Group Project 1

Preet Shah, Lea Jih-Vieira, Mitchell Whalen, & Blake Zimbardi

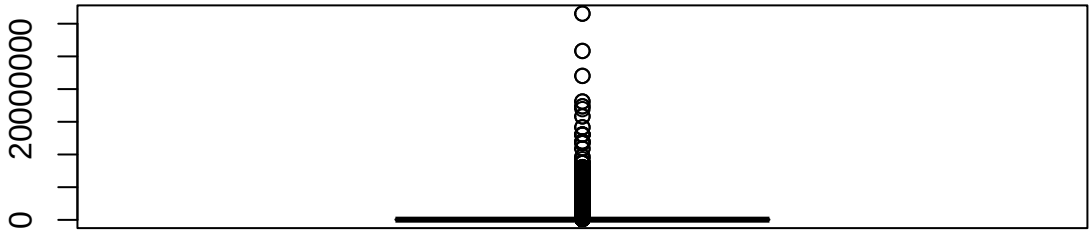
2023-10-29

Setup

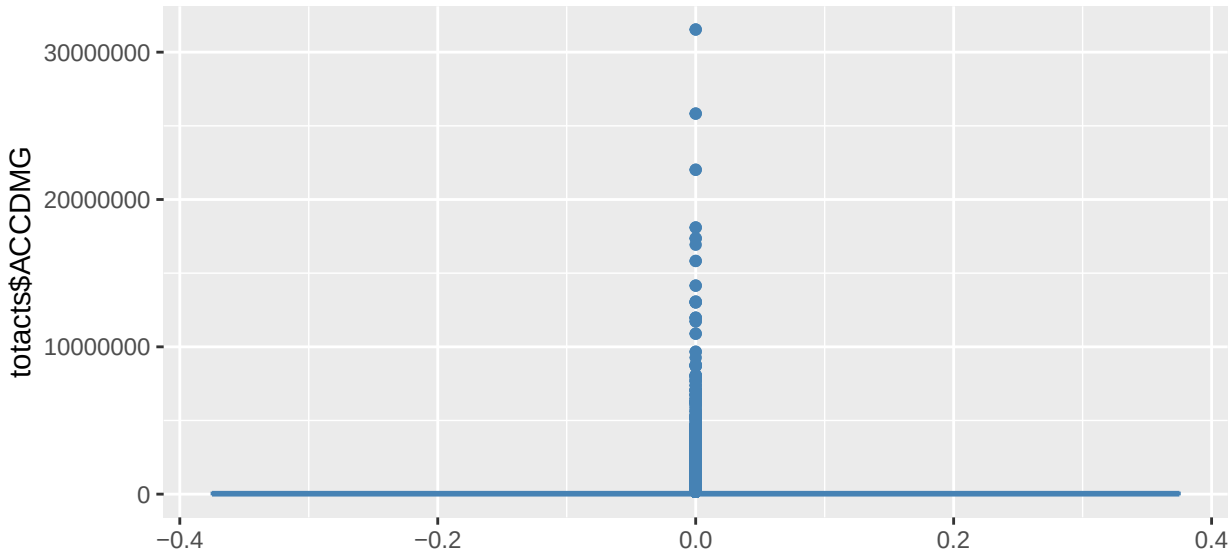
Load and filter data to extreme accidents

Here we are loading the accident data, assigning labels, and then filtering to only extreme accidents. Extreme accidents are all accidents with damages above the upper whisker

```
acts <- file.inputl(traindir)

totacts <- combine.data(acts)

# Convert Type to a factor and give more meaningful labels
totacts$Type <- factor(totacts$TYPE, labels = c("Derailment", "HeadOn", "Rearend", "Side", "Raking", "Bi

# Setup categorical variables
totacts$Cause <- rep(NA, nrow(totacts))

totacts$Cause[which(substr(totacts$CAUSE, 1, 1) == "M")] <- "M"
totacts$Cause[which(substr(totacts$CAUSE, 1, 1) == "T")] <- "T"
totacts$Cause[which(substr(totacts$CAUSE, 1, 1) == "S")] <- "S"
totacts$Cause[which(substr(totacts$CAUSE, 1, 1) == "H")] <- "H"
totacts$Cause[which(substr(totacts$CAUSE, 1, 1) == "E")] <- "E"

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

totacts$Cause <- factor(totacts$Cause)

# Now convert to factor with meaningful labels
totacts$TYPEQ <- factor(totacts$TYPEQ, labels = c("NA", "NA", "Freight", "Passenger", "Commuter","Work"

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

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




```
ggplot(as.data.frame(totacts$ACCDMG), aes(x=totacts$ACCDMG)) +
  geom_boxplot(col= "steelblue") + theme(plot.title = element_text(hjust = 0.5)) + coord_flip()
```




```
xdmg <- totacts[totacts$ACCDMG > dmgbox$stats[5],]

#remove 9/11
xdmg <- xdmg[-181,]

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

Generating Hypothesis

Accident Damage Hypotheses

ACCDMG Hypothesis 1 For our analysis of Hypothesis 1, we need to create three new binary columns: hError, crossing, and derail.

- hError: Indicates whether or not the primary cause of the accident was human error (0: No, 1: Yes)
- crossing: Indicates whether or not the accident occurred at a highway-rail or railroad grade crossing (0: No, 1: Yes)
- derail: Indicates whether or not the accident was a derailment (0: No, 1: Yes)

One thing to note regarding our use of railroad crossing data is that our data is potentially inaccurate in representing unprotected railroad crossings as we discussed in our research phase. As we mentioned earlier, our background research identified unprotected railroad crossings as a common cause of train accidents; however, our dataset does not contain information regarding unprotected railroad crossings. Instead, our dataset only provides information about highway-rail crossings and railroad grade crossings, while also posessing a code for "other" types of accidents (but no further detail on the "other" column). Highway-rail crossings and railroad grade crossings are known to have more signage and safety measures in place compared to unprotected crossings, which typically only have stop signs and do not possess gates (LA Times). However, we felt that since we still had data on some kinds of crossing-related accidents, it would still be informative to study this relationship. We do not have information as to what exact safety precautions were in place at the crossings in our dataset, as some may have had more in place than others, so we determined it would still be a useful to investigate.

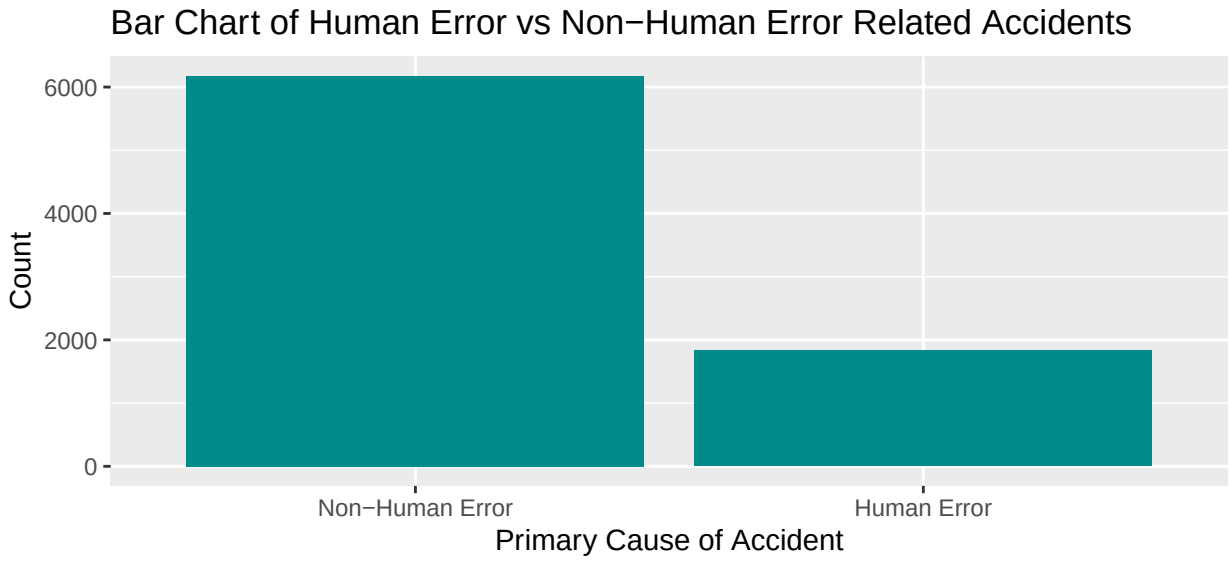
```
# hError
xdmgnd$hError <- 0

for (i in 1:nrow(xdmgnd)) {
```

```
  code <- substr(xdmgnd[i, "CAUSE"], 1, 1)

  if (code == "H") {
    xdmgnd[i, "hError"] <- 1
  }
}

# crossing
xdmgnd$crossing <- 0

for (i in 1:nrow(xdmgnd)) {
  # Hwy-rail crossing or RR Grade Crossing
  xdmgnd[i, "crossing"] <- xdmgnd$TYPE[i] == 7 || xdmgnd$TYPE[i] == 8
}

# derail
xdmgnd$derail <- 0

for (i in 1:nrow(xdmgnd)) {
  xdmgnd[i, "derail"] <- xdmgnd$TYPE[i] == 1 # Derailment
}
```

The only predictor variable that we did not need to create or transform was train speed, our only continuous predictor for Hypothesis 1.

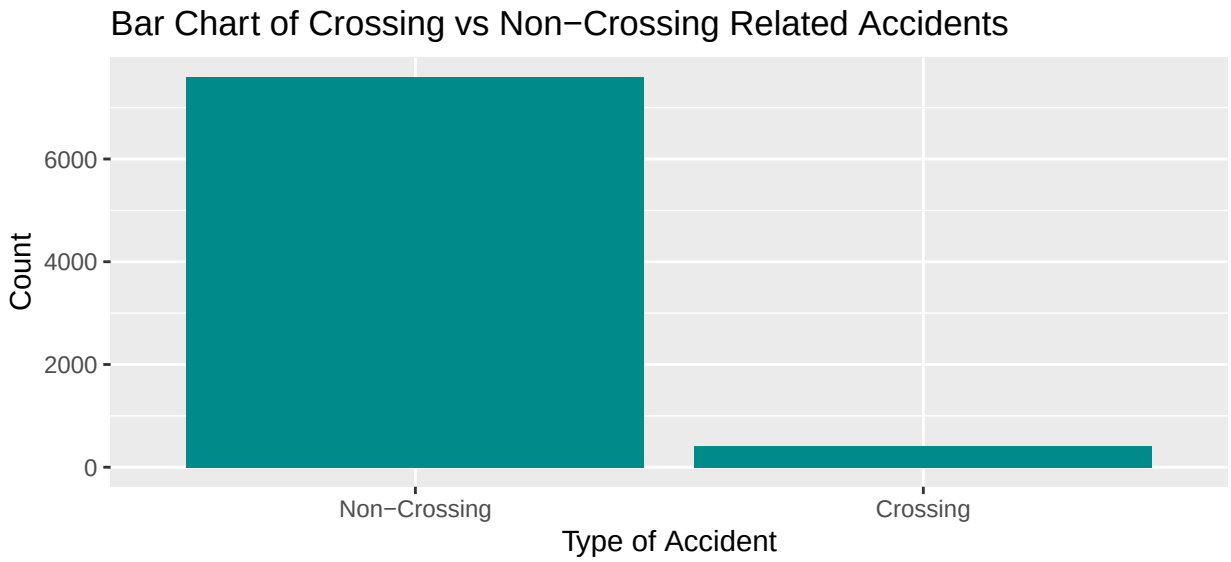
Next, we conducted some exploratory analysis of our predictor variables via visualizations and summary statistics to get a sense of how our model may look.

```
# hError
ggplot(xdmgnd, aes(x = as.factor(hError))) +
  geom_bar(fill= "cyan4") +
  ggtitle("Bar Chart of Human Error vs Non-Human Error Related Accidents") +
  labs(y= "Count", x = "Primary Cause of Accident") +
  scale_x_discrete(labels=c("Non-Human Error", "Human Error"))
```

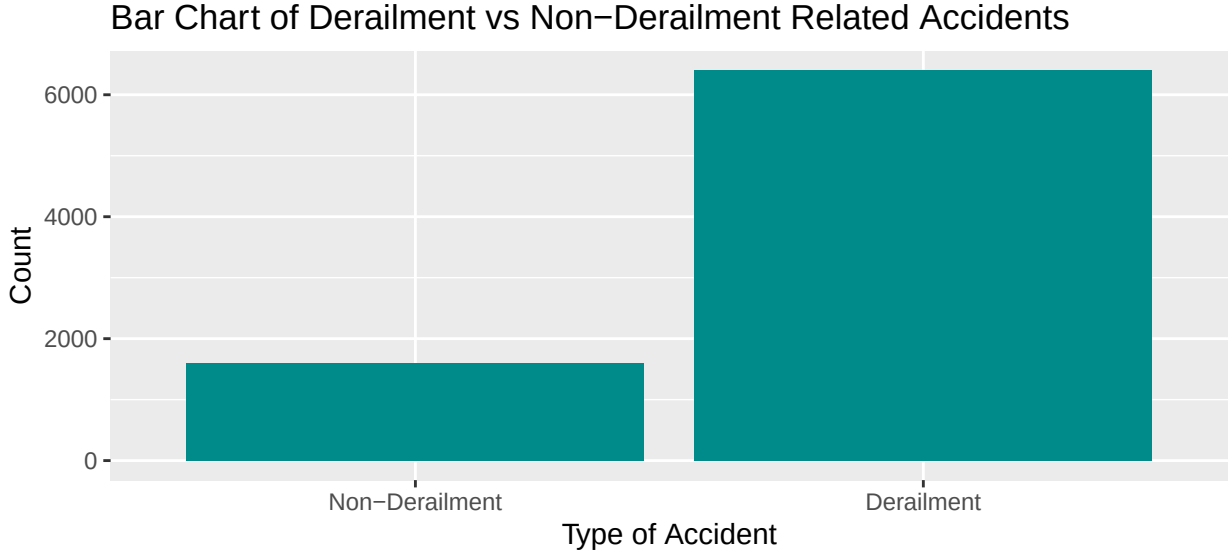



```
# crossing
ggplot(xdmgnd, aes(x = as.factor(crossing))) +
```

```
  geom_bar(fill= "cyan4") +
  ggtitle("Bar Chart of Crossing vs Non-Crossing Related Accidents") +
  labs(y= "Count", x = "Type of Accident") +
  scale_x_discrete(labels=c("Non-Crossing", "Crossing"))
```




```
# derail
ggplot(xdmgnd, aes(x = as.factor(derail))) +
  geom_bar(fill= "cyan4") +
  ggtitle("Bar Chart of Derailment vs Non-Derailment Related Accidents") +
  labs(y= "Count", x = "Type of Accident") +
  scale_x_discrete(labels=c("Non-Derailment", "Derailment"))
```

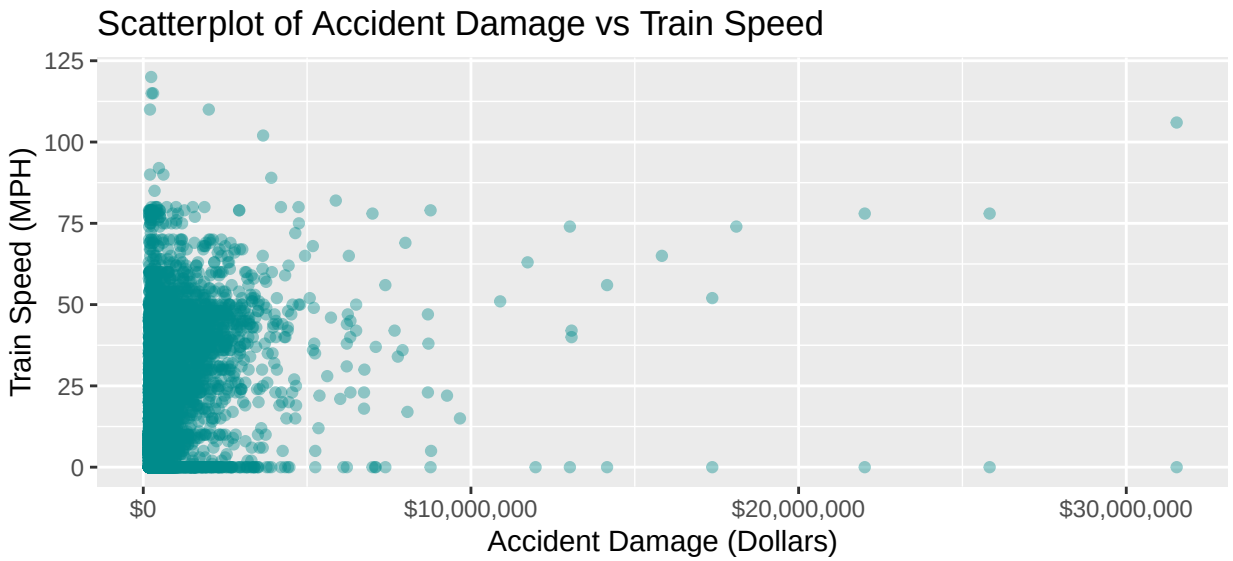


First we looked at bar charts of each of the binary predictor variables to get a sense of the frequency of each variable. For each predictor, it does appear to lean heavily in one direction vs the other, but still appears to have sufficient representation of both groups for most of the binary variables. Human error accidents are less frequent than non-human error accidents, but human error accidents still make up ~25% of the dataset. Derailments are more frequent than non-derailments, indicating that derailments may tend to incur high damage costs.

The only variable that seems to have less representation is crossing, as there is a noticeably lower frequency of crossing-related accidents compared to non-crossing. This observation could be indicative of the fact that there are not many crossing related accidents that incur high damage costs, since we are only looking at the subset of accidents above the upper whisker. Perhaps there are many more crossing-related accidents that are below the upper whisker threshold, but are excluded from our dataset due to our filtering guidelines. This highly skewed distribution of crossing-related accidents should not affect the model, and we have identified reason(s) why this skewedness makes sense in the context of this situation, but it is still important to draw attention to and keep in mind as we develop our regression model (Cross Validated).

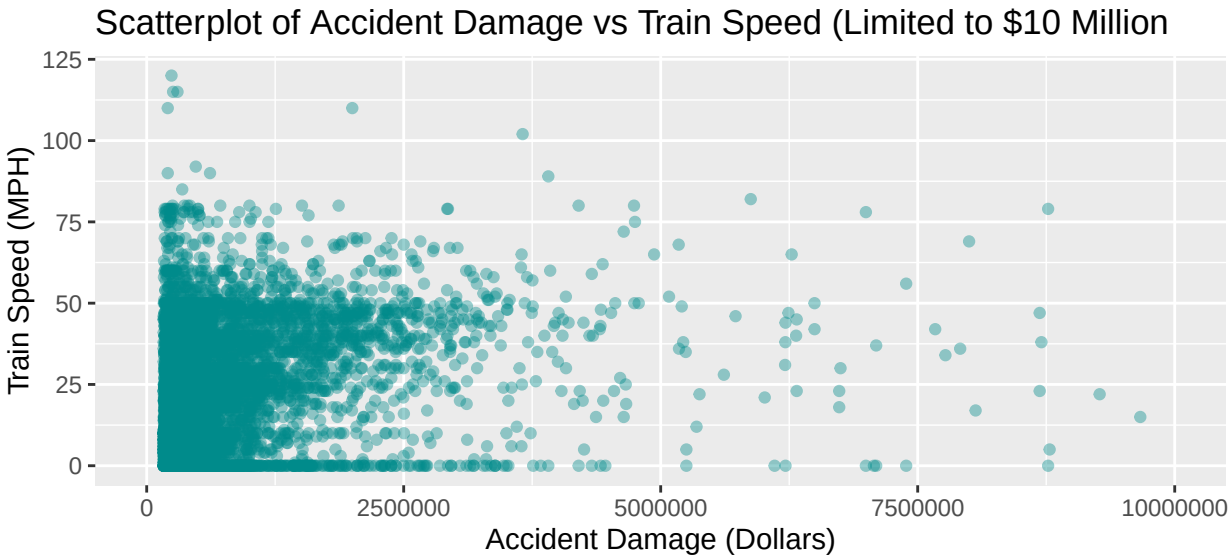
For our single continuous variable, train speed, we can make observations about how it interacts with damage costs by making a scatter plot of the two.

```
ggplot(xdmgnd, aes(x= ACCDMG, y= TRNSPD)) +
  geom_point(color= "cyan4", alpha = 0.4) +
  ggtitle("Scatterplot of Accident Damage vs Train Speed") +
  labs(y= "Train Speed (MPH)", x = "Accident Damage (Dollars)") +
  scale_x_continuous(labels = scales::dollar_format())
```



As we can observe from the figure above, most of the data is gathered on the left side of the plot, indicating that many of the accidents incur around the same damage costs regardless of the train speed. However, we do see a weak positive trend as we move left to right, showing that accident damages loosely seem to increase as train speed increases. Part of what is difficult about reading this figure is that the accident damage amounts are so spread out, with the majority of the data gathered on the left side of the graph. We can adjust our x-axis upper limit to get a better look at the behavior of the majority of the data, excluding the outliers.

```
ggplot(xdmgnd, aes(x= ACCDMG, y= TRNSPD)) +
  geom_point(color= "cyan4", alpha = 0.4) +
  ggtitle("Scatterplot of Accident Damage vs Train Speed (Limited to $10 Million") +
  labs(y= "Train Speed (MPH)", x = "Accident Damage (Dollars)") +
  scale_x_continuous(labels = scales::dollar_format()) + # FIX
  xlim(0, 10000000)
```



From this plot we can observe the interaction between accident damage and train speed in more detail. We can see more clearly here that the two have a positive linear relationship where accident damages tend to increase as train speed increases.

The last component of our exploratory analysis of Hypothesis 1 was to look at the Pearson correlation matrix to observe the variable correlations. In this case it does not make sense to plot a scatterplot matrix because most of the predictor variables are binary.

```
cor(xdmgnd[ , c("ACCDMG", "hError", "crossing", "derail", "TRNSPD")], method= "pearson")
```

```
##                ACCDMG       hError     crossing      derail      TRNSPD
## ACCDMG    1.000000000  0.003360779  0.009455751 -0.03635875  0.24901186
## hError    0.003360779  1.000000000 -0.111573359 -0.27157510 -0.24746550
## crossing  0.009455751 -0.111573359  1.000000000 -0.46335761  0.25007245
## derail   -0.036358748 -0.271575099 -0.463357612  1.00000000 -0.02488577
## TRNSPD    0.249011856 -0.247465499  0.250072454 -0.02488577  1.00000000
```

From this correlation matrix, we can see that train speed has the highest absolute correlation with accident damage with a correlation of 0.249. The other binary variables do not seem to have a high correlation with the outcome variable, which may indicate that our predictors will have limited inferential power of accident damages. However, we decided to move forward with building a full model with all four predictors to see what it looked like to then adjust accordingly after analyzing the initial model.

As per project instructions, we can only consider quantitative variables if they interact with a qualitative one. We chose to interact derailments with train speed, as they both seem to have the most strong relationship with accident damages.

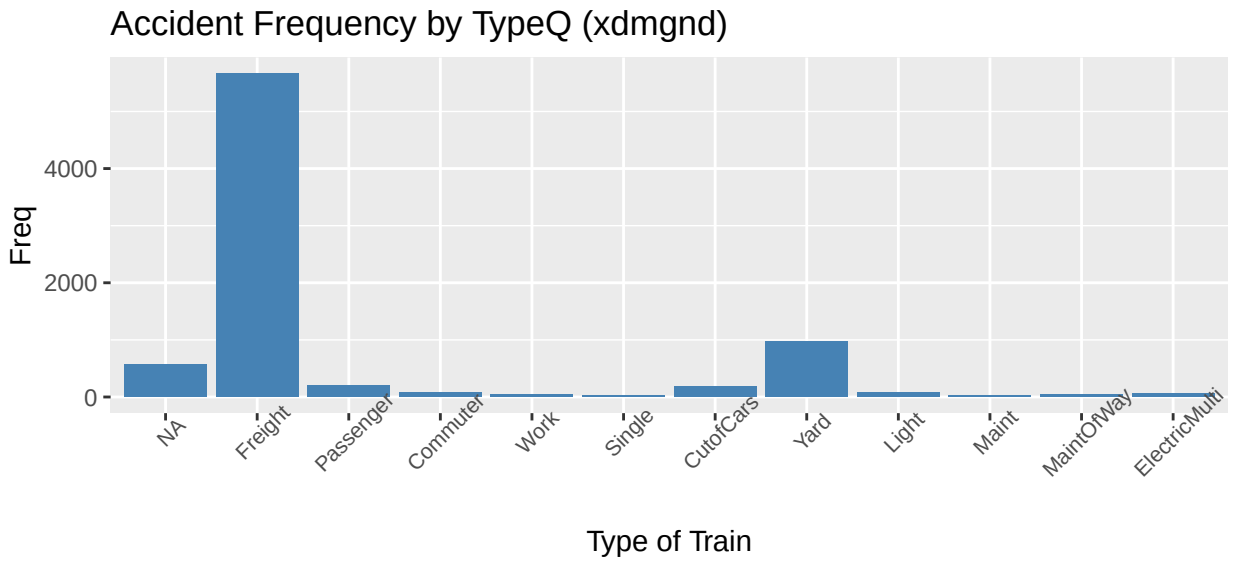
For the first hypothesis, we wanted to explore if common accident causes have any relationship with the amount of damage that occurs. From our outside research, we found that some of the most common causes of train accidents include human error, high speed trains, derailments, and unprotected railroad crossings (Gilreath & Associates). Knowing that these are some of the most common causes of train accidents, this gives us sufficient reason to study their relationship with accident damage costs. As per project instructions, we are studying only accidents with damages above the upper whisker, so we are also studying if the most common causes of train accidents result in accidents with high damages due to this filtering.

H0: Human error, crossing-related, high-speed, derailment accidents do not have a statistically significant relationship with accident damage costs.

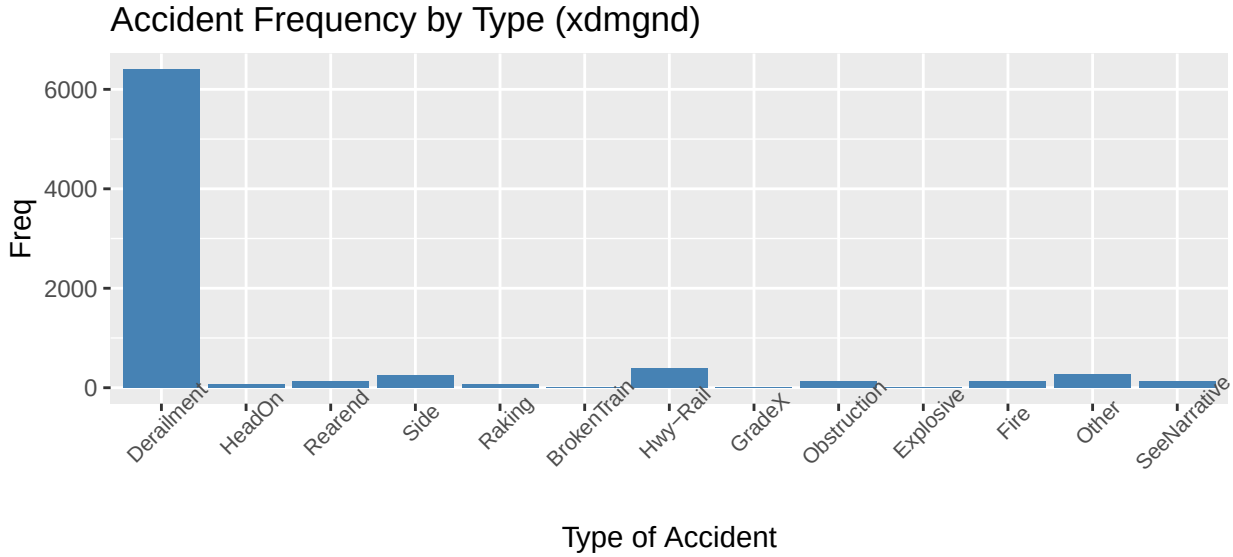
Ha: Human error, crossing-related, high-speed, derailment accidents have a statistically significant relationship with accident damage costs.

ACCDMG Hypothesis 2 Next we looked at the frequency of the different types of trains and types of accidents as well as their interaction with train speed in order to discern any notable patterns.

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

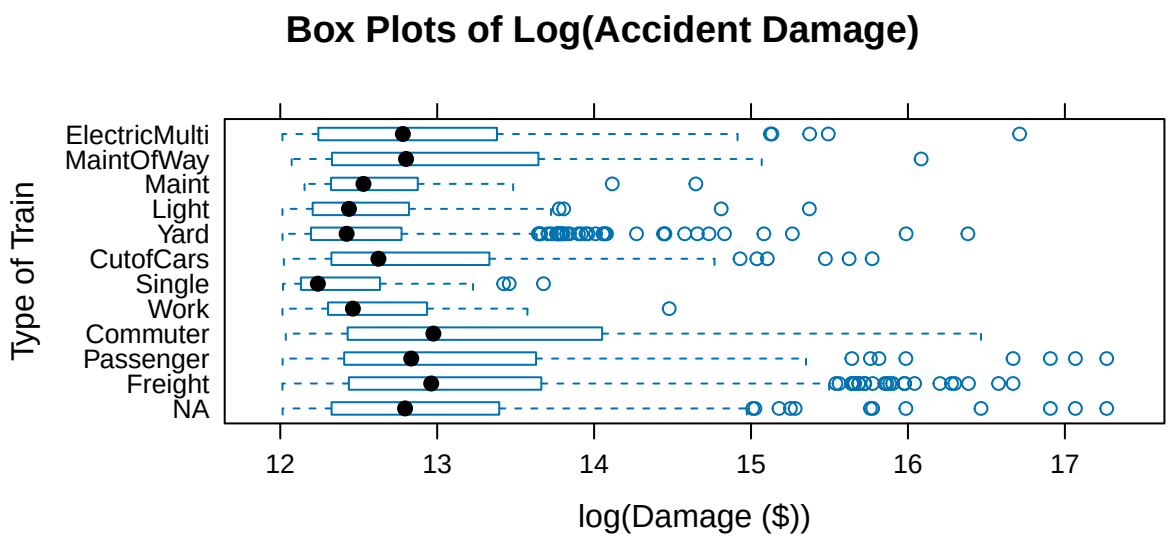



```
# Bar Graph of Type in xdmgnd
ggplot(as.data.frame(table(xdmgnd$Type)), aes(x = Var1, y= Freq)) +
  geom_bar(stat="identity",fill= "steelblue")+
  ggtitle("Accident Frequency by Type (xdmgnd)") +
  labs(x = "Type of Accident")+
  theme(axis.text.x = element_text(size = 8,  angle = 45))
```

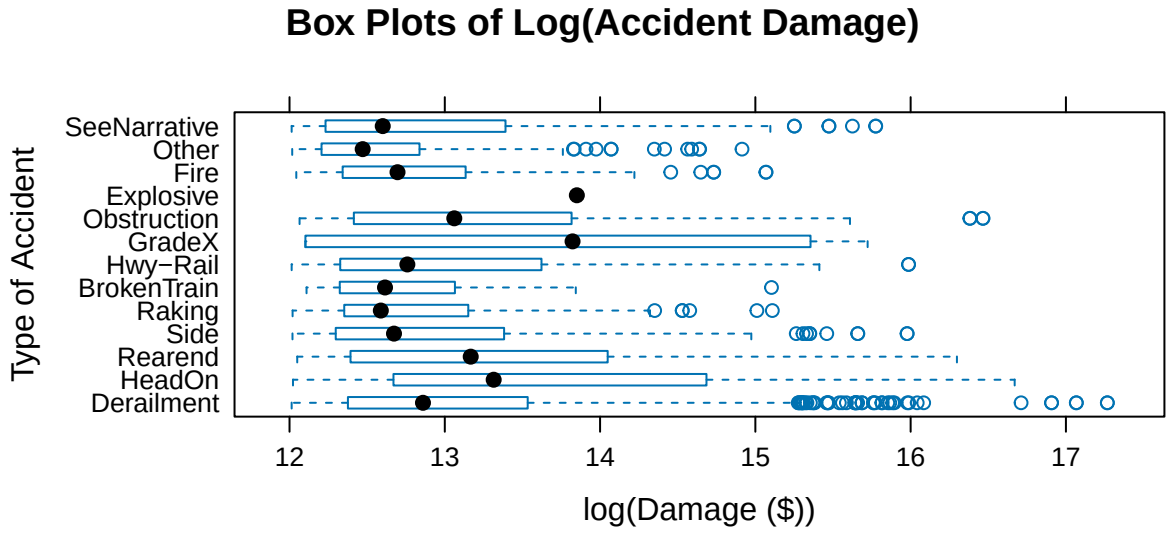


Clearly the most extreme accidents most frequently involve freight trains and derailment.

```
# Plot scaled (log) accident damage grouped by TypeQ using bwplot (from lattice package)
bwplot(
  TYPEQ~ log(ACCDMG+1),
  main = "Box Plots of Log(Accident Damage)",
  xlab = "log(Damage ($))", ylab = "Type of Train", data = xdmgnd
)
```



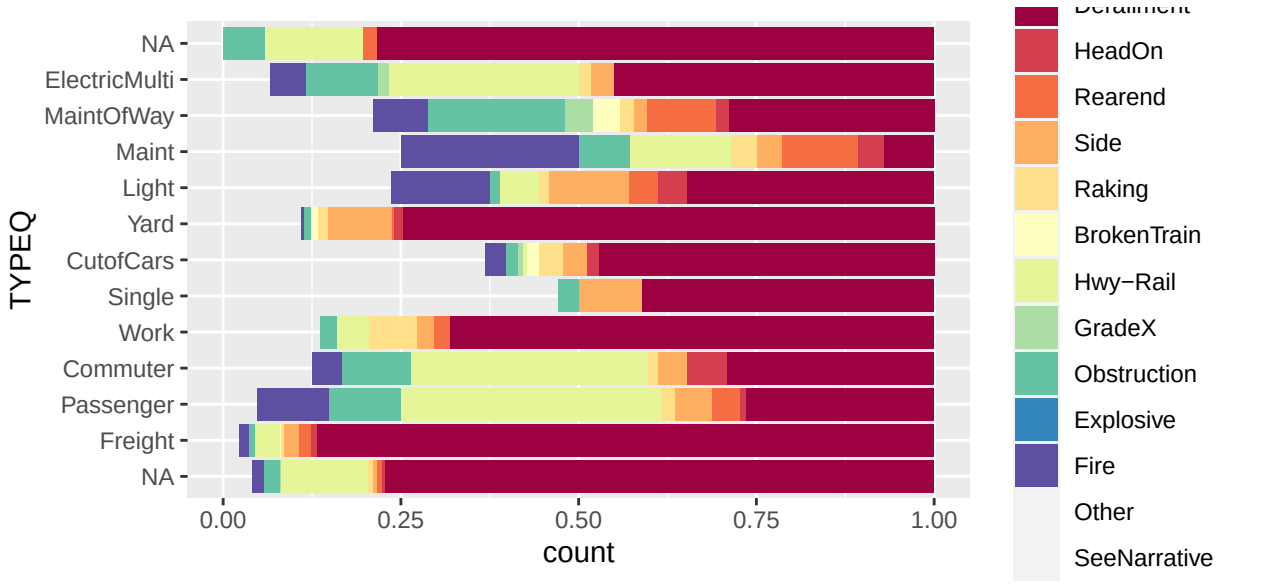

```
# Plot scaled (log) accident damage grouped by Type using bwplot (from lattice package)
bwplot(
  Type~ log(ACCDMG+1),
  main = "Box Plots of Log(Accident Damage)",
  xlab = "log(Damage ($))", ylab = "Type of Accident", data = xdmgnd
)
```



Freight, Commuter, & Passenger trains typically have the greatest damages, while the different types of accidents don't seem to vary much in terms of accident damage (GradeX has a very small sample size), however we can see an head on collisions seem to be especially damaging and derailment accidents have high potential for damage as shown by their outliers.

```
# types of each type of train's accidents
ggplot(xdmgnd, aes(x=TYPEQ, fill=Type), reorder(Type)) +
```

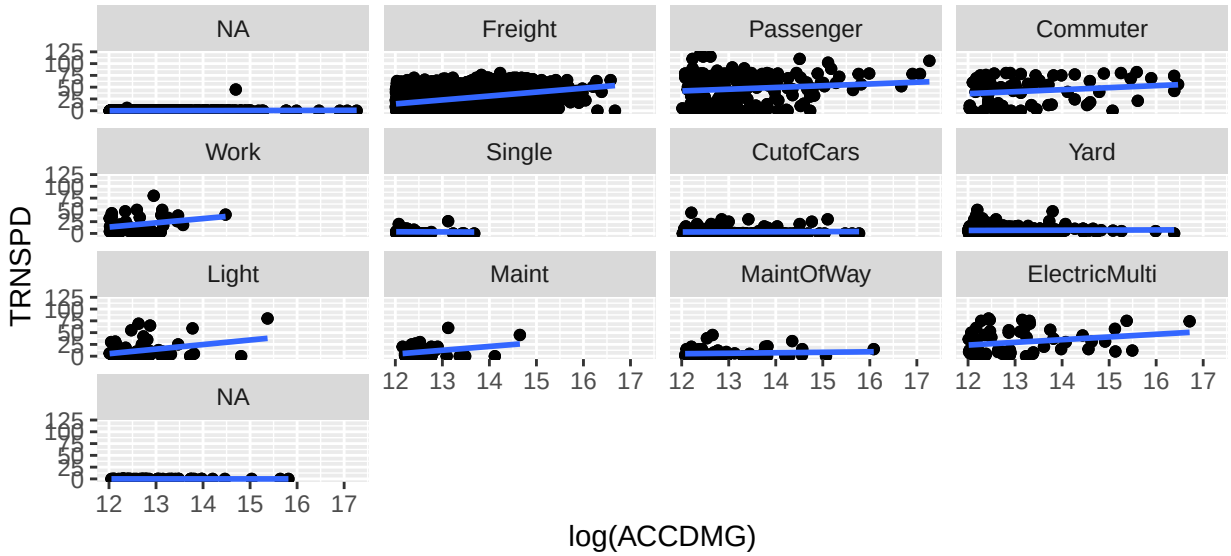
```
  geom_bar(position="fill") +
  scale_fill_brewer(palette = "Spectral") + #https://r-graph-gallery.com/38-rcolorbrewers-palettes.html
  coord_flip()
```



The most apparent pattern to us was the difference in types of accidents between Freight and Passenger/Commuter trains. AS you can see freight trains were typically involved in derailment accidents, while passenger/commuter trains were far more frequently involved in Hwy_Rail Accidents. NOTE: Moving forward we will group Passenger & Commuter trains since both serve a similar purpose and carry passengers.

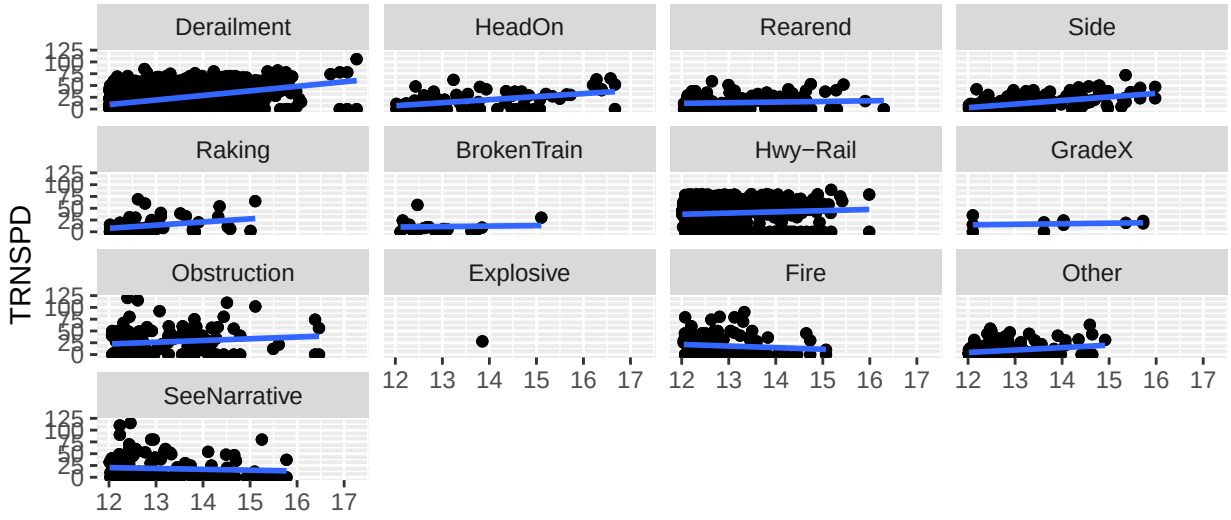
The last aspect we wanted to explore was train speed. Because different tyoes of trains may operate in different fashions, we suspect that this could influence their speed and potential to affect accidents.

```
# Looking for interactions between TRNSPD & ACCDMG by TYPEQ
qplot(log(ACCDMG), TRNSPD, data = xdmgnd) +  geom_point() +
  geom_smooth(method = "lm", se = FALSE) + facet_wrap(~ TYPEQ)
```




```
# Looking for interactions between TRNSPD & ACCDMG by Type
qplot(log(ACCDMG), TRNSPD, data = xdmgnd) +  geom_point() +
```

```
geom_smooth(method = "lm", se = FALSE) + facet_wrap(~ Type)
```



There appears to be some interaction here with train speed. It appears there is at least a slight interaction between speed and accident damage when it comes to freight trains and commuter trains. There also appears to be an interaction with some accident types including derailments, head-on, and side accidents.

Based on the visualizations, one can see that the Freight train accidents are far more likely to result in extreme outcomes than accidents involving other types of trains. Most extreme freight train accidents are derailments. This contrasts, passenger and commuter trains where a greater proportion accidents are of the highway-rail type. Given these observations, We hypothesize that the type of train (specifically the levels of freight or passenger/commuter), the type of accident (specifically the levels of derailment or Hwy-Rail), and their interaction can be used to predict accident damage. In addition, it is apparent that there is a significant interaction between train speed and accident damage, and if we group by type of accident and type of train we can see that the strength of the interaction appears to vary based on type of accident and type of train.

H0: Type of train (PassCom or Freight), type of accident (Derail or Hwy_Rail), and train speed have no significant relationship with accident damage Ha: Type of train (PassCom or Freight), type of accident (Derail or Hwy_Rail), and train speed have a significant relationship with accident damage

Casualty Hypotheses

Casualty Hypothesis 1 H0: The type of train and type of accident in a railroad accident does not have a statistically significant relationship with casualties. Ha: The type of train and type of accident in a railroad accident have a statistically significant relationship with casualties.

These hypothesis are actionable since testing this hypothesis will reveal the type of train and type of accidents that are involved with significant impacts on casualties in train accidents. Train safety can be improved by further analyzing accidents involving the most statistically significant type of trains in a train accident to determine what safety components of the train design and operation (i.e. conductor visibility, braking mechanisms, training, regulations) could be changed in favor of reducing severity of tail accidents.

For preliminary investigation of these hypothesis, we generate box plots comparing total number of casualties vs. different types of trains and vs. different types of accidents. Similarly, we observe that the type of train with largest number of accidents involving 1 or more casualties are freight trains. Finally, the East Palestine train derailment on February 3rd, 2023 peaked our interest in Freight Train accidents. While no casualties were reported, other notable freight train accidents like the 1986 Miamisburg train derailment (also in Ohio) and the 2005 Graniteville train crash (in South Carolina) are similar freight train accidents that not only caused multiple casualties, but also released toxic chemicals to the environment.

East Palestine: https://www.theguardian.com/us-news/2023/feb/11/ohio-train-derailment-wake-up-call Miamisburg: https://www.washingtonpost.com/archive/politics/1986/07/09/17000-evacuated-after-derailment/f6a0f635-bebb-4342-8c62-23304b12876e/ Graniteville: https://web.archive.org/web/20140719212353/http://www.wjbf.com/story/21686984/federal-prosecutors-say-norfolk-southern-should-be-fined-for-graniteville-pollution

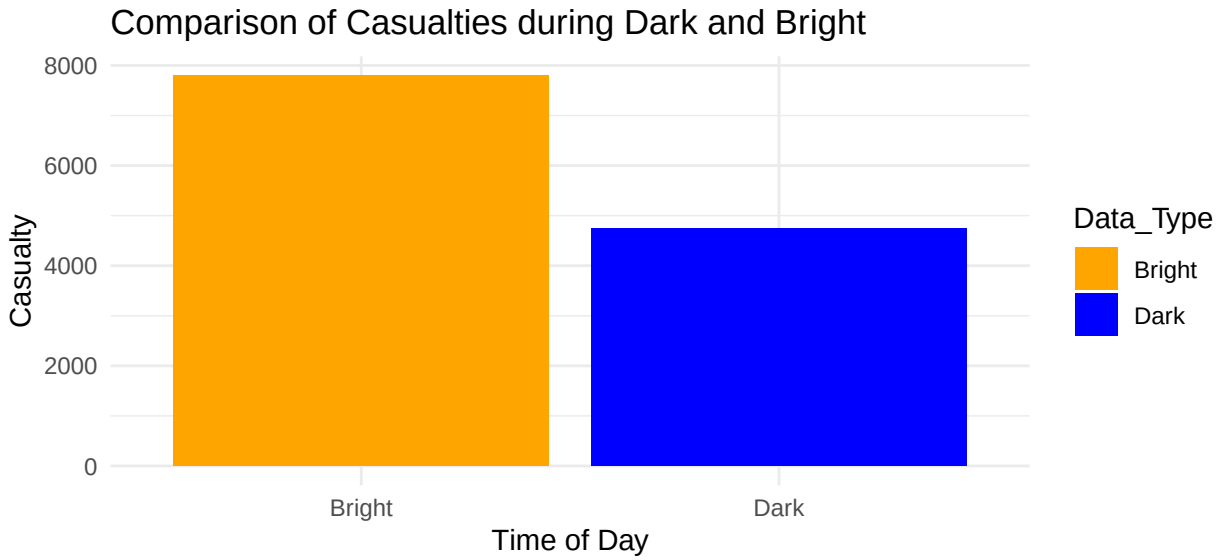
Casualty Hypothesis 2 For our 2nd hypothesis concerning casuaties, we explored whether the time of day influences the number of casualties in accidents.

To initiate an initial investigation, it is essential to begin by refining the training dataset. This entails the creation of a new variable named "Causality," which sums the total number of fatalities (TOTKLD) and injuries (TOTINJ).

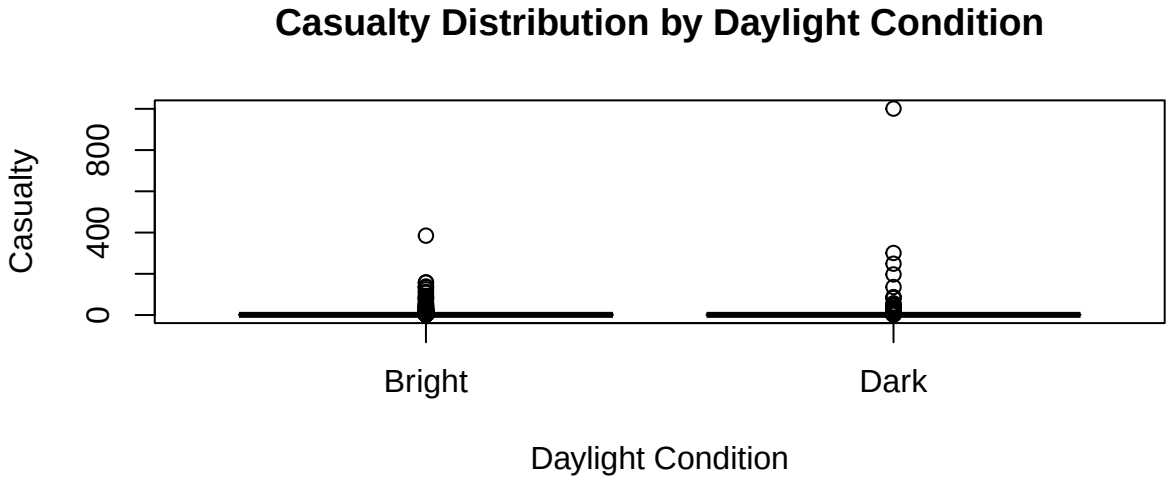
```
df2 <- totacts
df2$Casualty <- df2$TOTKLD + df2$TOTINJ

#We then remove data points without any casualties and filter out records with null, empty, or duplicat
df2 <- df2[!is.na(df2$Casualty), ]
df2 <- df2 %>% filter(Casualty > 0)
df2 <- df2 %>% distinct(INCDTNO, YEAR, MONTH, DAY, TIMEHR, TIMEMIN, .keep_all = TRUE)


#We then create two separate dataframes for "AM" and "PM" incidents
am_data <- df2 %>% filter(AMPM == "AM")
pm_data <- df2 %>% filter(AMPM == "PM")

#Assuming the sun rises at 6 am and sets at 6 pm
#we create a new dataframe "dark" for when the sun is down
dark <- am_data %>% filter(TIMEHR <= 6)
dark1 <- pm_data %>% filter(TIMEHR > 6)
dark <- rbind(dark, dark1)

#and "bright" for when the sun is up
bright <- pm_data %>% filter(TIMEHR <= 6)
bright1 <- am_data %>% filter(TIMEHR > 6)
bright <- rbind(bright, bright1)


# Create a binary variable 'AMPM' for 'dark'
dark$AMPM <- 1  # 1 represents "dark"
bright$AMPM <- 0  # 0 represents "bright"

# Combine "dark" and "bright"
combined_data <- rbind(dark, bright)


# Create a combined dataframe
combined_data <- rbind(dark, bright)
combined_data <- rbind(
  data.frame(Data_Type = "Dark", Casualty = dark$Casualty),
  data.frame(Data_Type = "Bright", Casualty = bright$Casualty)
)
```

Plots to visualize dark vs light extreme accidents

```r
# Create a grouped Bar Graph to compare the casualties in the dark vs in the light
ggplot(combined_data, aes(x = Data_Type, y = Casualty, fill = Data_Type)) +
  geom_bar(stat = "identity") +
  labs(
    x = "Time of Day",
    y = "Casualty",
    title = "Comparison of Casualties during Dark and Bright"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("Dark" = "blue", "Bright" = "orange"))
```




```r
# Box plot to visualize the distribution of casualties for each group
boxplot(Casualty ~ Data_Type, data = combined_data,
        col = c("Dark" = "blue", "Bright" = "orange"),
        main = "Casualty Distribution by Daylight Condition",
        xlab = "Daylight Condition",
        ylab = "Casualty")
```



H0: There is no significant difference in the number of casualties in daytime accidents compared to nighttime accidents. Ha: Daytime accidents have significantly different casualty counts compared to nighttime accidents.

ACCDMG Analysis

Testing ACCDMG Hypothesis 1

Our first model was a full main effects model with one interaction between derailments and train speed.

```
xdmgnd.main <- lm(ACCDMG ~ hError + crossing + derail*TRNSPD, data= xdmgnd)

summary(xdmgnd.main)
```

```
##
## Call:
## lm(formula = ACCDMG ~ hError + crossing + derail * TRNSPD, data = xdmgnd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1759252  -420167  -181935    70857 31154926
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     557084      45852  12.150 < 0.0000000000000002 ***
## hError          149922      34669   4.324 0.000015484728216841 ***
## crossing       -256805      78055  -3.290              0.00101 **
## derail         -323178      47900  -6.747 0.000000000016146683 ***
## TRNSPD           11723       1430   8.196 0.000000000000000287 ***
## derail:TRNSPD    10294       1686   6.107 0.000000001063720675 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1179000 on 7998 degrees of freedom
## Multiple R-squared:  0.07532,    Adjusted R-squared:  0.07474
## F-statistic: 130.3 on 5 and 7998 DF,  p-value: < 0.00000000000000022
```

Above is the summary of our main effects model for Hypothesis 1. The intercept is equal to 557084, meaning that when a train accident is not caused by human error, not a crossing-related accident, and not a derailment, the accident damage cost is $557,084. Human error, train speed, and the interaction between derailment and train speed have positive coefficients, meaning that damage costs increase if the binary variables are true and/or train speed increases. Crossing-related accidents and derailments have a negative relationship with damage costs, meaning that damage costs decrease if these terms are true. All of the parameters are significant at the $p < 0.001$ level.

The biggest downside to our main effects model is that it only has an adjusted R-squared value of 0.075. This means that our model explains very little of the variance of accident damage. Without even looking at other performance metrics like AIC or BIC, we decided to try alternative models since the R-squared value is so poor. We decided to start the improvement process by building a second order model that includes all pairwise interaction terms to see if the model improves at all.

```
xdmgnd.inter <- lm(ACCDMG ~ (hError + crossing + derail*TRNSPD)^2,
                   data= xdmgnd)

summary(xdmgnd.inter)
```

```
##
## Call:
## lm(formula = ACCDMG ~ (hError + crossing + derail * TRNSPD)^2,
##     data = xdmgnd)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4288503  -398394  -176196    85719 31394491
##
## Coefficients: (2 not defined because of singularities)
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)              683004.1    68702.4   9.941 < 0.0000000000000002 ***
## hError                  -345800.6    88297.2  -3.916       0.000090656605 ***
## crossing                 -71239.5   128407.6  -0.555              0.57905
## derail                  -381232.1    73518.5  -5.186       0.000000220690 ***
## TRNSPD                     3457.1     2256.4   1.532              0.12552
## hError:crossing         1049727.5   366041.0   2.868              0.00414 **
## hError:derail            188292.0   102917.7   1.830              0.06736 .
## hError:TRNSPD             42726.2     3807.6  11.221 < 0.0000000000000002 ***
## crossing:derail                NA         NA      NA                   NA
## crossing:TRNSPD             -549.8     3215.2  -0.171              0.86422
## derail:TRNSPD              15645.7     2458.1   6.365       0.000000000206 ***
## hError:derail:TRNSPD      -22056.8     4657.2  -4.736       0.000002216189 ***
## crossing:derail:TRNSPD         NA         NA      NA                   NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1162000 on 7993 degrees of freedom
## Multiple R-squared:  0.1014, Adjusted R-squared:  0.1002
## F-statistic: 90.16 on 10 and 7993 DF,  p-value: < 0.00000000000000022
```

Above is the summary of our second model that includes all of the pairwise interaction terms. The interaction model's intercept is 683004.1, meaning that when a train accident is not caused by human error, not a crossing-related accident, and not a derailment, the accident damage cost is $683,004.10. In comparison to the main effects model, the intercept of the interaction model is higher. Human error, crossing-related accidents, derailments, the interaction between crossing-related accidents and train speed, and the interaction between human error, derailment, and train speed all have negative relationships with accident damage due to their negative coefficient values. Alternatively, train speed and the remaining interaction coefficients are positive. Seven out of the eleven parameters are significant at the $p < 0.05$ level.

Our adjusted R-squared value only improved by 0.0255, making the improvement from the main effects model to the interactions model quite small. In a final effort to try and find a better model, we chose to conduct a stepwise regression to see what the algorithm would choose as the best predictors.

```
xdmgnd.inter.step <- step(xdmgnd.inter, trace= F)

summary(xdmgnd.inter.step)
```

```
##
## Call:
## lm(formula = ACCDMG ~ hError + crossing + derail + TRNSPD + hError:crossing +
##     hError:derail + hError:TRNSPD + derail:TRNSPD + hError:derail:TRNSPD,
##     data = xdmgnd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4288194  -398671  -176248    85691 31394491
##
## Coefficients:
##                       Estimate Std. Error t value             Pr(>|t|)
## (Intercept)             688174      61690  11.155 < 0.0000000000000002 ***
```

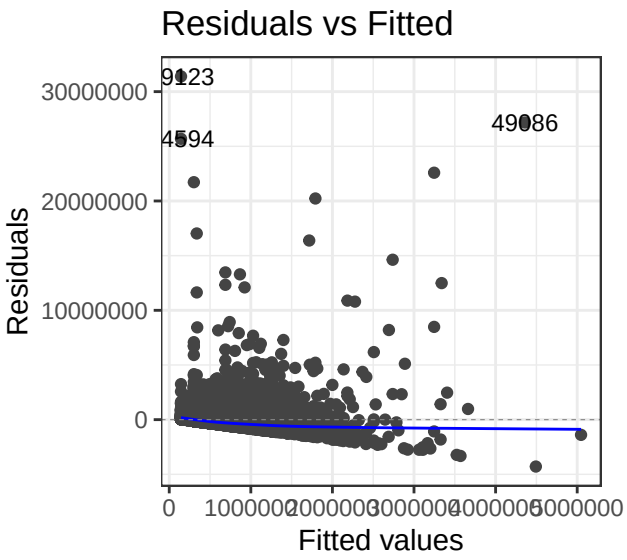
```
## hError                -350926      83051  -4.225        0.00002411208 ***
## crossing               -87578      85786  -1.021              0.30733
## derail                -386402      67012  -5.766        0.00000000841 ***
## TRNSPD                   3187       1614   1.976              0.04825 *
## hError:crossing       1056642     363779   2.905              0.00369 **
## hError:derail          193417      98452   1.965              0.04950 *
## hError:TRNSPD           42992       3476  12.369 < 0.0000000000000002 ***
## derail:TRNSPD           15915       1885   8.442 < 0.0000000000000002 ***
## hError:derail:TRNSPD   -22323       4390  -5.085        0.00000037597 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1162000 on 7994 degrees of freedom
## Multiple R-squared:  0.1014, Adjusted R-squared:  0.1003
## F-statistic: 100.2 on 9 and 7994 DF,  p-value: < 0.00000000000000022
```

Above is the final stepwise regression model. Out of the ten coefficients available from the interaction model, the stepwise model chose nine of them. The only term left out was the interaction term between crossing-related accidents and train speed, probably because it had the highest p-value of all the coefficients in the interaction model. All of the parameters are significant at the $p < 0.05$ level except for crossing-related accidents.

The stepwise model produced the highest adjusted R-squared value at 0.1003, but this is still only a 0.0001 improvement from the full interaction model, making this improvement almost negligible.

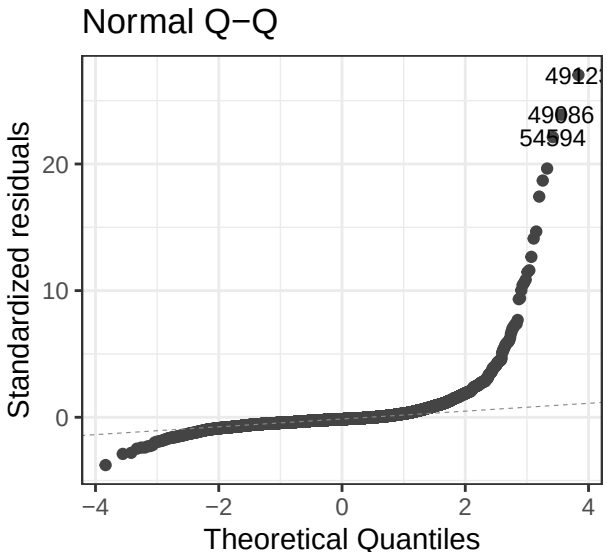
Before finalizing the model, we also need to check the diagnostics plots to identify if linear regression assumptions are satsfied and if we need to make any transformations or adjustments.

```
autoplot(xdmgnd.inter.step, which=1, label.size = 3) + theme_bw()
```



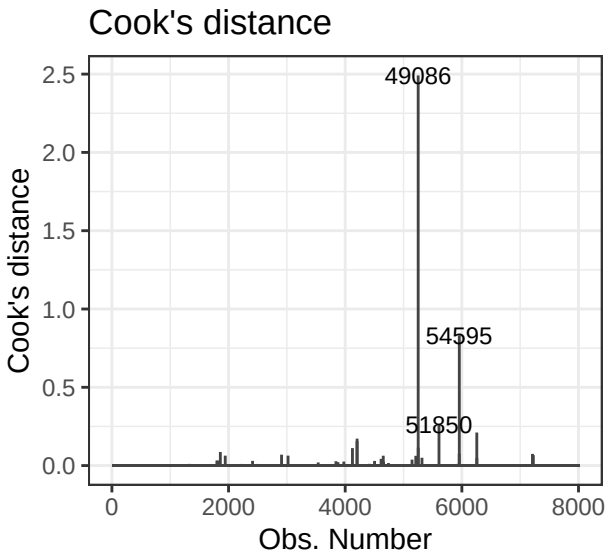
The residuals vs fitted plot is used to test for a constant mean of 0 and constant variance. The lack of a constant mean of 0 is expected here because our model is not a good fit for the data with such a low adjusted R-squared value. We can also see that the variance is not homoscedastic.

```
autoplot(xdmgnd.inter.step, which=2, label.size = 3) + theme_bw()
```



The QQ plot is used to look for close alignment with the quantiles of a normal distribution. Here we can see that the QQ plot violates the assumption and displays heavy-tailed behavior. The QQ plot combined with the heteroscedastic variance indicates to us that we should transform the response variable.

```
autoplot(xdmgnd.inter.step, which=4, label.size = 3) + theme_bw()
```



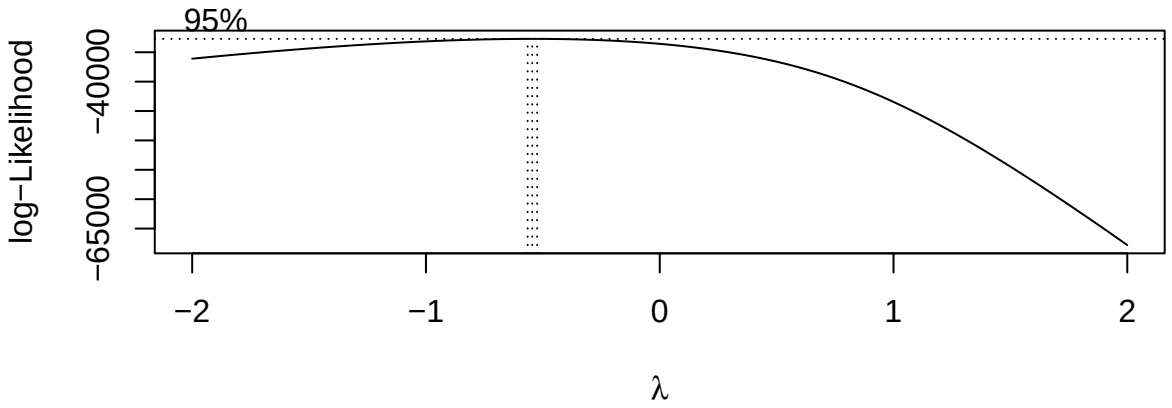
Cook's distance is used to identify influential points in the dataset for our model. Here we can see that there are two points above 0.5, making them notably influential and indicating to us that we should remove them.

After analyzing the diagnostics plots, we identified that we need to transform the response variable and remove two influential points. First we will remove the two influential points.

```
cooks.rm <- c(5251, 5956)

xdmgnd.rm <- xdmgnd[-cooks.rm, ]

rownames(xdmgnd.rm) <- NULL
```

After removing the two influential points, we ran a boxcox plot to identify how we should transform accident damage.

```
boxcox(xdmgnd.inter.step)
```



From the boxcox plot above, we can see that the brackets are between -1 and 0. This means that we should perform a boxcox transformation with the optimal lambda value.

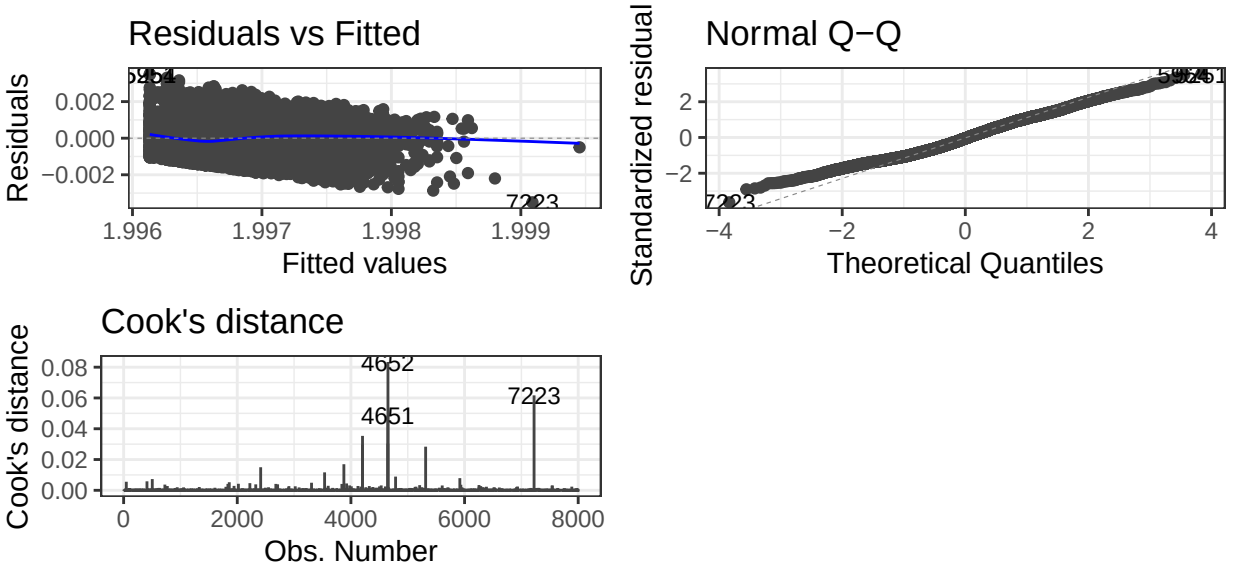
```
# Optimal lambda
xval <- which.max(boxcox(xdmgnd.inter.step, plotit = F)$y)
lam <- boxcox(xdmgnd.inter.step, plotit = F)$x[xval]

# Run the new interaction regression model with the transformation
xdmgnd.inter.boxcox <- lm((ACCDMG^lam-1)/lam ~ (hError + crossing + derail*TRNSPD)^2,
                          data= xdmgnd.rm)

# Run a stepwise regression with the new interaction model
xdmgnd.inter.step.boxcox <- step(xdmgnd.inter.boxcox, trace= F)

summary(xdmgnd.inter.step.boxcox)
```

```
##
## Call:
## lm(formula = (ACCDMG^lam - 1)/lam ~ hError + crossing + derail +
##     TRNSPD + hError:crossing + hError:derail + hError:TRNSPD +
##     derail:TRNSPD + hError:derail:TRNSPD, data = xdmgnd.rm)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0035048 -0.0007773 -0.0000547  0.0007354  0.0035102
##
## Coefficients:
##                           Estimate   Std. Error   t value             Pr(>|t|)
## (Intercept)           1.996628267  0.000052485 38041.819 < 0.0000000000000002
## hError               -0.000268743  0.000070658    -3.803             0.000144
## crossing              0.000014067  0.000072985     0.193             0.847163
## derail               -0.000361728  0.000057012    -6.345     0.00000000023494
## TRNSPD                0.000003916  0.000001373     2.853             0.004348
## hError:crossing       0.000537068  0.000309496     1.735             0.082726
## hError:derail         0.000135869  0.000083920     1.619             0.105478
## hError:TRNSPD         0.000026419  0.000002957     8.934 < 0.0000000000000002
## derail:TRNSPD         0.000025891  0.000001604    16.142 < 0.0000000000000002
## hError:derail:TRNSPD -0.000025886  0.000003774    -6.859     0.00000000000745
##
## (Intercept)          ***
```

```
## hError                ***
## crossing
## derail                ***
## TRNSPD                **
## hError:crossing       .
## hError:derail
## hError:TRNSPD         ***
## derail:TRNSPD         ***
## hError:derail:TRNSPD ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0009887 on 7992 degrees of freedom
## Multiple R-squared:  0.184,  Adjusted R-squared:  0.1831
## F-statistic: 200.2 on 9 and 7992 DF,  p-value: < 0.00000000000000022
```

Above is the final stepwise interaction model after removing influential points and performing a boxcox transformation. Similar to the first stepwise interaction model, the only term left out of this model is the interaction term between crossing-related accidents and train speed. Seven of the ten coefficients are signficant at the $p < 0.001$ level, showing less overall significance among the predictors compared to the previous model. However, looking at the adjusted R-squared value we can see that we have obtained our highest value yet at 0.183, which is 0.083 higher than the model before we made adjustments. Additionally, we have an f-statistic with a p-value < 0.001, indicating that the model is significant.

```
autoplot(xdmgnd.inter.step.boxcox, which=c(1, 2, 4), label.size = 3) + theme_bw()
```



We can also take another look at the diagnostics plots above. After performing a boxcox transformation and removing the two influential points, we can see that our diagnostics plots look much better than before the adjustments. The residual vs fitted plot has a more constant mean of 0 and more homoscedasticity than the previous model. The QQ plot displays more linear behavior than the previous model, and the Cook's distance values are all below 0.09.

While the model has improved a lot compared to previous iterations, there are certainly still problems present with the model. The first is that the adjusted R-squared value, while improved, is still quite low. Despite the low f-statistic p-value indicating that the model is statistically significant, the low adjusted R-squared value makes us very skeptical that this model has much predictive or inferential power. Additionally, while the variance improved from the adjustments we made, it still displays some heteroscedastic behavior and therefore still violates the constant variance assumption. Violating this assumption also makes us skeptical of the

model's power since lack of constant variance affects the precision of the model. In addition, heteroscedasticity tends to produce p-values that are deceptively significant, which may be misleading us to think that the model and its coefficients are more signficant than they actually are (Statistics by Jim).

Testing ACCDMG Hypothesis 2

Prepping to test hypothesis by coding categorical variables as dummy variables:

```
# Remove rows in dmgnd where there is a null value for TYPEQ or Type
xdmgnd <- xdmgnd[complete.cases(xdmgnd[c("TYPEQ", "Type")]), ]

# Create Necessary Dummy Variables
xdmgnd$Derail <- (xdmgnd$Type == "Derailment")
xdmgnd$Hwy_Rail <- (xdmgnd$Type == "Hwy-Rail")
xdmgnd$Freight <- (xdmgnd$TYPEQ == "Freight")
xdmgnd$PassCom <- ifelse(xdmgnd$Type %in% c("Passenger", "Commuter"), TRUE, FALSE) #Grouped because com
```

Multiple Linear Regression using TYPEQ to Predict ACCDMG (LM1)

```
xdmgnd.lm1<-lm(ACCDMG~Freight+PassCom, data=xdmgnd)
summary(xdmgnd.lm1)
```

```
##
## Call:
## lm(formula = ACCDMG ~ Freight + PassCom, data = xdmgnd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -581284  -458388  -335902    25155 30903761
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   634993      25557  24.846 < 0.0000000000000002 ***
## FreightTRUE   111278      30292   3.673             0.000241 ***
## PassComTRUE       NA         NA      NA                   NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1224000 on 7951 degrees of freedom
## Multiple R-squared:  0.001694,   Adjusted R-squared:  0.001569
## F-statistic: 13.49 on 1 and 7951 DF,  p-value: 0.0002408
```

Multiple Linear Regression using Type to Predict ACCDMG (LM2)

```
xdmgnd.lm2<-lm(ACCDMG~Derail+Hwy_Rail, data=xdmgnd)
summary(xdmgnd.lm2)
```

```
##
## Call:
## lm(formula = ACCDMG ~ Derail + Hwy_Rail, data = xdmgnd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -665047  -476312  -336320    40156 30847510
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
```

```
## (Intercept)    830047      35295  23.518 < 0.0000000000000002 ***
## DerailTRUE    -138803      38486  -3.607             0.000312 ***
## Hwy_RailTRUE   -98649      71172  -1.386             0.165767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1224000 on 7950 degrees of freedom
## Multiple R-squared:  0.001644,   Adjusted R-squared:  0.001393
## F-statistic: 6.544 on 2 and 7950 DF,  p-value: 0.001446
```

Multiple Linear Regression using Type & TYPEQ to Predict ACCDMG (LM3)

```
xdmgnd.lm3<-lm(ACCDMG~Freight+PassCom+Derail+Hwy_Rail,data=xdmgnd)
summary(xdmgnd.lm3)
```

```
##
## Call:
## lm(formula = ACCDMG ~ Freight + PassCom + Derail + Hwy_Rail,
##     data = xdmgnd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -746317  -467639  -319605    38556 30964100
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    761444      38039  20.017 < 0.0000000000000002 ***
## FreightTRUE    150751      31441   4.795           0.00000166 ***
## PassComTRUE        NA         NA      NA                   NA
## DerailTRUE    -186790      39715  -4.703           0.00000260 ***
## Hwy_RailTRUE  -105421      71088  -1.483                0.138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1222000 on 7949 degrees of freedom
## Multiple R-squared:  0.004523,   Adjusted R-squared:  0.004147
## F-statistic: 12.04 on 3 and 7949 DF,  p-value: 0.00000007362
```

Multiple Linear Regression using Type & TYPEQ plus interactions to Predict ACCDMG (LM4)

```
xdmgnd.lm4<-lm(ACCDMG~(Freight+PassCom+Derail+Hwy_Rail)^2,data=xdmgnd)
summary(xdmgnd.lm4)
```

```
##
## Call:
## lm(formula = ACCDMG ~ (Freight + PassCom + Derail + Hwy_Rail)^2,
##     data = xdmgnd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -861946  -450851  -320831    31790 30930405
##
## Coefficients: (5 not defined because of singularities)
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                664880      47717  13.934 < 0.0000000000000002 ***
## FreightTRUE                362944      70734   5.131          0.000000295 ***
```

```
## PassComTRUE                    NA         NA      NA                 NA
## DerailTRUE                 -56531      57548  -0.982            0.32597
## Hwy_RailTRUE                66117      99428   0.665            0.50609
## FreightTRUE:PassComTRUE        NA         NA      NA                 NA
## FreightTRUE:DerailTRUE    -255762      79633  -3.212            0.00132 **
## FreightTRUE:Hwy_RailTRUE  -362141     142201  -2.547            0.01089 *
## PassComTRUE:DerailTRUE         NA         NA      NA                 NA
## PassComTRUE:Hwy_RailTRUE       NA         NA      NA                 NA
## DerailTRUE:Hwy_RailTRUE        NA         NA      NA                 NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1221000 on 7947 degrees of freedom
## Multiple R-squared:  0.006011,   Adjusted R-squared:  0.005385
## F-statistic: 9.611 on 5 and 7947 DF,  p-value: 0.00000000369
```

Models LM1-4 have almost no predictive value based on their R^2. We will explore their interaction

Will now add in train speed to hopefully improve model performance. Multiple Linear Regression main effects using Type, TYPEQ, & TRNSPD to Predict ACCDMG (LM5)

```
xdmgnd.lm5<-lm(ACCDMG~Freight+PassCom+Derail+Hwy_Rail+TRNSPD,data=xdmgnd)
summary(xdmgnd.lm5)
```

```
##
## Call:
## lm(formula = ACCDMG ~ Freight + PassCom + Derail + Hwy_Rail +
##     TRNSPD, data = xdmgnd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2782824  -392625  -194196    77154 31134897
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    602846      37250  16.184 < 0.0000000000000002 ***
## FreightTRUE   -144411      32628  -4.426           0.00000973 ***
## PassComTRUE        NA         NA      NA                   NA
## DerailTRUE    -198989      38302  -5.195           0.00000021 ***
## Hwy_RailTRUE  -617178      71668  -8.612 < 0.0000000000000002 ***
## TRNSPD          20178        824  24.487 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1178000 on 7948 degrees of freedom
## Multiple R-squared:  0.07436,    Adjusted R-squared:  0.07389
## F-statistic: 159.6 on 4 and 7948 DF,  p-value: < 0.00000000000000022
```

Multiple Linear Regression 2nd order model using Type, TYPEQ, & TRNSPD to Predict ACCDMG (LM6)

```
xdmgnd.lm6 <- lm(ACCDMG ~ (Freight + PassCom + Derail + Hwy_Rail + TRNSPD)^2, data = xdmgnd)
summary(xdmgnd.lm6)
```

```
##
## Call:
## lm(formula = ACCDMG ~ (Freight + PassCom + Derail + Hwy_Rail +
##     TRNSPD)^2, data = xdmgnd)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3466761  -406673  -157295    95186 31185755
##
## Coefficients: (6 not defined because of singularities)
##                           Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                 421533      50507   8.346 < 0.0000000000000002 ***
## FreightTRUE                 516675      77520   6.665  0.00000000002822465 ***
## PassComTRUE                     NA         NA      NA                   NA
## DerailTRUE                  -68534      58849  -1.165                0.244
## Hwy_RailTRUE                102992     134135   0.768                0.443
## TRNSPD                       22330       2006  11.131 < 0.0000000000000002 ***
## FreightTRUE:PassComTRUE         NA         NA      NA                   NA
## FreightTRUE:DerailTRUE    -703057      80979  -8.682 < 0.0000000000000002 ***
## FreightTRUE:Hwy_RailTRUE   206130     147906   1.394                0.163
## FreightTRUE:TRNSPD         -17576       2215  -7.936  0.00000000000000237 ***
## PassComTRUE:DerailTRUE          NA         NA      NA                   NA
## PassComTRUE:Hwy_RailTRUE        NA         NA      NA                   NA
## PassComTRUE:TRNSPD              NA         NA      NA                   NA
## DerailTRUE:Hwy_RailTRUE         NA         NA      NA                   NA
## DerailTRUE:TRNSPD            18373       2364   7.773  0.00000000000000863 ***
## Hwy_RailTRUE:TRNSPD         -17060       3033  -5.625  0.00000001916186310 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1165000 on 7943 degrees of freedom
## Multiple R-squared:  0.09548,    Adjusted R-squared:  0.09445
## F-statistic: 93.16 on 9 and 7943 DF,  p-value: < 0.00000000000000022
```

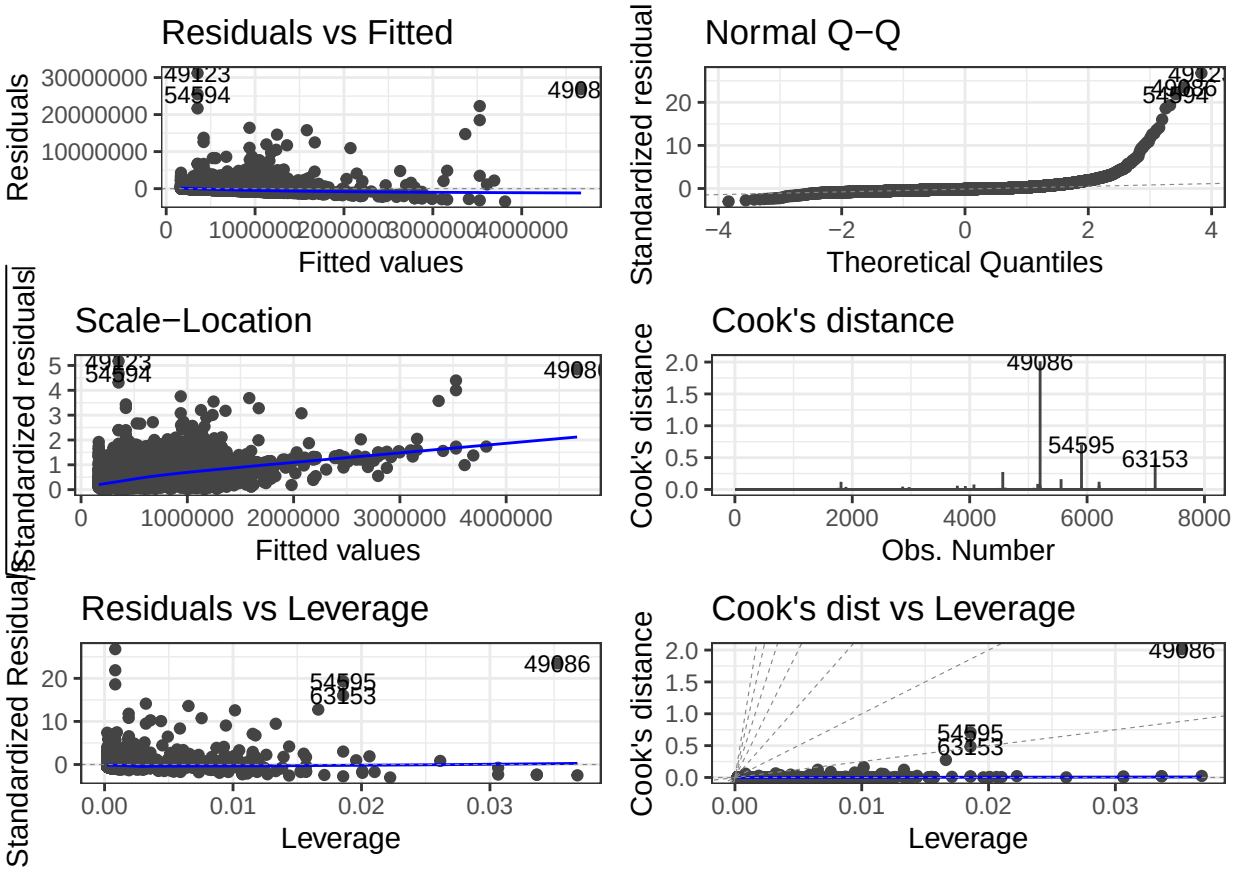
```
AIC(xdmgnd.lm6)
```

```
## [1] 244763.7
```

Adding TrnSpd to the model instantly boosted the model demonstrating its importance in predicting accident damage. The 2nd order model provided incremental additional improvements.

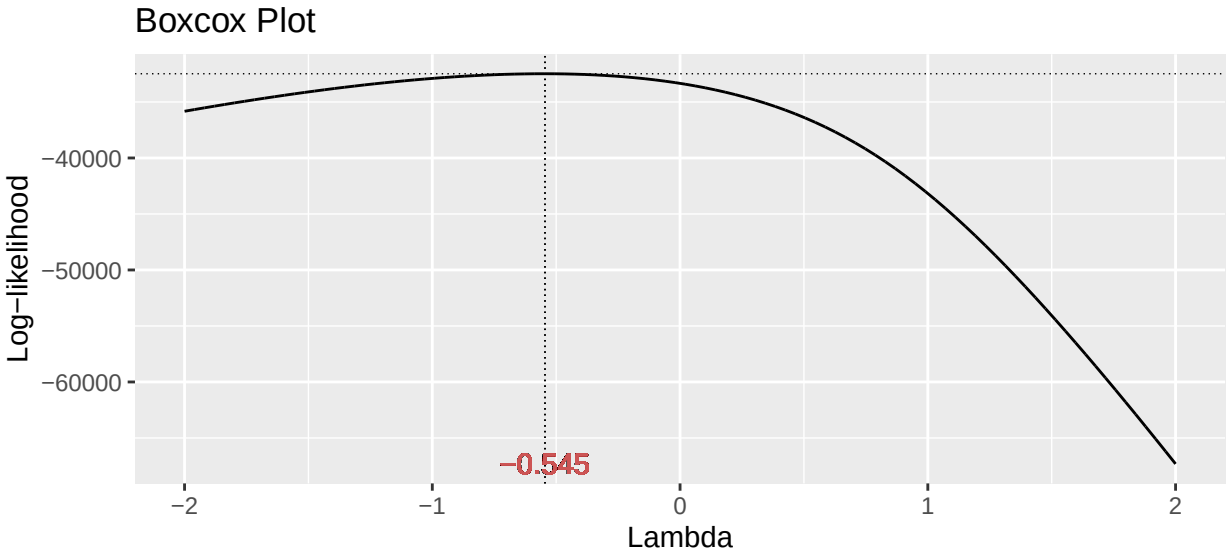
Diagnostics, Transformations, & Variable Selection We first need to examine our diagnostic plots to see if LM6 is meeting basic regression assumptions

```
autoplot(xdmgnd.lm6, which=1:6, label.size = 3) + theme_bw()
```



Immediately we can see clear heteroscedasticity as evidenced by the residual v fitted and QQ plot. We cna also, see at least 3 observations with concerning Cook's distance values. Therefore we know we will need to transform our response variable. For this we will use a Box Cox transformation

```
# Box Cox
gg_boxcox(xdmgnd.lm6)
```




```
#The best lambda and store in L
L<-boxcox(xdmgnd.lm6, plotit = F)$x[which.max(boxcox(xdmgnd.lm6, plotit = F)$y)]
L
```

```
## [1] -0.5
```

Now that we have an optimal lambda value of -0.5, we can use this to transform accident damage and improve LM6

```
# The model with the best lamda transformation (LM6.boxcox)
xdmgnd.lm6.boxcox<-lm((ACCDMG^L-1)/L~(Freight + PassCom + Derail + Hwy_Rail + TRNSPD)^2,data=xdmgnd)
# Display regression results for boxcox model
summary(xdmgnd.lm6.boxcox)
```

```
##
## Call:
## lm(formula = (ACCDMG^L - 1)/L ~ (Freight + PassCom + Derail +
##     Hwy_Rail + TRNSPD)^2, data = xdmgnd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0027080 -0.0007782 -0.0000457  0.0007308  0.0033501
##
## Coefficients: (6 not defined because of singularities)
##                               Estimate   Std. Error   t value
## (Intercept)                1.996446533  0.000042783 46664.637
## FreightTRUE                0.000136993  0.000065665     2.086
## PassComTRUE                         NA           NA        NA
## DerailTRUE                -0.000152749  0.000049849    -3.064
## Hwy_RailTRUE               0.000205001  0.000113622     1.804
## TRNSPD                     0.000005898  0.000001699     3.471
## FreightTRUE:PassComTRUE             NA           NA        NA
## FreightTRUE:DerailTRUE    -0.000249520  0.000068595    -3.638
## FreightTRUE:Hwy_RailTRUE -0.000530795  0.000125287    -4.237
## FreightTRUE:TRNSPD         0.000013441  0.000001876     7.165
## PassComTRUE:DerailTRUE              NA           NA        NA
## PassComTRUE:Hwy_RailTRUE            NA           NA        NA
## PassComTRUE:TRNSPD                  NA           NA        NA
## DerailTRUE:Hwy_RailTRUE             NA           NA        NA
## DerailTRUE:TRNSPD          0.000013193  0.000002002     6.589
## Hwy_RailTRUE:TRNSPD       -0.000004288  0.000002569    -1.669
##                                    Pr(>|t|)
## (Intercept)               < 0.0000000000000002 ***
## FreightTRUE                          0.036988 *
## PassComTRUE                                NA
## DerailTRUE                           0.002190 **
## Hwy_RailTRUE                         0.071232 .
## TRNSPD                               0.000522 ***
## FreightTRUE:PassComTRUE                    NA
## FreightTRUE:DerailTRUE               0.000277 ***
## FreightTRUE:Hwy_RailTRUE    0.000022945422015 ***
## FreightTRUE:TRNSPD          0.000000000000849 ***
## PassComTRUE:DerailTRUE                     NA
## PassComTRUE:Hwy_RailTRUE                   NA
## PassComTRUE:TRNSPD                         NA
## DerailTRUE:Hwy_RailTRUE                    NA
## DerailTRUE:TRNSPD           0.000000000047134 ***
## Hwy_RailTRUE:TRNSPD                  0.095124 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.000987 on 7943 degrees of freedom
## Multiple R-squared:  0.188,  Adjusted R-squared:  0.1871
## F-statistic: 204.3 on 9 and 7943 DF,  p-value: < 0.00000000000000022
```

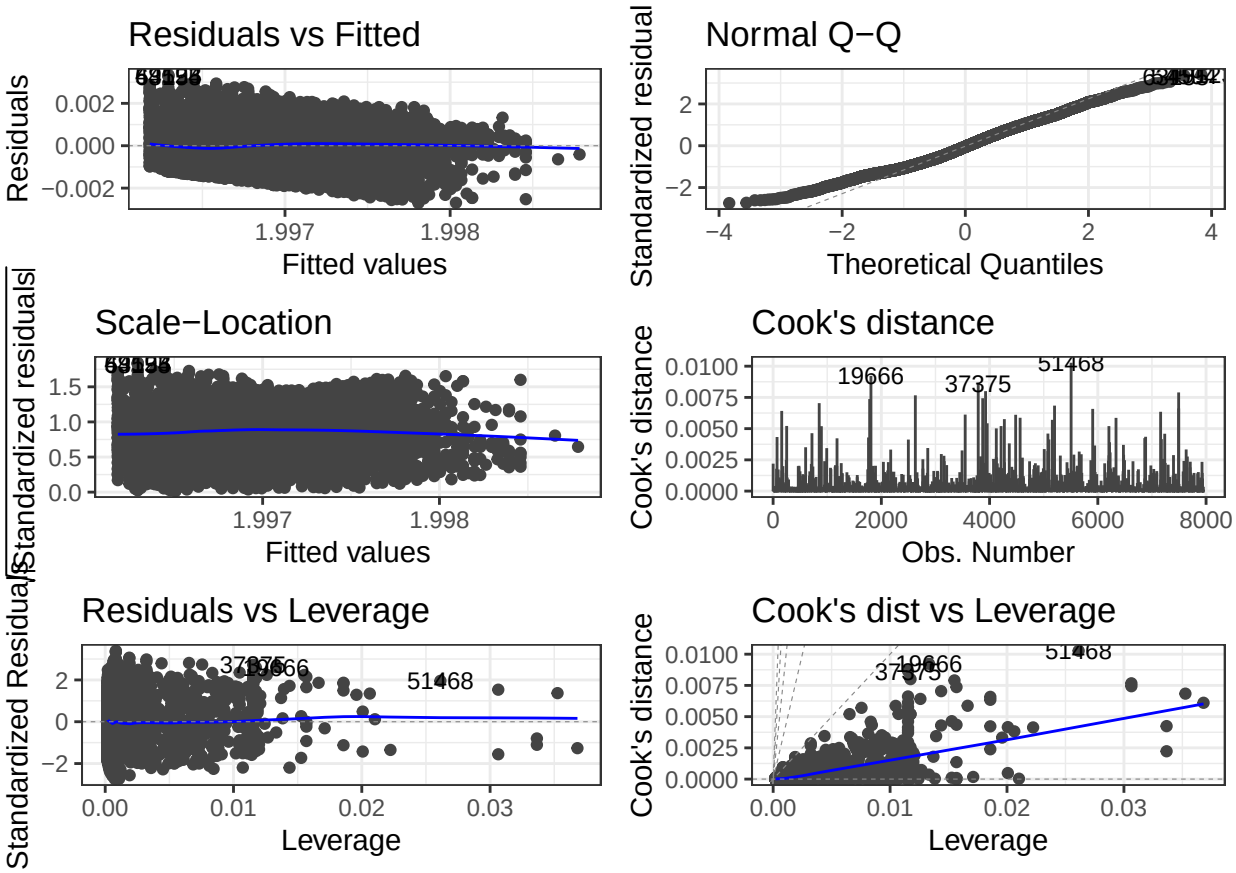
```
AIC(xdmgnd.lm6.boxcox)
```

```
## [1] -87500.48
```

This transformation also greatly improved the predictive value of the model.

Now we will recheck the diagnostics plots

```
autoplot(xdmgnd.lm6.boxcox, which=1:6, label.size = 3) + theme_bw()
```



As we can see we have addressed the galring heteroscedasticity issue. Some heteroscedasticity remains as we can see in the residuals vs fitted plot, but we can see it is clearly much improved. Also, there are no concerning Cook's distance values.

We next aimed to trim our model using stepwise regression We used backwards elimination to trim our model and remove insignificant predictors.

```
xdmgnd.lm6.step <- step(xdmgnd.lm6.boxcox, direction = "backward", trace = F)
summary(xdmgnd.lm6.step)
```

```
##
## Call:
## lm(formula = (ACCDMG^L - 1)/L ~ Freight + Derail + Hwy_Rail +
##     TRNSPD + Freight:Derail + Freight:Hwy_Rail + Freight:TRNSPD +
```

```
##     Derail:TRNSPD + Hwy_Rail:TRNSPD, data = xdmgnd)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0027080 -0.0007782 -0.0000457  0.0007308  0.0033501
##
## Coefficients:
##                               Estimate   Std. Error   t value
## (Intercept)                1.996446533  0.000042783 46664.637
## FreightTRUE                0.000136993  0.000065665     2.086
## DerailTRUE                -0.000152749  0.000049849    -3.064
## Hwy_RailTRUE               0.000205001  0.000113622     1.804
## TRNSPD                     0.000005898  0.000001699     3.471
## FreightTRUE:DerailTRUE    -0.000249520  0.000068595    -3.638
## FreightTRUE:Hwy_RailTRUE -0.000530795  0.000125287    -4.237
## FreightTRUE:TRNSPD         0.000013441  0.000001876     7.165
## DerailTRUE:TRNSPD          0.000013193  0.000002002     6.589
## Hwy_RailTRUE:TRNSPD       -0.000004288  0.000002569    -1.669
##                                      Pr(>|t|)
## (Intercept)              < 0.0000000000000002 ***
## FreightTRUE                          0.036988 *
## DerailTRUE                           0.002190 **
## Hwy_RailTRUE                         0.071232 .
## TRNSPD                               0.000522 ***
## FreightTRUE:DerailTRUE               0.000277 ***
## FreightTRUE:Hwy_RailTRUE    0.000022945422015 ***
## FreightTRUE:TRNSPD          0.000000000000849 ***
## DerailTRUE:TRNSPD           0.000000000047134 ***
## Hwy_RailTRUE:TRNSPD                  0.095124 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.000987 on 7943 degrees of freedom
## Multiple R-squared:  0.188,  Adjusted R-squared:  0.1871
## F-statistic: 204.3 on 9 and 7943 DF,  p-value: < 0.00000000000000022
```

```
AIC(xdmgnd.lm6.step)
```

```
## [1] -87500.48
```

```
anova(xdmgnd.lm6.boxcox,xdmgnd.lm6.step)
```

```
## Analysis of Variance Table
##
## Model 1: (ACCDMG^L - 1)/L ~ (Freight + PassCom + Derail + Hwy_Rail + TRNSPD)^2
## Model 2: (ACCDMG^L - 1)/L ~ Freight + Derail + Hwy_Rail + TRNSPD + Freight:Derail +
##     Freight:Hwy_Rail + Freight:TRNSPD + Derail:TRNSPD + Hwy_Rail:TRNSPD
##   Res.Df       RSS Df Sum of Sq F Pr(>F)
## 1   7943 0.0077386
## 2   7943 0.0077386  0         0
```

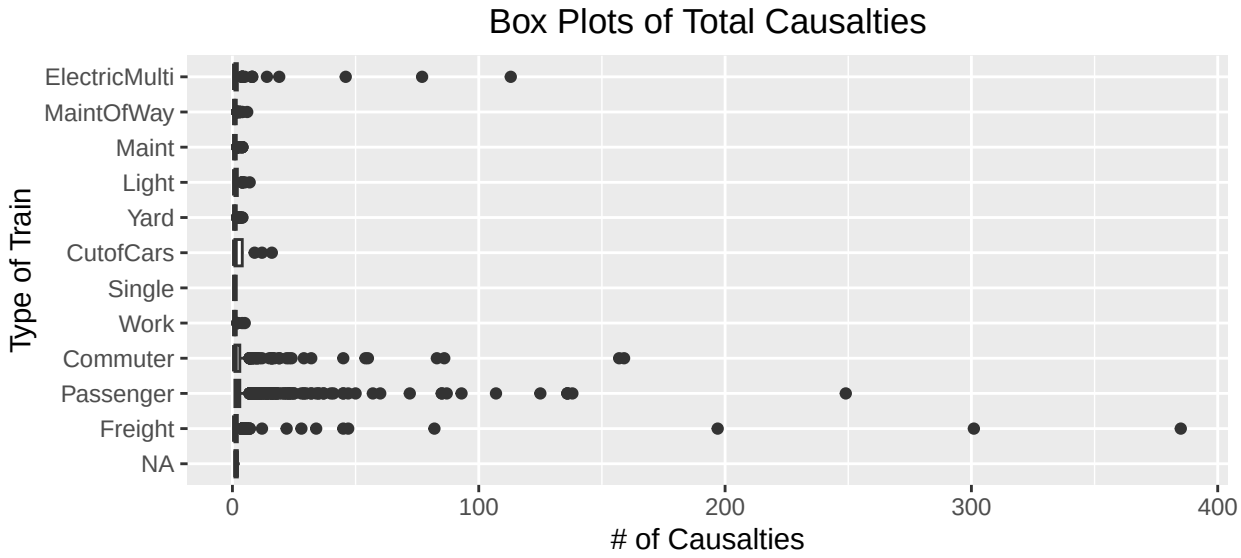
This backward elimination process removed the PassCom variable and its interactions as well as Hwy_Rail:TRNSPD and FreightTRUE:Hwy_RailTRUE The AIC is slightly lower than the full model but Adj. R^2 is the same. Because the partial f-test has a p-value > 0.05, we can be confident that none of the removed predictors were significant and we will move forward with the stepwise mode

Casualties Analysis

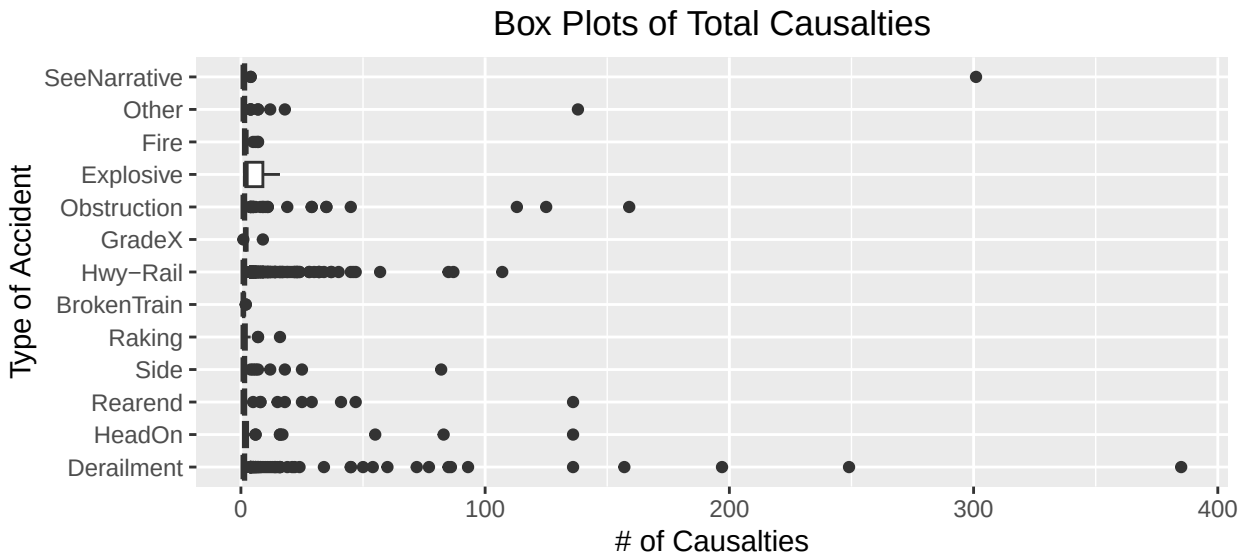
Testing Casualities Hypothesis 1

Treatment of variables In order to conduct preliminary investigation, we must first cleanup the train data set. This involves creating a new "Causality" variable with total killed (TOTKLD) and total injured (TOTINJ) together. We remove the data with no casualties from our data set as well as null, empty, and duplicated data. Finally, a severe outlier was removed in order to prevent high leveraged points appearing in our model.

```
# Create Casualty variable
totacts["Casualty"] = totacts["TOTKLD"] + totacts["TOTINJ"]
# Remove data without a Casualty
totacts_posCas <- filter(totacts, Casualty > 0)
# Remove data with Null or empty
totacts_posCas_null <- filter(totacts_posCas, TYPEQ != "NULL" & TYPEQ != "")
# Remove duplicate reports
totacts_posCas_nd <- totacts_posCas_null %>% distinct(INCDTNO, YEAR, MONTH, DAY, TIMEHR, TIMEMIN, .keep
# Remove outlier
totacts_posCas_nd <- totacts_posCas_nd[-c(which.max(totacts_posCas_nd$Casualty)), ]

# Box Plots of Type of Train vs. Total Casualties per accident
ggplot(data = totacts_posCas_nd, aes(x = TYPEQ, y = Casualty)) +
  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 Total Causalties") +
  labs(y = "# of Causalties", x = "Type of Train")
```




```
# Box Plots of Type of Accident vs. Total Casualties per accident
ggplot(data = totacts_posCas_nd, aes(x = Type, y = Casualty)) +
  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 Total Causalties") +
  labs(y = "# of Causalties", x = "Type of Accident")
```




```
# Type of train and accident w/ largest number of accidents w/ >=1 casualties
table(totacts_posCas_nd$TYPEQ)
```

```
##
##            NA       Freight     Passenger      Commuter          Work
##            12          2097           827           222            28
##        Single     CutofCars          Yard         Light         Maint
##             5            14           217           165           115
##    MaintOfWay ElectricMulti
##           112            90
```

```
table(totacts_posCas_nd$Type)
```

```
##
##   Derailment       HeadOn      Rearend         Side       Raking  BrokenTrain
##          489           90          141          132           28            8
##     Hwy-Rail       GradeX  Obstruction    Explosive         Fire        Other
##         2457            5          242            3           32          219
## SeeNarrative
##           58
```

```
# Reset row names
rownames(totacts_posCas_nd) <- NULL
```

Based on the results from the box plots and data tables, we observe that most of the data groups around the left side of the box plots, demonstrating that the majority of severe accidents occur regardless of the type of accident of type of train involved. Despite this, we observe a higher frequency of large number of casualties for Derailment accidents and Freight and Passenger train accidents. It's thus implied there may be an interaction between freight and/or passenger trains and derailment accidents, so we next build a full main effects model with interactions between derailments, freight trains, passenger trains, and casualties.

To accomplish this, we must create three new binary columns: Freight, Passenger, and Derailment, where: Freight = whether the accident has a Freight train involved (1 if yes, 0 if no) Passenger = whether the accident has a Passenger train involved (1 if yes, 0 if no) Derailment = whether the accident is due to the derailment of the train (1 if yes, 0 if no)

Transforming these predictor variables was necessary as all these variables were categorical with different levels to them. Afterwards, the next step is to build a full main effects model with interactions between derailments and passenger trains and derailments and freight trains. ###

```r
# Freight
totacts_posCas_nd$Freight <- 0

for (i in 1:nrow(totacts_posCas_nd)) {
  totacts_posCas_nd[i, "Freight"] <- totacts_posCas_nd$TYPEQ[i] == "Freight"
}

# Passenger
totacts_posCas_nd$Passenger <- 0

for (i in 1:nrow(totacts_posCas_nd)) {
  # Hwy-rail crossing or RR Grade Crossing
  totacts_posCas_nd[i, "Passenger"] <- totacts_posCas_nd$TYPEQ[i] == "Passenger"
}

# Derailment
totacts_posCas_nd$Derail <- (totacts_posCas_nd$Type == "Derailment")
```

```r
causdmg.lm1<-lm(Casualty ~ Derail + Passenger + Freight + Derail*Passenger + Derail*Freight,data=totacts
summary(causdmg.lm1)
```

```
##
## Call:
## lm(formula = Casualty ~ Derail + Passenger + Freight + Derail *
##     Passenger + Derail * Freight, data = totacts_posCas_nd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.44  -1.64  -0.72  -0.51 381.08
##
## Coefficients:
##                       Estimate Std. Error t value    Pr(>|t|)
## (Intercept)             2.5083     0.4280   5.860 0.000000005 ***
## DerailTRUE              1.7845     1.1324   1.576    0.115144
## Passenger               2.1274     0.6221   3.420    0.000633 ***
## Freight                -0.7897     0.5174  -1.526    0.127078
## DerailTRUE:Passenger    9.0242     1.9036   4.741 0.000002206 ***
## DerailTRUE:Freight      0.4138     1.3865   0.298    0.765404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.41 on 3898 degrees of freedom
## Multiple R-squared:  0.02728,    Adjusted R-squared:  0.02603
## F-statistic: 21.86 on 5 and 3898 DF,  p-value: < 0.00000000000000022
```

```r
AIC(causdmg.lm1)
```

```
## [1] 30748.47
```

Above is the summary of the main effects model for our hypothesis. The intercept is equal to 2.5083, meaning that when a train accident is not caused by derailment and does not involve a freight train or passenger train, the number of casualties is 2.5083. Derailments, Passenger, the interaction between derailments and passenger, and the interaction between derailment and freight have positive coefficients, meaning that casualties increase if the binary variables are true. Freight trains have a negative relationship with casualties, meaning that casualties decrease if this variable is true. Only the Passenger variable and interaction between derailment

and passenger trains are significant at the $p < 0.001$ level, and the model is overall significant at this level as well.

However, the biggest downside to our main effects model is that it only has an adjusted R-squared value of 0.026 and a high AIC of 30748.47. Since a high adjusted R-squared and low AIC indicates a model is a good fit for the data, this means that this linear model explains very little of the variance of casualties. We investigate alternative models since the R-squared value and AIC are so poor. We decided to adjust our model by building a second order model that includes all pairwise interaction terms to see if the model improves at all. ###

```
causdmg.lm2 <- lm(Casualty ~ (Derail + Passenger + Freight + Derail*Passenger + Derail*Freight)^2,
                  data= totacts_posCas_nd)
summary(causdmg.lm2)
```

```
##
## Call:
## lm(formula = Casualty ~ (Derail + Passenger + Freight + Derail *
##     Passenger + Derail * Freight)^2, data = totacts_posCas_nd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.44  -1.64  -0.72  -0.51 381.08
##
## Coefficients: (2 not defined because of singularities)
##                               Estimate Std. Error t value    Pr(>|t|)
## (Intercept)                     2.5083     0.4280   5.860 0.000000005 ***
## DerailTRUE                      1.7845     1.1324   1.576    0.115144
## Passenger                       2.1274     0.6221   3.420    0.000633 ***
## Freight                        -0.7897     0.5174  -1.526    0.127078
## DerailTRUE:Passenger            9.0242     1.9036   4.741 0.000002206 ***
## DerailTRUE:Freight              0.4138     1.3865   0.298    0.765404
## Passenger:Freight                   NA         NA      NA          NA
## DerailTRUE:Passenger:Freight        NA         NA      NA          NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.41 on 3898 degrees of freedom
## Multiple R-squared:  0.02728,    Adjusted R-squared:  0.02603
## F-statistic: 21.86 on 5 and 3898 DF,  p-value: < 0.00000000000000022
```

```
AIC(causdmg.lm2)
```

```
## [1] 30748.47
```

Above is the summary of the second model with all pairwise interaction terms. The results of this model (intercept value, significant variables, coefficients) are identical to the above model, demonstrating that there is no improvement by including all pairwise interaction terms. We attempt to remedy this by utilizing a stepwise regression to determine what the best predictors would be. ###

```
causdmg.lm2.step <- step(causdmg.lm2, trace= F)
summary(causdmg.lm2.step)
```

```
##
## Call:
## lm(formula = Casualty ~ Derail + Passenger + Freight + Derail:Passenger,
##     data = totacts_posCas_nd)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.44  -1.64  -0.74  -0.47 381.20
##
## Coefficients:
##                       Estimate Std. Error t value      Pr(>|t|)
## (Intercept)             2.4689     0.4071   6.065 0.00000000144 ***
## DerailTRUE              2.0605     0.6534   3.154      0.001624 **
## Passenger               2.1669     0.6078   3.565      0.000369 ***
## Freight                -0.7320     0.4800  -1.525      0.127333
## DerailTRUE:Passenger    8.7482     1.6636   5.259 0.00000015290 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.4 on 3899 degrees of freedom
## Multiple R-squared:  0.02725,    Adjusted R-squared:  0.02626
## F-statistic: 27.31 on 4 and 3899 DF,  p-value: < 0.00000000000000022
```

```
AIC(causdmg.lm2.step)
```

```
## [1] 30746.56
```

We see that there is minimal improvement in this model as the adjusted R-squared value is now 0.02626, the AIC is 30746.56, and the overall model is still significant at the same level. Interestingly, while Passenger and the interaction between derailment and passenger are still significant at the $p < 0.001$ level, the derailment variable is now significant at the $p < 0.01$ level. Also of note is the interaction term between Derail and Freight is removed from the model.

In another attempt to create an improved model, we add additional binary variables from different levels of the TYPE and TYPEQ variables to further explain accident causualties. Looking back at the box plots, we see that adding Commuter and Electric Multicar trains as well as Obstruction, Rear End, and Head On accidents to the dataset could improve the model. We model these predictor variables as follows:

Commuter = whether the accident has a Commuter train involved (1 if yes, 0 if no) ElectricMulti = whether the accident has a Electric Multicar train involved (1 if yes, 0 if no) Obstruction = whether the accident is due to the obstruction on the track (1 if yes, 0 if no) Rearend = whether the accident is due to the train rear ending something on the track (1 if yes, 0 if no) HeadOn = whether the accident is due to the train colliding head on with something on the track (1 if yes, 0 if no)

After coding these categorical variables, we create a new main effects model with full interactions between each type of train and type of collision for our created binary variables. ###

```
# Commuter
totacts_posCas_nd$Commuter <- (totacts_posCas_nd$TYPEQ == "Commuter")
# ElectricMulti
totacts_posCas_nd$ElectricMulti <- (totacts_posCas_nd$TYPEQ == "ElectricMulti")
# HeadOn
totacts_posCas_nd$Obstruction <- (totacts_posCas_nd$Type == "Obstruction")
# Rearend
totacts_posCas_nd$Rearend <- (totacts_posCas_nd$Type == "Rearend")
# HeadOn
totacts_posCas_nd$HeadOn <- (totacts_posCas_nd$Type == "HeadOn")
```

```
causdmg.lm3 <- lm(Casualty ~ Derail + Passenger + Freight + Commuter + ElectricMulti + Obstruction + Rea
                  data= totacts_posCas_nd)
summary(causdmg.lm3)
```

```
##
```

```
## Call:
## lm(formula = Casualty ~ Derail + Passenger + Freight + Commuter +
##     ElectricMulti + Obstruction + Rearend + HeadOn + Derail *
##     Passenger + Derail * Freight + Derail * Commuter + Derail *
##     ElectricMulti + Obstruction * Passenger + Obstruction * Freight +
##     Obstruction * Commuter + Obstruction * ElectricMulti + Rearend *
##     Passenger + Rearend * Freight + Rearend * Commuter + Rearend *
##     ElectricMulti + HeadOn * Passenger + HeadOn * Freight + HeadOn *
##     Commuter + HeadOn * ElectricMulti, data = totacts_posCas_nd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46.67  -1.81  -0.70  -0.17 381.08
##
## Coefficients:
##                                   Estimate Std. Error t value      Pr(>|t|)
## (Intercept)                        1.40547    0.60634   2.318      0.020504 *
## DerailTRUE                        -0.24026    1.28563  -0.187      0.851766
## Passenger                          2.68847    0.76915   3.495      0.000479 ***
## Freight                            0.29349    0.67696   0.434      0.664649
## CommuterTRUE                       1.40378    1.10542   1.270      0.204197
## ElectricMultiTRUE                  1.46295    1.52064   0.962      0.336078
## ObstructionTRUE                    0.18276    1.59409   0.115      0.908728
## RearendTRUE                       -0.11976    1.73403  -0.069      0.944943
## HeadOnTRUE                         0.26119    2.41693   0.108      0.913947
## DerailTRUE:Passenger              11.59076    1.98229   5.847 0.00000000541 ***
## DerailTRUE:Freight                 2.45827    1.50898   1.629      0.103374
## DerailTRUE:CommuterTRUE           16.38101    3.14594   5.207 0.00000020184 ***
## DerailTRUE:ElectricMultiTRUE      14.97183    5.75818   2.600      0.009355 **
## Passenger:ObstructionTRUE          0.34830    2.14767   0.162      0.871177
## Freight:ObstructionTRUE           -0.50293    2.20705  -0.228      0.819755
## CommuterTRUE:ObstructionTRUE      10.28072    3.18016   3.233      0.001236 **
## ElectricMultiTRUE:ObstructionTRUE 16.94882    5.39615   3.141      0.001697 **
## Passenger:RearendTRUE             19.77582    3.94298   5.015 0.00000055279 ***
## Freight:RearendTRUE                0.70651    2.28229   0.310      0.756909
## CommuterTRUE:RearendTRUE           5.31051   12.31491   0.431      0.666328
## ElectricMultiTRUE:RearendTRUE     -1.24866    8.87972  -0.141      0.888178
## Passenger:HeadOnTRUE              43.31153    7.43847   5.823 0.00000000626 ***
## Freight:HeadOnTRUE                 0.03985    2.95310   0.013      0.989235
## CommuterTRUE:HeadOnTRUE           28.26289    5.59719   5.049 0.00000046338 ***
## ElectricMultiTRUE:HeadOnTRUE      -1.12962   12.47326  -0.091      0.927845
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.16 on 3879 degrees of freedom
## Multiple R-squared:  0.07033,    Adjusted R-squared:  0.06458
## F-statistic: 12.23 on 24 and 3879 DF,  p-value: < 0.00000000000000022
```

```
AIC(causdmg.lm3)
```

```
## [1] 30609.72
```

We observe that we additional predictor variables involved, the intercept changes to 1.35203 (so approximately 1.35203 casualties per accident), a decrease from our previous models. Passenger, the interaction between Passenger and Rearend, the interaction between passenger and head on, and the interaction between passenger

and commuter as significant at the $p < 0.001$ level. The Commuter, interaction between Commuter and Obstruction, and interaction between ElectricMulti and Obstruction are significant at the $p < 0.01$ level. There is significance for ElectricMulti at the $p < 0.1$ level, and overall our model is significant at the $p < 0.05$ level.

This model's adjusted R-squared value is vastly improved to 0.04098, albeit still incredibly small overall, as well as the AIC to 30702.03. We again run a backwards stepwise regression to see which predictors are significant

```
causdmg.lm3.step <- step(causdmg.lm3, trace= F)
summary(causdmg.lm3.step)
```

Variable Selection, Diagnostics, & Transformations

```
##
## Call:
## lm(formula = Casualty ~ Derail + Passenger + Freight + Commuter +
##     ElectricMulti + Obstruction + Rearend + HeadOn + Derail:Passenger +
##     Derail:Freight + Derail:Commuter + Derail:ElectricMulti +
##     Commuter:Obstruction + ElectricMulti:Obstruction + Passenger:Rearend +
##     Passenger:HeadOn + Commuter:HeadOn, data = totacts_posCas_nd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46.67  -1.84  -0.69  -0.13 381.08
##
## Coefficients:
##                                   Estimate Std. Error t value      Pr(>|t|)
## (Intercept)                         1.3637     0.5463   2.496       0.01260 *
## DerailTRUE                         -0.1985     1.2576  -0.158       0.87458
## Passenger                           2.7705     0.6961   3.980 0.00007017247 ***
## Freight                             0.3289     0.6002   0.548       0.58376
## CommuterTRUE                        1.4736     1.0700   1.377       0.16854
## ElectricMultiTRUE                   1.4476     1.4689   0.985       0.32445
## ObstructionTRUE                     0.1580     0.8733   0.181       0.85644
## RearendTRUE                         0.3077     1.1098   0.277       0.78162
## HeadOnTRUE                          0.2921     1.3774   0.212       0.83207
## DerailTRUE:Passenger               11.5087     1.9537   5.891 0.00000000417 ***
## DerailTRUE:Freight                  2.4229     1.4751   1.642       0.10057
## DerailTRUE:CommuterTRUE            16.3112     3.1313   5.209 0.00000019971 ***
## DerailTRUE:ElectricMultiTRUE       14.9872     5.7402   2.611       0.00906 **
## CommuterTRUE:ObstructionTRUE       10.2774     2.8839   3.564       0.00037 ***
## ElectricMultiTRUE:ObstructionTRUE  17.0307     5.2171   3.264       0.00111 **
## Passenger:RearendTRUE              19.3081     3.7044   5.212 0.00000019625 ***
## Passenger:HeadOnTRUE               43.2403     7.1607   6.039 0.00000000170 ***
## CommuterTRUE:HeadOnTRUE            28.2039     5.2283   5.394 0.00000007284 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.15 on 3886 degrees of freedom
## Multiple R-squared:  0.07022,    Adjusted R-squared:  0.06615
## F-statistic: 17.26 on 17 and 3886 DF,  p-value: < 0.00000000000000022
```

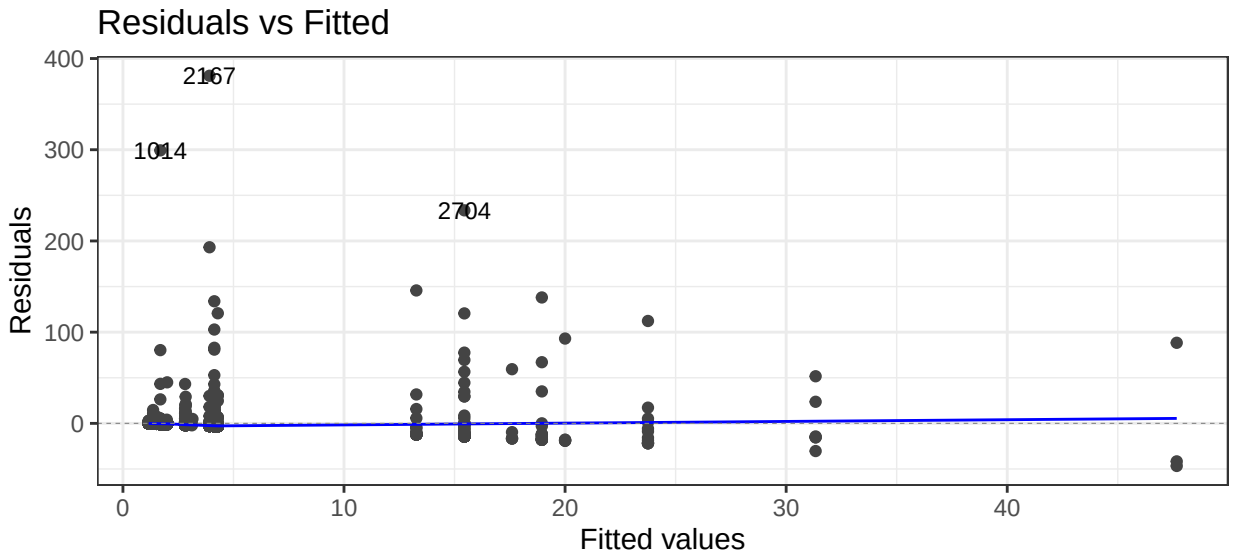
```
AIC(causdmg.lm3.step)
```

```
## [1] 30596.2
```

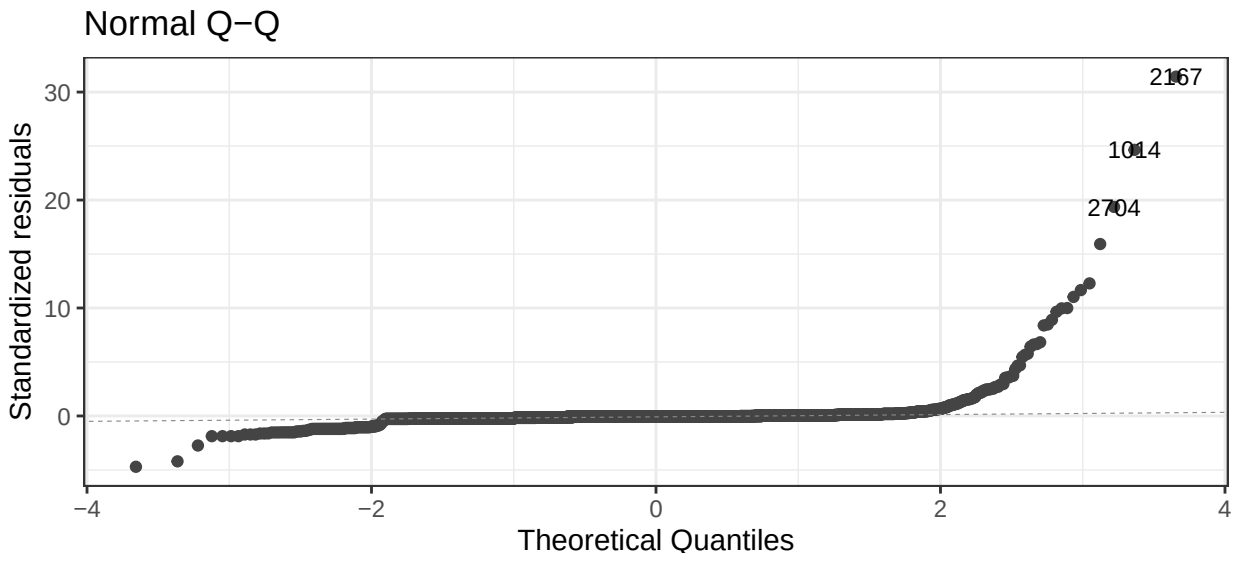
There is marginal improvement in the adjusted R-squared value to 0.06615, the AIC to 30596.2, and intercept casualty predictor to 1.3637. The model is still significant at $p < 0.01$ level albeit slightly more significant in p value..

Despite improvements, we must check the diagnostic plots to verify linear regression assumptions are satisfied and if any transformations are required.

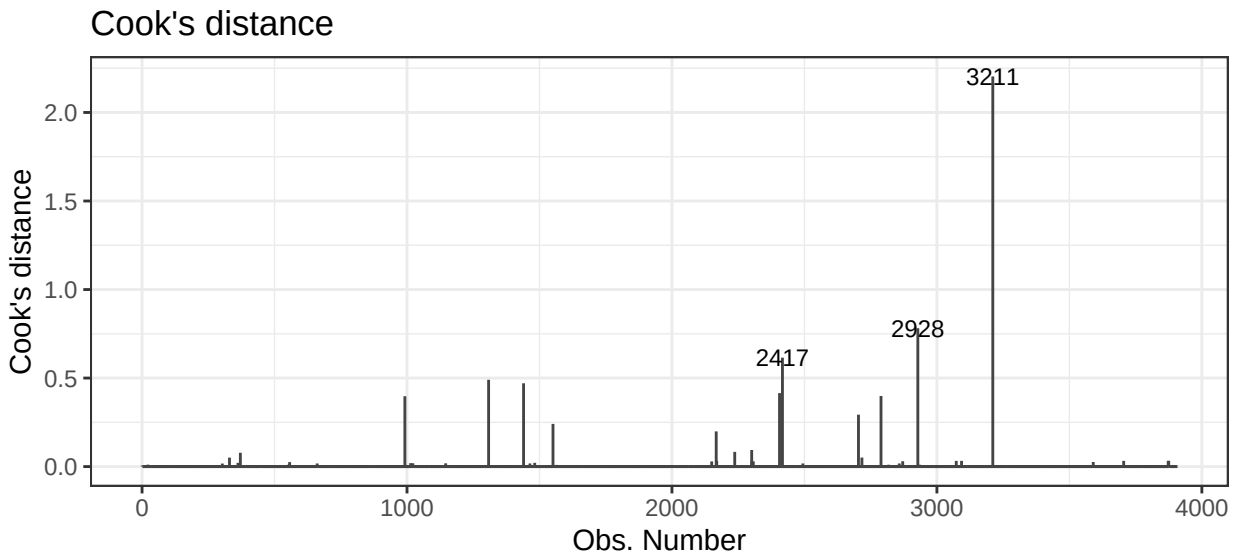
```
autoplot(causdmg.lm3.step,which=1, ncol = 1, label.size = 3) + theme_bw() #Residual vs. Fitted
```




```
autoplot(causdmg.lm3.step,which=2, ncol = 1, label.size = 3) + theme_bw() #QQ
```



```
autoplot(causdmg.lm3.step, which=4, ncol = 1, label.size = 3) + theme_bw() #Cook's distance
```



The residuals vs. fitted plot tests for a constant mean of 0 and constant variance. The lack of a constant mean of 0 is expected here since the chosen model is a poor fit for the data with a low adjusted R-squared value. The variance is also heteroscedastic.

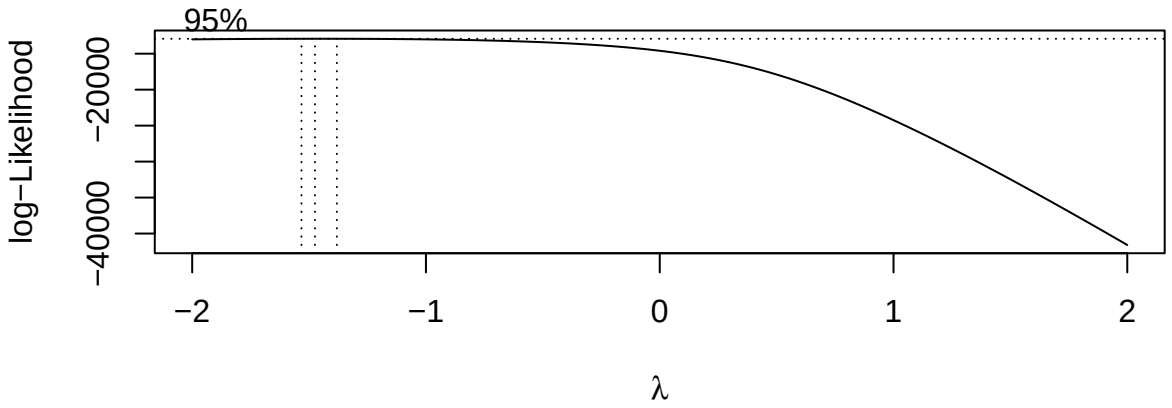
The QQ plot is used to look for close alignment with the quantiles of a normal distribution. Here we can see that the QQ plot violates the assumption and displays heavy-tailed behavior for both ends, but especially for the right tail, going to standard residuals upwards of 30. The QQ plot combined with the heteroscedastic variance indicates that the response variable should be transformed.

Cook's distance is used to identify influential points in the dataset for our model. Here we can see that there are six points above 0.5, making them notably influential and indicating to us that we should remove them.

After analyzing the diagnostics plots, we identified that we need to transform the response variable and remove six influential points. We will remove the six influential points and then create a boxcox plot to see how to transform the casualty data.

Remove cook's distance points and boxcox plot

```
totacts_posCas_nd.rm <- totacts_posCas_nd[-c(1440, 1551, 2417, 2789, 2928, 3211), ]
boxcox(causdmg.lm3.step)
```



The bracket locations are located between -2 and -1. Because of this, a boxcox transformation with the optimal lambda value is necessary.

```r
# Optimal lambda
xval <- which.max(boxcox(causdmg.lm3.step, plotit = F)$y)
lam <- boxcox(causdmg.lm3.step, plotit = F)$x[xval]

# Run the new interaction regression model with the transformation
causdmg.lm3.boxcox <- lm((Casualty^lam-1)/lam ~ (Derail + Passenger + Freight + Commuter + ElectricMulti
                         data= totacts_posCas_nd.rm)

# Run a stepwise regression with the new interaction model
causdmg.lm3.boxcox.step <- step(causdmg.lm3.boxcox, trace= F)
summary(causdmg.lm3.boxcox.step)
```

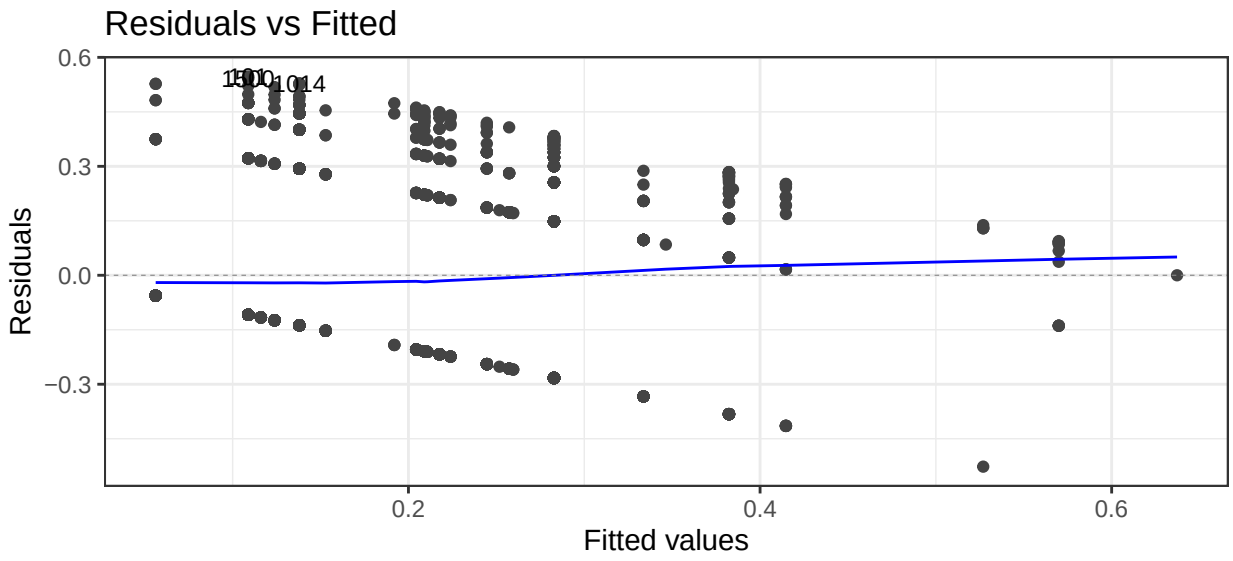
```
##
## Call:
## lm(formula = (Casualty^lam - 1)/lam ~ Derail + Passenger + Freight +
##     Commuter + ElectricMulti + Obstruction + Rearend + HeadOn +
##     Derail:Passenger + Derail:Freight + Derail:Commuter + Passenger:Obstruction +
##     Passenger:Rearend + Freight:Rearend + Freight:HeadOn + Commuter:Rearend +
##     Commuter:HeadOn, data = totacts_posCas_nd.rm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5269 -0.1379 -0.1379  0.2556  0.5474
##
## Coefficients:
##                            Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                0.108871   0.011013   9.886 < 0.0000000000000002 ***
## DerailTRUE                -0.052651   0.023791  -2.213              0.02695 *
## Passenger                  0.173908   0.014251  12.203 < 0.0000000000000002 ***
## Freight                    0.029052   0.012234   2.375              0.01761 *
## CommuterTRUE               0.100094   0.019848   5.043          0.000000479 ***
## ElectricMultiTRUE          0.135720   0.026630   5.096          0.000000363 ***
## ObstructionTRUE            0.014973   0.019346   0.774              0.43900
## RearendTRUE                0.007188   0.032417   0.222              0.82454
## HeadOnTRUE                 0.101825   0.044488   2.289              0.02214 *
## DerailTRUE:Passenger       0.152189   0.037443   4.065          0.000049083 ***
## DerailTRUE:Freight         0.132386   0.028116   4.709          0.000002582 ***
## DerailTRUE:CommuterTRUE    0.258403   0.059614   4.335          0.000014975 ***
## Passenger:ObstructionTRUE -0.093381   0.033690  -2.772              0.00560 **
## Passenger:RearendTRUE      0.279950   0.077903   3.594              0.00033 ***
## Freight:RearendTRUE        0.112164   0.043070   2.604              0.00924 **
## Freight:HeadOnTRUE         0.093949   0.055066   1.706              0.08807 .
## CommuterTRUE:RearendTRUE   0.421052   0.235844   1.785              0.07429 .
## CommuterTRUE:HeadOnTRUE    0.216147   0.114509   1.888              0.05915 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.233 on 3880 degrees of freedom
## Multiple R-squared:  0.09995,    Adjusted R-squared:  0.09601
## F-statistic: 25.35 on 17 and 3880 DF,  p-value: < 0.00000000000000022
```

```r
AIC(causdmg.lm3.boxcox.step)
```

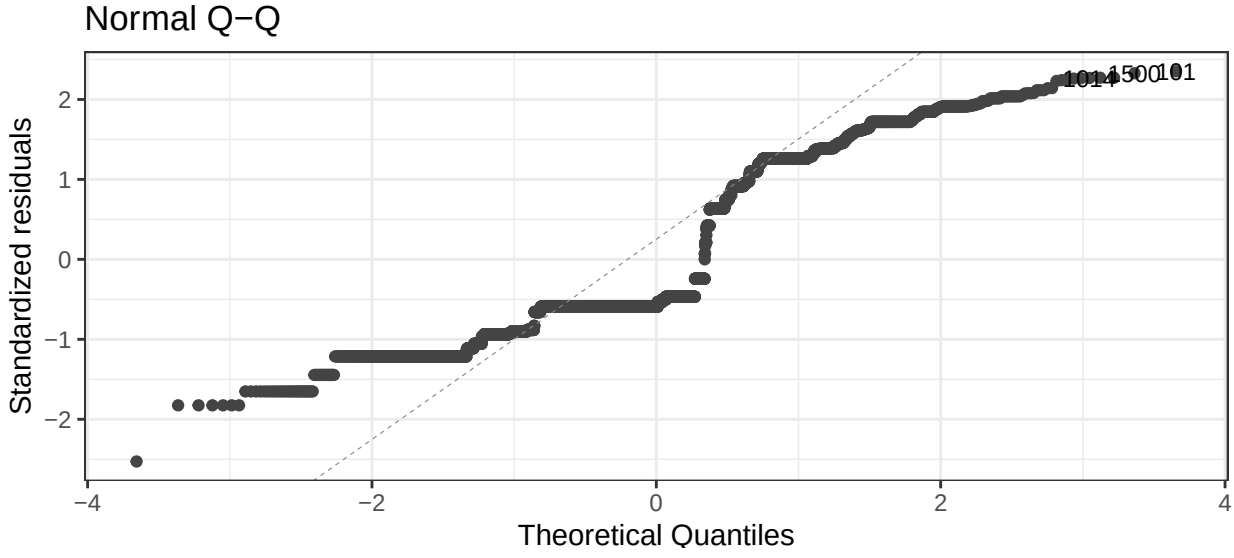
```
## [1] -273.8746
```

Above is the final stepwise interaction model after removing influential points and performing a boxcox transformation. Seven of the seventeen coefficients are significant at the $p < 0.001$ level, two terms are significant at the $p < 0.01$ level, three terms are significant at the $p < 0.05$ level, and three are significant at the $p < 0.1$ level, showing greater overall significance among the predictors compared to the previous model. Additionally, looking at the adjusted R-squared value we can see that we have obtained our highest value yet at 0.09601, almost double than the previous model. We also have a much lower AIC of -273.8746, orders of magnitude lower than any of the previous models. Finally, we have an f-statistic with a p-value < 0.001, indicating that the model is significant.

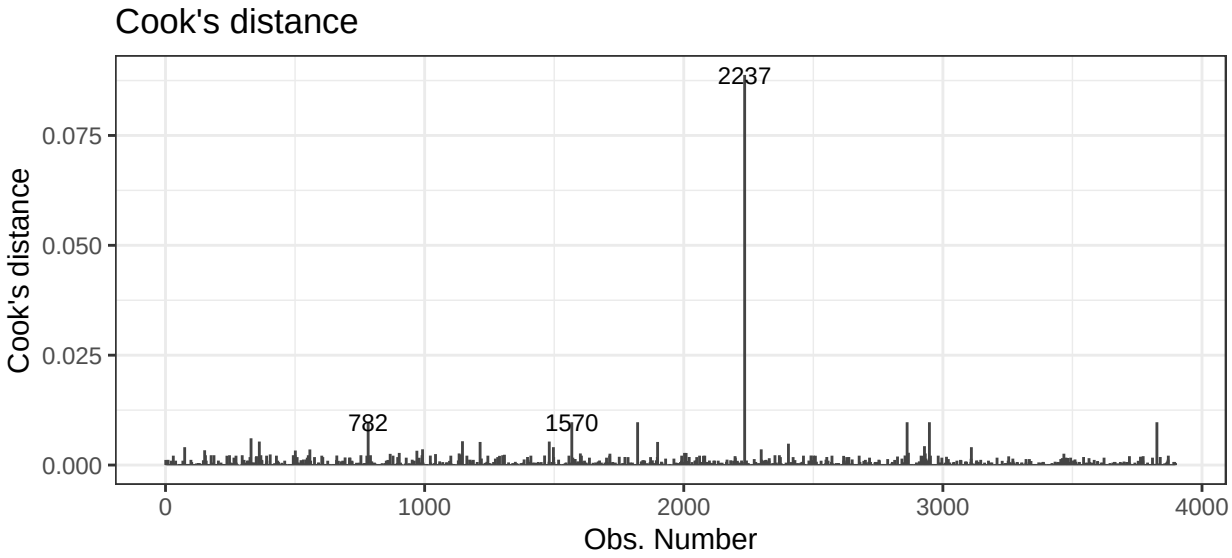
```
autoplot(causdmg.lm3.boxcox.step,which=1, ncol = 1, label.size = 3) + theme_bw() #Residual vs. Fitted
```




```
autoplot(causdmg.lm3.boxcox.step,which=2, ncol = 1, label.size = 3) + theme_bw() #QQ
```




```
autoplot(causdmg.lm3.boxcox.step, which=4, ncol = 1, label.size = 3) + theme_bw() #Cook's distance
```



We can also take another look at the diagnostics plots above. After performing a boxcox transformation and removing the six influential points, we can see that our diagnostics plots show significant improvements than before the adjustments. The residual vs. fitted plot has a more constant mean of 0 and more homoscedasticity than the previous model. The QQ plot displays more linear behavior than the previous model, and the Cook's distance values are all below 0.005.

Testing Casualty Hypothesis 2

Perform a t-test to test the significance of the hypothesis

```
t_test_result <- t.test(dark$Casualty, bright$Casualty)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  dark$Casualty and bright$Casualty
## t = 0.75294, df = 1488, p-value = 0.4516
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.077639  2.420330
## sample estimates:
## mean of x mean of y
##  3.659459  2.988113
```

The p-value associated with the test is 0.3074. This p-value is greater than the significance level of 0.05. The 95 percent confidence interval for the difference in means is [-0.8371525, 2.6548912]. This interval includes zero, further suggesting no statistically significant difference in means.

```
#Create Binary Variable AMPM
combined_data$AMPM <- ifelse(combined_data$Data_Type == "Dark", 1, 0)
# Linear Model
lm_model <- lm(Casualty ~ AMPM, data = combined_data)
summary(lm_model)
```

```
##
## Call:
## lm(formula = Casualty ~ AMPM, data = combined_data)
##
```

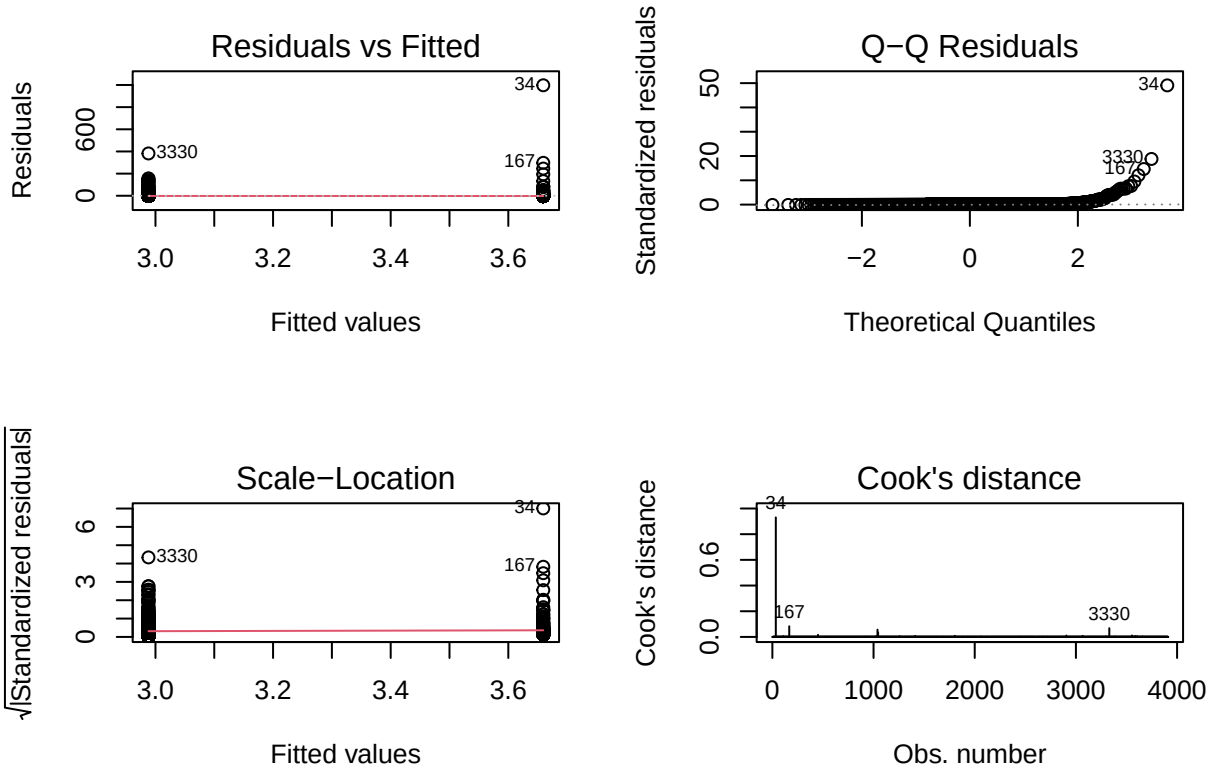
```
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2.66  -1.99  -1.99  -0.99 997.34
##
## Coefficients:
##             Estimate Std. Error t value           Pr(>|t|)
## (Intercept)   2.9881     0.3981   7.507 0.0000000000000748 ***
## AMPM          0.6713     0.6911   0.971              0.331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.33 on 3901 degrees of freedom
## Multiple R-squared:  0.0002419,  Adjusted R-squared:  -1.442e-05
## F-statistic: 0.9437 on 1 and 3901 DF,  p-value: 0.3314
```

```r
AIC(lm_model)
```

```
## [1] 34592.17
```

The F-statistic tests whether the addition of the "AMPM" variable as a predictor is significant. In this case, the F-statistic is 1.026, and the associated p-value is 0.3314, which is greater than 0.05. The Adjusted R-squared and the Multiple R-squared values are close to zero. This indicates that the model explains very little of the variance in the number of casualties.

```r
#Diagnostic Plots
par(mfrow = c(2, 2))
plot(lm_model, which = 1)
plot(lm_model, which = 2)
plot(lm_model, which = 3)
plot(lm_model, which = 4)
```



From the diagnostic plots we can conclude non-linearity, non-normality, and heteroscedasticity. The Cook's

Distance Plot suggests non-influential observations.

```
#Adding more Predictor Variables
trans_model <- lm(Casualty ~ AMPM + WEATHER + TEMP + VISIBLTY, data = df2)
summary(trans_model)
```

Transformations, Diagnostics, & Variable Selection

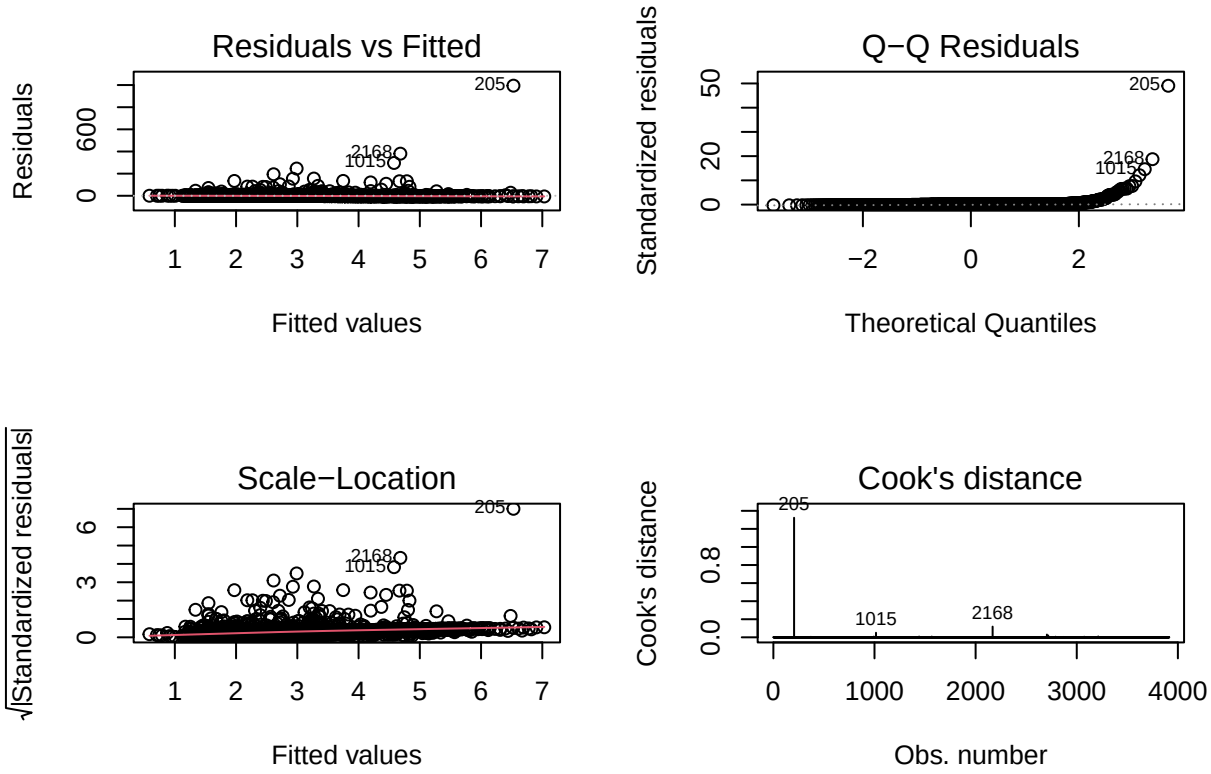
```
##
## Call:
## lm(formula = Casualty ~ AMPM + WEATHER + TEMP + VISIBLTY, data = df2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6.03  -2.56  -1.66  -0.77 994.47
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.54088   14.42712   0.245   0.8061
## AMPMAM       2.76479   14.37458   0.192   0.8475
## AMPMPM       2.02061   14.37672   0.141   0.8882
## WEATHER     -0.37643    0.37510  -1.004   0.3157
## TEMP        -0.04230    0.01661  -2.546   0.0109 *
## VISIBLTY     0.19112    0.36025   0.531   0.5958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.31 on 3899 degrees of freedom
## Multiple R-squared:  0.002482,   Adjusted R-squared:  0.001203
## F-statistic:  1.94 on 5 and 3899 DF,  p-value: 0.08445
```

```
AIC(trans_model)
```

```
## [1] 34607.16
```

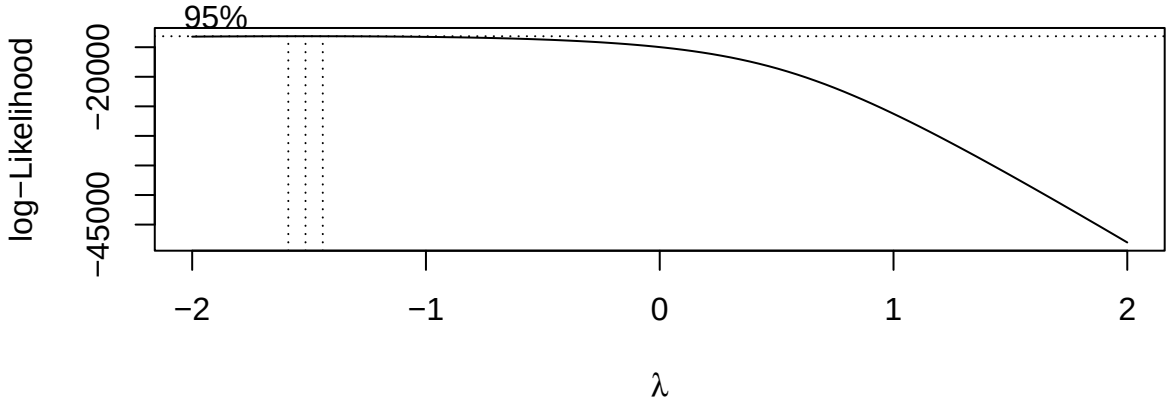
From the Adjusted R-squared,F-statistic, AIC and p-value: we can observe that the influence of time of the day is still not significant. But the influence of TEMP is marginally statistically significant.

```
par(mfrow = c(2, 2))
plot(trans_model, which = 1)
plot(trans_model, which = 2)
plot(trans_model, which = 3)
plot(trans_model, which = 4)
```



From the diagnostic plots we can conclude non-linearity, non-normality, and heteroscedasticity. The Cook's Distance Plot suggests some influential observations.

```
#Box Cox
boxcox_result <- boxcox(lm(Casualty ~ 1, data = df2))
```




```
lambda_optimal <- boxcox_result$x[which.max(boxcox_result$y)]
df2$CasualtyBoxCox <- (df2$Casualty^lambda_optimal - 1) / lambda_optimal

bc_model <- lm(CasualtyBoxCox ~ AMPM, data = df2)
summary(bc_model)
```

```
##
## Call:
## lm(formula = CasualtyBoxCox ~ AMPM, data = df2)
##
## Residuals:
```

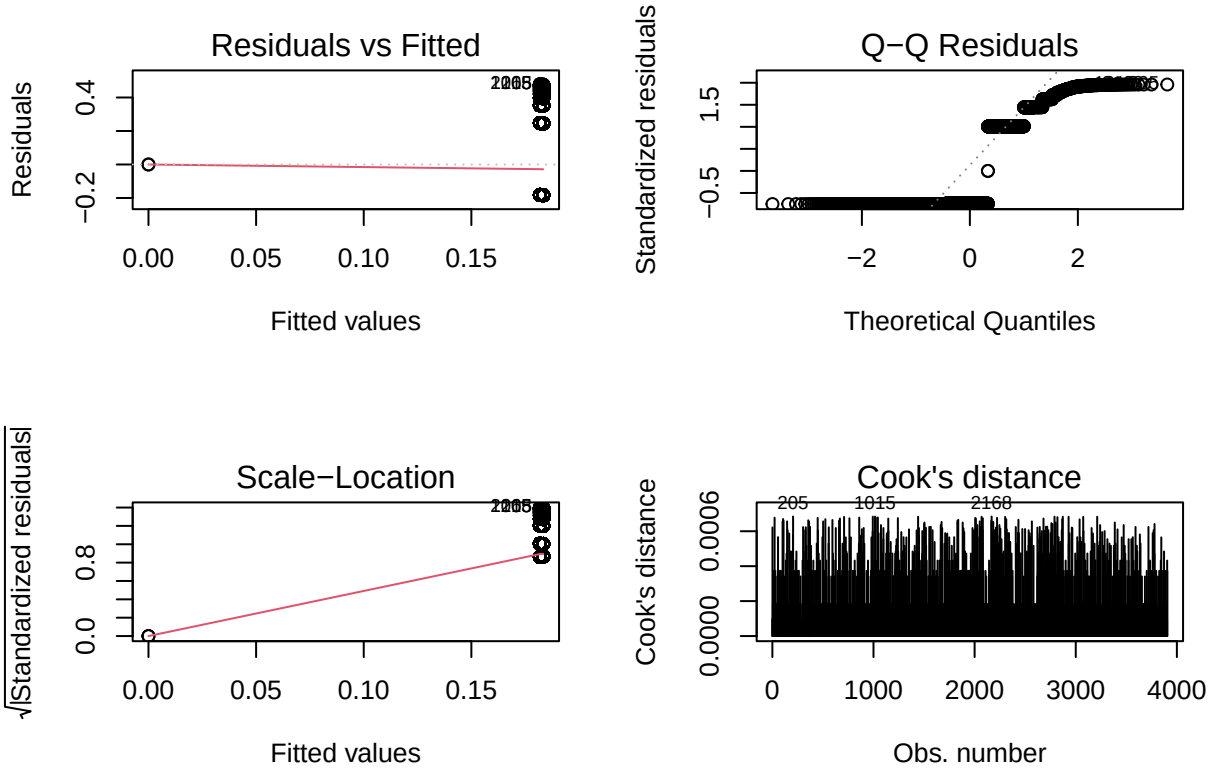
```
##     Min      1Q  Median      3Q     Max
## -0.1834 -0.1834 -0.1818  0.2473  0.4782
##
## Coefficients:
##                         Estimate         Std. Error t value Pr(>|t|)
## (Intercept) 0.0000000000003035 0.1725788449721833   0.000    1.000
## AMPMAM      0.1818360117912917 0.1726710099159876   1.053    0.292
## AMPMPM      0.1834084655027417 0.1726637964146870   1.062    0.288
##
## Residual standard error: 0.2441 on 3902 degrees of freedom
## Multiple R-squared:  0.0002972,  Adjusted R-squared:  -0.0002152
## F-statistic:  0.58 on 2 and 3902 DF,  p-value: 0.5599
```

```
AIC(bc_model)
```

```
## [1] 72.25114
```

From the Adjusted R-squared,F-statistic, AIC and p-value: we can observe that the influence of time of the day is still not significant.

```
par(mfrow = c(2, 2))
plot(bc_model, which = 1)
plot(bc_model, which = 2)
plot(bc_model, which = 3)
plot(bc_model, which = 4)
```



From the diagnostic plots we can conclude non-linearity, non-normality, and heteroscedasticity. The Cook's Distance Plot suggests no-influential observations.

```
#Log Transformation
df2$LogCasualty <- log(df2$Casualty + 1)
lg_model <- lm(LogCasualty ~ AMPM, data = df2)
```

```r
summary(lg_model)
```

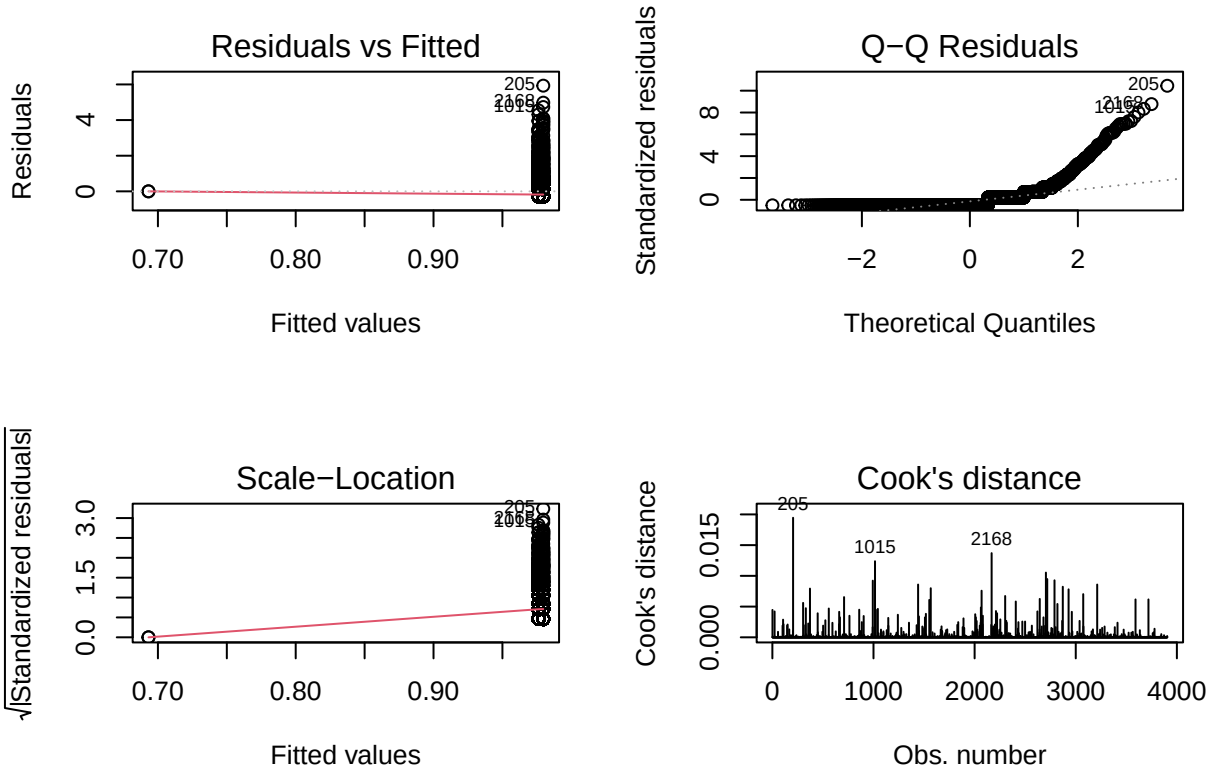
```
##
## Call:
## lm(formula = LogCasualty ~ AMPM, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2865 -0.2865 -0.2830  0.1224  5.9301
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6931     0.4014   1.727   0.0843 .
## AMPMAM        0.2865     0.4016   0.713   0.4757
## AMPMPM        0.2830     0.4016   0.705   0.4810
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5677 on 3902 degrees of freedom
## Multiple R-squared:  0.0001381,  Adjusted R-squared:  -0.0003744
## F-statistic: 0.2694 on 2 and 3902 DF,  p-value: 0.7638
```

```r
AIC(lg_model)
```

```
## [1] 6664.734
```

From the Adjusted R-squared,F-statistic, AIC and p-value: we can observe that the influence of time of the day is still not significant.

```r
par(mfrow = c(2, 2))
plot(lg_model, which = 1)
plot(lg_model, which = 2)
plot(lg_model, which = 3)
plot(lg_model, which = 4)
```



From the diagnostic plots we can conclude non-linearity, non-normality, and heteroscedasticity.#The Cook's Distance Plot suggests no-influential observations.

Evidence & Recommendation to FRA

ACCDMG Recommendations

Conclusion for ACCDMG Hypothesis 1 Based on our findings above, we fail to reject the null hypothesis. Despite the model producing a significant f-statistic p-value (p-value < 0.00000000000000022), our violation of the constant variance assumption leads us to believe our model is not reliable enough to reject the null hypothesis. Our low adjusted R-squared value of 0.183 also supports this conclusion.

Recommendations based on Hypothesis 1 The purpose of Hypothesis 1 was to study the relationship between common train accident causes and the level of accident damages with the intention of providing the FRA with recommendations to improve railroad safety. Based on our findings, we have found that human-error, crossing-related, high-speed, and derailment accidents do not have a statistically significant impact on accident damage costs. Some of these factors such as train speed and derailments showed potential to influence accident damage more, but we could not gather enough evidence to definitively support that claim. Perhaps the common train accident causes studied in this project result in more low-cost accidents, but that relationship was not studied due to our filtering for accident damages above the upper whisker.

We recommend investigating alternative factors to improve railroad safety as per the FRA's primary goal, but we also believe some of these factors should still be monitored for their potential significance in the future.

Conclusion for ACCDMG Hypothesis 2 Similar to our first hypothesis, despite having a p-value lower than the alpha ($.05 > 0.00000000000000022$). This is once again due to the presence of heterscedasticity where we can see a downward trend in the residuals vs fitted diagnostic plot. Similarly, our relativeley low adjsuted R^2 value of 0.2337 supports our conclusion, while also providing some evidence that there is potential for this model to explain accident damage to some degree.

Recommendations based on ACCDMG Hypothesis 2 Hypothesis 2 was intended to test whether different types of trains which get in different types of accidents results in significantly different damage profiles. We can see that freight trains are the most common type of train when it comes to extreme accidents, while derailment is the most common type of accident. However, passenger and commuter trains are more likely to get into highway-rail accidents. This information could be useful for training conductors and staff on the hazards and accidents they are most likely to encounter based on the type of train they are working on.

We want to interpret our model with a degree of skepticism since we failed to reject the null hypothesis, however exploring a few notable elements could be informative and useful for future studies. A clear takeaway is the role of speed in accidents. Speed has a positive interaction with freight trains and derailment accidents and its main effect predict greater accident damage. A potential implication of this is that the FRA could perhaps enforce greater safety regulations or be more prepared to respond to accidents involving trains with higher max speeds.

In addition, derailment accidents are the most common among extreme accidents but derailment accidents actually predict less damage than the base case (non derailmetn or highway-rail accidents). The fact derailment accidents are the most frequent but not the most severe, makes the FRA choose to direct their focus on either the most common accidentt type or the most severe accident type. We suggest the FRA continues to focus onhazard and damage reduction for derailment accidents, because they are disproportionately common among extreme accidents. Reducing the volume of derailment accidents we believe would be more effective than reducing the rarer, more extreme types of accidents.

Casualties Recommendations

Conclusion for Casualties Hypothesis 1 While our final model improved a lot compared to previous iterations (such as having a great AIC value), there are certainly still problems present with the model. The first is that the adjusted R-squared value, while improved, is still quite low. Despite the low f-statistic p-value indicating that the model is statistically significant, the low adjusted R-squared value makes us very skeptical that this model has much predictive or inferential power. Additionally, while the variance improved from the adjustments we made, it still displays some heteroscedastic behavior and therefore still violates the constant variance assumption. Violating this assumption also makes us skeptical of the model's power since lack of constant variance affects the precision of the model. In addition, heteroscedasticity tends to produce p-values that are deceptively significant, which may be misleading us to think that the model and its coefficients are more significant than they actually are.

Based on our findings above, we fail to reject the null hypothesis. Despite the model producing a significant f-statistic p-value (p-value < 0.0000000000000002), our violation of the constant variance assumption leads us to believe our model is not reliable enough to reject the null hypothesis. And despite having a low AIC of -273.8746, our low adjusted R-squared value of 0.09601 also supports this conclusion.

Recommendations for Casualties Hypothesis 1 The purpose of this hypothesis was to study the relationship between common train accident causes by type of train and type of accident and the amount of casualties with the intention of providing the FRA with recommendations to improve railroad safety. Based on our findings, we have found the type of train and type of accident in a railroad accident does not have a statistically significant relationship with casualties. Some of these factors such as Passenger trains and Commuter trains showed potential to influence the number of casualties more, but we could not gather enough evidence to definitively support that claim. Perhaps the common train accident causes studied in this project result in more no casualty accidents, but that relationship was not studied due to our filtering for train accidents with at least one casualty reported.

We recommend investigating alternative factors to improve railroad safety as per the FRA's primary goal,while continuing to monitor these factors due to their potential significance in the future.

Conclusion for Casualties Hypothesis 2 We fail to reject the null hypothesis. therefore, there is no strong evidence to suggest that the mean number of casualties in the dark is significantly different from the mean number of casualties in the day light.

Recommendations for Casualties Hypothesis 2 The objective of this hypothesis was to investigate the potential relationship between sunlight and casualties, aiming to provide safety-related recommendations to the Federal Railroad Administration (FRA). However, our findings indicate that there is no statistically significant correlation between the time of day, particularly sunlight, and the number of casualties in railroad incidents.

In light of these results, we recommend exploring alternative variables and factors to improve railroad safety, aligning with the primary mission of the FRA. Nevertheless, we also suggest that certain factors be continuously monitored, as they may still hold significance in influencing safety outcomes in the future.

https://uknowledge.uky.edu/ktc_researchreports/1069/