Project 1

Reese Quillian, Sofia Zajec, Cecilia Smith, TJ Gwilliam

2022-10-24

Load data / packages

```
# 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)
library(MASS)

## 
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
## 
##     select
```

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

setwd(sourcedir)

# files for analysis
source("AccidentInput.R")
source("SPM_Panel.R")
source("PCAplots.R")
```

```
## 
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
## 
##     between, first, last
```

```
## -------------------------------------------------------------------------
----
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first,
then dplyr:
## library(plyr); library(dplyr)
```

```
## -------------------------------------------------------------------------
----
```

```
## 
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
## 
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## 
## Attaching package: 'scales'
```

```
## The following objects are masked from 'package:psych':
## 
##     alpha, rescale
```

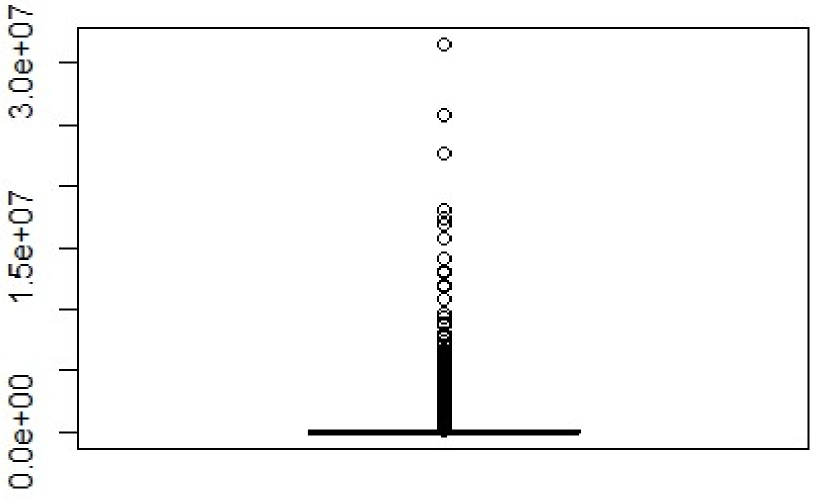
```
## 
## Attaching package: 'ggpubr'
## The following object is masked from 'package:plyr':
##
##     mutate
```

```
# load data
acts <- file.inputl(traindir)

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

Cleaning

For this analysis, we only consider extreme accidents, as we are not concerned with accidents that do not lead to significant damages/casualties. For the ACCDMG metric, extreme accidents are those whose ACCDMG are above the upper whisker. For the Casualties metric, extreme accidents are those with at least 1 casualty.

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

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



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

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

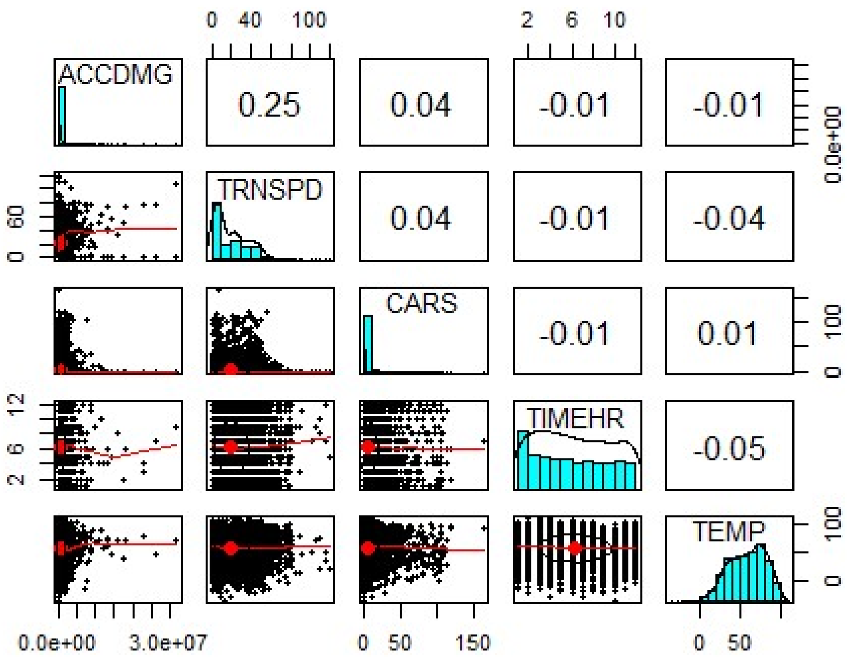
```r
## Remove duplicates from xdmg and call new data frame xdmgnd
xdmgnd <- xdmg[!(duplicated(xdmg[, c("INCDTNO", "YEAR", "MONTH", "DAY",
"TIMEHR", "TIMEMIN")])),]

xdmgnd$Type <- factor(xdmgnd$TYPE, labels = c("Derailment", "HeadOn",
"Rearend", "Side", "Raking", "BrokenTrain", "Hwy-Rail", "GradeX",
"Obstruction", "Explosive", "Fire","Other","SeeNarrative"))

# casualties = TOTINJ + TOTKLD
xdmgnd <- xdmgnd %>% mutate(casualties = TOTKLD + TOTINJ)

# Setup cause variable
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"

xdmgnd$Cause <- factor(xdmgnd$Cause)

# Speed variable
xdmgnd$Speed <- cut(xdmgnd$TRNSPD,
c(min(xdmgnd$TRNSPD),median(xdmgnd$TRNSPD),max(xdmgnd$TRNSPD)),
include.lowest = T, labels = c("low speed", "high speed"))

# 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)

# dataframe for casualties analysis

#Create  a new dataframe with only 1 or more casualties
xdmgnd_cas <- xdmgnd %>% filter(casualties > 0)

# remove max (outlier)
xdmgnd_cas <- xdmgnd_cas %>% filter(casualties != max(casualties))
```

1. Hypotheses

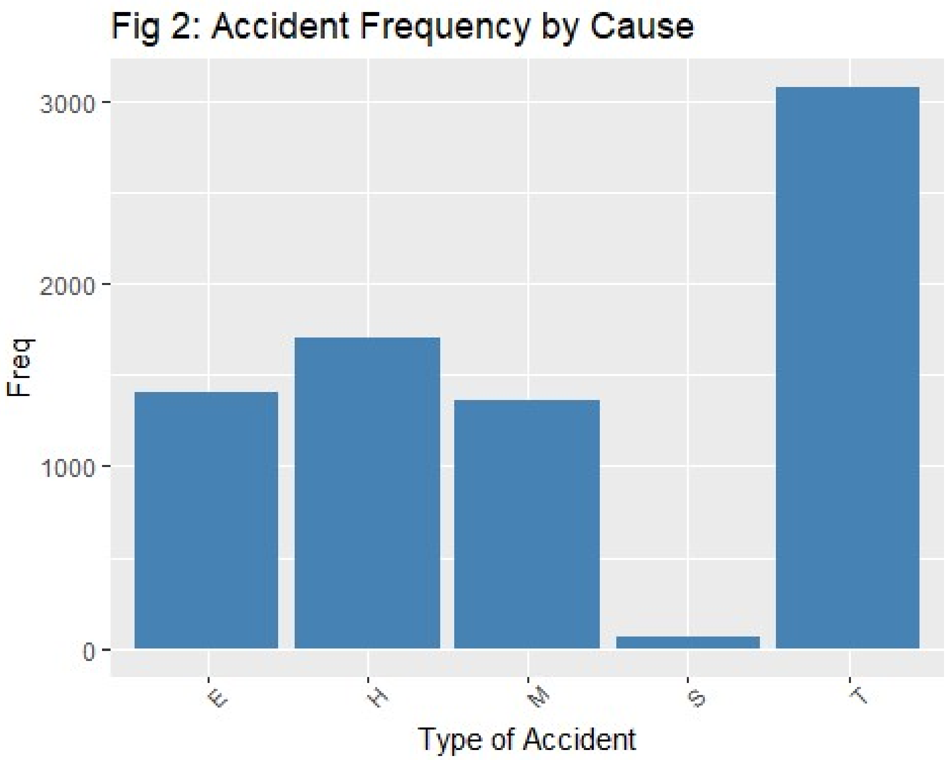
1.1 ACCDMG

We first began with a correlation matrix (Fig. 1) and saw that train speed (TRNSPD) was the most highly correlated with accident damage (ACCDMG). In exploring a categorical variable not included in the correlation matrix, we decided to make the following bar charts: Accident Frequency by Cause (Fig. 2) and Mean Accident Damage by Cause (Fig. 3). These two visualizations showed that H, which corresponds to an accident attributed to human factors, is the second most common accident cause as well as the second highest mean accident cost, leading to human factors being a variable of interest because it is both common and costly per accident. We then recoded the train speed (TRNSPD) into a categorical variable (Speed) with two levels, high and low speeds and created an interaction plot as shown below:

```
pairs.panels(xdmgnd[,c("ACCDMG", "TRNSPD", "CARS", "TIMEHR", "TEMP")])
```

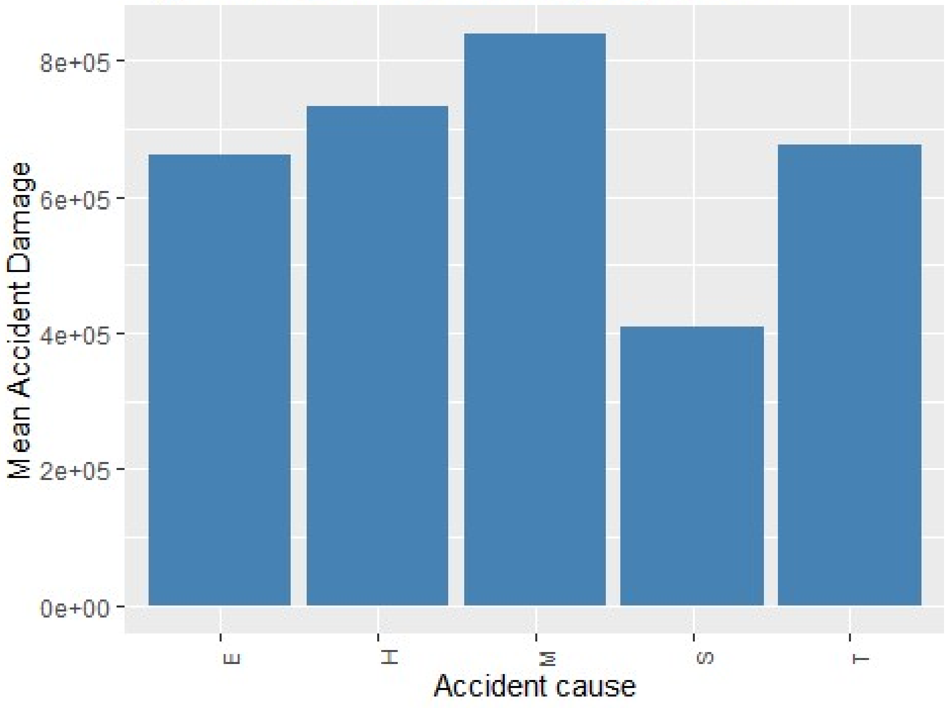


```
ggplot(as.data.frame(table(xdmgnd$Cause)), aes(x = Var1, y= Freq)) +
  geom_bar(stat="identity",fill= "steelblue")+
  ggtitle("Fig 2: Accident Frequency by Cause") +
  labs(x = "Type of Accident")+
  theme(axis.text.x = element_text(size = 8,  angle = 45))
```

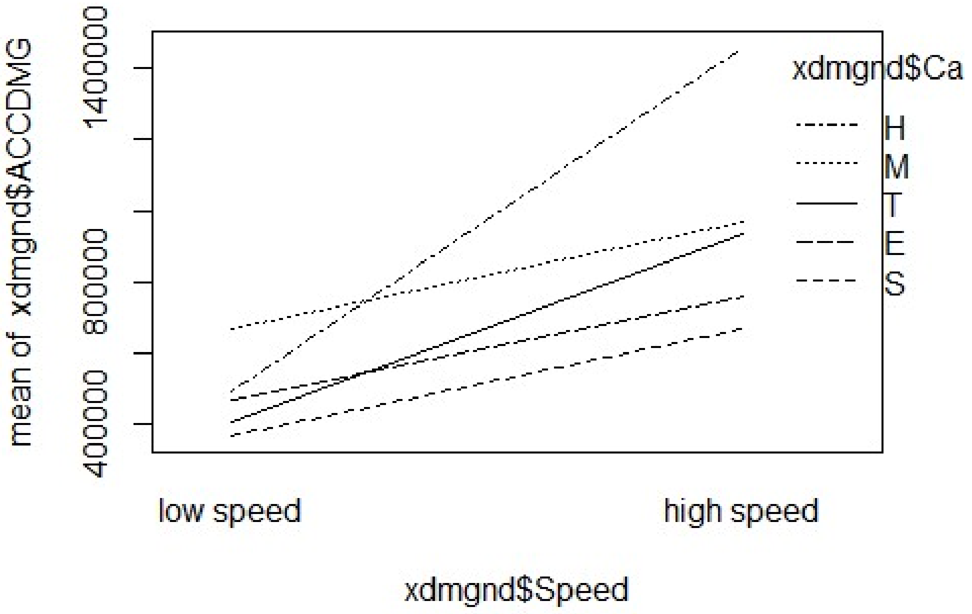


```
df_causes<- xdmgnd %>% group_by(Cause) %>%
dplyr::summarise(Damage=mean(ACCDMG),n=n())

ggplot(df_causes, aes(x = Cause, y=Damage)) +
  geom_col(fill= "steelblue")+
  ggtitle("Fig 3: Mean Accident Damage by Cause") +
  labs(x = "Accident cause", y = "Mean Accident Damage")+
  theme(axis.text.x = element_text(size = 8,  angle = 90))
```

Fig 3: Mean Accident Damage by Cause



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



The interaction plot (Fig. 4) shows that human factors have the greatest discrepancy in accident cost between low and high speeds compared to any other cause, leading to the generation of our first hypothesis: At high train speeds, human error is the most costly cause of accident.

Hypothesis 1: At high train speeds, human errors are the most costly cause of accidents.

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

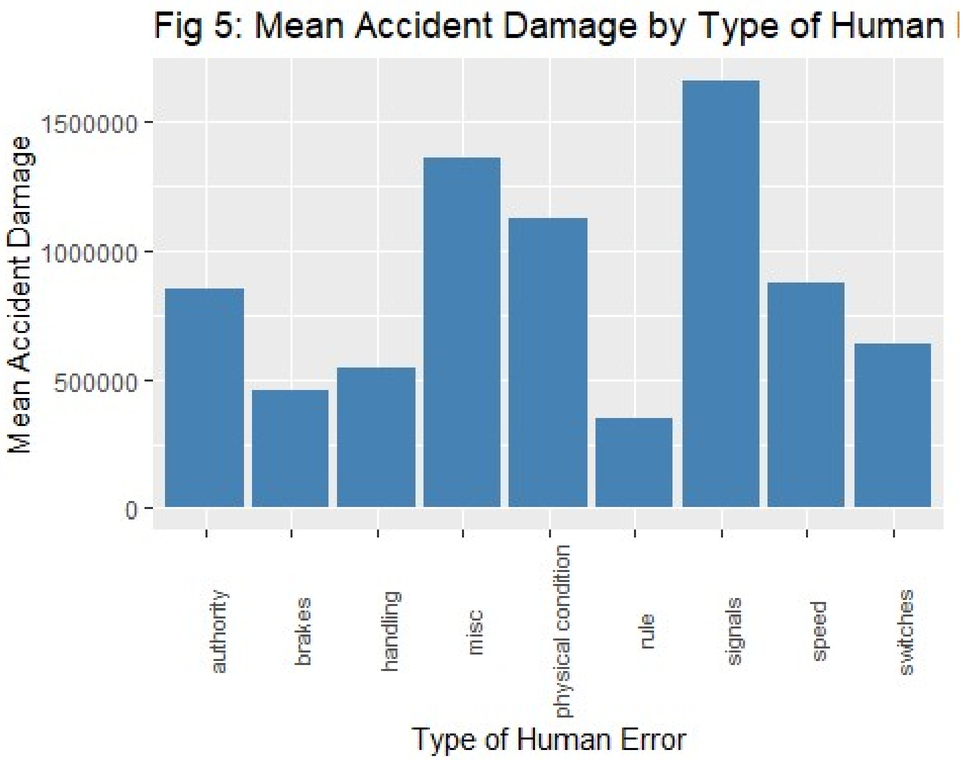
HA: ACCDMG for human factors is not equal to other causes at high speeds.

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 cause a disproportionate amount of damage.

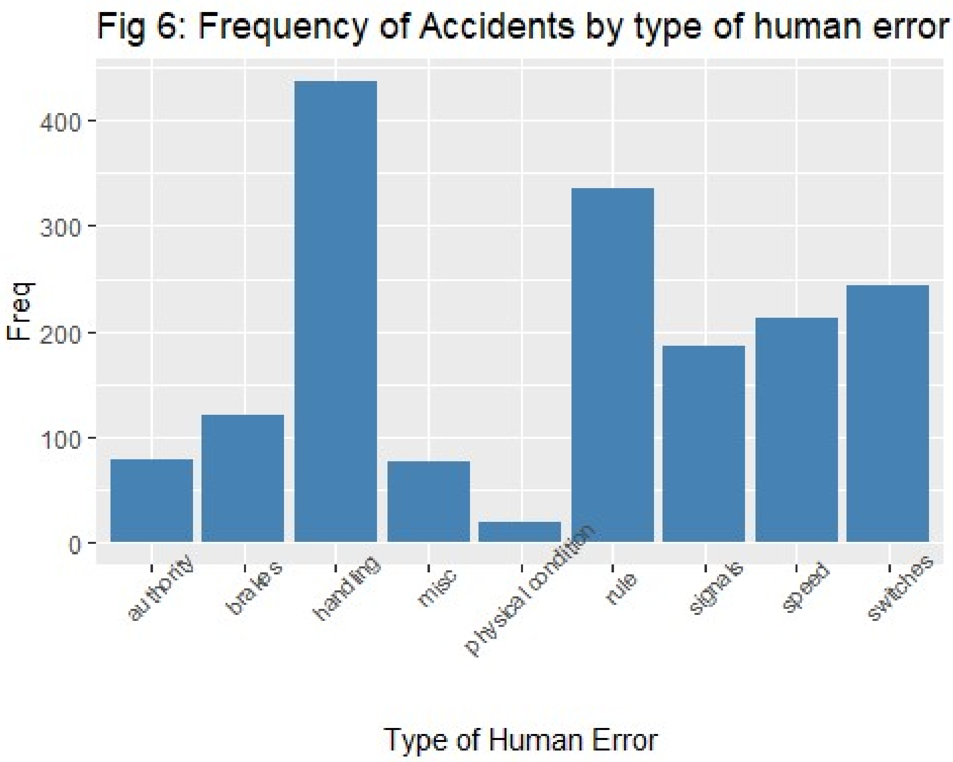
We then turned to exploring the more specific cause (CAUSE) of the accident. Exploration of our first hypothesis showed human factors were overrepresented in the accident damage severity metric, however in creating a bar chart Mean Accident Damage by Type of Human Error, Flagging, Fixed, Hand and Radio Signals (signals)(Fig. 5) were the most costly type of human error. The Frequency of Accidents by Type of Human Error (Fig. 6) bar chart showed that signals were not the most common, but still made up ~25% of the cost of extreme accidents caused by human factors.

```r
# recode human factor 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)

df<- xdmgnd %>% filter(Cause == "H") %>% group_by(human_factor_level) %>%
dplyr::summarise(Damage=mean(ACCDMG),n=n())
```

```
ggplot(df, aes(x = human_factor_level, y=Damage)) +
  geom_col(fill= "steelblue")+
  ggtitle("Fig 5: Mean Accident Damage by Type of Human Error") +
  labs(x = "Type of Human Error", y = "Mean Accident Damage")+
  theme(axis.text.x = element_text(size = 8,  angle = 90))
```



```
df_hf<- xdmgnd %>% filter(Cause == "H")

ggplot(as.data.frame(table(df_hf$human_factor_level)), aes(x = Var1, y=
Freq)) +
  geom_bar(stat="identity",fill= "steelblue")+
  ggtitle("Fig 6: Frequency of Accidents by type of human error") +
  labs(x = "Type of Human Error")+
  theme(axis.text.x = element_text(size = 8,  angle = 45))
```



```
# Total cost of human error accidents by type as a proportion of total
accident damage
sumbytype<- as.numeric(tapply(as.numeric(df_hf$ACCDMG),
as.factor(df_hf$human_factor_level), sum))
proptype <- sumbytype / sum(as.numeric(df_hf$ACCDMG))

proptype

## [1] 0.05305501 0.04376580 0.18849434 0.08280430 0.01708031 0.09407907
0.24698521
## [8] 0.14849948 0.12523647
```

The high cost and proportion of total accident damage that is made up by signals led to the development of our second hypothesis: Signaling errors (a type of human error) lead to disproportionately more costly accidents.

Hypothesis 2: Signaling errors lead to disproportionately more costly accidents.

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

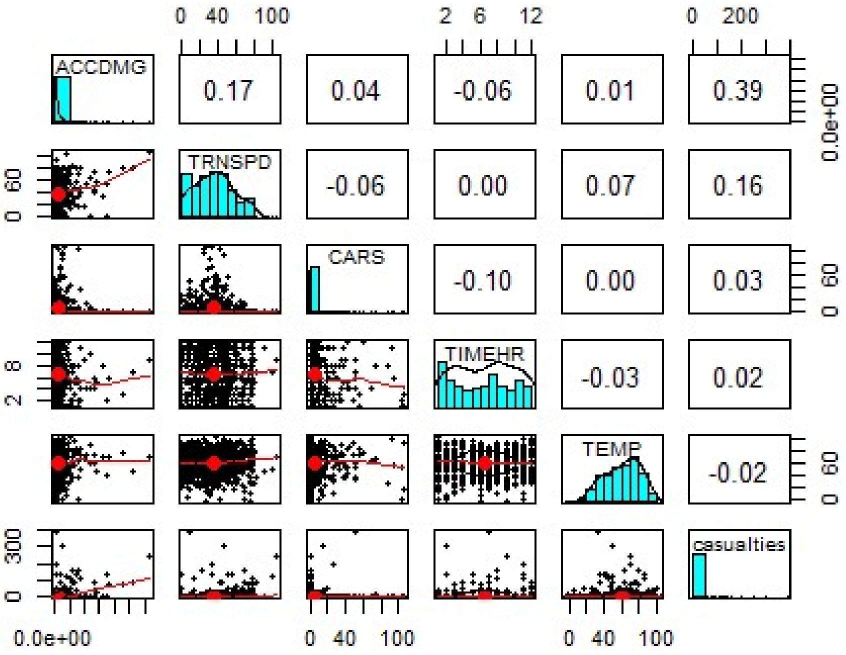
HA = ACCDMG caused by signaling errors is higher than other errors.

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.

1.2 Casualties

Next we want to explore the cause of casualties by first making a correlation matrix (Fig. 7) that includes the quantitative variables from our previous exploration and includes casualties. The matrix showed that casualties is most highly correlated with accident damage with a correlation coefficient of 0.39 and as a result, led us to continue our analysis of casualties by exploring the cause. A bar plot of Casualty Frequency by Cause (Fig. 8) showed that miscellaneous was the most frequent and the third highest in mean, as seen in the bar plot of Mean Casualties by Cause (Fig. 9). Consideration of these plots together lead us to believe it warrants further consideration, especially given the vague overarching definition of the category.

```
pairs.panels(xdmgnd_cas[,c("ACCDMG", "TRNSPD", "CARS", "TIMEHR", "TEMP",
"casualties")])
```



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

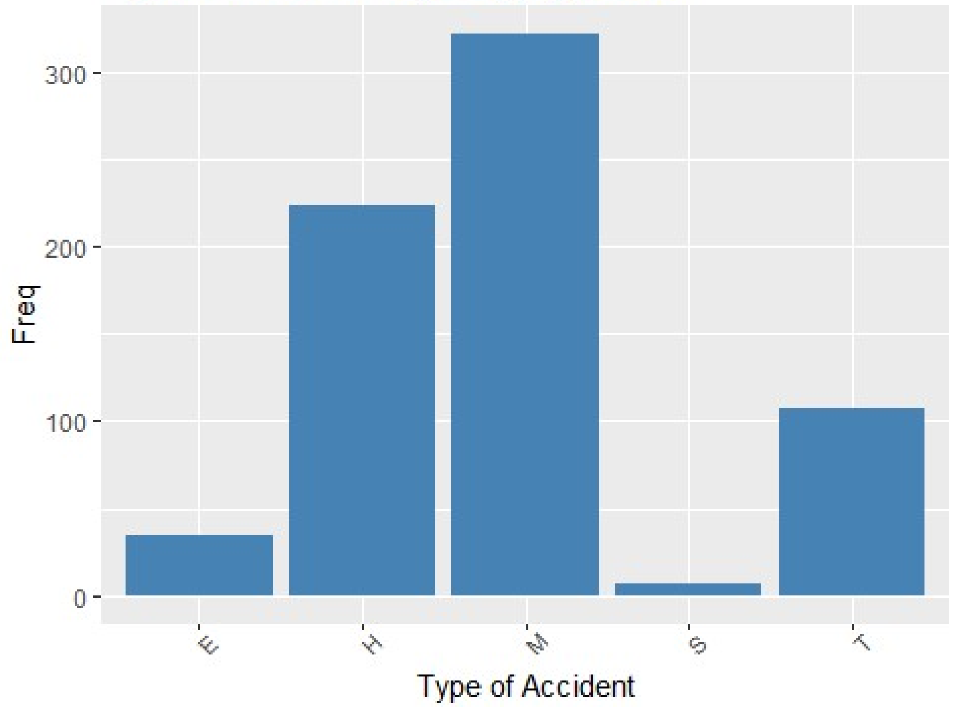
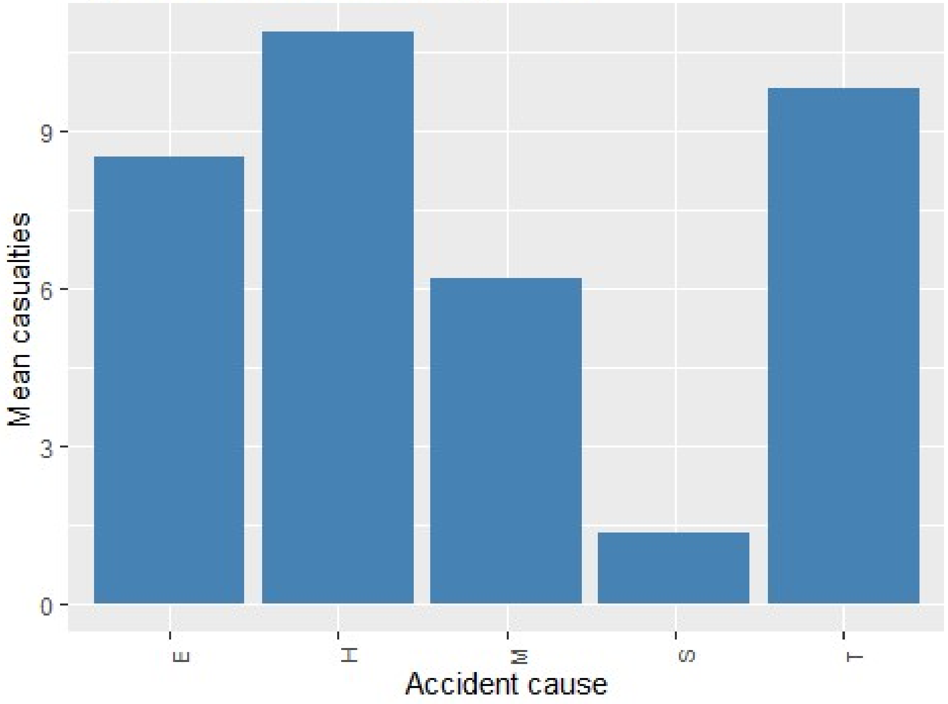


Fig 8: Accident Frequency by Cause

```
df_causes_cas<- xdmgnd_cas %>% group_by(Cause) %>%
dplyr::summarise(average_casualties=mean(casualties),n=n())

ggplot(df_causes_cas, aes(x = Cause, y=average_casualties)) +
  geom_col(fill= "steelblue")+
  ggtitle("Fig 9: Mean Casualties by Cause") +
  labs(x = "Accident cause", y = "Mean casualties")+
  theme(axis.text.x = element_text(size = 8,  angle = 90))
```

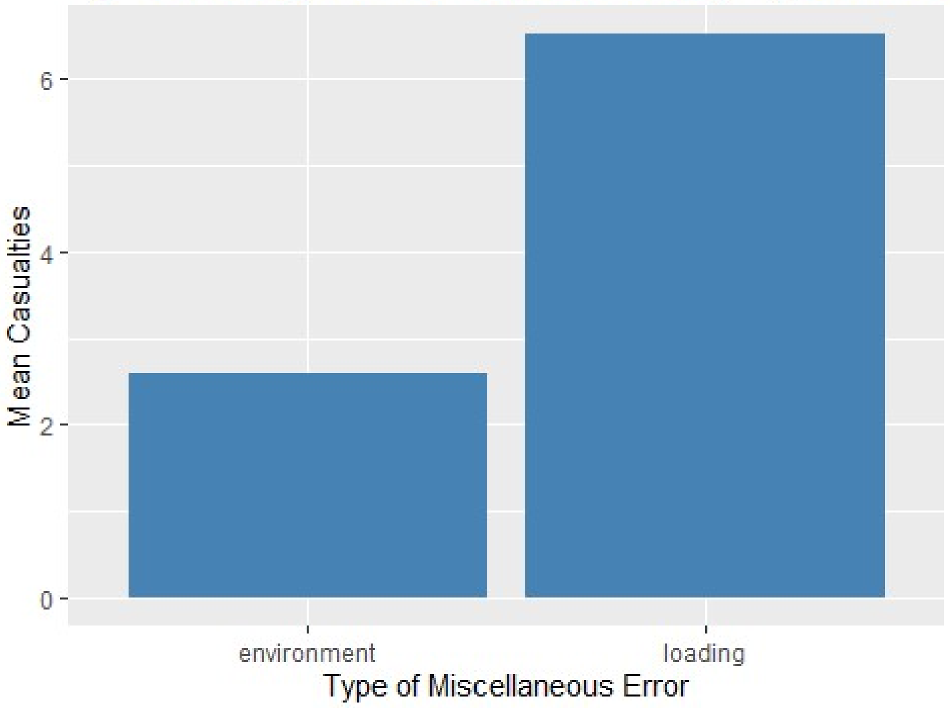
Fig 9: Mean Casualties by Cause



Casualties are most frequently attributed to miscellaneous causes. Now, we want to drill down further to analyze causes within the miscellaneous category. Breakdowns of miscellaneous causes discussed in the data dictionary were applied through the recoding of the CAUSE variable. Next, this breakdown was visualized to understand the number of fatal or injury-inducing accidents by cause, as seen in the Average Number of Casualties by Type of Miscellaneous Error (Fig. 10) bar chart.

```
xdmgnd_cas$misc_type <- rep(NA, nrow(xdmgnd_cas))

xdmgnd_cas$misc_type[which(substr(xdmgnd_cas$CAUSE,1,2)=="M1")] <-
"environment"
xdmgnd_cas$misc_type[which(substr(xdmgnd_cas$CAUSE,1,2)=="M2")] <- "loading"
xdmgnd_cas$misc_type[which(substr(xdmgnd_cas$CAUSE,1,2)=="M3")] <- "loading"
xdmgnd_cas$misc_type[which(substr(xdmgnd_cas$CAUSE,1,2)=="M4")] <- "loading"
xdmgnd_cas$misc_type[which(substr(xdmgnd_cas$CAUSE,1,2)=="M5")] <- "loading"

xdmgnd_cas$misc_type <- factor(xdmgnd_cas$misc_type)

df2<- xdmgnd_cas %>% filter(Cause == "M") %>% group_by(misc_type) %>%
dplyr::summarize(average_casualties = mean(casualties),n=n())

ggplot(df2, aes(x = misc_type, y=average_casualties)) +
  geom_col(fill= "steelblue")+
  ggtitle("Fig 10: Average Number of Casualties by Type of Miscellaneous
Error") +
  labs(x = "Type of Miscellaneous Error", y = "Mean Casualties")
```

Fig 10: Average Number of Casualties by Type of Misce



```r
# loading procedures variable (will use in model)
xdmgnd_cas$loading <- rep(0, nrow(xdmgnd_cas))
xdmgnd_cas$loading[which(xdmgnd_cas$misc_type == "loading")] <- 1
xdmgnd_cas$loading <- factor(xdmgnd_cas$loading)
```

The overrepresentation of loading procedure errors as a cause under miscellaneous accident causes led us to our third recommendation: Errors in loading procedures lead to disproportionately more casualties.

Hypothesis 3: Errors in loading procedures lead to disproportionately more casualties.

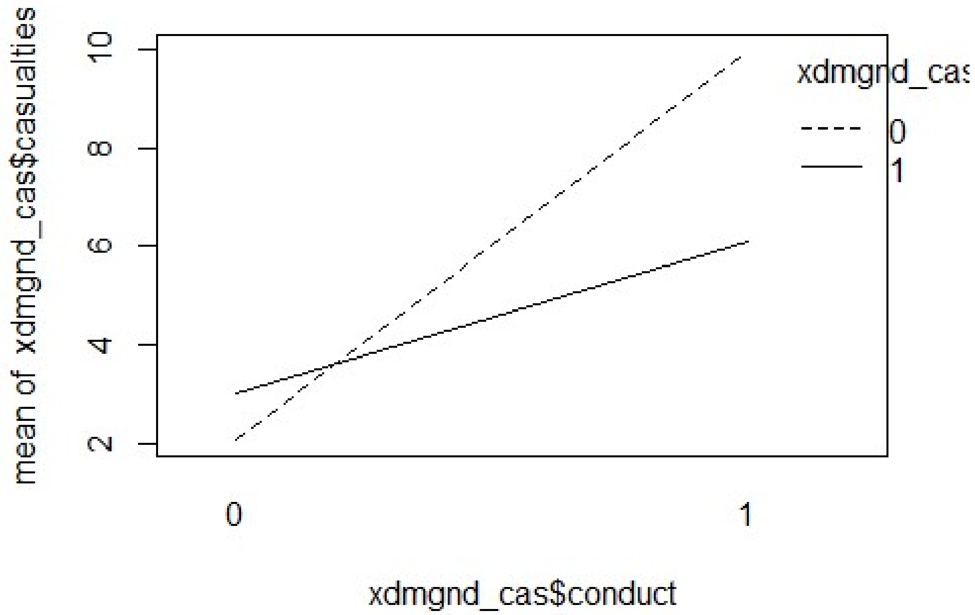
H0 = casualties in accidents caused by loading errors are the same as all other causes

HA = casualties in accidents caused by loading errors are higher than other human errors.

This hypothesis is actionable because loading procedures can be improved or modified if the evidence rejects the null hypothesis. This hypothesis was selected because miscellaneous causes resulted in the most frequent number of casualties in comparison to all other causes. Upon further investigation, miscellaneous cause is defined primarily by loading activities. This classification includes a range of activities, and if the null hypothesis is rejected, then action can be taken to further investigate the root cause of these activities and any patterns that exist in their occurrence.

According to the Association of American Railroads, collisions at grade crossings and incidents involving trespassers on railroad property account for well over 95% of rail-related fatalities [1]. Based on this statistic, we decided to look into highway-rail accidents (TYPE) as being a predictor for casualties. Furthermore, we believed that having no conductors on board the train would likely result in more casualties as there would be no supervision on the train itself. Given this relationship, we created two binary variables to represent if an accident was a highway-rail and if there was a conductor on board and made an interaction plot (Fig. 11) with the two variables alongside casualties. The interaction plot showed that for highway-rail accidents, having a conductor seems to reduce the average number of casualties.

```
# hwyrail variable
xdmgnd_cas$hwyrail <- rep(0, nrow(xdmgnd_cas))
xdmgnd_cas$hwyrail[which(xdmgnd_cas$Type == "Hwy-Rail")] <- 1
xdmgnd_cas$hwyrail <- factor(xdmgnd_cas$hwyrail)

# conductor variable
xdmgnd_cas$conduct <- rep(0, nrow(xdmgnd_cas))
xdmgnd_cas$conduct[which(xdmgnd_cas$CONDUCTR > 0)] <- 1
xdmgnd_cas$conduct <- factor(xdmgnd_cas$conduct)

# interaction
interaction.plot(xdmgnd_cas$conduct, xdmgnd_cas$hwyrail,
xdmgnd_cas$casualties)
```



Hypothesis 4: Trains with conductors are more likely to reduce the number of casualties in highway-rail accidents.

H0 = For highway-rail type accidents, the number of casualties on trains with conductors is equal to the number of casualties on trains without conductors.

HA = For highway-rail type accidents, the number of casualties on trains with conductors is less than the number of casualties on trains without conductors.

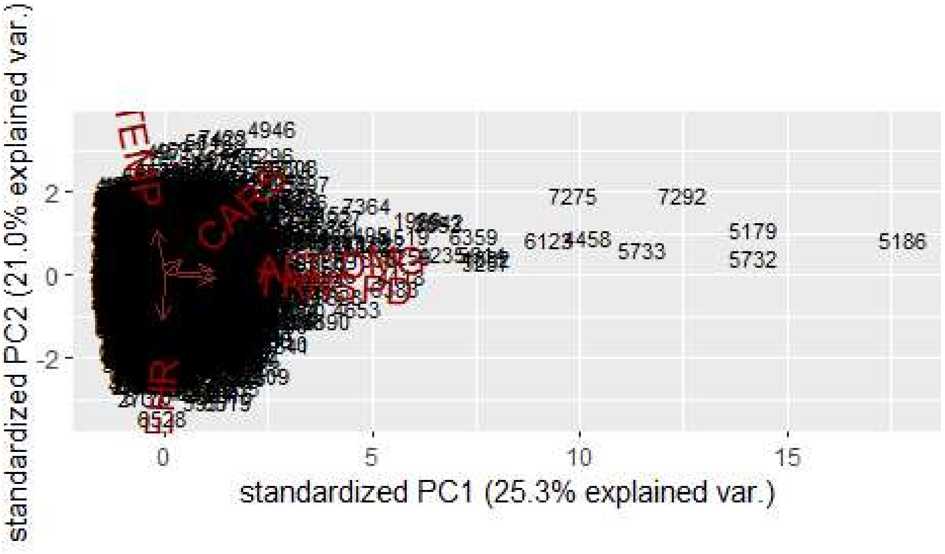
We believe this is easily actionable as trains traveling on routes with a high volume of highway rail crossings should be allocated at least one conductor. Similarly, reducing the number of trains with zero conductors by using conductors staffed on trains with more than one conductor would be valuable if this analysis is shown to be statistically significant.

2. ACCDMG Analysis

a) feature and model selection techniques

We start by using PCA (Fig. 12) to determine which quantitative variables are most correlated with ACCDMG and therefore can be used as predictors.

```
predictors.accdmg.pca <- princomp(xdmgnd[,c("ACCDMG", "TRNSPD", "CARS",
"TIMEHR", "TEMP")], cor = T )

ggbiplot(predictors.accdmg.pca, varname.size = 5, labels=row(xdmgnd)[,1])
```



TRNSPD and ACCDMG are very correlated, so we will use it as a predictor to our models.

b) treatment of ordinal and categorical variables

As seen in the initial Cleaning section, we transformed the CAUSE variable by recoding according to the first letter of each CAUSE code, as they represent cause categories according to the data dictionary. From this, a binary variable for whether an accident was caused by human factors (human_factors) or not was created as well as the variable human factors level which recodes the ten different causes (human_factor_level) with more easily understandable words.

As seen in Hypothesis 2 analysis below, a binary variable (signal) was created and defined according to whether an accident was caused by signaling errors. (This is a level of human factors.)

Hypothesis 1: At high train speeds, human errors are the most costly cause of accidents.

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.

```
# interaction model because our hypothesis is about human errors AT high
speeds
accdmg.lm1 <- lm(ACCDMG~(TRNSPD+human_factors)^2,data=xdmgnd)
summary(accdmg.lm1)

## 
## Call:
## lm(formula = ACCDMG ~ (TRNSPD + human_factors)^2, data = xdmgnd)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -3770704  -366831  -201007    69903 31304687 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           388093.8    24669.8  15.732  < 2e-16 ***
## TRNSPD                 14212.7      851.7  16.686  < 2e-16 ***
## human_factors1       -154027.1    45250.0  -3.404 0.000668 ***
## TRNSPD:human_factors1  27363.2     2214.9  12.354  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1194000 on 7597 degrees of freedom
## Multiple R-squared:  0.08355,    Adjusted R-squared:  0.08319 
## F-statistic: 230.9 on 3 and 7597 DF,  p-value: < 2.2e-16
```

Hypothesis 2: Signaling errors lead to disproportionately more costly accidents.

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

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

```
# signals variable
xdmgnd$signals <- rep(0, nrow(xdmgnd))
xdmgnd$signals[which(xdmgnd$human_factor_level == "signals")] <- 1
xdmgnd$signals <- factor(xdmgnd$signals)
```

```
accdmg.lm2 <- lm(ACCDMG~signals+TRNSPD,data=xdmgnd)
summary(accdmg.lm2)

##
## Call:
## lm(formula = ACCDMG ~ signals + TRNSPD, data = xdmgnd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2168664  -353422  -182995    60797 31197312
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  341442.4    20813.0   16.41   <2e-16 ***
## signals1    1018594.6    89040.2   11.44   <2e-16 ***
## TRNSPD        17238.5      765.9   22.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1199000 on 7598 degrees of freedom
## Multiple R-squared:  0.07609,    Adjusted R-squared:  0.07585
## F-statistic: 312.9 on 2 and 7598 DF,  p-value: < 2.2e-16
```

c) model assessment

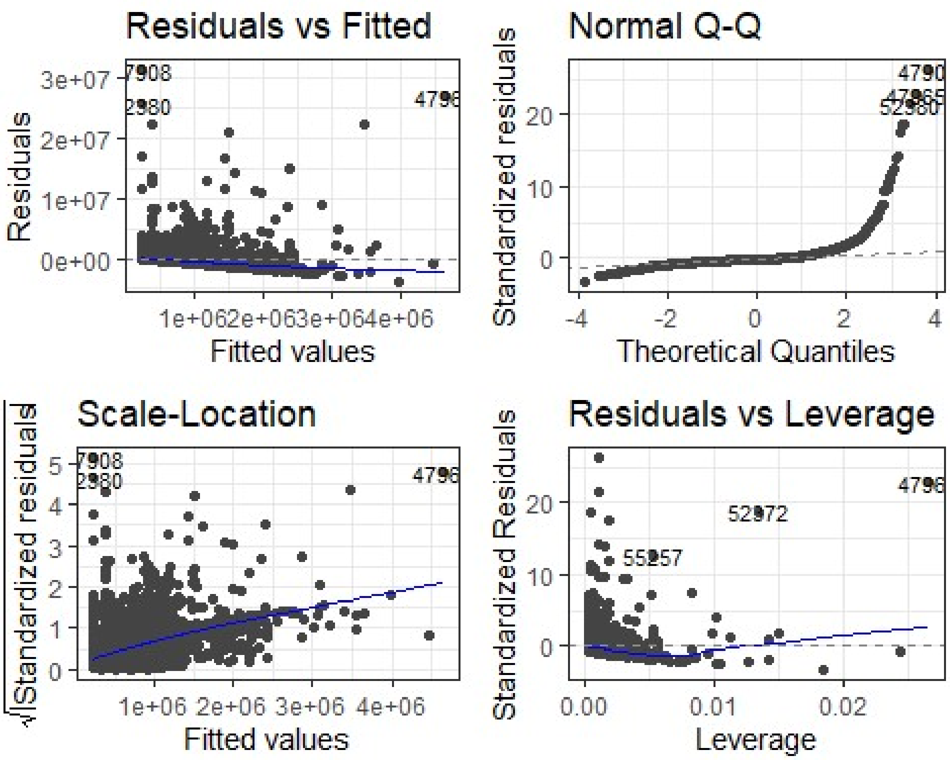
The above results show that signals are significant when controlling for speed.

We will now use AIC to investigate which model is better. Since signaling errors are a subset of human factors, we cannot include both in our final model due to multicollinearity. Based on the model assessment below, accdmg.lm1 is the better model. We will move forward with using human factors as a predictor and not signals.

```
AIC(accdmg.lm1)

## [1] 234299.4

AIC(accdmg.lm2)

## [1] 234359
```

diagnosing issues

```
autoplot(accdmg.lm1, which = c(1,2,3,5), ncol = 2, label.size = 3) +
theme_bw()
```

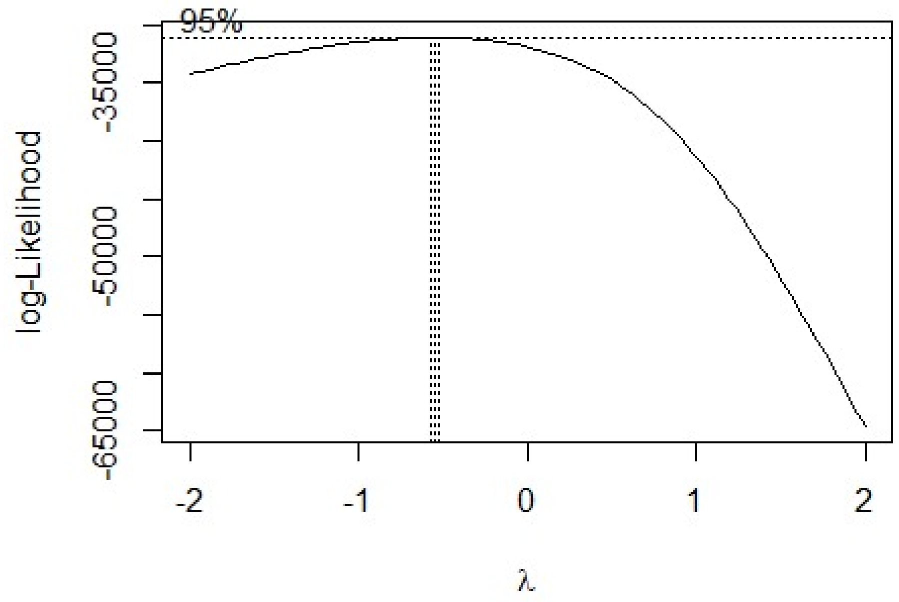


These plots (Fig. 13) show a violation of constant variance from the downward sloping trendline in the Residuals vs. Fitted graph. The error term is non-Gaussian, as seen in the skewed lower tail of the QQ Plot. The other plots show consistent assumption violations, which will be addressed in future transformations.

e) adjustments

To address the assumption violations discovered in the diagnostic plots, we first checked if a transformation is needed by creating a Box-Cox plot (Fig. 14) and to determine what transformation (log, exponential, or none) was required. As neither zero nor one was within the parameters, we opted for an exponential transformation with the optimal lambda value of -0.5 determined from the plot.

```
#Box-Cox Transformation
boxcox(accdmg.lm1) #box-cox plot
```

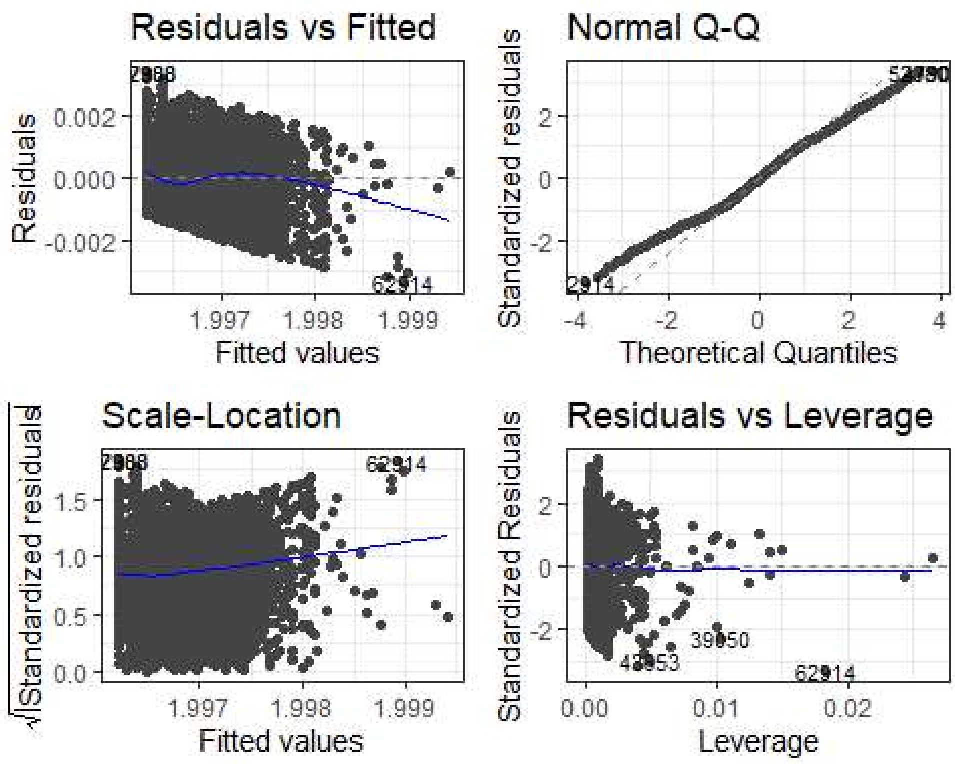


The plot above suggests that we need to use a boxcox transformation for ACCDMG. The optimal lambda value is:

```
L<-boxcox(accdmg.lm1, plotit = F)$x[which.max(boxcox(accdmg.lm1, plotit =
F)$y)]
L

## [1] -0.5

# boxcox transformation
accdmg.lm1.boxcox <- lm((ACCDMG^L-1)/L~(TRNSPD+human_factors)^2,data=xdmgnd)
summary(accdmg.lm1.boxcox)

##
## Call:
## lm(formula = (ACCDMG^L - 1)/L ~ (TRNSPD + human_factors)^2, data = xdmgnd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0033451 -0.0008175 -0.0000368  0.0007836  0.0034365
##
## Coefficients:
##                         Estimate Std. Error   t value Pr(>|t|)
## (Intercept)            1.996e+00  2.104e-05 94895.401  < 2e-16 ***
## TRNSPD                 2.160e-05  7.263e-07    29.739  < 2e-16 ***
## human_factors1        -1.724e-04  3.859e-05    -4.467 8.03e-06 ***
## TRNSPD:human_factors1  8.650e-06  1.889e-06     4.580 4.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001018 on 7597 degrees of freedom
## Multiple R-squared:  0.1458, Adjusted R-squared:  0.1454
## F-statistic: 432.1 on 3 and 7597 DF,  p-value: < 2.2e-16
```

After performing our transformation, we check model diagnostics. The new diagnostic plots (Fig. 15) show the model performs better regarding the assumptions. The Residuals vs. Fitted plot shows a slope of smaller magnitude than the previous plot. The error term fits a Gaussian distribution with a linear regression line on the QQ Plot. Both Scale-Location and Residuals vs. Leverage show similar improvements to the assumptions. These plots confirm the transformations performed above.

```
autoplot(accdmg.lm1.boxcox, which = c(1,2,3,5), ncol = 2, label.size = 3) +
theme_bw()
```

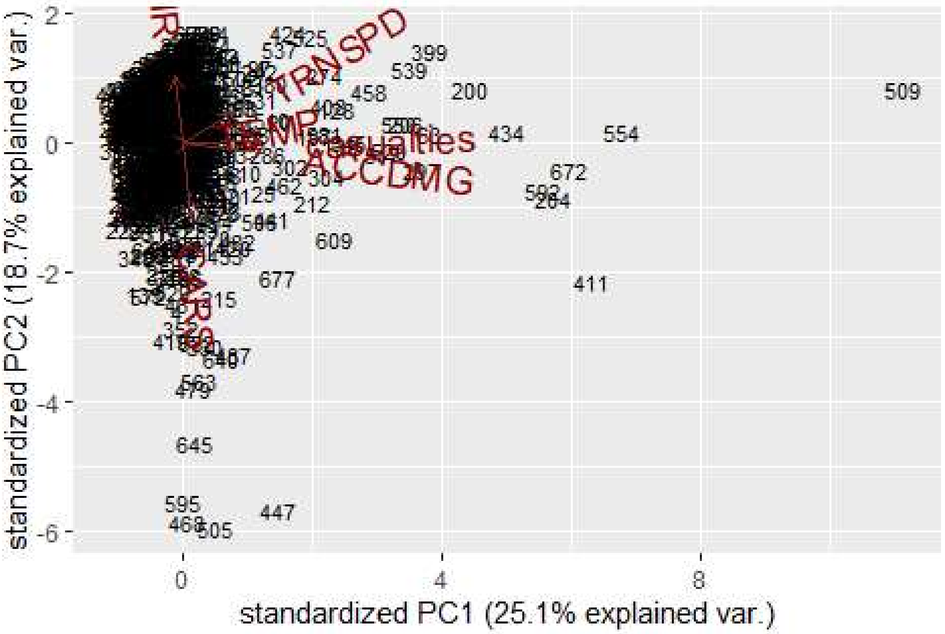


3. Casualties Analysis

a) feature and model selection techniques

Our first step in feature selection was to set up the casualties variable by adding total killed and total injured for every accident, as seen in the previous Cleaning section. We then defined a new data frame with only one or more casualties to use in our analysis. To identify quantitative predictors that we could use in our model, we use PCA analysis (Fig. 16):

```
predictors.casualties.pca <- princomp(xdmgnd_cas[,c("casualties", "TRNSPD",
"CARS", "TIMEHR", "TEMP","ACCDMG")], cor = T )

ggbiplot(predictors.casualties.pca, varname.size = 5,
labels=row(xdmgnd_cas)[,1])
```



Based on the biplot displayed above (Fig. 16), we can see that TRNSPD and ACCDMG are correlated with casualties (though the temperature variable is pointing in the same direction, the arrow is very small). The correlation with ACCDMG suggests that the two severity metrics may have similar predictors. We can use TRNSPD in our casualties model and add predictors from our ACCDMG analysis if others found are not predictive enough. We discuss feature selection and treatment of ordinal and categorical variables in the following section.

b) treatment of ordinal and categorical variables

The first, and main categorical variable we investigate in our analysis is the CAUSE variable. We recoded CAUSE to a factor with 5 levels that represent the 5 broad classes of accident causes (M: miscellaneous, T: rack, roadbed and structures, S: signal and communication, H: human factors, E: mechanical and electrical failures). A bar plot of Casualty Frequency by Cause (Fig. 8) showed the relative frequencies of accidents with casualties among each cause, and demonstrated that accidents caused by “miscellaneous - other factors” were the most common in our casualties dataframe. This led us to break down the variable further to examine specific errors that occurred within accidents due to “miscellaneous” errors, using the first two indices of the cause code (M1XX, M2XX, etc.). Further breakdown is more useful so that we can actually understand what is leading to a higher number of casualties. We find that most of these accidents were cited as caused by loading procedures. Based on this, we created a loading procedures binary variable to use in our model for our third hypothesis: that errors in loading procedures are associated with disproportionately more casualties.

For our second hypothesis, we recode the TYPE variable into a hwyrail binary dummy variable to select out and use in our model. This then determines accidents of type hwyrail and group the rest of the other types (derail, Head on, Rear End, etc…). We then recode the conductor variable to a binary dummy variable as our notion surrounded the idea of having "at least one conductor on duty." We are then able to use an interaction plot to see that having a conductor on board was associated with lower casualties in highway rail accidents, thereby informing our hypothesis (Fig. 11).

Hypothesis 3: Errors in loading procedures lead to disproportionately more casualties.

H0 = casualties in accidents caused by loading errors are the same as all other causes. HA = casualties in accidents caused by loading errors are higher than other human errors.

```
casualties.lm1 <- lm(casualties~loading+TRNSPD,data=xdmgnd_cas)
summary(casualties.lm1)

##
## Call:
## lm(formula = casualties ~ loading + TRNSPD, data = xdmgnd_cas)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23.46  -9.04  -4.79  -0.50 381.14
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.50453    2.03411   0.740 0.459763
## loading1    -8.96036    2.38124  -3.763 0.000182 ***
## TRNSPD       0.29430    0.05309   5.543 4.23e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.85 on 689 degrees of freedom
## Multiple R-squared:  0.04562,    Adjusted R-squared:  0.04285
## F-statistic: 16.47 on 2 and 689 DF,  p-value: 1.032e-07
```

Hypothesis 4: Trains with conductors are more likely to reduce the number of casualties in highway-rail accidents.

H0 = For highway-rail type accidents, the number of casualties on trains with conductors is equal to the number of casualties on trains without conductors.

HA = For highway-rail type accidents, the number of casualties on trains with conductors is less than the number of casualties on trains without conductors.

```
casualties.lm2 <- lm(casualties~(conduct+hwyrail+TRNSPD)^2,data=xdmgnd_cas)
summary(casualties.lm2)

##
## Call:
## lm(formula = casualties ~ (conduct + hwyrail + TRNSPD)^2, data =
xdmgnd_cas)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31.52  -8.70  -3.89   0.19 383.05
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.07174    6.35350   0.169    0.866
## conduct1           -2.02791    6.66918  -0.304    0.761
## hwyrail1            4.64990   13.39353   0.347    0.729
## TRNSPD              0.09035    0.27465   0.329    0.742
## conduct1:hwyrail1  -7.99409   13.60418  -0.588    0.557
## conduct1:TRNSPD     0.27352    0.27381   0.999    0.318
## hwyrail1:TRNSPD    -0.16013    0.11866  -1.349    0.178
##
## Residual standard error: 27.83 on 685 degrees of freedom
## Multiple R-squared:  0.05303,    Adjusted R-squared:  0.04473
## F-statistic: 6.393 on 6 and 685 DF,  p-value: 1.452e-06
```

c) model assessment

Performing a partial-f between casualties.lm2 and casualties.lm3 test shows that there is no significant difference between the bigger (casualties.lm3) and smaller (casualties.lm2) models, so we move forward with casualties.lm2.

Next, we performed a stepwise regression on casualties.lm2 (casualties.lm2.step) to reduce insignificant parameters. Following this, we compared the AIC between casualties.lm1 and casualties.lm2.step. Given its lower AIC value of 6570.99 compared to 6573.322, we choose casualties.lm2.step.

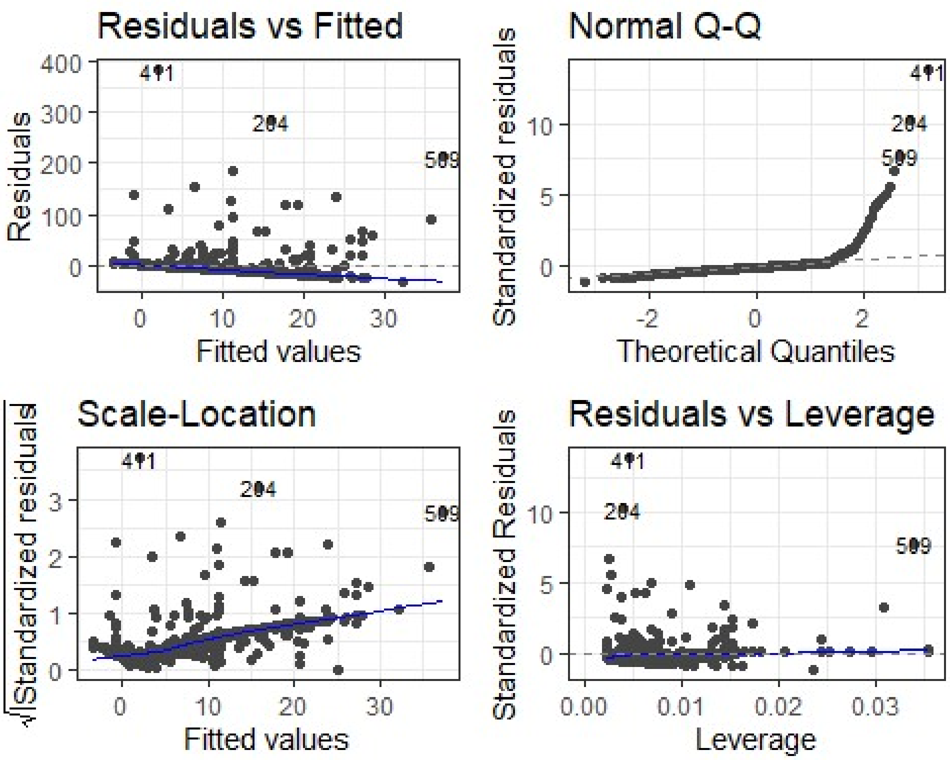
```
# all predictors from lm1 and lm2
casualties.lm3 <-
lm(casualties~(conduct+hwyrail+TRNSPD+loading)^2,data=xdmgnd_cas)

anova(casualties.lm2,casualties.lm3)

## Analysis of Variance Table
##
## Model 1: casualties ~ (conduct + hwyrail + TRNSPD)^2
## Model 2: casualties ~ (conduct + hwyrail + TRNSPD + loading)^2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    685 530398
## 2    681 526673  4    3724.2 1.2039 0.3078
```

```
casualties.lm2.step <- step(casualties.lm2, trace=T)

## Start:  AIC=4610.12
## casualties ~ (conduct + hwyrail + TRNSPD)^2
## 
##                     Df Sum of Sq    RSS    AIC
## - conduct:hwyrail  1    267.37 530665 4608.5
## - conduct:TRNSPD   1    772.65 531170 4609.1
## - hwyrail:TRNSPD   1   1410.12 531808 4610.0
## <none>                          530398 4610.1
## 
## Step:  AIC=4608.47
## casualties ~ conduct + hwyrail + TRNSPD + conduct:TRNSPD + hwyrail:TRNSPD
## 
##                    Df Sum of Sq    RSS    AIC
## - conduct:TRNSPD  1    505.71 531171 4607.1
## <none>                         530665 4608.5
## - hwyrail:TRNSPD  1   1711.45 532376 4608.7
## 
## Step:  AIC=4607.13
## casualties ~ conduct + hwyrail + TRNSPD + hwyrail:TRNSPD
## 
##                    Df Sum of Sq    RSS    AIC
## - conduct         1      36.8 531207 4605.2
## <none>                         531171 4607.1
## - hwyrail:TRNSPD  1    1735.8 532906 4607.4
## 
## Step:  AIC=4605.18
## casualties ~ hwyrail + TRNSPD + hwyrail:TRNSPD
## 
##                    Df Sum of Sq    RSS    AIC
## <none>                         531207 4605.2
## - hwyrail:TRNSPD  1    1764.5 532972 4605.5

AIC(casualties.lm1)

## [1] 6573.322

AIC(casualties.lm2.step)

## [1] 6570.99
```

d) diagnosing issues

The following diagnostic plots (Fig. 17) show a violation of constant variance from the downward sloping trendline in the Residuals vs. Fitted graph. The error term is non-Gaussian, as seen in the skewed lower tail of the QQ Plot. The other plots show consistent assumption violations, which will be addressed in future transformations.

```
autoplot(casualties.lm2.step, which = c(1,2,3,5), ncol = 2, label.size = 3) +
theme_bw()
```

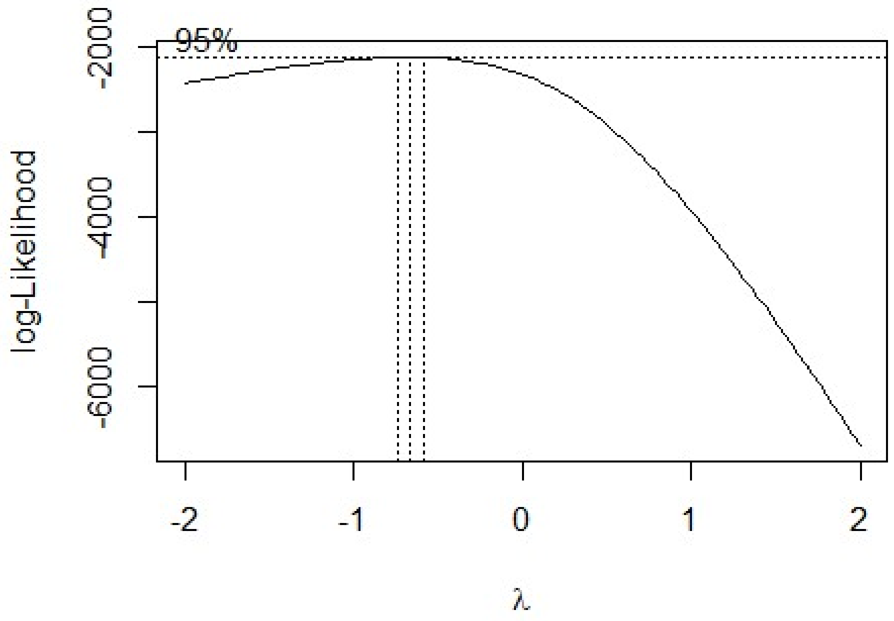


e) adjustments

Exactly the same as the accident damage model, the assumption violations discovered in the diagnostic plots warranted a transformation. As such, we created a Box-Cox plot (Fig. 18) to determine what transformation is best. As the confidence interval does not include zero or one, we used an exponential transformation using the optimal lambda value obtained from the box cox plot.

Diagnostics above show we need to transform the casualties variable:

```r
#Box-Cox Transformation
boxcox(casualties.lm2.step) #box-cox plot
```



```
# boxcox transformation
L<-boxcox(accdmg.lm1, plotit = F)$x[which.max(boxcox(accdmg.lm1, plotit =
F)$y)]

casualties.lm2.boxcox <- lm((casualties^L-
1)/L~(conduct+hwyrail+TRNSPD)^2,data=xdmgnd_cas)
summary(casualties.lm2.boxcox)

##
## Call:
## lm(formula = (casualties^L - 1)/L ~ (conduct + hwyrail + TRNSPD)^2,
##     data = xdmgnd_cas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07002 -0.42997 -0.00192  0.31091  1.52636
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.4112589  0.1177482   3.493 0.000509 ***
## conduct1          -0.1060546  0.1235985  -0.858 0.391161
## hwyrail1           0.1151219  0.2482196   0.464 0.642945
## TRNSPD             0.0005843  0.0050900   0.115 0.908636
## conduct1:hwyrail1 -0.2474165  0.2521235  -0.981 0.326776
## conduct1:TRNSPD    0.0077289  0.0050744   1.523 0.128194
## hwyrail1:TRNSPD    0.0005574  0.0021991   0.253 0.799970
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5157 on 685 degrees of freedom
## Multiple R-squared:  0.09559,    Adjusted R-squared:  0.08767
## F-statistic: 12.07 on 6 and 685 DF,  p-value: 6.483e-13
```

Create a step model with casualties transformed:

```
casualties.lm2.boxcox.step <- step(casualties.lm2.boxcox, trace=T)
```

```
## Start:  AIC=-909.56
## (casualties^L - 1)/L ~ (conduct + hwyrail + TRNSPD)^2
##
##                    Df Sum of Sq    RSS     AIC
## - hwyrail:TRNSPD   1   0.01709 182.19 -911.50
## - conduct:hwyrail  1   0.25611 182.43 -910.59
## <none>                         182.17 -909.56
## - conduct:TRNSPD   1   0.61695 182.79 -909.22
##
## Step:  AIC=-911.5
## (casualties^L - 1)/L ~ conduct + hwyrail + TRNSPD + conduct:hwyrail +
##     conduct:TRNSPD
##
##                    Df Sum of Sq    RSS     AIC
## - conduct:hwyrail  1   0.24032 182.43 -912.59
## <none>                         182.19 -911.50
## - conduct:TRNSPD   1   0.60136 182.79 -911.22
##
## Step:  AIC=-912.59
## (casualties^L - 1)/L ~ conduct + hwyrail + TRNSPD + conduct:TRNSPD
##
##                   Df Sum of Sq    RSS     AIC
## - conduct:TRNSPD  1   0.36552 182.80 -913.20
## <none>                        182.43 -912.59
## - hwyrail         1   1.18317 183.61 -910.11
##
## Step:  AIC=-913.2
## (casualties^L - 1)/L ~ conduct + hwyrail + TRNSPD
##
##           Df Sum of Sq    RSS     AIC
## - conduct  1    0.0182 182.81 -915.13
## <none>                 182.80 -913.20
## - hwyrail  1    1.2075 184.00 -910.65
## - TRNSPD   1   17.3269 200.12 -852.53
##
## Step:  AIC=-915.13
## (casualties^L - 1)/L ~ hwyrail + TRNSPD
##
##           Df Sum of Sq    RSS     AIC
## <none>                 182.81 -915.13
## - hwyrail  1    1.1947 184.01 -912.63
## - TRNSPD   1   17.7338 200.55 -853.07
```

```
summary(casualties.lm2.boxcox.step)
```

```
## 
## Call:
## lm(formula = (casualties^L - 1)/L ~ hwyrail + TRNSPD, data = xdmgnd_cas)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.05874 -0.43372  0.00269  0.30151  1.52117 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.3119680  0.0374882   8.322 4.65e-16 ***
## hwyrail1    -0.0989881  0.0466494  -2.122   0.0342 *  
## TRNSPD       0.0081171  0.0009929   8.175 1.42e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5151 on 689 degrees of freedom
## Multiple R-squared:  0.09241,    Adjusted R-squared:  0.08977 
## F-statistic: 35.08 on 2 and 689 DF,  p-value: 3.114e-15
```

```
AIC(casualties.lm2.boxcox.step)
```

```
## [1] 1050.678
```

The AIC shows that the model transformed according to the recommendation of the Box-Cox transformation (casualties.lm2.box.step).

4. Recommendation

a) Assess whether or not you can reject your null hypotheses

As seen by our final accident damage model (accdmg.lm1.boxcox), we are able to determine that human factors, train speed, and the interaction between the two are significant. As such, we can reject the first null hypothesis, that accident damage for all causes are equal at high train speeds. With this, we concur that accident damage associated with human factor errors is not equivalent to other error types. Furthermore, with our initial testing of models, we can see that signals and train speed are significant. While this was not as strong of a model (according to AIC), we can still reject the second null hypothesis and conclude that signaling errors are associated with a greater amount of accident damage.

In our first preliminary model for casualties, we reject the null hypothesis that the number of casualties in accidents with loading errors are equivalent to other errors. Similarly, our initial model for the fourth hypothesis is significant as a whole, meaning we reject our null hypothesis that the number of casualties on trains with conductors is equal to the number of casualties on trains without conductors. However, this is likely an area for improved data as specific predictors were not significant themselves.

b) Summarize your findings and your recommendations for safety improvements based on the evidence you have discovered and include next steps the FRA should take to reduce the severity of rail accidents.

While all of our null hypotheses were able to be rejected through models generated through the course of our analysis, the final optimized models represent much more actionable recommendations. These other models represent several areas where data collection and experiments may be of high value in generating important conclusions in the future.

The actionable recommendations from our finalized models primarily surround the train speed and human factors, as well as train speed and highway rail incidents. As such, we believe that there is increased value to be had by allocating resources towards human factor associated areas, such as training for employees and implementing greater oversight for new employees. With the importance of train speed considered, we believe that it is critical that employees tasked with train operation of high speed trains have extensive experience and are not overworked. Furthermore, we recognize that high speed trains are significantly important in the prevalence of highway rail accidents. As such, we recommend imposing strict rules around speed reduction within 1000 meters of any highway or road crossing. We believe there may also be value in investing in greater safety measures around rail and highway crossing and spaces where the two are in close proximity.

Lastly, we believe that our analysis shows areas where experimentation and greater data collection may provide critical information and conclusions in the future. Our analysis shows that, while not the most critical we found, loading and signaling should be understood better in the future as they likely are impactful to some extent. Experiments or data collection around loading and signaling practices would be likely to play a role in developing greater knowledge around these practices, improving them and potentially saving lives and money in the long run.

5. References

[1] "Freight Rail Safety Record," Association of American Railroads. https://www.aar.org/issue/freight-rail-safety-record/