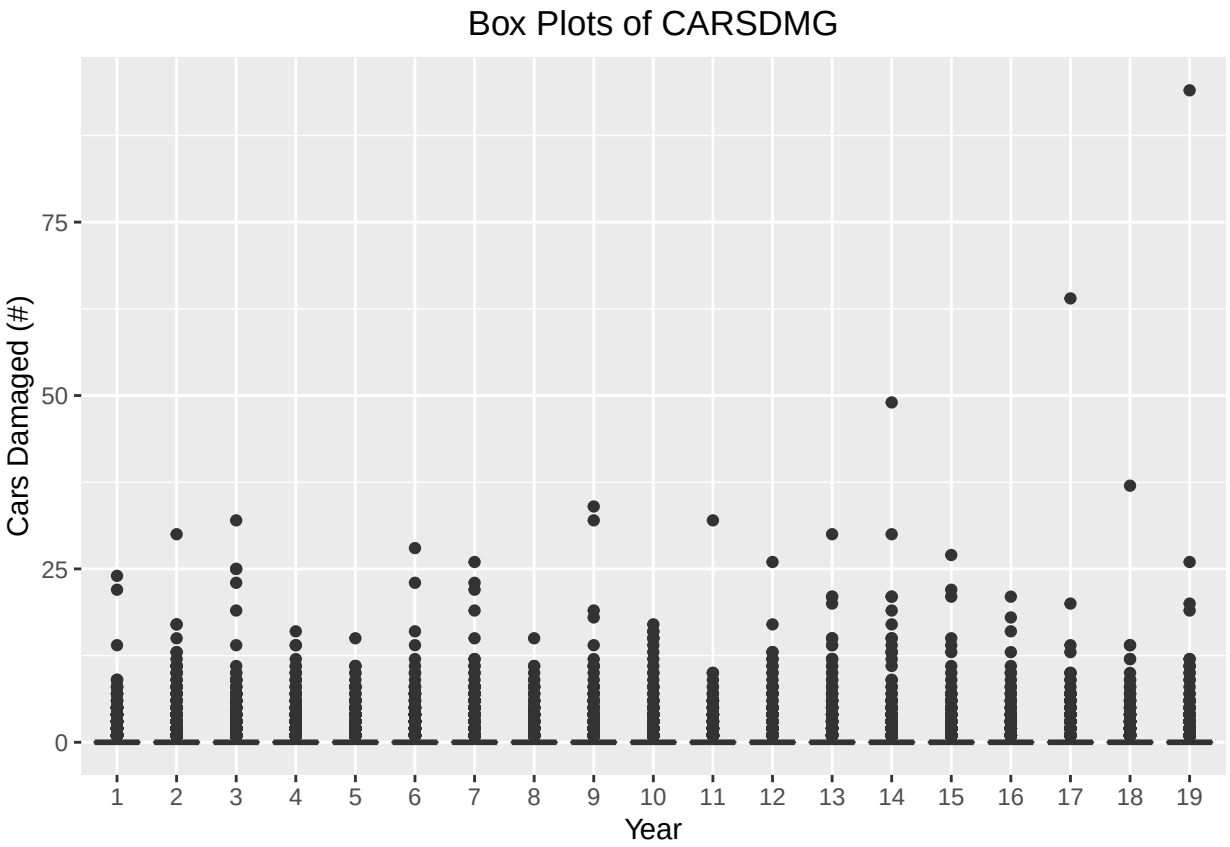
The above code shows a bar plot of the frequency of accidents per year over 19 years.

```
# library(ggplot2)
# library(lattice)
ggplot(data = totacts, aes(x = as.factor(YEAR), y = ACCDMG)) +
  geom_boxplot() +
  # coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of ACCDMG") +
  labs(x = "Year", y = "Accident Damage ($)")
```



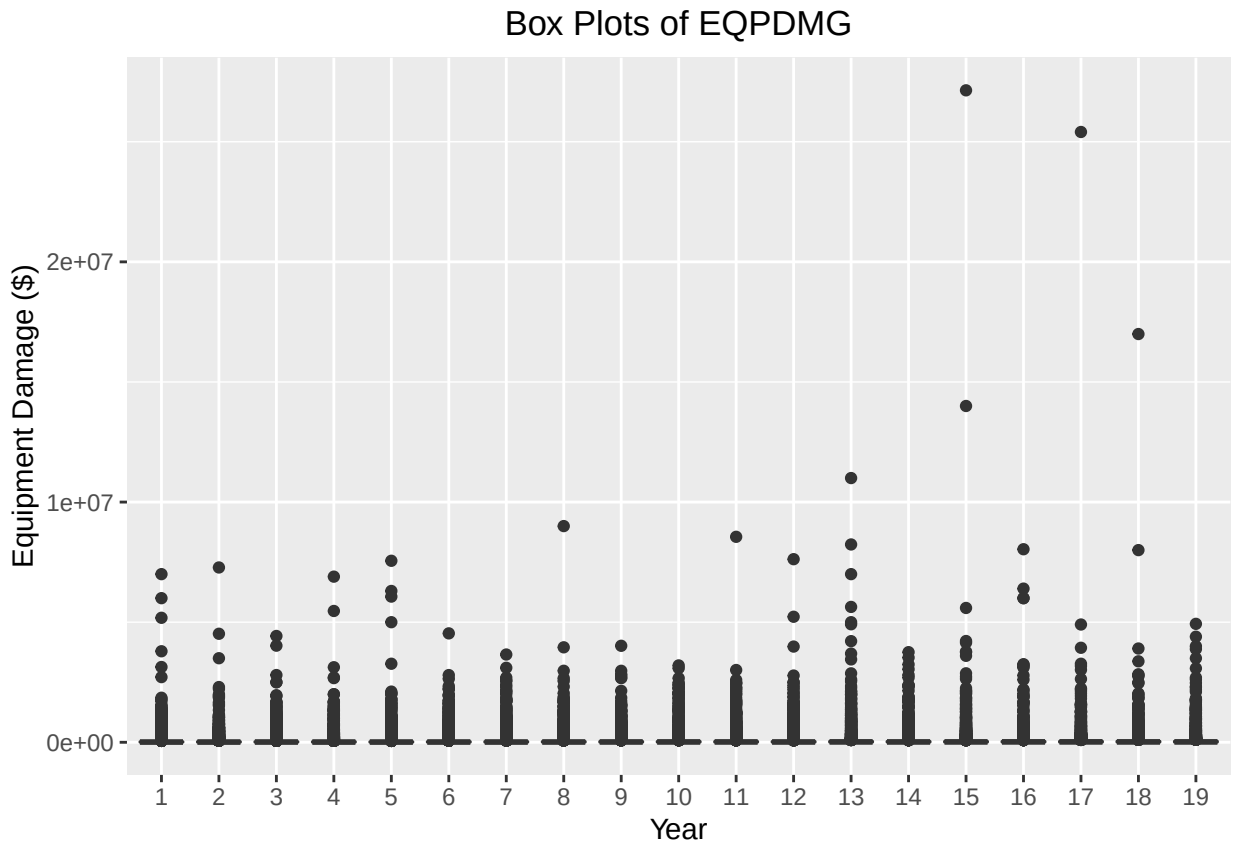
The above code shows a box plot of the cost of accidents per accident over 19 years.

```
ggplot(data = totacts, aes(x = as.factor(YEAR), y = CARSDMG)) +
  geom_boxplot() +
  # coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of CARSDMG") +
  labs(x = "Year", y = "Cars Damaged (#)")
```



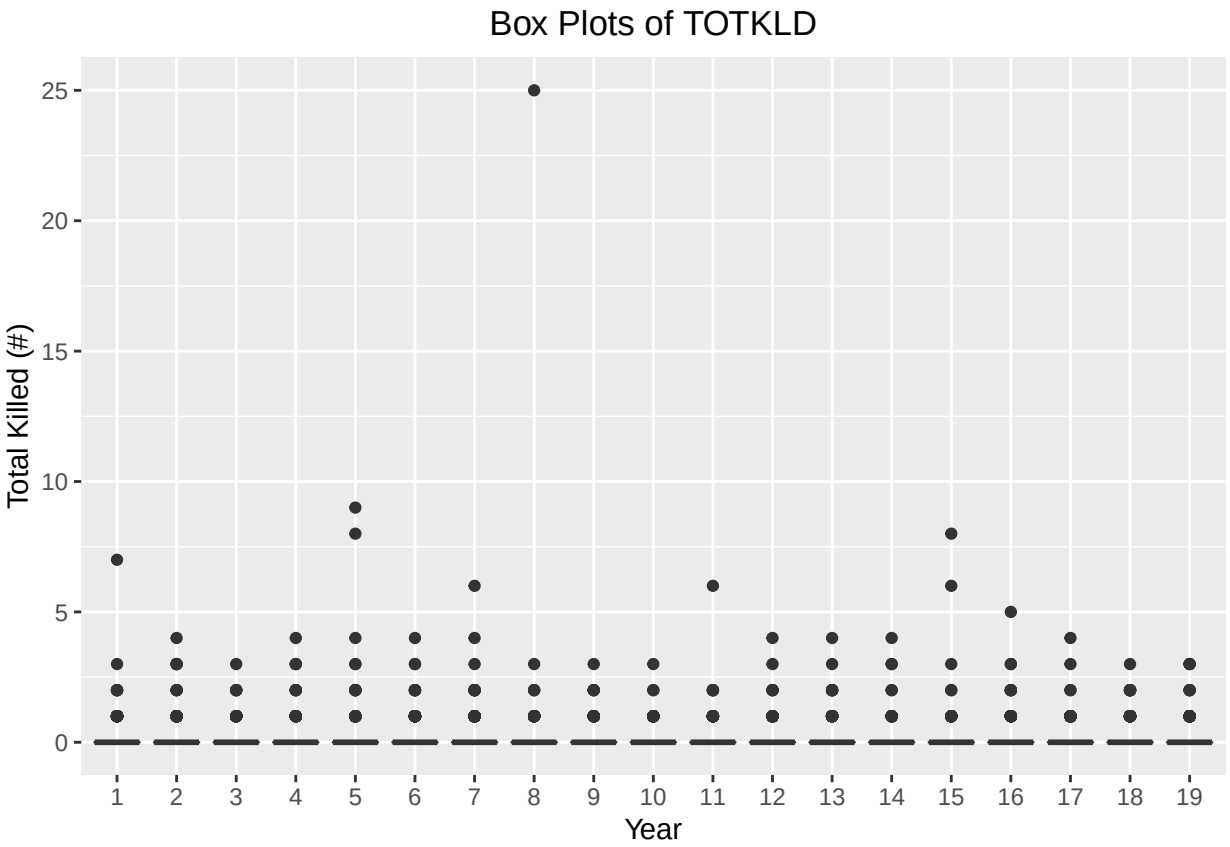
The above code shows a box plot of the number of cars damaged per accident over 19 years.

```
ggplot(data = totacts, aes(x = as.factor(YEAR), y = EQPDMG)) +
  geom_boxplot() +
  # coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of EQPDMG") +
  labs(x = "Year", y = "Equipment Damage ($)")
```



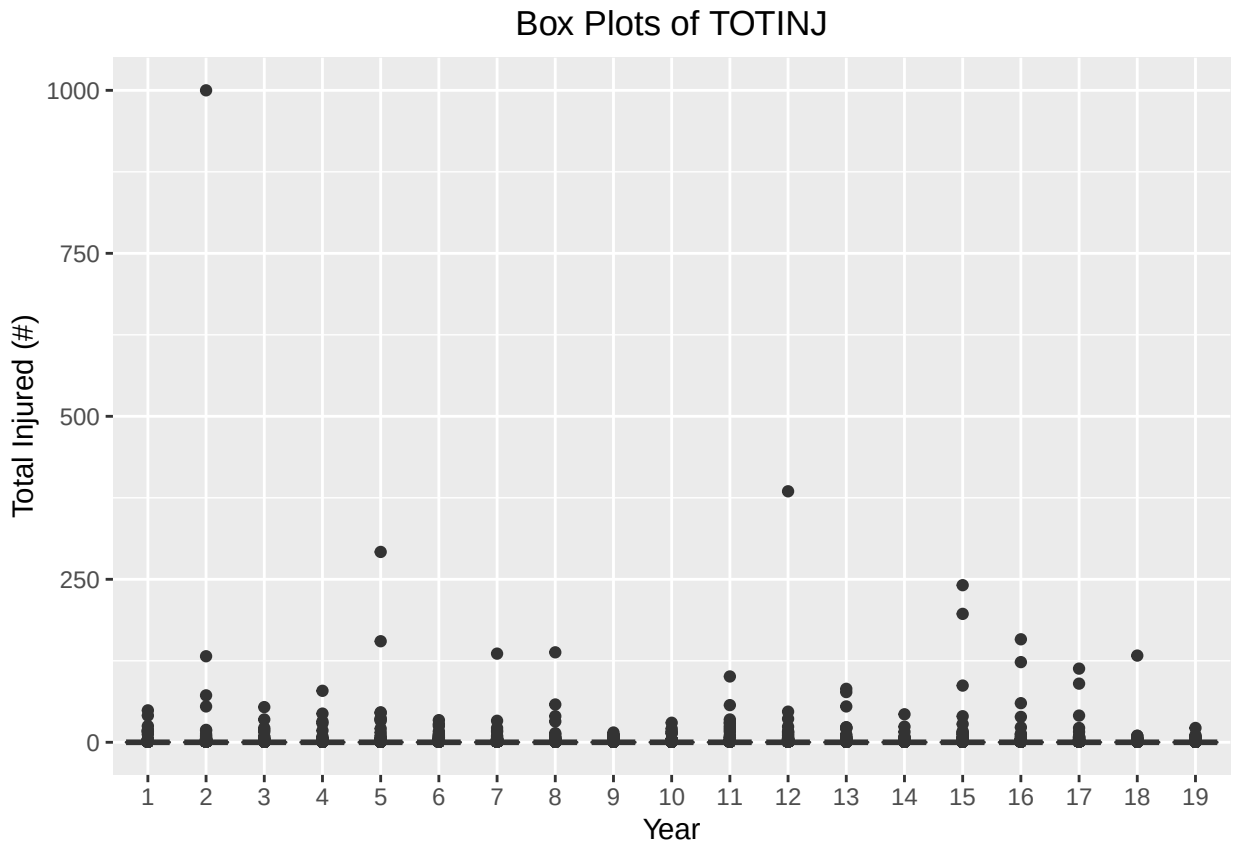
The above code shows a box plot of the cost of equipment damage per accident over 19 years.

```
ggplot(data = totacts, aes(x = as.factor(YEAR), y = TOTKLD)) +
  geom_boxplot() +
  # coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of TOTKLD") +
  labs(x = "Year", y = "Total Killed (#)")
```



The above code shows a box plot of the total number of people killed per accident over 19 years.

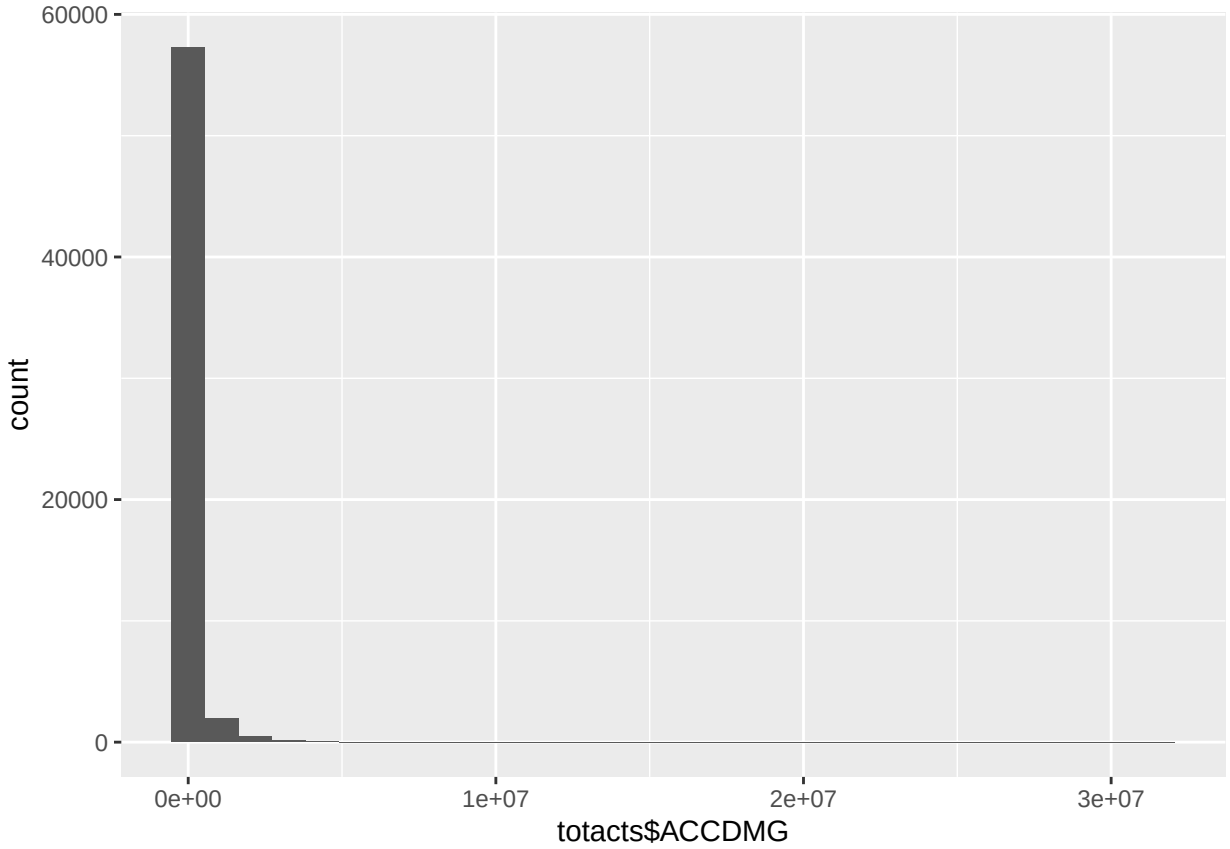
```
ggplot(data = totacts, aes(x = as.factor(YEAR), y = TOTINJ)) +
  geom_boxplot() +
  # coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of TOTINJ") +
  labs(x = "Year", y = "Total Injured (#)")
```



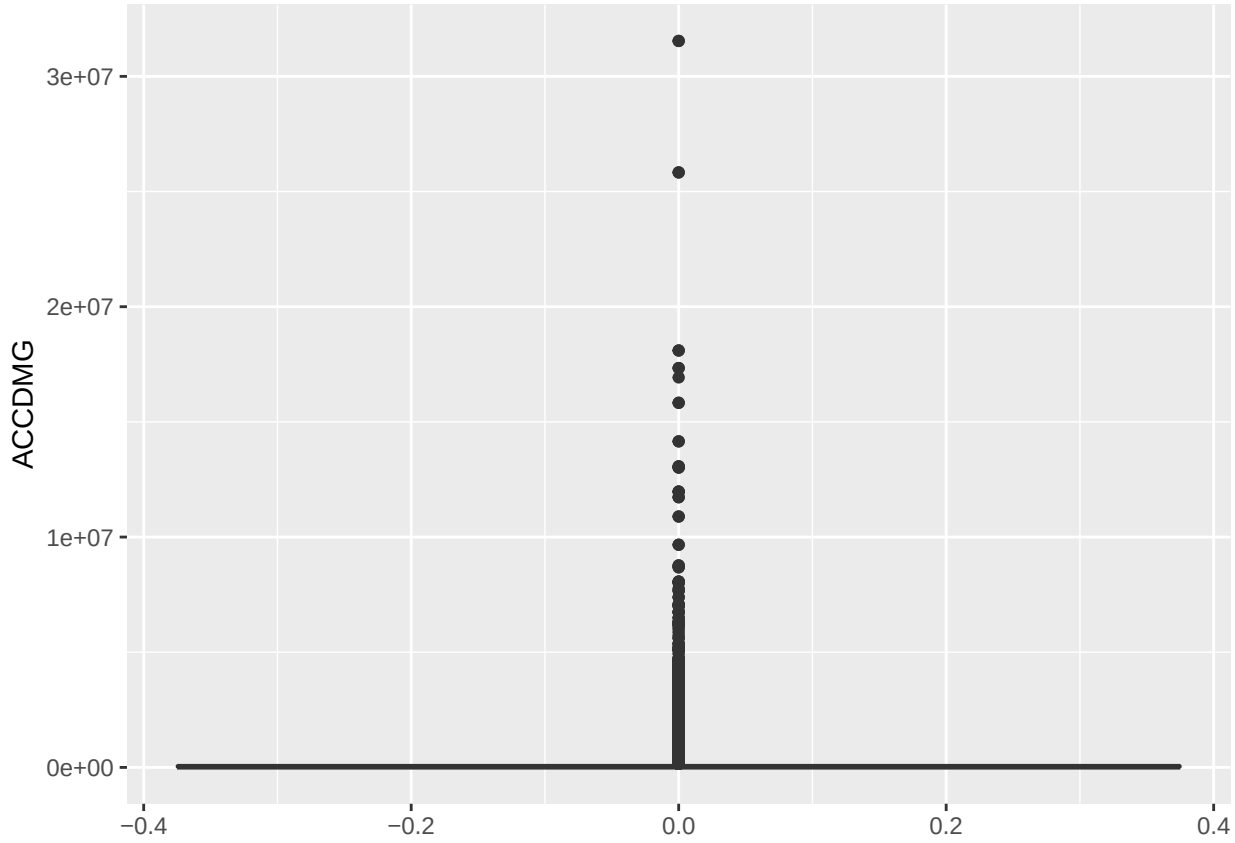
The above code shows a box plot of the total number of people injured per accident over 19 years.

```
# library(ggplot2)
ggplot(as.data.frame(totacts$ACCDMG), aes(x=totacts$ACCDMG)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
dmgbox <- ggplot(totacts, aes(y=ACCDMG)) + geom_boxplot()
dmgbox
```




```
# names(ggplot_build(dmgbox)$data[[1]])
upper <- ggplot_build(dmgbox)$data[[1]]$ymax
xdmg <- totacts[totacts$ACCDMG > upper,]
# Number of Outliers
nrow(xdmg)
```

```
## [1] 7888
```

```
# Proportion of Outliers
nrow(xdmg)/nrow(totacts)
```

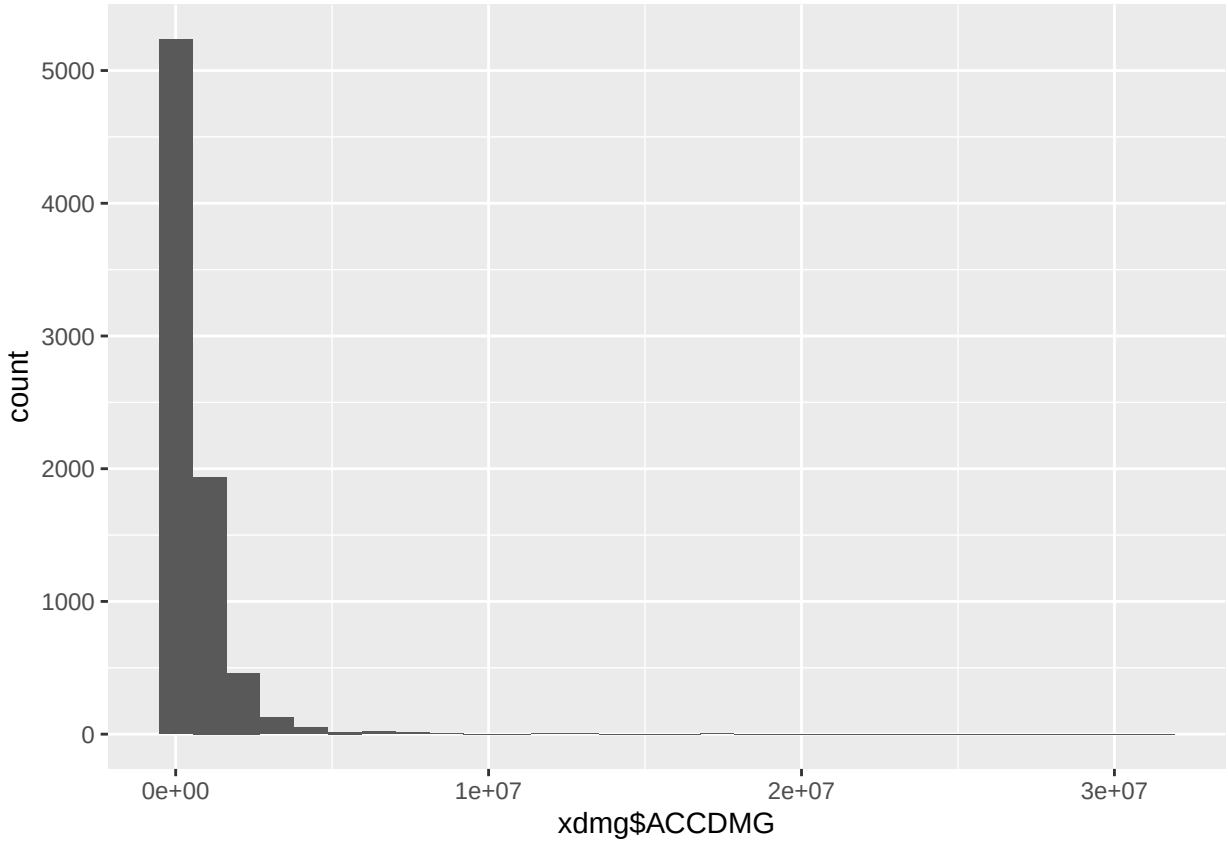
```
## [1] 0.1316334
```

```
# Proportion of Costs
sum(as.numeric(totacts$ACCDMG[which(totacts$ACCDMG > ggplot_build(dmgbox)$data[[1]]$ymax)]))/sum(as.num
```

```
## [1] 0.747106
```

```
ggplot(as.data.frame(xdmg$ACCDMG), aes(xdmg$ACCDMG)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Current accident situation:

The graphs above show that the overall frequency of accidents has been decreasing since 2001, even though there are some 3-4 year periods (e.g., 2016-2019) where the frequency of accidents has been increasing. However, from a cost perspective, the cost of accidents since 2001 appears to have no trend in either direction.

Why the analysis should focus on the extreme accidents:

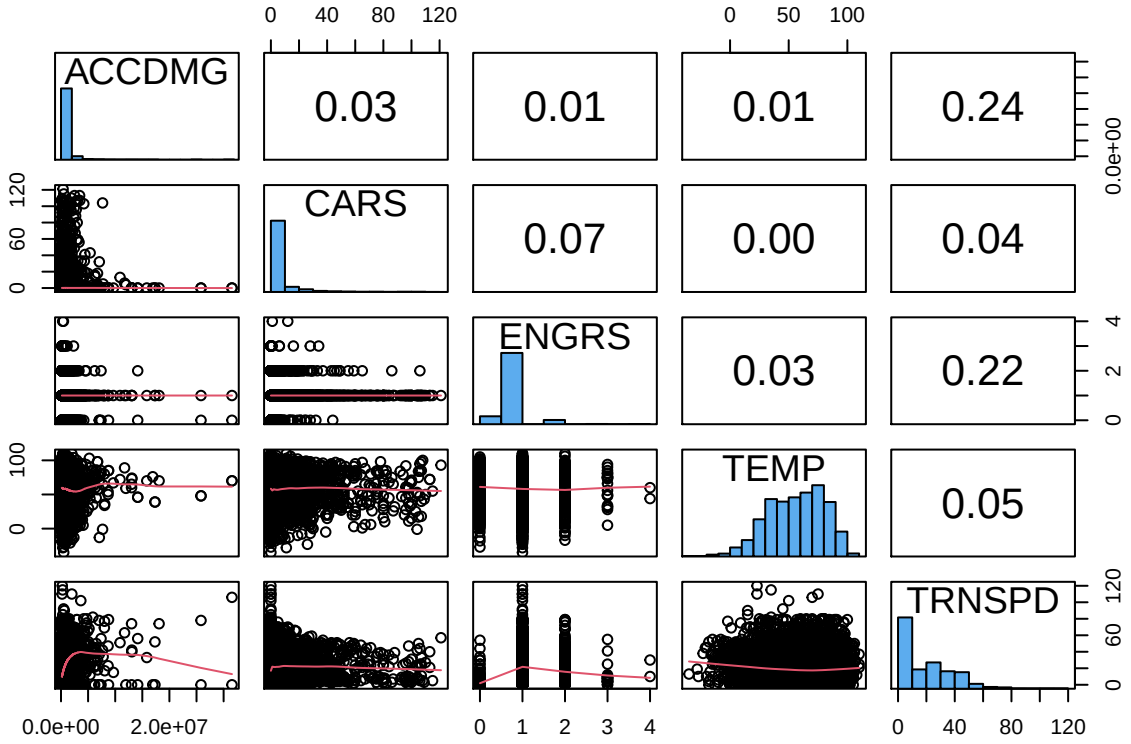
The analysis should focus on the extreme accidents for two reasons: 1) the proportion of outliers relative to the total number of accidents and 2) the proportion of costs relative to the total costs of accidents. That is, 13% of the total number of accidents accounts for nearly 75% of the total costs of accidents.

2. (b)

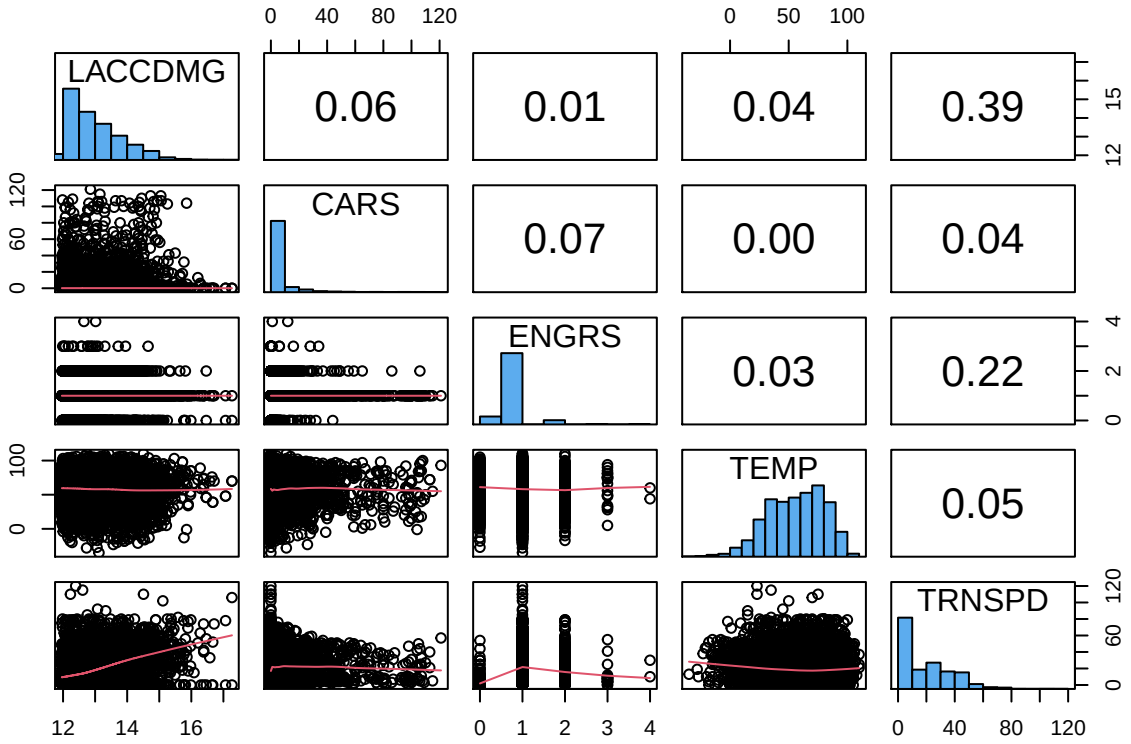
```r
# Remove the duplicates
xdmgnd <- xdmg[!(duplicated(xdmg[, c("INCDTNO", "YEAR", "MONTH", "DAY", "TIMEHR", "TIMEMIN")])),]

# Accident Cause
xdmgnd$Cause <- rep(NA, nrow(xdmgnd))
xdmgnd$Cause[which(substr(xdmgnd$CAUSE, 1, 1) == "M")] <- "M"
xdmgnd$Cause[which(substr(xdmgnd$CAUSE, 1, 1) == "T")] <- "T"
xdmgnd$Cause[which(substr(xdmgnd$CAUSE, 1, 1) == "S")] <- "S"
```

```
xdmgnd$Cause[which(substr(xdmgnd$CAUSE, 1, 1) == "H")] <- "H"
xdmgnd$Cause[which(substr(xdmgnd$CAUSE, 1, 1) == "E")] <- "E"

# Cause = factor
xdmgnd$Cause <- factor(xdmgnd$Cause)
```

```
uva.pairs(xdmgnd[,c("ACCDMG", "CARS", "ENGRS", "TEMP", "TRNSPD")])
```



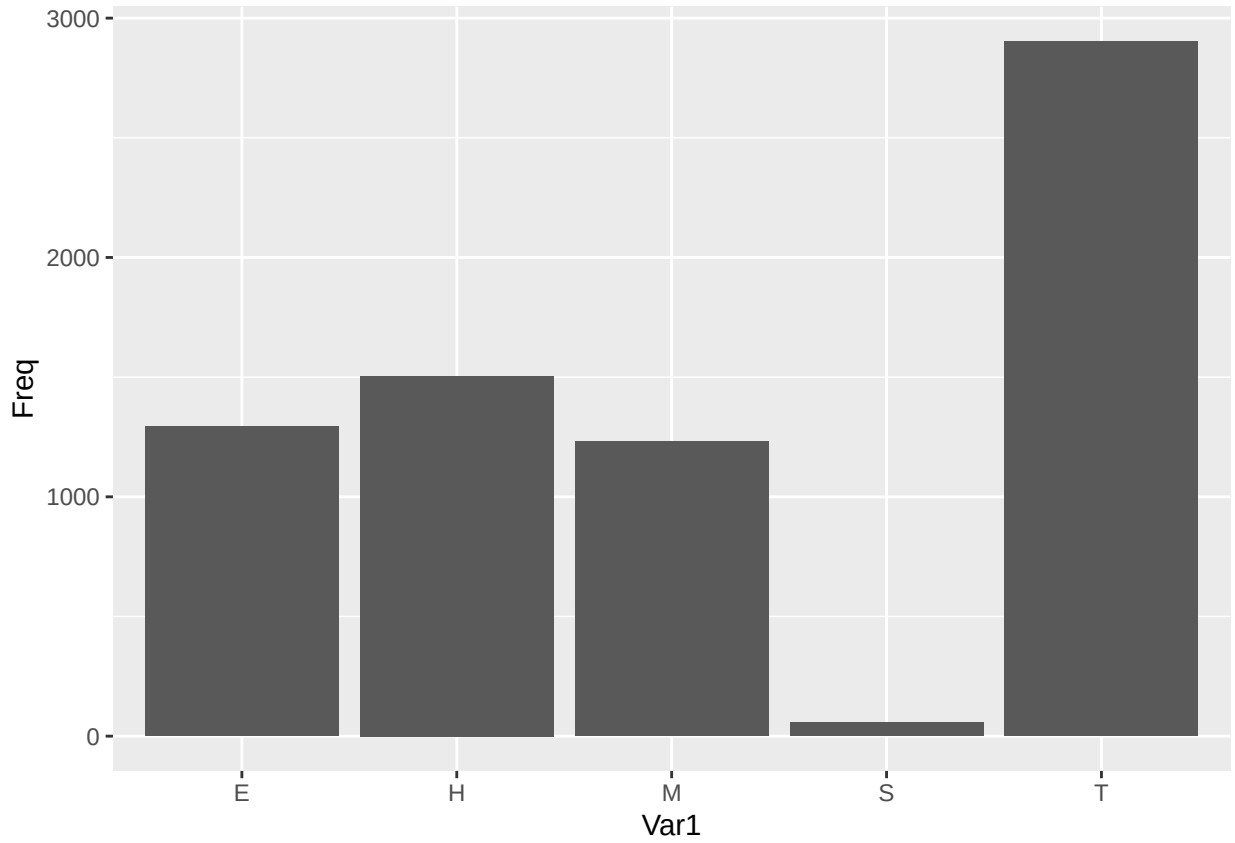

```
xdmgnd$LACCDMG <- log(xdmgnd$ACCDMG)
uva.pairs(xdmgnd[,c("LACCDMG", "CARS", "ENGRS", "TEMP", "TRNSPD")])
```



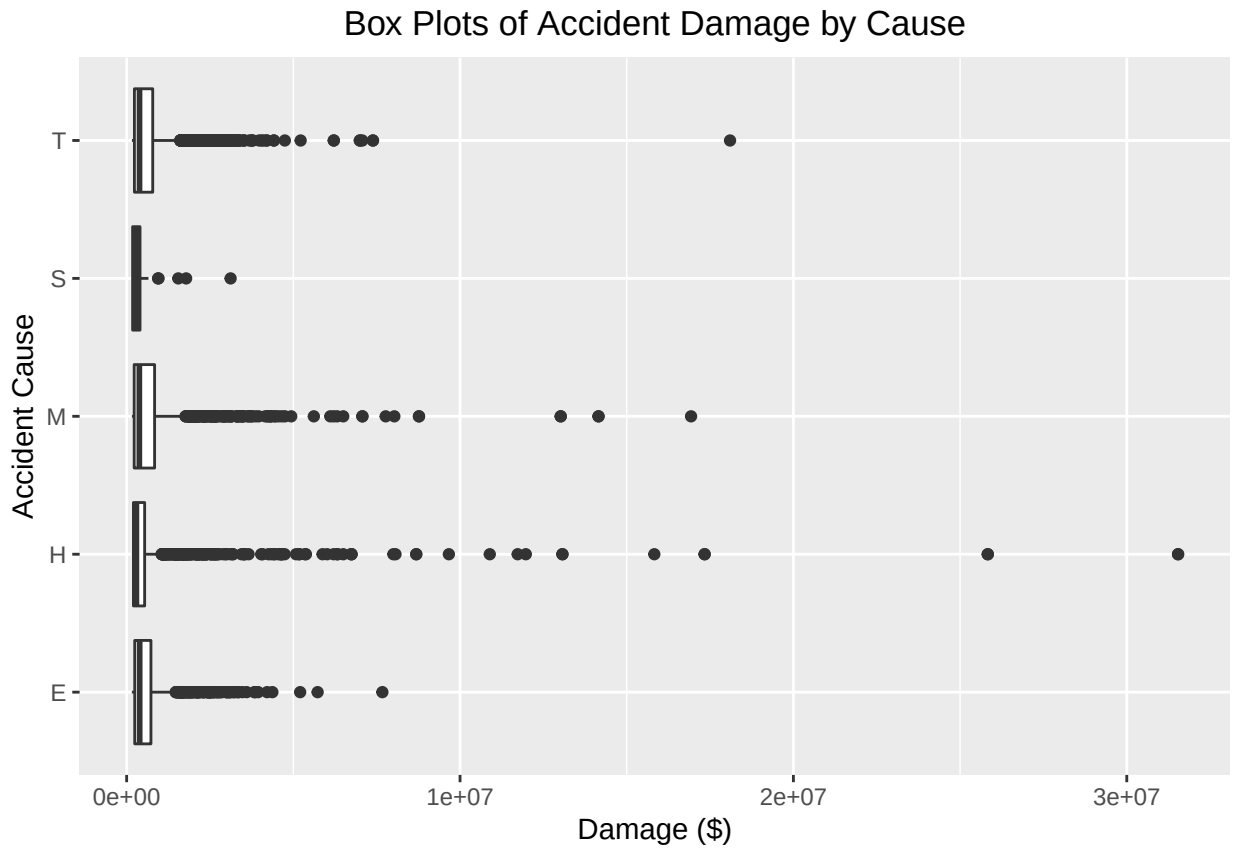
The code above shows two graphs:

1. A linear scctter plot matrix that shows the relationships between the selected metrics.
2. A logarithmic scatter plot matrix that shows the relationship between the selected metrics.

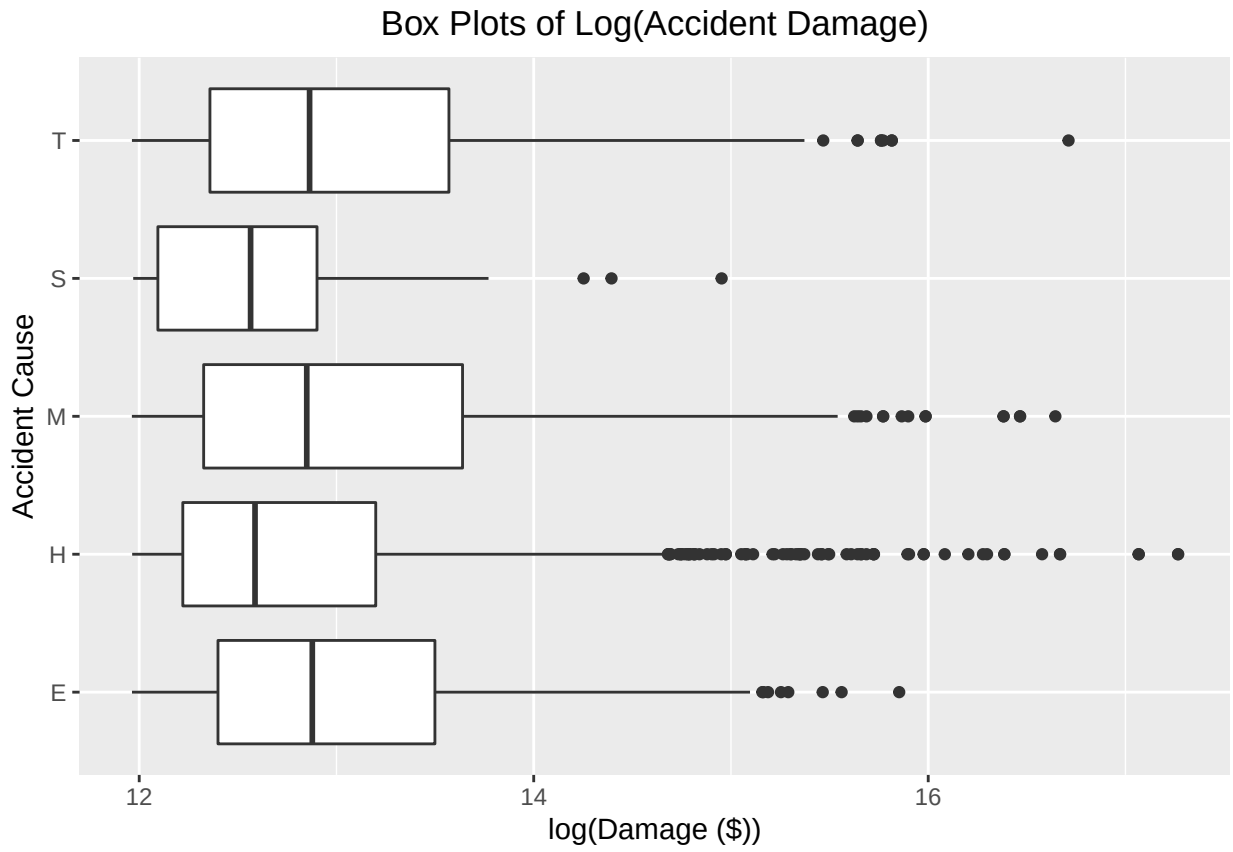
```
ggplot(as.data.frame(table(xdmgnd$Cause)), aes(x = Var1, y= Freq)) + geom_bar(stat="identity")
```




```
ggplot(data = xdmgnd, aes(x = Cause, y = ACCDMG)) +
  geom_boxplot() +
  coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of Accident Damage by Cause") +
  labs(x = "Accident Cause", y = "Damage ($)")
```



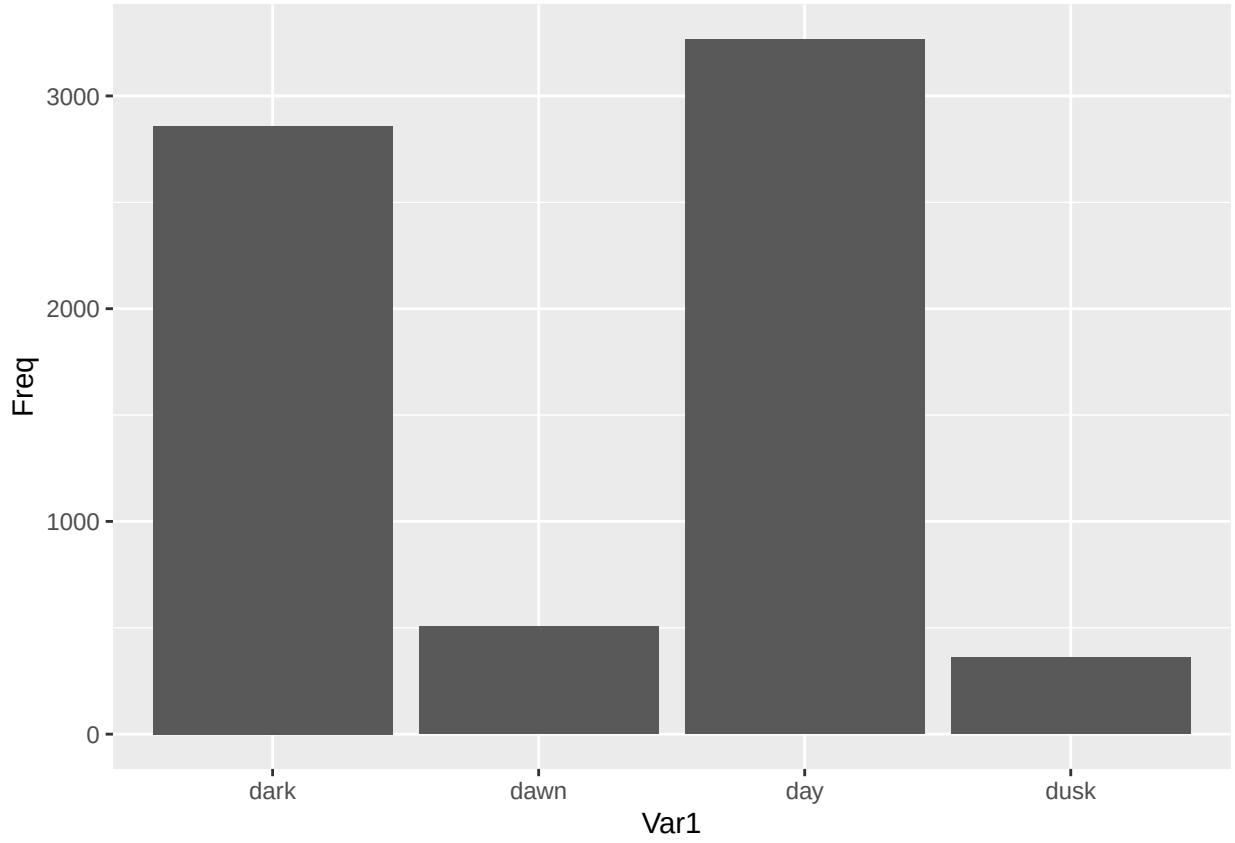

```
ggplot(data = xdmgnd, aes(x = Cause, y = log(ACCDMG+1))) +
  geom_boxplot() +
  coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of Log(Accident Damage)") +
  labs(x = "Accident Cause", y = "log(Damage ($))")
```



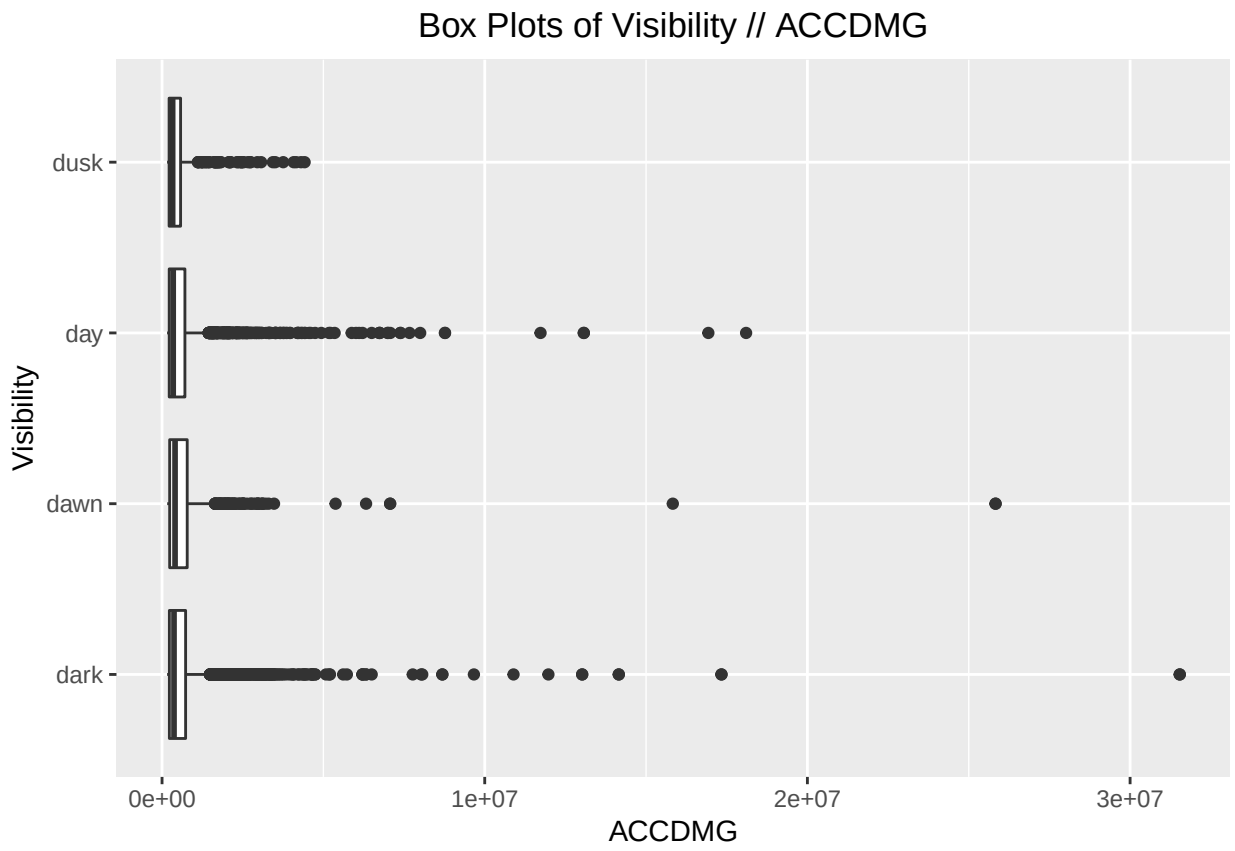
The code above shows three graphs:

1. A bar graph that shows the frequency of accidents in terms of their cause.
2. A linear box plot that shows the cost of accidents in terms of their cause.
3. A logarithmic box plot that shows the cost of accidents in terms of their cause.

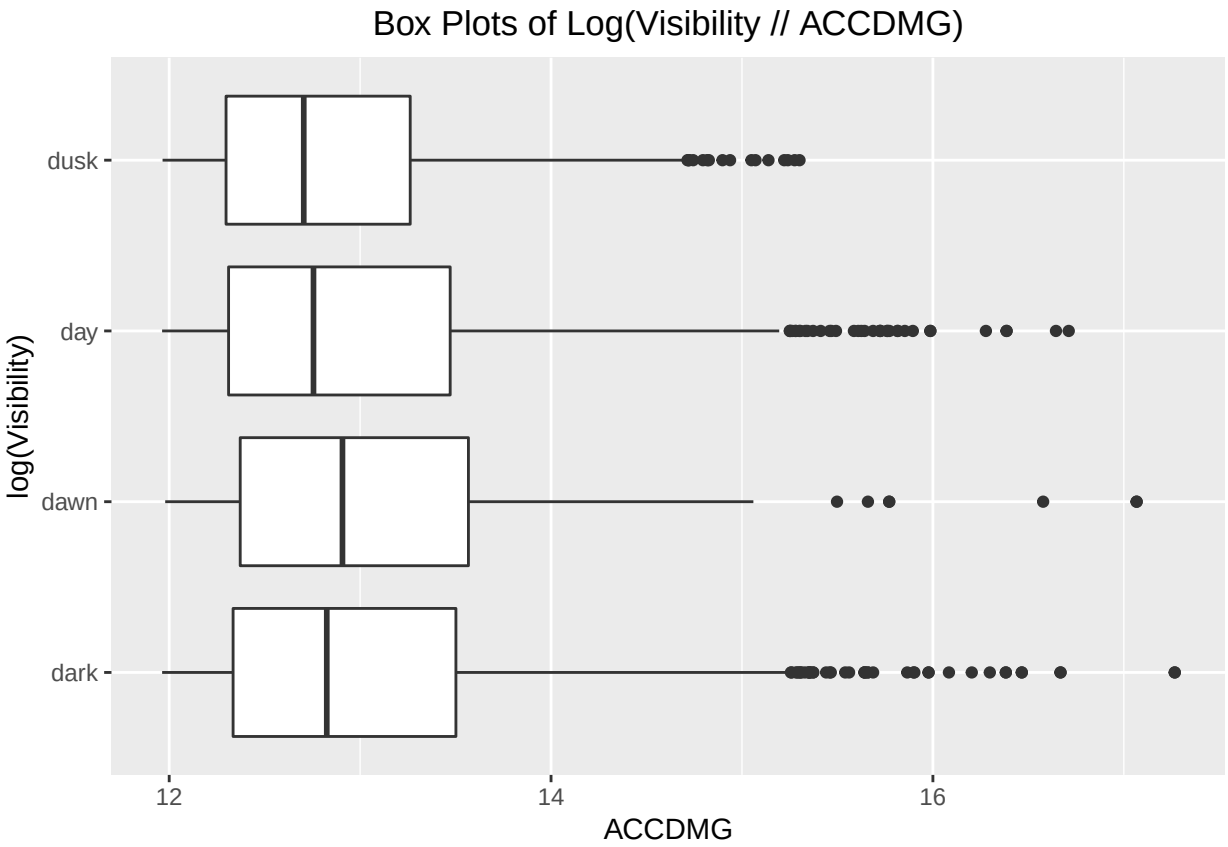
```
xdmgnd <- xdmg[!(duplicated(xdmg[, c("INCDTNO", "YEAR", "MONTH", "DAY", "TIMEHR", "TIMEMIN")])),]

xdmgnd$Visibility <- rep(NA, nrow(xdmgnd))
xdmgnd$Visibility[which(substr(xdmgnd$VISIBLTY, 1, 1) == "1")] <- "dawn"
xdmgnd$Visibility[which(substr(xdmgnd$VISIBLTY, 1, 1) == "2")] <- "day"
xdmgnd$Visibility[which(substr(xdmgnd$VISIBLTY, 1, 1) == "3")] <- "dusk"
xdmgnd$Visibility[which(substr(xdmgnd$VISIBLTY, 1, 1) == "4")] <- "dark"

xdmgnd$Visibility <- factor(xdmgnd$Visibility)

# Visibility // ACCDMG

ggplot(as.data.frame(table(xdmgnd$Visibility)), aes(x = Var1, y= Freq)) + geom_bar(stat="identity")
```




```
ggplot(data = xdmgnd, aes(x = Visibility, y = ACCDMG)) +
  geom_boxplot() +
  coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of Visibility // ACCDMG") +
  labs(x = "Visibility", y = "ACCDMG")
```



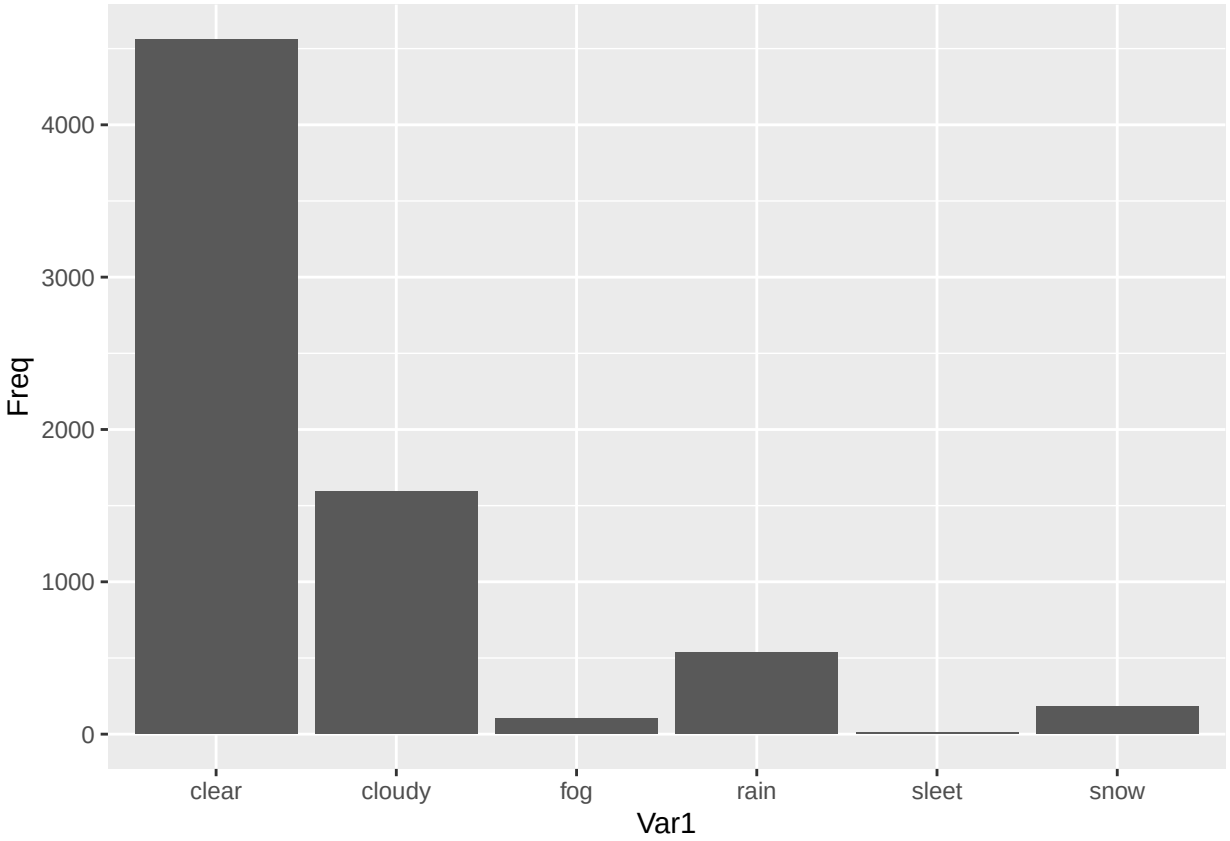

```
ggplot(data = xdmgnd, aes(x = Visibility, y = log(ACCDMG+1))) +
  geom_boxplot() +
  coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of Log(Visibility // ACCDMG)") +
  labs(x = "log(Visibility)", y = "ACCDMG")
```



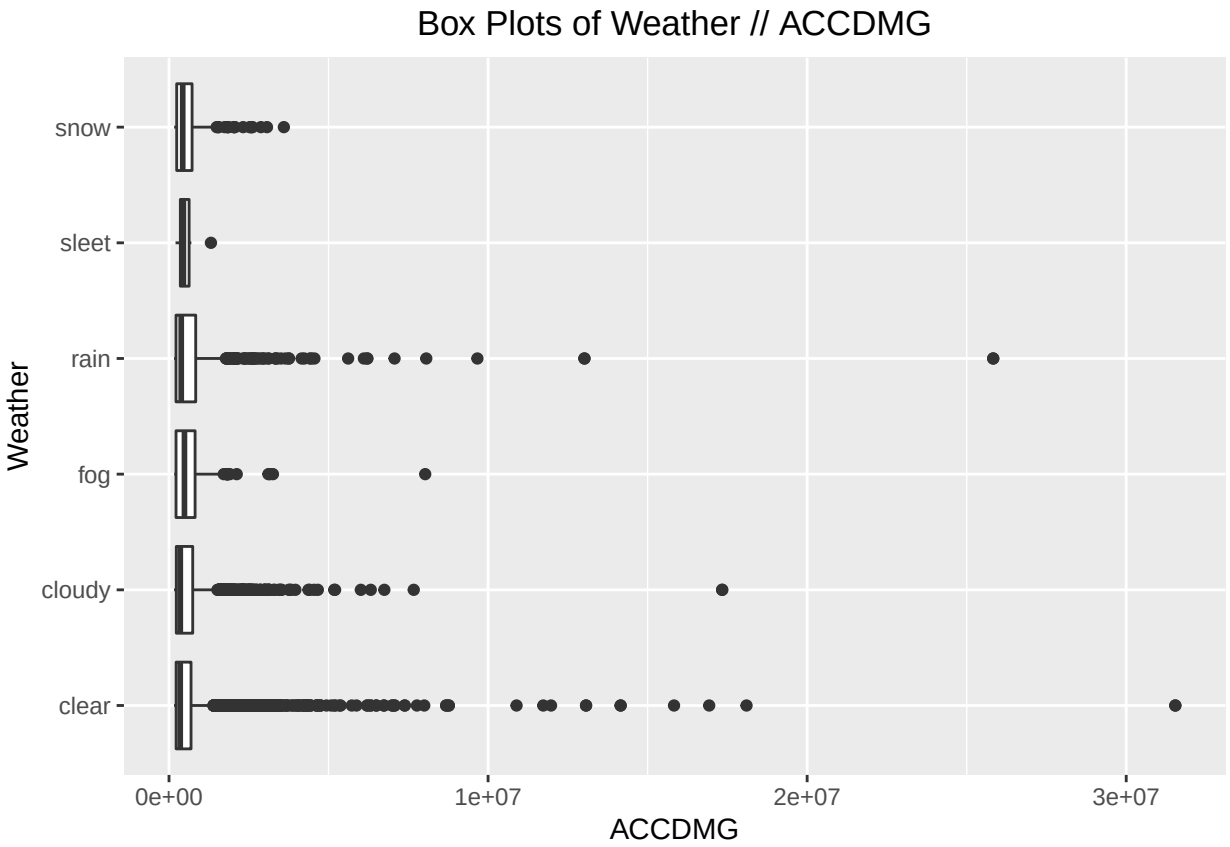
The code above shows three graphs:

1. A bar graph that shows the frequency of accidents in terms of visibility.
2. A linear box plot that shows the cost of accidents in terms of visibility.
3. A logarithmic box plot that shows the cost of accidents in terms of visibility.

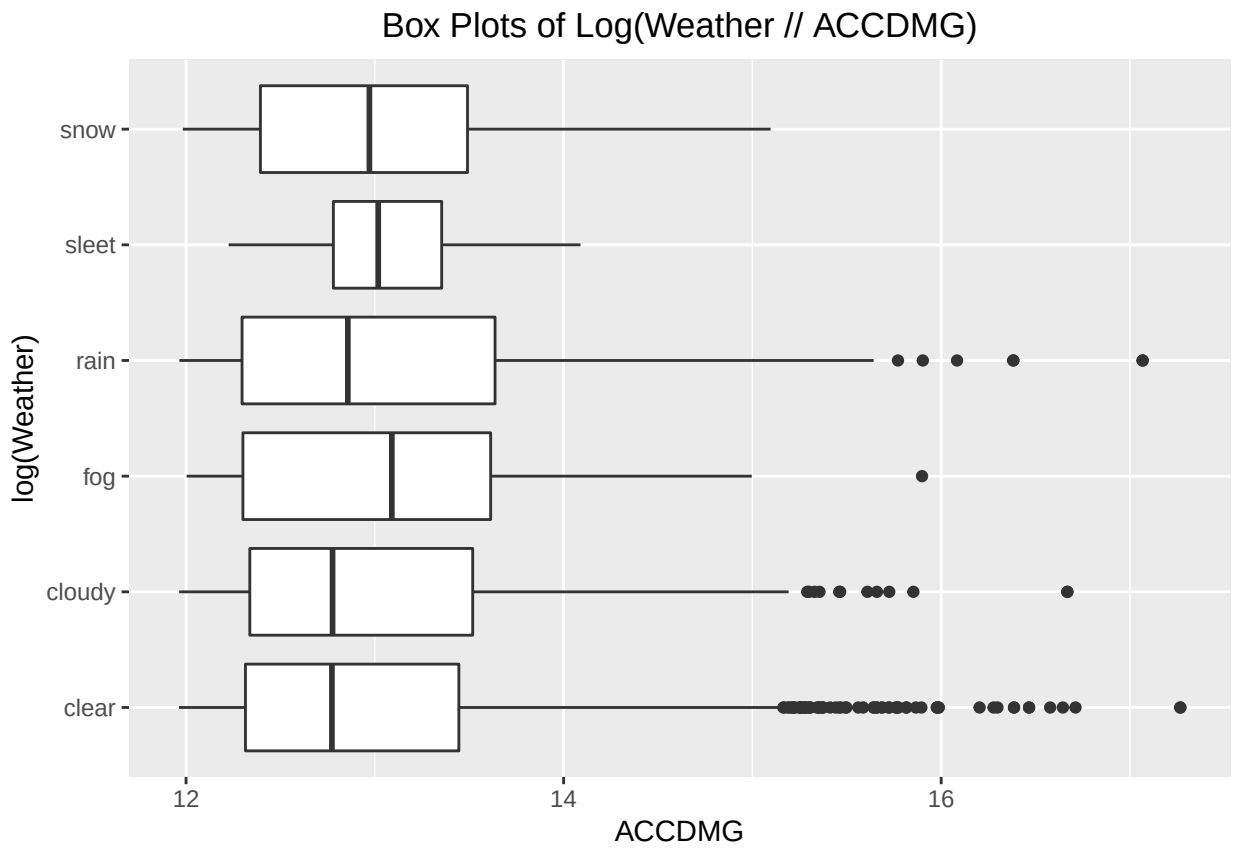
```
xdmgnd$Weather <- rep(NA, nrow(xdmgnd))
xdmgnd$Weather[which(substr(xdmgnd$WEATHER, 1, 1) == "1")] <- "clear"
xdmgnd$Weather[which(substr(xdmgnd$WEATHER, 1, 1) == "2")] <- "cloudy"
xdmgnd$Weather[which(substr(xdmgnd$WEATHER, 1, 1) == "3")] <- "rain"
xdmgnd$Weather[which(substr(xdmgnd$WEATHER, 1, 1) == "4")] <- "fog"
xdmgnd$Weather[which(substr(xdmgnd$WEATHER, 1, 1) == "5")] <- "sleet"
xdmgnd$Weather[which(substr(xdmgnd$WEATHER, 1, 1) == "6")] <- "snow"

xdmgnd$Weather <- factor(xdmgnd$Weather)

# Weather // ACCDMG

ggplot(as.data.frame(table(xdmgnd$Weather)), aes(x = Var1, y= Freq)) + geom_bar(stat="identity")
```




```
ggplot(data = xdmgnd, aes(x = Weather, y = ACCDMG)) +
  geom_boxplot() +
  coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of Weather // ACCDMG") +
  labs(x = "Weather", y = "ACCDMG")
```




```
ggplot(data = xdmgnd, aes(x = Weather, y = log(ACCDMG+1))) +
  geom_boxplot() +
  coord_flip() +
  scale_fill_grey(start = 0.5, end = 0.8) +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Box Plots of Log(Weather // ACCDMG)") +
  labs(x = "log(Weather)", y = "ACCDMG")
```



The code above shows three graphs:

1. A bar graph that shows the frequency of accidents in terms of weather
2. A linear box plot that shows the cost of accidents in terms of weather
3. A logarithmic box plot that shows the cost of accidents in terms of weather

Quantitative variables:

Based on the scatter plot matrices, we can see that there is a relatively strong correlation between ACCDMG and TRNSPD, and between ENGRS and TRNSPD. However, when we take the log of the scatter plot matrices, we can see that there is an even stronger correlation between ACCDMG and TRNSPD. There appears to be minimal correlation between ACCDMG and CARS, ENGRS, and TEMP.

Categorical variables:

Cause:

Based on the box plot, we can see that H (Train operation - Human Factors) yields the most extreme accidents, while T (Rack, Roadbed and Structures) yields the highest frequency of accidents.

Visibility:

Based on the box plot, we can see that the most extreme accidents occur while it is dark, while the highest frequency of accidents occur during the day.

Weather:

Based on the box plot, we can see that both the most extreme accidents and the highest frequency of accidents occur when the weather is clear.

3. (a)

Based on the observations from the graphs above, I have formulated 2 hypotheses that relate Human Factors and TRNSPD to the severity of ACCDMG. From the data, we can see that Human Factors is associated with the most extreme cases of ACCDMG, and that TRNSPD is associated with the Cause of Human Factors-related accidents and is associated with ACCDMG in general.

1. Hypothesis 1

H0. Human Factors do not increase the severity of ACCDMG relative to other types of causes.

HA. Human Factors do increase the severity of ACCDMG relative to other types of causes.

2. Hypothesis 2

H0. The relationship between TRNSPD and accidents related to Human Factors does not increase the severity of ACCDMG.

HA. The relationship between TRNSPD and accidents related to Human Factors does increase the severity of ACCDMG.

BONUS

The 2008 Chatsworth train collision supports the hypothesis that Human Factors do increase the severity of ACCDMG relative to other types of causes. Due to the number and cost of accidents caused by human factors, Congress has even passed legislation and mandated the use of technology to attempt to combat this cause of accidents.

References

2008 Chatsworth train collision. (2020). In Wikipedia. https://en.wikipedia.org/w/index.php?title=2008_Chatsworth_train_collision&oldid=974943375

Federal Register / Vol. 73, No. 195 / Tuesday, October 7, 2008 / Notices. (2008, October 7). US Department of Transportation. https://web.archive.org/web/20081028210416/http://www.fra.dot.gov/downloads/PubAffairs/EmergencyOrder26.pdf

Positive train control. (2020). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Positive_train_control&oldid=978045357

Rail safety improvement act of 2008. (2020). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Rail_Safety_Improvement_Act_of_2008&oldid=976374082

Role of human factors in rail accidents. (2017, June 6). US Department of Transportation. https://www.transportation.gov/testimony/role-human-factors-rail-accidents