In-Class Assignment - Generating Hypotheses

Reese Quillian, TJ Gwilliam, Sofia Zajec, Cecilia Smith

2022-10-12

Work with your group to obtain an understanding of the contributors to accident severity (ACCDMG) in your extreme accidents data using the multivariate visualization techniques discussed in class.

This analysis should explore the relationship between categorical variables and your response to describe the severe accidents, and inform at least two actionable hypotheses.

Explain why the hypotheses are actionable (i.e., how they could lead to meaningful recommendations) and explain how you arrived at each hypothesis. Write out your null and alternative hypotheses.

```
# Libraries and files
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(lattice)
library(GGally)
```

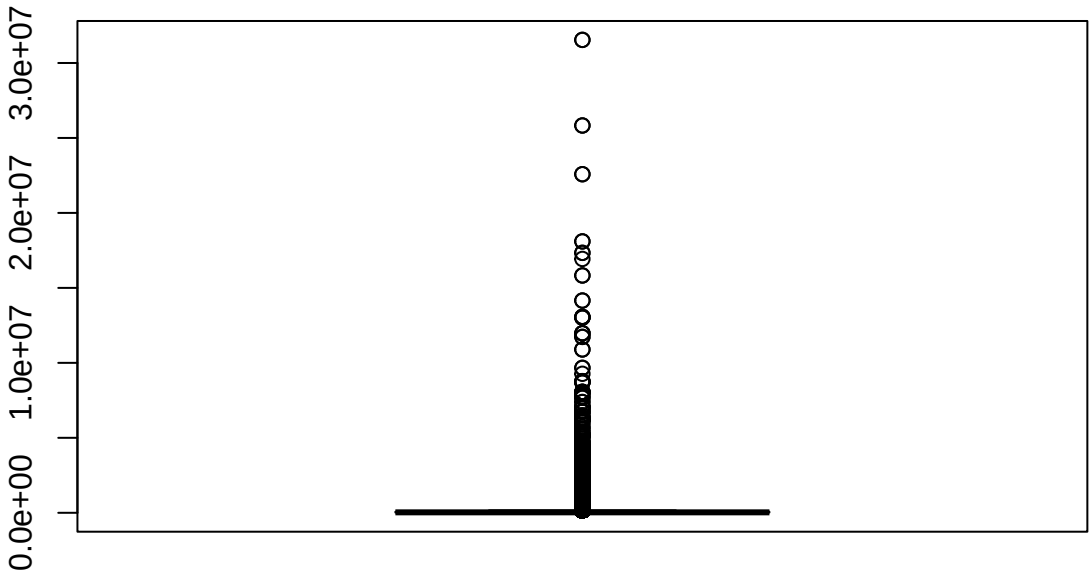
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(devtools)
```

```
## Loading required package: usethis
```

```
library(ggfortify)

traindir <- "C:/Users/Student/OneDrive - University of Virginia/Documents/SYS4021/In Class/Data/Train Da
sourcedir <-"C:/Users/Student/OneDrive - University of Virginia/Documents/SYS4021/In Class"

# load data
setwd(sourcedir)
source("AccidentInput.R")

acts <- file.inputl(traindir)

totacts <- combine.data(acts)
```

```
# extreme accidents
# Build a data frame with only extreme accidents for ACCDMG

dmgbox <-boxplot(totacts$ACCDMG)
```



```r
# accidents above upper whisker
xdmg <- totacts[totacts$ACCDMG > dmgbox$stats[5],]

#remove 9/11
xdmg <- xdmg[-183,]

## Remove duplicates from xdmg and call new data frame xdmgnd
xdmgnd <- xdmg[!(duplicated(xdmg[, c("INCDTNO", "YEAR", "MONTH", "DAY", "TIMEHR", "TIMEMIN")])),]
# xdmgnd = dataframe to use
```

Hypothesis 1: Human errors, specifically signaling errors, lead to disproportionately more costly accidents.

H0 = ACCDMG for signaling errors is the same as other human errors.

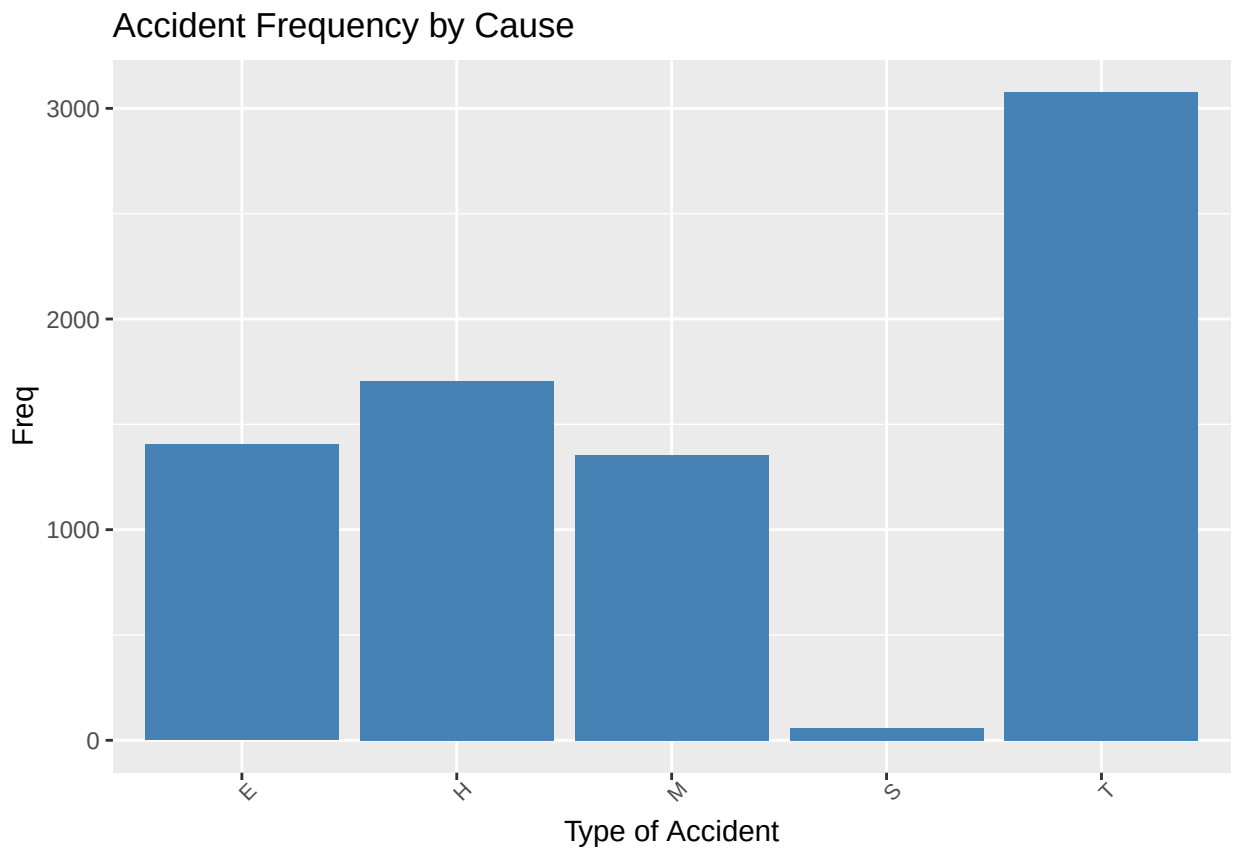
HA = ACCDMG caused by signaling errors is higher than other human errors.

Actions:

This hypothesis is actionable because training of conductors can be updated or improved if the evidence supports rejection of H0. We arrived at this hypothesis by first looking into the overall frequency of accidents by cause and found that human factors was the second most common cause of train accidents. This lead us to look into specific types of human errors, and found that signaling errors incur the most damage despite being 5/10 in terms of frequency.

```r
# Setup categorical variables
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"

# This new variable, Cause, has to be a factor

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

```r
ggplot(as.data.frame(table(xdmgnd$Cause)), aes(x = Var1, y= Freq)) +
  geom_bar(stat="identity",fill= "steelblue")+
  ggtitle("Accident Frequency by Cause") +
  labs(x = "Type of Accident")+
  theme(axis.text.x = element_text(size = 8,  angle = 45))
```




```
# T = rack, roadbed, structures
```

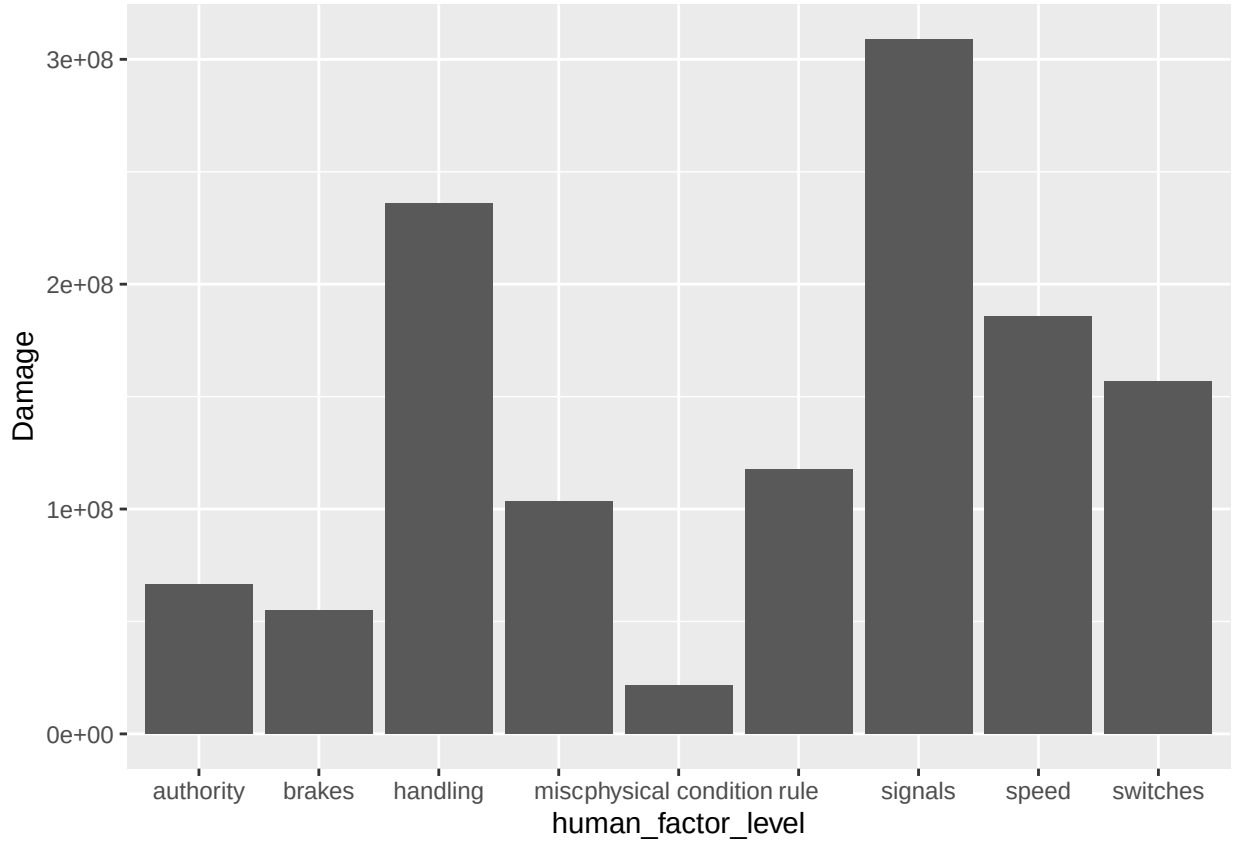
```
# recode CAUSE

xdmgnd$human_factor_level <- rep(NA, nrow(xdmgnd))

xdmgnd$human_factor_level[which(substr(xdmgnd$CAUSE,1,2)=="H0")] <- "brakes"
xdmgnd$human_factor_level[which(substr(xdmgnd$CAUSE,1,2)=="H1")] <- "physical condition"
xdmgnd$human_factor_level[which(substr(xdmgnd$CAUSE,1,2)=="H2")] <- "signals"
xdmgnd$human_factor_level[which(substr(xdmgnd$CAUSE,1,2)=="H3")] <- "rule"
xdmgnd$human_factor_level[which(substr(xdmgnd$CAUSE,1,2)=="H4")] <- "authority"
xdmgnd$human_factor_level[which(substr(xdmgnd$CAUSE,1,2)=="H5")] <- "handling"
xdmgnd$human_factor_level[which(substr(xdmgnd$CAUSE,1,2)=="H6")] <- "speed"
xdmgnd$human_factor_level[which(substr(xdmgnd$CAUSE,1,2)=="H7")] <- "switches"
xdmgnd$human_factor_level[which(substr(xdmgnd$CAUSE,1,2)=="H8")] <- "cab"
xdmgnd$human_factor_level[which(substr(xdmgnd$CAUSE,1,2)=="H9")] <- "misc"

xdmgnd$human_factor_level <- factor(xdmgnd$human_factor_level)
```

```
table(xdmgnd$human_factor_level)
```

```
##
##          authority             brakes           handling               misc
##                 78                120                436                 76
## physical condition               rule            signals              speed
```

```
##                  19                 335                 186                 212
##            switches
##                 244
```

```
df<- xdmgnd %>% filter(Cause == "H") %>% select(human_factor_level, ACCDMG) %>% group_by(human_factor_l

ggplot(data=df, aes(x=human_factor_level,y=Damage)) + geom_col()
```



Hypothesis 2: At high train speeds, human errors is the most costly cause of accident.

H0: ACCDMG for all causes are equal at high train speeds.

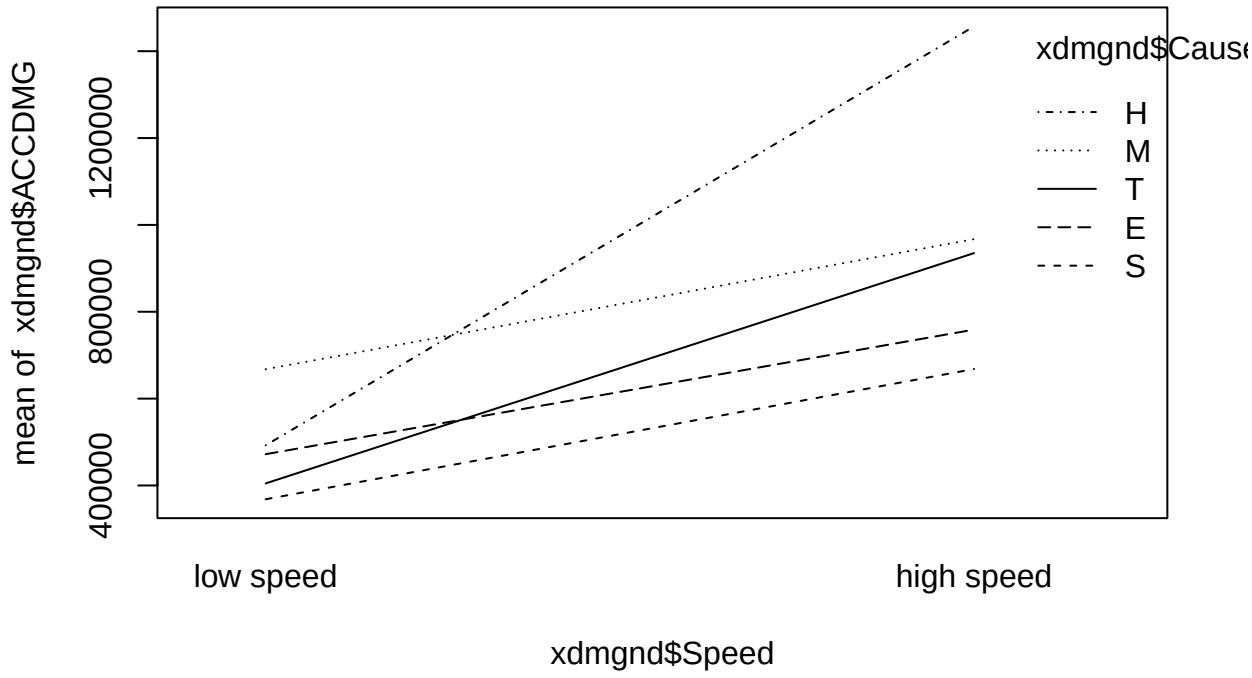
HA: ACCDMG for human factors is not equal to other causes at high speeds.

Actions:

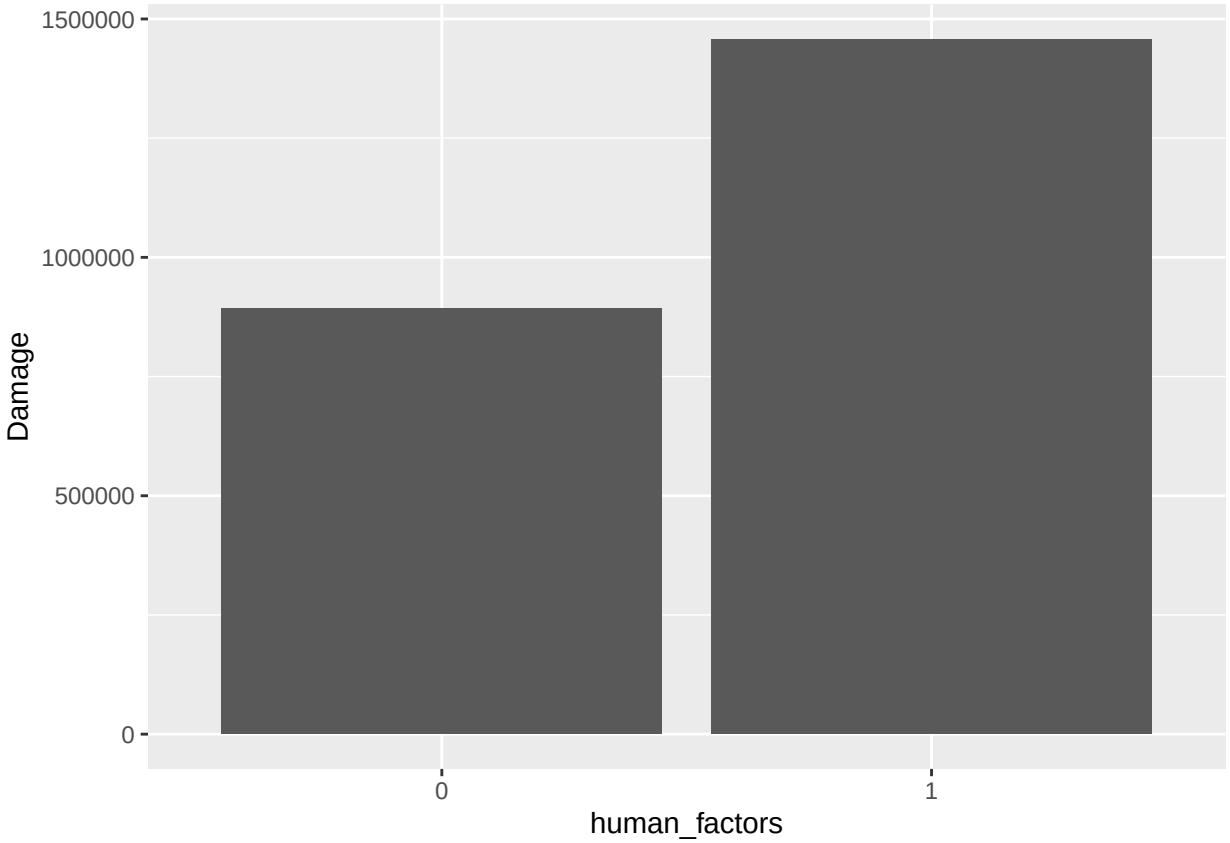
This hypothesis is actionable because trains known to go at higher speeds can be paid more attention to. For example, additional or more senior staff can be assigned to high speed trains. To arrive at this hypothesis, we looked at the interaction plot between cause and speed and saw that human factors had the steepest slope between low and high speeds. We then looked at the interaction between only human errors and speed and it appears that at higher speeds, human errors causes a disproportionate amount of damage.

```r
# Speed variable
xdmgnd$Speed <- cut(xdmgnd$TRNSPD, c(min(xdmgnd$TRNSPD),median(xdmgnd$TRNSPD),max(xdmgnd$TRNSPD)), inclu

# Create human factors variable
xdmgnd$human_factors <- rep(0, nrow(xdmgnd))
xdmgnd$human_factors[which(xdmgnd$Cause == "H")] <- 1
xdmgnd$human_factors <- factor(xdmgnd$human_factors)
```

```r
interaction.plot(xdmgnd$Speed, xdmgnd$Cause, xdmgnd$ACCDMG)
```




```r
df1<- xdmgnd %>% filter(Speed == "high speed") %>% select(human_factors, ACCDMG) %>% group_by(human_fac

ggplot(data=df1, aes(x=human_factors,y=Damage)) + geom_col()
```


1500000
1000000
500000
0
Damage
0
1
human_factors