Linear Statistical Models

SYS 4021

Project 1

Analysis of Train Accidents in the U.S. During 2001-2016

Fan Feng
ff9sd@virginia.edu

Steve Sandry
ss7tf@virginia.edu

Lauren Schmeider
les8ae@virginia.edu

Summary

We have decided on two metrics to reduce the severity of train accidents: Accident Damage and Casualties (Total Killed plus Total Injured). In order to reduce Accident Damage (ACCDMG), trains should be slowed down on occasions where there is likely to be a derailment. Holding all else constant, the interaction between High Speed and Derailments is positive and statistically significant at the 5% level, with a p-value of 2*10^-16. Our best model for accident damage has a F-statistic of 180.8 and a p-value of less than 2.2*10^-16. Our three models are compared in **Table 4**. A one unit increase in HIGHSPD, in derailment accidents, results in a -8.431*10^-6 change in ACCDMG^(-0.5), meaning that there is a higher accident damage.

In order to reduce the number of casualties, trains that are travelling at higher speeds should have more advanced training of the employees who work on the train in order to reduce the number of accidents that are caused by human factors. The interaction between the cause of human factors and high speed is statistically significant and positive, meaning that higher speed accidents cause more casualties, holding all else constant, for crashes caused by human factors compared to crashes with another cause. This interaction is statistically significant at the 5% level, with a p-value of 0.00396. Our overall model is significant with an F-statistic of 21.78 and a p-value of 2.2*10^-16. Our three models are compared in **Table 6**. A one unit increase in HIGHSPD, in accidents caused by human factors, results in an increase in casualties of 0.074.

In summary, our ACCDMG model has a positive and statistically significant interaction between high speed and accident type derailment, meaning that decreasing speed on tracks that are likely to cause derailments and reducing derailments in general will decrease accident damage. Also, our Casualty model has a positive and statistically significant interaction between high speed and human factors, meaning that decreasing speed on tracks that are subject to human factors accidents and reducing human errors in general would decrease casualties. We are confident in the validity of both of our selected models due to their extremely low p-values, as listed above. As a result, we feel comfortable providing recommendations based on our results. Therefore, these different actions can help reduce overall train accident severity.

Honor Pledge: On my honor, I pledge that I am the sole author of this paper and I have accurately cited all help and references used in its completion.

x Lauren Schmeider, Steve Sandry, Fan Feng

1 Problem Description

1.1 Situation

This project [1] will employ the specified template [2]. The data set indicates that between 2001-2016, there are 45,507 different reported incidents of train accidents[3]. The motivation for this report is to investigate factors that result in severe train accidents and provide recommendations on how to improve these conditions. **Figure 1** demonstrates the monetary damage landscape in each of these years. **Figure 2** presents a log transformation of the accident damage ($) data. As can be seen in **Figure 2**, while there is some variety between years, there is a consistent problem with train accidents resulting in high monetary damage which we aim to reduce.

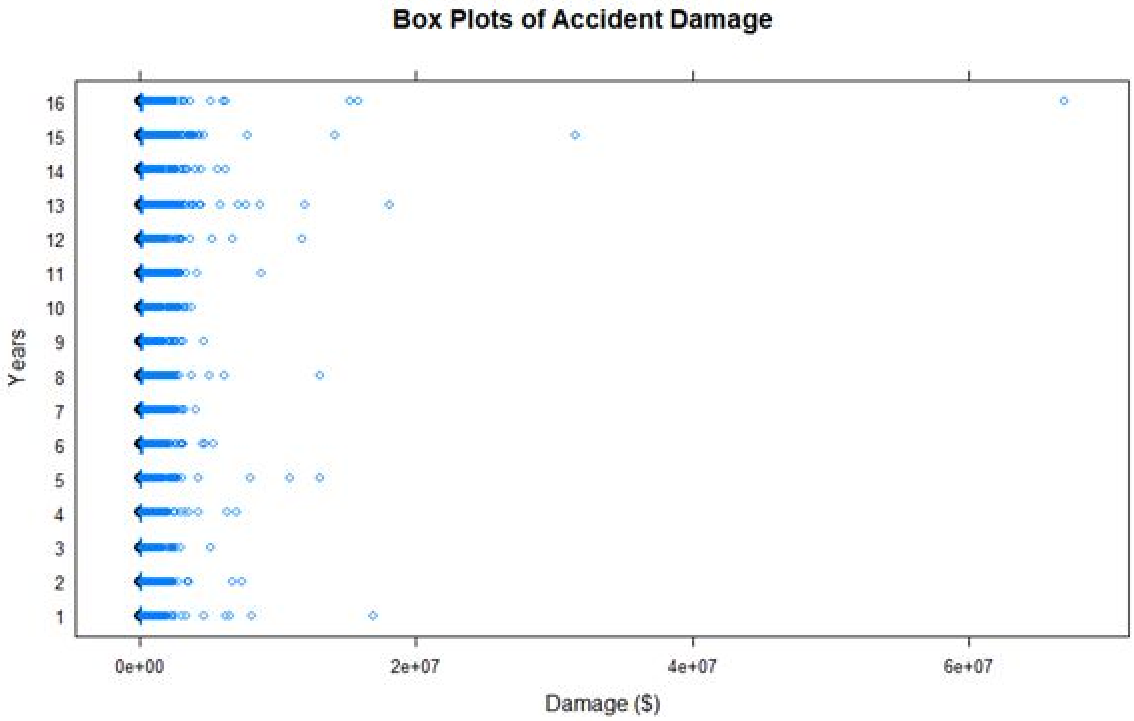

Figure 1: Boxplots of accident damage for the total accidents data

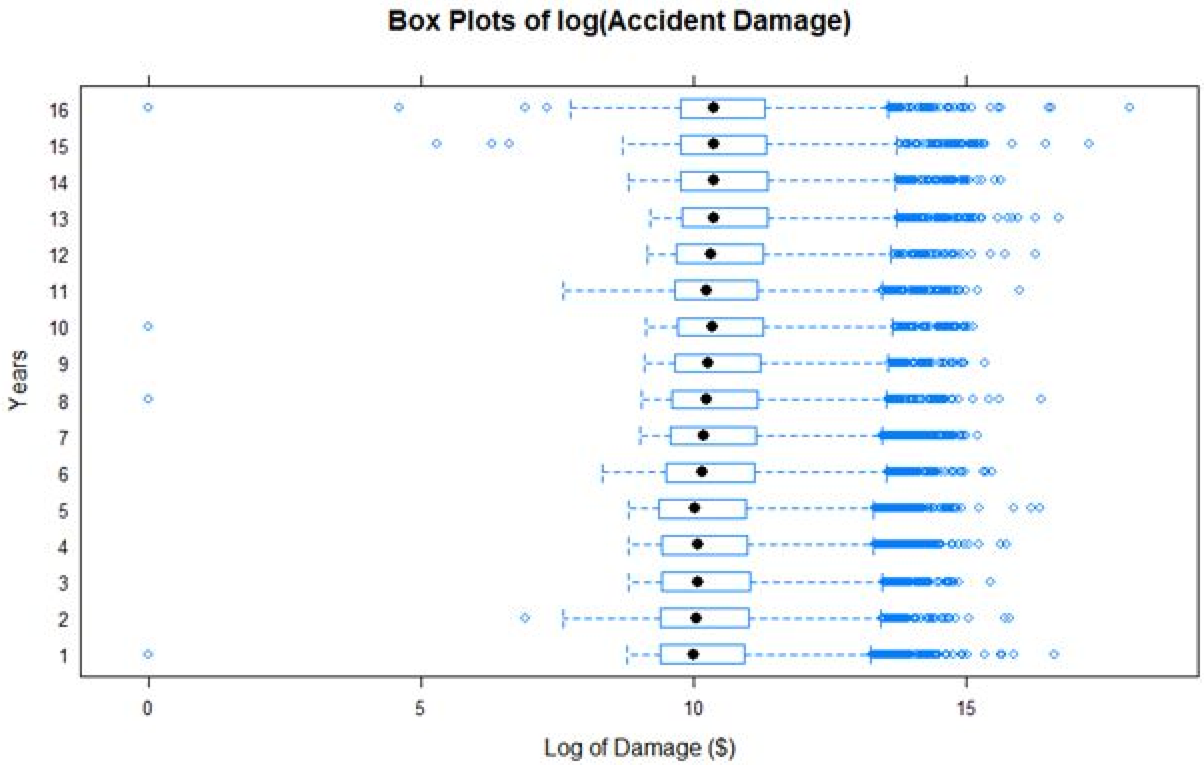

Figure 2: Boxplots of (log) accident damage for the total accidents data

These accidents also prove to negatively impact human safety issues. A total of 831 people have been killed in train accidents over the past 16 years while 9,134 people have been injured. **Figure 3** demonstrates that there are significantly more derailment accidents than there are of any other type of accidents. Therefore, safety is greatly impacted by factors that cause derailed train cars. Within the data, there are five different identified causes of accidents: mechanical and electrical failures (E), miscellaneous causes (M), rack, roadbed, and structures causes (T), signal and communication causes (S), and train operation and human factors causes (H). **Figure 4** demonstrates that human factors is the most frequent cause of train accidents, followed by rack, roadbed, and structures causes. Unlike the third most likely cause, miscellaneous causes, the top two are causes that can be examined through analysis as the problem is clearly identified.

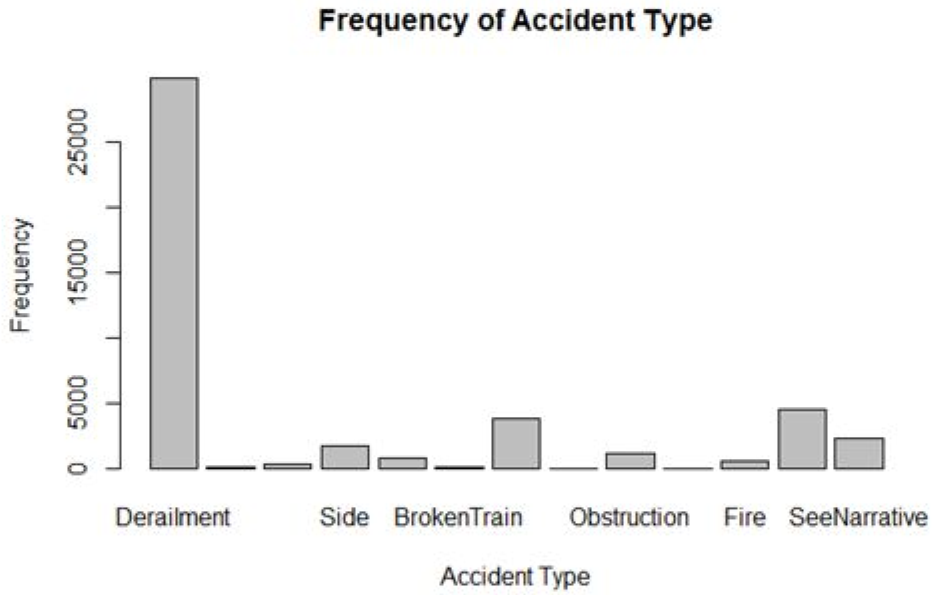

Figure 3: Frequency of the different types of accident damage

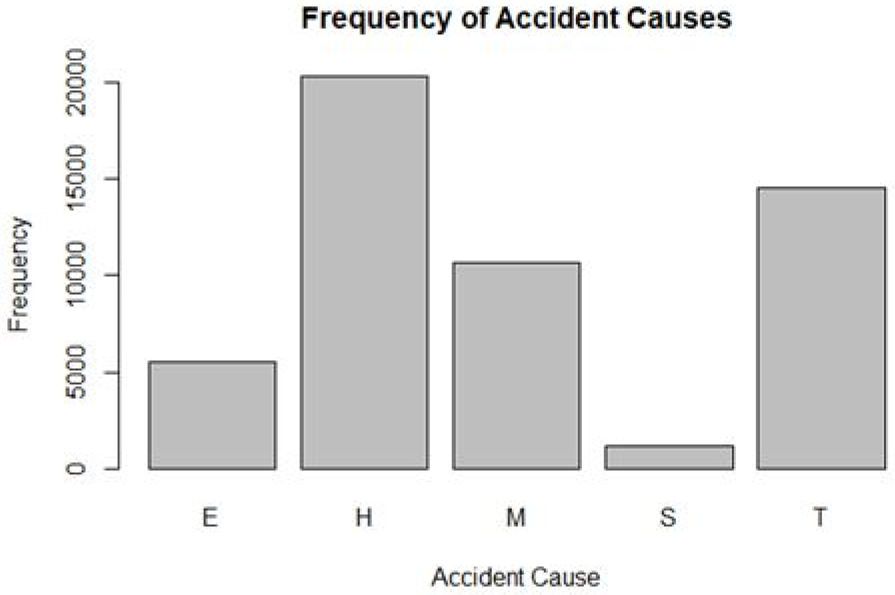

Figure 4: Frequency of accident causes

While there are many accidents, only 18.7% of accidents result in monetary damage that is greater than $100,000. In the collection of accidents, there are relatively few that cause severe impacts. Therefore, in order to focus on reducing the severity of train accidents, only those accidents labeled as "severe" should be considered. An "extreme accident" should be considered anything that falls above the top whisker of the boxplot in **Figure 5**. This whisker indicates that an extreme accident is an accident resulting in greater than $153,129 in monetary accident damage.

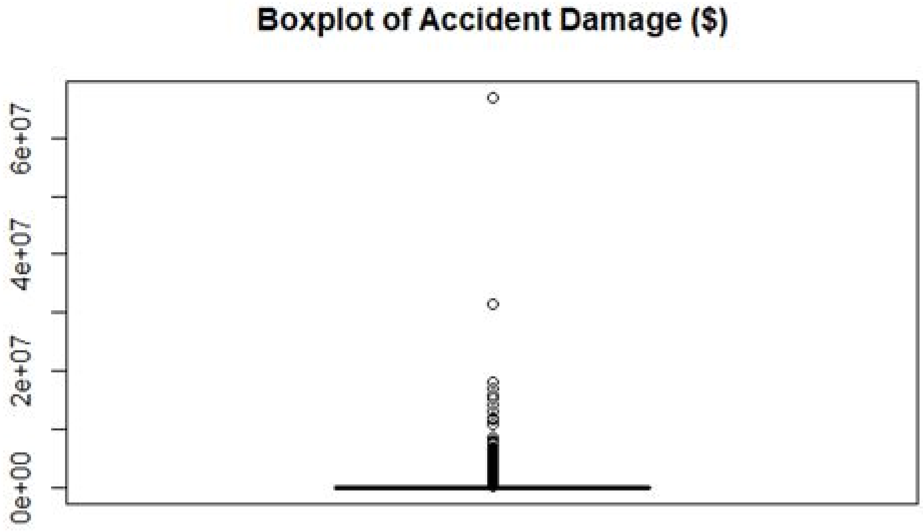

Figure 5: Boxplot of accident damage

In order to assess variables that correlate with the monetary accident damage, the scatter plots in **Figure 6** were generated. In order to select which values were to be included, intuition was used to infer potentially relevant factors. Note that other factors that proved not to be significant were also considered during the analysis. This graphic demonstrates that HIGHSPD and TRNSPD show strong correlation with monetary accident damage, which suggests a potential relationship between the train's speed and the cost of the resulting accident. Also, TRKDNSTY, or the amount of weight that the track bears each year, has a correlation with accident damage of 0.15. This implies that TRKDNSTY may impact the severity of accidents. Also, this plot suggests that TypeD (derailments) may also correlate with monetary accident damage. Further, it can be noticed that Cause and TypeD have a correlation of 0.27, so it may be beneficial to explore interactions with Cause in our analysis, as it pertains to our hypotheses. Since this is the most common type of accident, this variable was explored and it appears that derailment accidents may also influence the severity of an accident monetarily. Some other intuitively selected variables, such as CARS and TEMP, demonstrated fairly low correlation within this plot (0.02, 0.01).

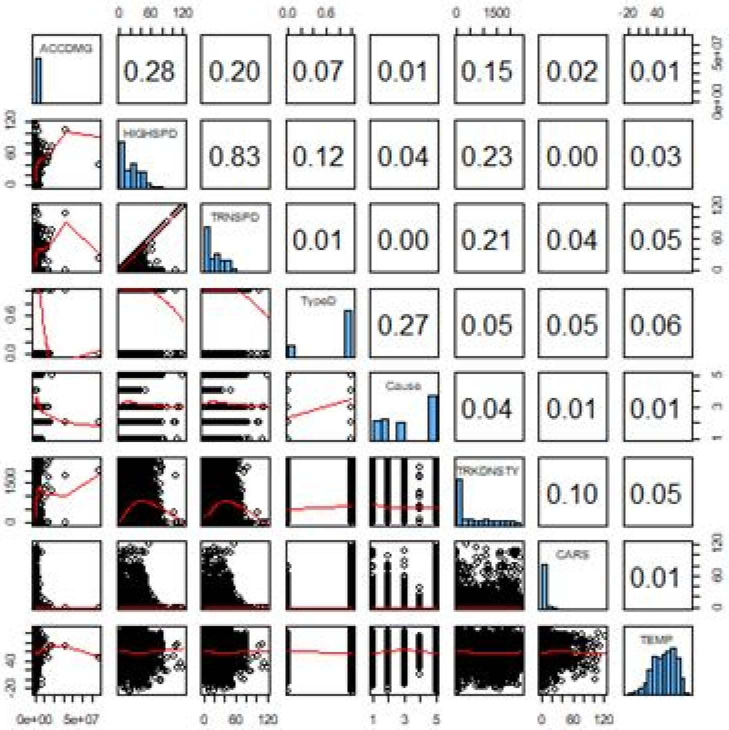

Figure 6: Scatter plot matrix for severe accidents for accident damage

If we consider a casualty to be an individual that has been injured or killed, we found that among all accidents, 6.6% of them have at least 1 human casualty. Similar to the case where we use monetary damage as the metric, only cases that are considered “severe” are used here. An accident is “severe” if it falls above the upper whisker of the boxplot in **Figure 7**, which means at least 1 human being is killed or injured in the accident.

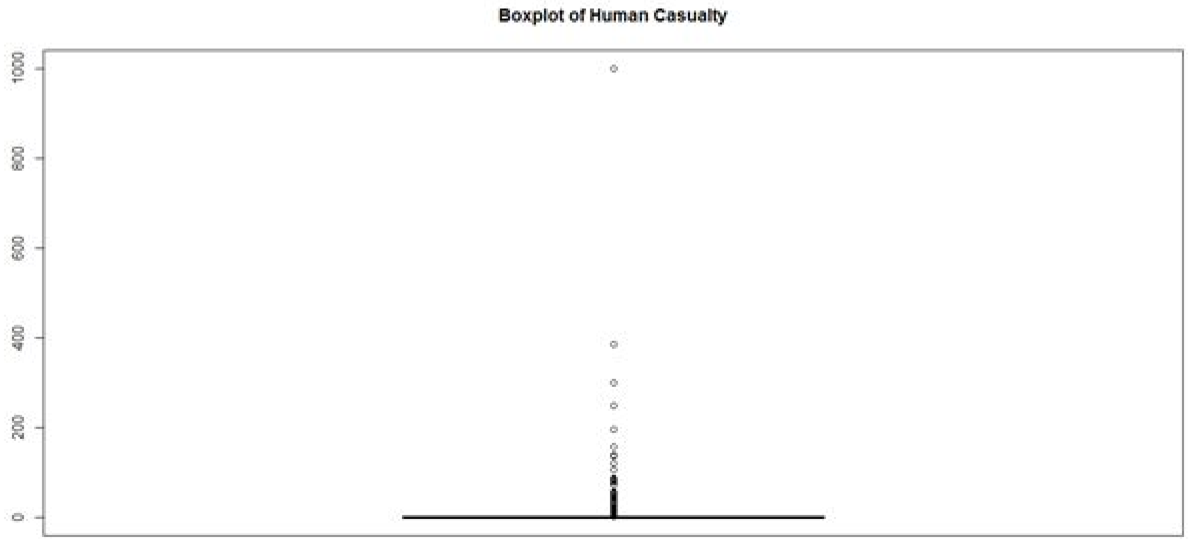

Figure 7: Boxplot of human casualty

In order to choose predictors for Casualty, we created the scatter plots of potentially impactful variables (**Figure 8**). Most of these variables were selected for this plot by intuition or due to results from other previous investigatory plots. Again, note that other variables outside this range were considered, but not deemed appropriate for our approach. This graph shows potential involvement of HIGHSPD and TRNSPD with casualties, therefore implying that the speed of a train may impact the number of casualties at the time of the accident. Also, LOADP1, or the number of loaded passenger cars, may play a role in the number of casualties. This concept is logical because if there are more passengers on the train, it is more likely that many people may be impacted physically by the accident. Along with this, LOADP1 appears to be correlated with train speed, implying that trains with more passengers may travel faster typically than those with less passengers. Also, while CauseH (human factors cause) may not appear directly impactful to the number of casualties here, it is interesting to note that CauseH may have a relationship with TRNSPD/HIGHSPD; since these speed variables may relate to the number of casualties, it is relevant to keep in mind what may be interacting with speed as well. Other variables that were explored, such as Badweather (foggy, snowy, rainy, sleet), do not appear to demonstrate significant impact in terms of the number of Casualties produced from an accident. However, it may still be interesting to observe variables such as Badweather within some models to ensure that it is having no impact. Lastly, TRKDNSTY shows a relevant correlation with HIGHSPD (0.24), so it may be worthwhile to consider this variable when working with HIGHSPD; therefore, we will better be able to understand the driving forces between certain variables and their interactions.

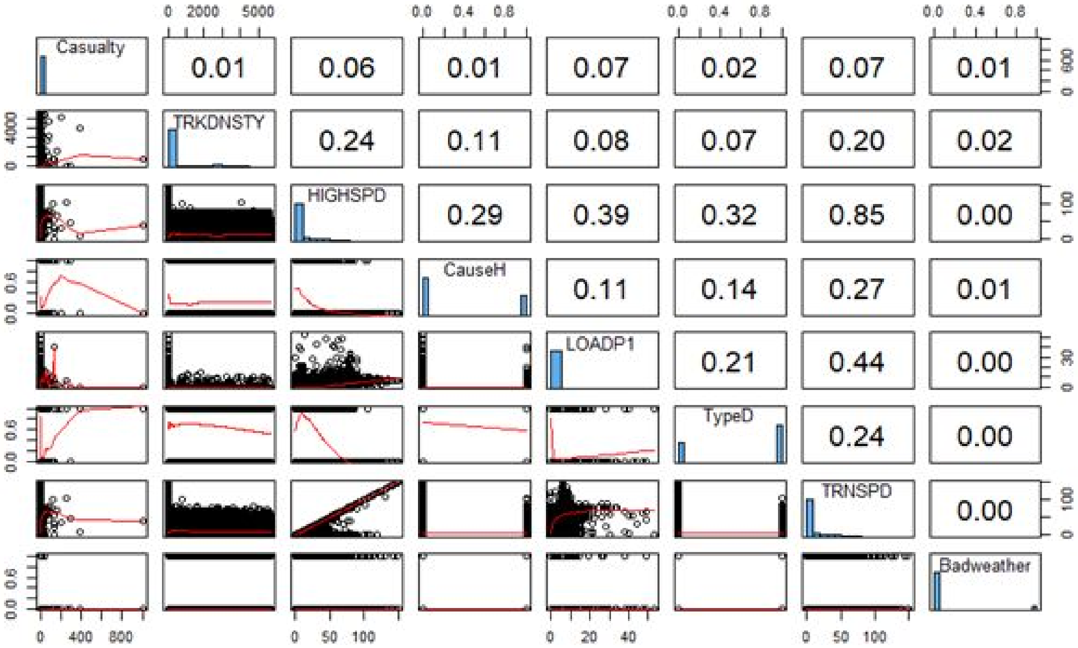

Figure 8: Scatter plot matrix for severe accidents for casualties

In order to further investigate some of the variables that were previously discussed and deemed potentially relevant in the scatter plots, we generated interaction plots. Note that other interaction plots were investigated beyond those present in the report. In order to differentiate between "low" and "high" speeds in the HIGHSPD variable, we decided to categorize anything 40 mph or greater as "high speed," and anything below this margin as "low speed." When first observing **Figure 9**, we found something relevant to consider: at high speeds, derailments appear to cause more monetary accident damage than other types of accidents. Therefore, this plot demonstrates potential relationships between the likely significant variables of HIGHSPD, TYPED, and ACCDMG. Then, when observing **Figure 10**, it appears that among high speed accidents, those of the human factors cause (CauseH) appear to cause far more casualties than accidents of other causes. Therefore, this plot again demonstrates budding relationships between variables that were deemed potentially significant in the scatter plots. We intend to explore these relationships further as we generate our hypotheses.

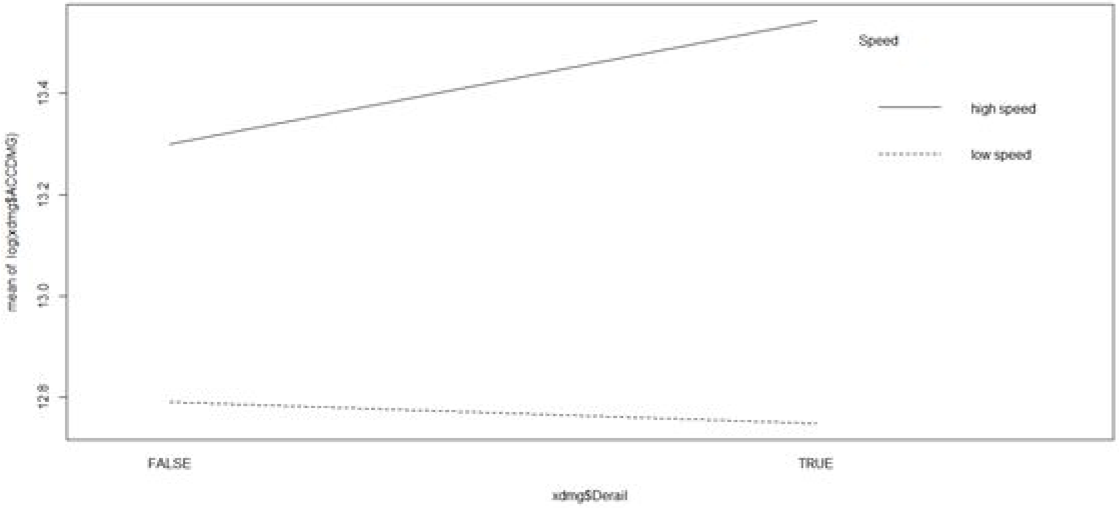

Figure 9: Interaction plot for building hypothesis with respect to ACCDMG

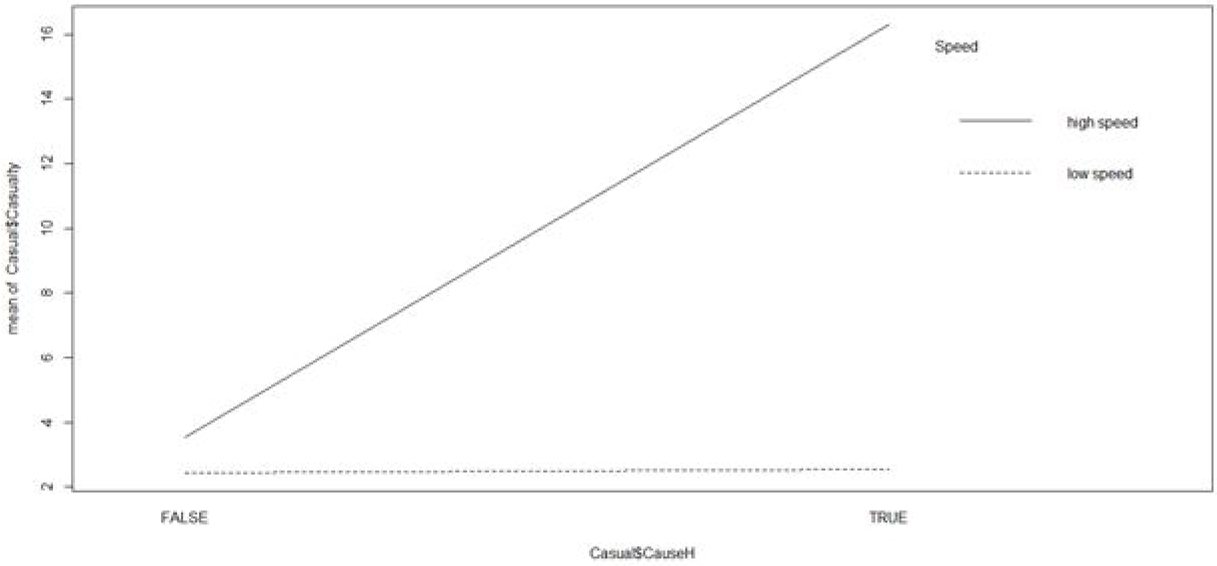

Figure 10: Interaction plot for building hypothesis with respect to Casualty

1.2 Goal

General Goal: Reduce the severity of train accidents

1) *Subgoal #1:* Aid the FRA in the reduction of monetary damage resulting from train accidents by providing them with the information outlined in the objectives
    a) <u>Objective</u>: Determine which factors impact the accumulated damage of extreme train accidents
    b) <u>Objective</u>: Determine what implementations may help reduce the severity of accidents in terms of monetary damage
2) *Subgoal #2:* Aid the FRA in the reduction of human casualties (injured and killed) resulting from train accidents by providing them with the information outlined in the objectives
    a) <u>Objective</u>: Determine which factors impact the number of human casualties in extreme train accidents
    b) <u>Objective</u>: Determine what implementations may help reducing the severity of accidents in terms of human casualties

1.3 Metrics

- Metric for Subgoal #1: **ACCDMG**

ACCDMG, TRKDMG, and EQPDMG were compared as potential metrics for monetary damage. ACCDMG was decided upon in part due to the biplot in **Figure 11**. It shows ACCDMG being highly significant, with the longest arrow. Also, this variable had the most loading in Component 1 and 2 of our PCA analysis, as shown in **Figure 12** and **Figure 13**.

- Metric for Subgoal #2: **Casualty**

Instead of using TOTINJ or TOTKLD, these variables were combined in the same weight to determine the total amount of human security damages that these accidents caused. After consideration of a metric that quantify human life monetarily (Casualty Cost), Casualty was still deemed the best option because most accidents have very few total deaths, hence making it a much higher weight than total injured in the response variable does not result in statistically acceptable models. For example, if accident A has 50 injured and accident B has 1 killed, the casualty cost of accident B will be much higher than that of accident A, while it is not the case in terms of accident severity. This variable was deemed significant as well based on principal component analysis. As can be seen in the biplot in **Figure 11**, Casualty appears to be a significantly stronger metric than TOTKLD. Also, Casualty has a higher loading for Component 2 of PCA (**Figure 13**). Therefore, Casualty is a comprehensive variable that appears to be a suitable metric for a hypothesis involving human safety in accidents.

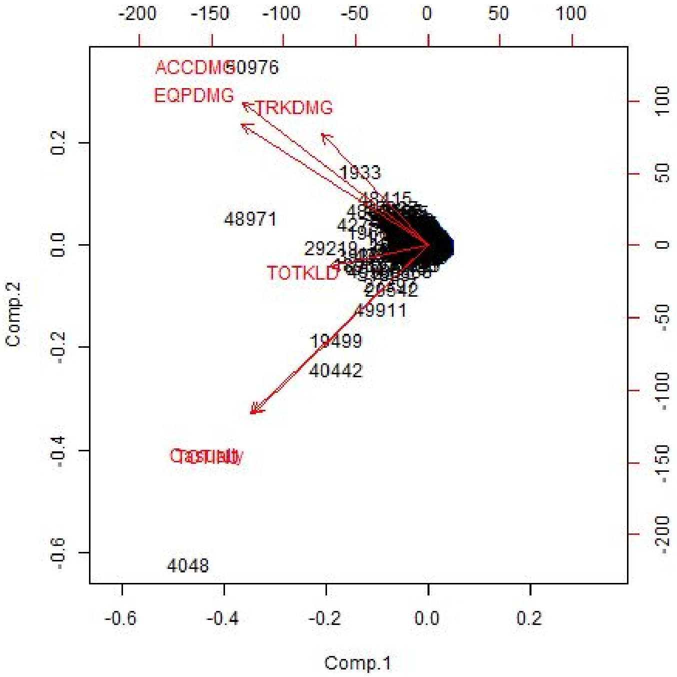

Figure 11: Biplot of principal components for determining metrics

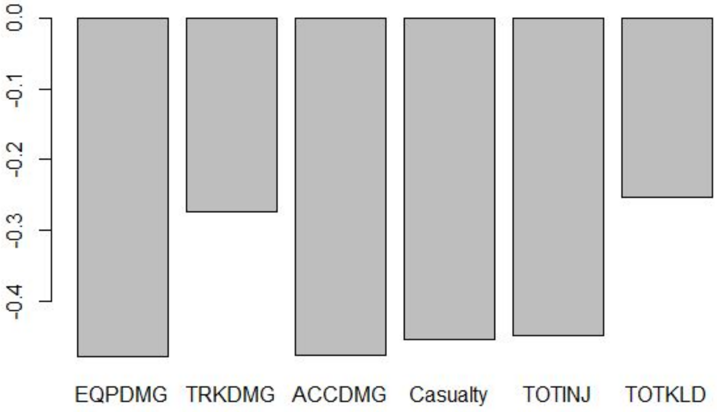

Figure 12: Loading 1 for PCA analysis

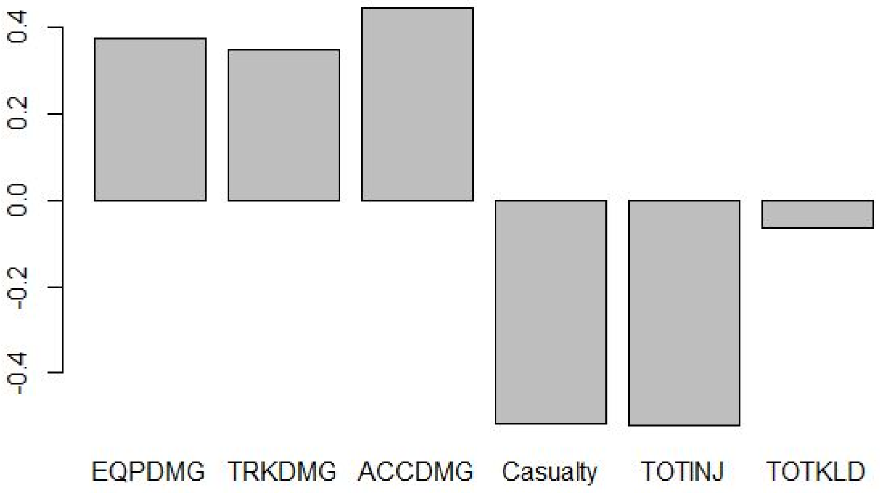

Figure 13: Loading 2 for PCA analysis

1.4 Hypothesis(es)

Hypothesis for Subgoal #1, pertaining to the metric ACCDMG
High speed derailments cause more monetary damage than high speed accidents of other types.

Hypothesis for Subgoal #2, pertaining to the metric Casualty
Accidents of trains traveling at high speeds that are of accident cause “H” (human factors) result in significant increases in human casualties, as compared to accidents of other causes.

2 Approach

2.1 Data

The database for this project originates from the Federal Railroad Administration; the data utilized is the total database for all traffic accidents happened in the U.S from 2001 and 2016 [3]. For the purpose of analysis, we removed all the duplicated observations with identical Incident Number, Year, Month, Day and Time. Since our objective is to reduce the severity of accidents, we built 2 subsets, each corresponding to accidents with extreme monetary damage and human casualties. Our first set that accounts for extreme monetary damage, xdmgnd, is a dataset that only includes the upper whisker of accident damage ($153,129 or more) from the total accidents data set without duplicates. Within this data set, a categorical variable that indicates if the accident is a derailment, TypeD, was created.

The second dataset that accounts for human casualties, Casual, includes all accidents (no duplicates) where the Casualty variable is greater than 0. TRKDNSTY was also originally coded as a categorical variable, so we changed that into a quantitative variable in R. After removing the 4 missing TRKDNSTY observations from xdmgnd and the 2 missing TRKDNSTY observations from Casual, the variable could be used properly. We also removed the 9/11/2001 observation from our data set because it is very unlikely for something like this to happen again, and if something were to happen, it would most likely not cause a different level of damage than we have in the scope of this project. During our model creation for Casualty, we discovered a few points that had a Cook's distance above 0.5. We considered removing any values that had this Cook's distance, and as a result we removed 9 data points from the Casual data set.

There are potential biases for this data. For example, a person would most likely have to report that an accident had occurred. This could allow for the potential of accidents going unreported because the person didn't want to be disciplined, or simply because the person did not believe what damage had occurred would be justified as an accident. There is also the possibility for someone to report something that should not have been classified as an accident. To correct for these biases, we split our data into two sets and only used the extreme values of accident damage and casualties in these data frames. This can account for accidents being reported that shouldn't have been, but we have no way of correcting for accidents going unreported.

There could also be biases in our data analysis. We removed some data points from the Casual data set because they were outliers for our linear model. However, it is possible that they could lead to insights in other models that we will no longer be able to find. To correct for this potential bias, we removed only the data points that we believed were necessary to remove and had a high cook's distance.

2.2 Analysis

Before we investigated our hypotheses, we created a few dummy variables to account for categorical variables in the dataset. Since derailment is the most common type of accident, we created TypeD to indicate if an accident is a derailment case. Also, we figured human factors, the most common cause among all the accident causes, would be the most controllable cause, so we created CauseH to indicate if the cause of an accident is human factors, and whose base case is all other causes of accidents. Last, we considered the weather condition might be useful in our analysis, so we created Badweather to indicate if the weather condition when the accident happened was hazardous (specifically, rain, snow, sleet and fog). The codomain of all 3 dummy variables is {0, 1}, with 1 means True and 0 means False.

Hypothesis 1

We mainly used scatter plots and correlation coefficients from uva.pairs to decide on the variables that we use in our models, as well as some intuition. In order to consider predictors for Accident Damage, we created the scatter plots of potentially impactful variables (**Figure 6**). HIGHSPD was selected as the measurement for speed because it is highly correlated with TRNSPD (0.83), but more strongly relates to ACCDMG (0.28 > 0.20). We think this is appropriate because TRNSPD is the speed when the accident occurred, while HIGHSPD is the maximum speed recorded up to the accident. Also compared to any other metrics, HIGHSPD is simpler to regulate and provide recommendations on. This figure also demonstrates some variables worth exploring as it pertains to ACCDMG, such as TypeD, TRKDNSTY, and some of the other interactions that were deemed potentially impactful (**Figure 6**). For example,

For the verification of our first hypothesis, three different compilations of different predictors were created in order to find the most appropriate model. The variables in these models were selected based on analysis of the scatter plots; detailed discussion of these variables can be found in the Situation section. It was determined to create the first model with the variables necessary to test the hypothesis; then, the other two incorporate variables and interactions that could potentially be impacting the hypothesis variables.

Model Number	Response variable	Predictors
1	ACCDMG	TypeD:HIGHSPD, HIGHSPD, TypeD
2	ACCDMG	TypeD:HIGHSPD, HIGHSPD, TypeD, Cause
3	ACCDMG	TypeD:HIGHSPD, HIGHSPD, TypeD, Cause, Cause:HIGHSPD, TRKDNSTY

Table 1: Variables utilized in ACCDMG linear models

*Model 1: ACCDMG = 433477 + 15774*HIGHSPD + -377521*TypeD + 10063*TypeD:HIGHSPD +* ε

*Model 2: ACCDMG = 120000 + 17688*HIGHSPD + -276467*TypeD + 397006*CauseH + 152602*CauseM + 221094*CauseS + 182413*CauseT + 9920*TypeD:HIGHSPD +* ε

*Model 3: ACCDMG = 4.803e+05 + -9.456e+02*HIGHSPD + -2.328e+05*TypeD + -5.861e+05*CauseH + 3.772e+04*CauseM + -1.743e+05*CauseS + -1.268e+05*CauseT + 6.823e+03*CauseD:HIGHSPD + 5.041e+04*HIGHSPD:CauseH + 4.751e+03*HIGHSPD:CauseM + 1.632e+04*HIGHSPD:CauseS + 1.137e+04*HIGHSPD:CauseT + 7.886*HIGHSPD:TRKDNSTY +* ε

After constructing the three models, the adjusted-R^2 values, AIC values, and BIC values were observed, found in **Table 3** in the Evidence Section (Section 3.1).

After observation of these values, linear model 3 exhibited a higher adjusted-R^2 and lower AIC and BIC values than the other two models. As a result, it was decided to create a stepwise regression of the linear models. However, the stepwise models resulted in the same exact resulting models and criterion values as the linear models beforehand.

Hypothesis 2

Again, we used scatter plots and intuition to choose variables that we are using to model Casualty (**Figure 8**). HIGHSPD was selected as the measurement for speed because it is highly correlated with TRNSPD (0.83), and still relatively correlated with Casualty. Compared to any other metrics, HIGHSPD is easier to regulate and give recommendations on. This figure also demonstrates some variables worth exploring as it pertains to Casualty, such as LOADP1, TypeD, and some of the other interactions that were deemed potentially impactful.. Also from the scatter plot matrix, we could identify interaction terms we would be using in the model, because if two predictors are correlated with each other, it is necessary to put the product of them as a parameter in the model otherwise the assumption of linearly independent predictors would be violated. Some of the ones implemented in these models are discussed in detail in the Situation section.

In order to test hypothesis 2, three different linear models were created and compared, similarly to hypothesis 1. After evaluating the “Situation” information, another method of determining predictors was creating scatter plots of potentially impactful variables (**Figure 8**). HIGHSPD was again selected instead of TRNSPD under the same rationale. See **Figure 8** in Section 1.1 for more information from the scatter plots relating the predictors.

Model Number	Response variable	Predictors
1	Casualty	LOADP1 + CauseH + HIGHSPD + TypeD + CauseH*HIGHSPD + TypeD*HIGHSPD + TRKDNSTY
2	Casualty	(HIGHSPD + CauseH + TRKDNSTY + Badweather)^2
3	Casualty	HIGHSPD + CauseH + TRKDNSTY + HIGHSPD*CauseH

Table 2: Variables utilized in Casualty linear models

After running the linear models, we obtained the following formula:

Model 1: Casualty = 1.610663 + 0.300130* LOADP1 + 0.048023*CauseH + 0.003757*HIGHSPD
+ -2.741850*TypeD + 0.074050*CauseH:HIGHSPD + 0.171774*HIGHSPD:TypeD + ε

Model 2: Casualty = 1.196 + 3.941e-02*HIGHSPD + 2.850e-01*CauseH + 2.950e-04*TRKDNSTY
+ -1.693e-01*Badweather + 9.928e-02*HIGHSPD:CauseH + -2.915e-05*HIGHSPD:TRKDNSTY
+ 3.063e-03*HIGHSPD:Badweather + -2.096e-03*CauseH:TRKDNSTY +
-1.119*CauseH*Badweather + -1.426e-05*TRKDNSTY*Badweather + ε

Model 3: Casualty = Casualty ~ 1.4794788 + 0.0345444*HIGHSPD + -0.2245897*CauseH +
-0.0011460*TRKDNSTY + 0.0916421*HIGHSPD:CauseH + ε

After constructing the three models, the adjusted-R^2 values, AIC values, and BIC values were observed and the results in **Table 6** in Section 3.2 (Evidence) were found.

After observation of these values, linear model 1 exhibited a higher adjusted-R^2 and lower AIC and BIC values than the other two models (Section 3.2). As a result, it was decided to create a stepwise regression of the linear models. The stepwise model for Linear Model 1 slightly improves this model in terms of the criterion in the table above, so we will proceed with this stepwise. The criterion values can be seen for this updated model in **Table 7** in Section 3.2 (Evidence). However, performing a stepwise regression on Linear Model 2 makes this model equivalent to Linear Model 3 and performing a stepwise regression on Linear Model 3 results in an identical model. Therefore, it is determined that it is best to leave these models in their original form in order to provide more variety in potential models.

3 Evidence

Hypothesis 1

In order to compare the three models that were created for this hypothesis, the adjusted R^2, AIC, and BIC were calculated for each and can be found in **Table 3**.

	Linear Model 1	**Linear Model 2**	**Linear Model 3**
Adjusted R^2	0.08432	0.09052	0.1458
AIC	185538.6	185502	185132.8
BIC	188572	185562.2	185226.5

Table 3: Criterion of the three models with the best result highlighted

Table 3 exemplifies that Linear Model 3 has the highest adjusted-R^2 of the three models, as well as the lowest AIC and BIC. Therefore, based on these different criteria, it can be stated that Linear Model 3 is the dominant model. However, it is important to check the assumptions to ensure that these models are valid before drawing conclusions.

Figures 14, **15**, and **16** show the assumptions for each linear model, respectively, and are found in Section 3.1. As can be seen for the plots of all three models, the QQ plot of each model indicates some failure due to the trend on the left side of each plot as well as deviation from line. The Residuals vs. Leverage plots also indicate potentially influential points that deviate from the rest of the data. These influential points can also be seen on the other residual plots. There is also some heteroscedasticity shown in the residual vs fitted plots, with higher variance above zero compared to below zero. Overall, it appears that the assumptions could be better met and transformations of the response variable should be explored.

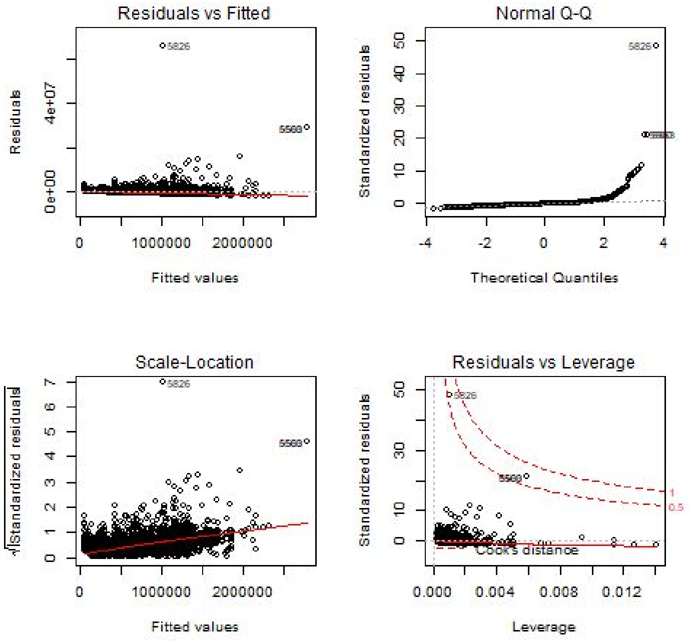

Figure 14: Diagnostic plots for Linear Model 1

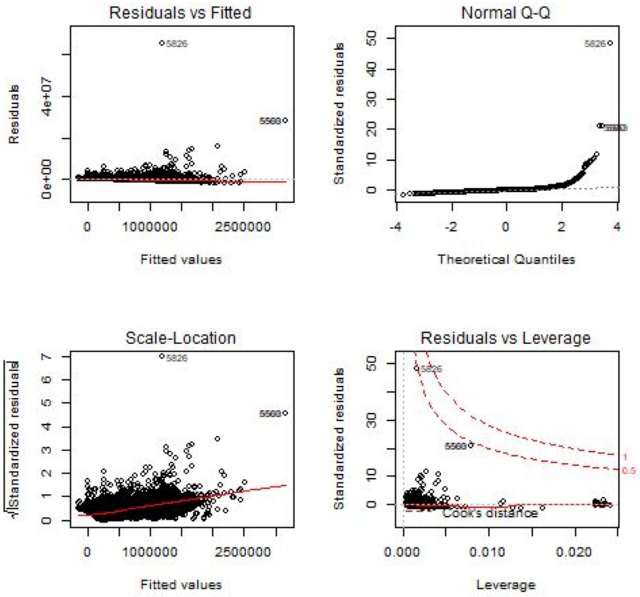

Figure 15: Diagnostic plots for Linear Model 2

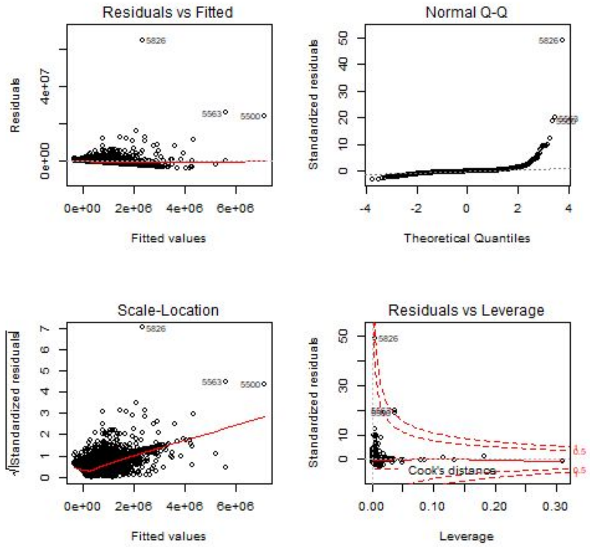

Figure 16: Diagnostic Plots for Linear Model 3

In order to determine response variable transformation, Boxcox plots were generated. An example Boxcox plot for Linear Model 3 is presented in **Figure 17**. **Figure 17** demonstrates that a transformation of -0.5 should be utilized. All three of our models required the same transformation action, hence graphs for only one model are shown as an example. The diagnostic plots for the transformed models are improved for each model, and are shown in **Figures 18**, **19**, and **20**, respectively. The normality is greatly improved, the residuals vs. fitted and scale-location plots demonstrate better scatter, and the influential points do not appear to be created a significant impact on the data. Therefore, the transformed models will be utilized for the rest of the model comparison. Note that for interpretation purposes, the original linear models will be interpreted in terms of the transformation, for simplicity. We feel that showing the transformed models may cause difficulty in understanding the results, as they are more complicated.

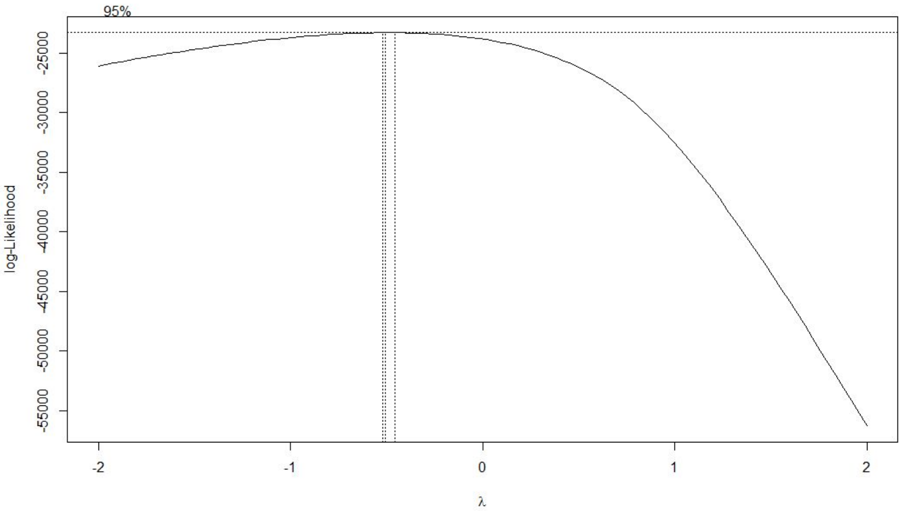

Figure 17: Boxcox plot for Linear Model 3, before transformation

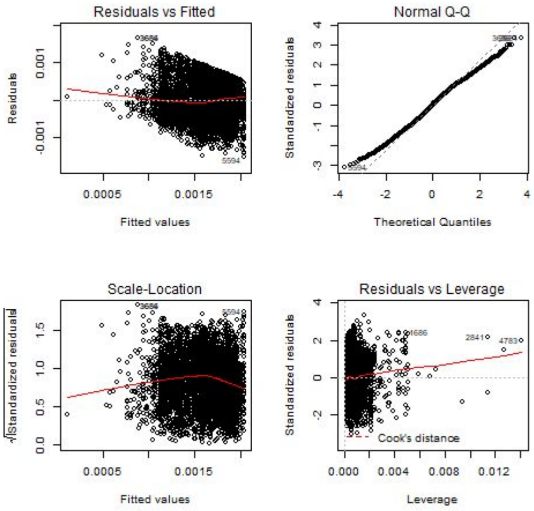

Figure 18: Diagnostic plots for Linear Model 1 - Transformed

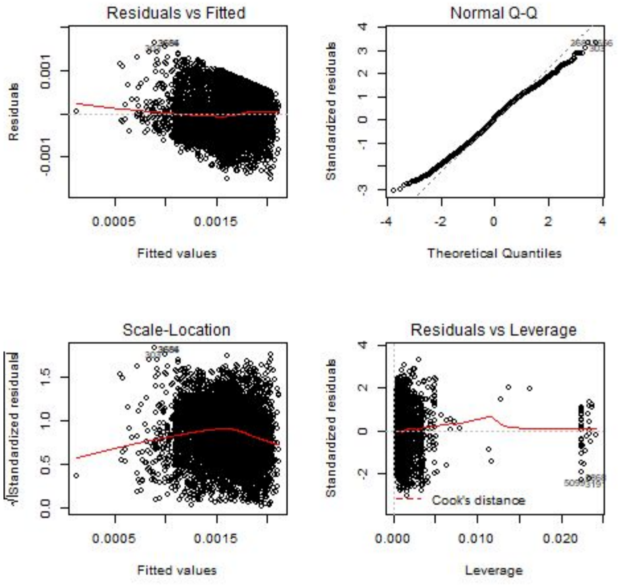

Figure 19: Diagnostic plots for Linear Model 2 - Transformed

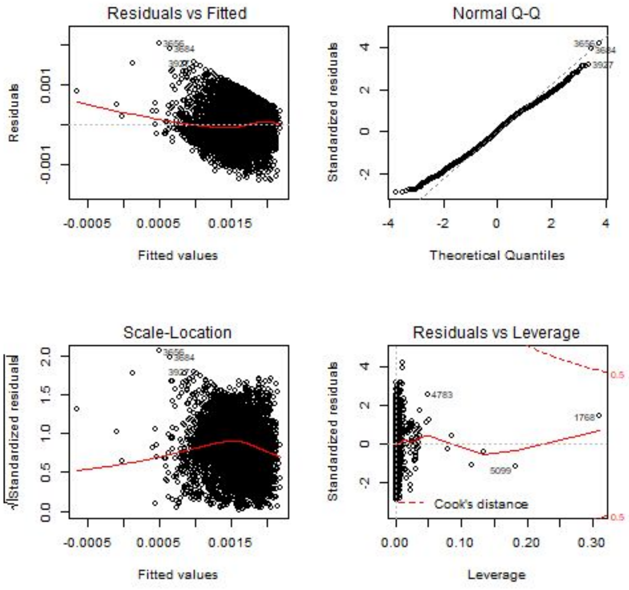

Figure 20: Diagnostic plots for Linear Model 3 - Transformed

In order to compare the newly transformed plots, the criterion values are again evaluated, as can be seen in **Table 4**. **Table 4** demonstrates how the transformation improved all three of the values for each model. Also, this information highlights how Linear Model 3 - Transformed, is again the strongest model for all three assessment values.

	Linear Model 1 - Transformed	**Linear Model 2 - Transformed**	**Linear Model 3 - Transformed**
Adjusted R^2	0.228	0.232	0.266
AIC	-73819.85	-73858.69	-74107.87
BIC	-73786.38	-73788.44	-74014.15

Table 4: Criterion values for the three transformed linear models

Another way to compare the three generated models is through the creation of test sets. The test sets will be run on the transformed models that were previously generated. In order to evaluate these test sets, the predicted mean square error (PMSE) values will be generated on a graph of 20 different test runs, to account for the values fluctuating with each run. However, when running the test sets for our three models, the PMSE graphs for each transformed model were all identical. Therefore, test sets do not provide any further information when trying to decipher the preferred model. As a result, we will not consider the test sets results when selecting our model; we will focus on the criterion metrics and employ cross-validation.

Then, we used cross-validation to compare our models. After running cross validation on each model, the predicted mean square error (MSE) for each transformed model can be analyzed. After running the test a few times, it seems that the trends are about the same. **Table 5** below shows an example set of values. It can be seen that Linear Model 3 - Transformed, has the smallest predicted MSE, followed by Linear Model 2 - Transformed and then Linear Model 1 - Transformed. Therefore, Linear Model 3 - Transformed exhibits the best predicted MSE compared to the other two models; based on cross-validation, that would make Linear Model 3 - Transformed the preferred model.

Linear Model 1 - Transformed	Linear Model 2 - Transformed	Linear Model 3 - Transformed
2.485e-07	2.475e-07	2.369e-07

Table 5: Predicted MSE from cross-validation for each transformed model

As a result, these analytic findings suggest that Linear Model 3 - Transformed should be utilized as evidence for Hypothesis 1.

As it pertains to the first subgoal's first objective, it can be stated that some factors that impact the accumulated damage of extreme accidents are: TypeDTRUE, CauseH, CauseS, CauseT, TypeDTRUE:HIGHSPD, HIGHSPD:CAUSEH, HIGHSPD:CauseM, HIGHSPD:CausesS, HIGHSPD:CauseT, and HIGHSPD:TRKDNSTY. These are variables that appear significant in this model by a 0.001 significance level. Specifically shifting to the hypothesis for this section, this model clearly demonstrates how high speed derailments cause an increase in ACCDMG. For every 1 unit increase in HIGHSPD:TypeD, there is a -8.431*10^-6 change in ACCDMG^(-0.5). This means that ACCDMG increases as HIGHSPD:TypeD increases[5]. We are confident in this model as it has a p-value of 2.2*10^-16. Therefore, we feel that this information can be used in the recommendation section to address the second objective of subgoal #1.

Hypothesis 2 (Section 3.2)

In order to compare the three models that were created for this hypothesis, the adjusted R^2, AIC, and BIC were calculated for each and can be found in **Table 6**. These results exemplify that Linear Model 1 has the highest adjusted-R^2 and the lowest AIC and BIC of all the models.

	Linear Model 1	**Linear Model 2**	**Linear Model 3**
Adjusted R^2	0.05056	0.01659	0.01739
AIC	19250	19349.06	19340.84
BIC	19303.22	19420.02	19376.32

Table 6: Criterion for the 3 linear models generated for Hypothesis 2

Therefore, based on these different criteria, it can be stated that Linear Model 1 is the dominant model. Also, it was determined in the Analysis section that Linear Model 1 - Step will be employed, so the criterion for this improved model are demonstrated below. Again, these criterion are all dominant as compared to the other models.

	Linear Model 1 - Step
Adjusted R^2	0.05086
AIC	19248.12
BIC	19310.37

Table 7: Criterion for the stepwise regression of Linear Model 1 for Hypothesis 2

Diagnostic plots for each model were created and evaluated. **Figures 21**, **22**, and **23** show the assumptions for each linear model, respectively. The most problematic assumption appears to be Gaussian distribution of the error term in the Q-Q Plot. The right tail is deviant from the expected trend; also, the spread in Residuals vs. Fitted and Scale-Location are not as random as one would expect. Therefore, the next step would be to run a transformation on the response variable.

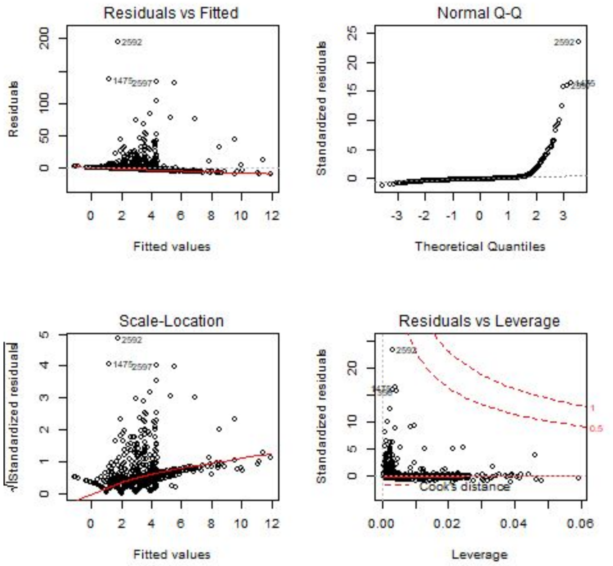

Figure 21: Diagnostic plots for Linear Model 1 - Step

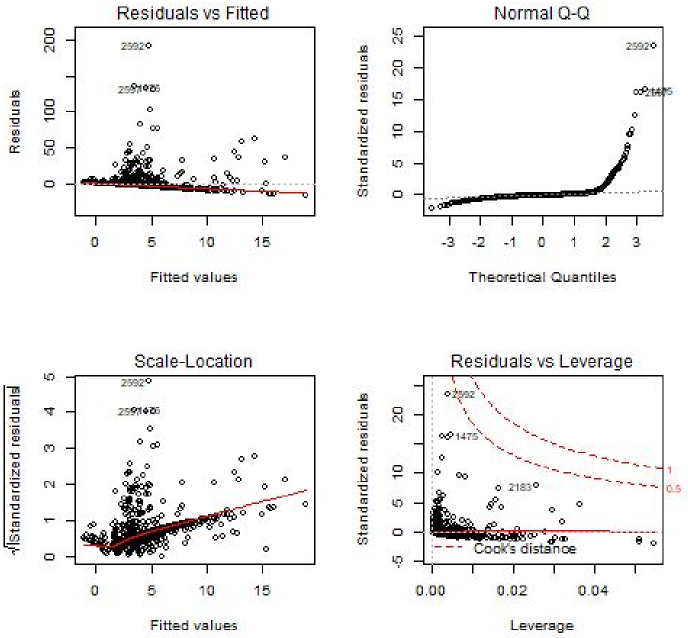

Figure 22: Diagnostic plots for Linear Model 2

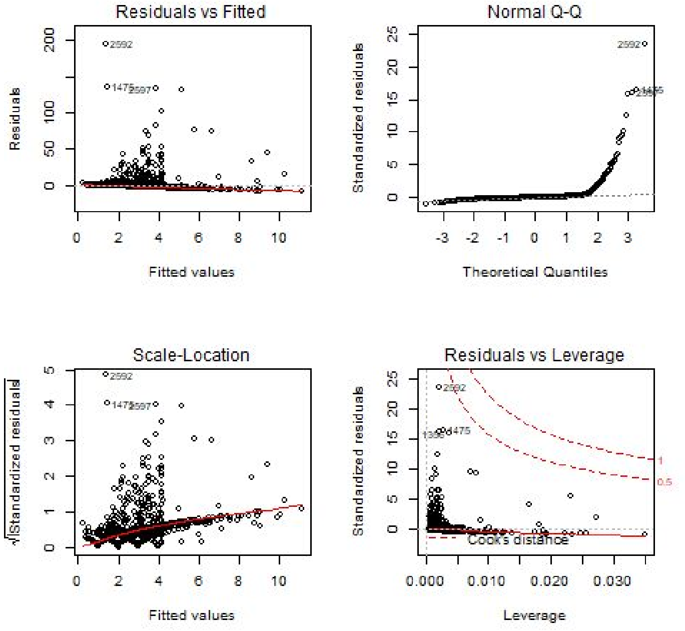

Figure 23: Diagnostic plots for Linear Model 3

As can be seen in **Figure 24**, the Boxcox plot demonstrates the necessary transformation. However, the transformed linear models worsens the diagnostic plots. An example can be seen for the transformation of Linear Model 2 - Transformed in **Figure 25**. The transformation appears to make the assumptions very poorly met, especially as compared to the assumptions prior to the transformation. The first and third model also follow this trend when transformed. Therefore, it was decided by us to move forward with the models that were not transformed because they best meet the assumptions, despite some downfalls. Also, the original model meets the objective of working with the most simple model possible. The reason that these transformations don't aid in our analysis is because another method for predicting casualties needs to be utilized. This is caused by the factor that Casualty data is counts, which isn't the same as typical continuous quantitative data. Transformations on count data doesn't improve the normality of linear models. Instead, a more appropriate model would be a Poisson regression model[4]. However, we are going to continue analysis with linear regression because this is the best method we currently have at our disposal.

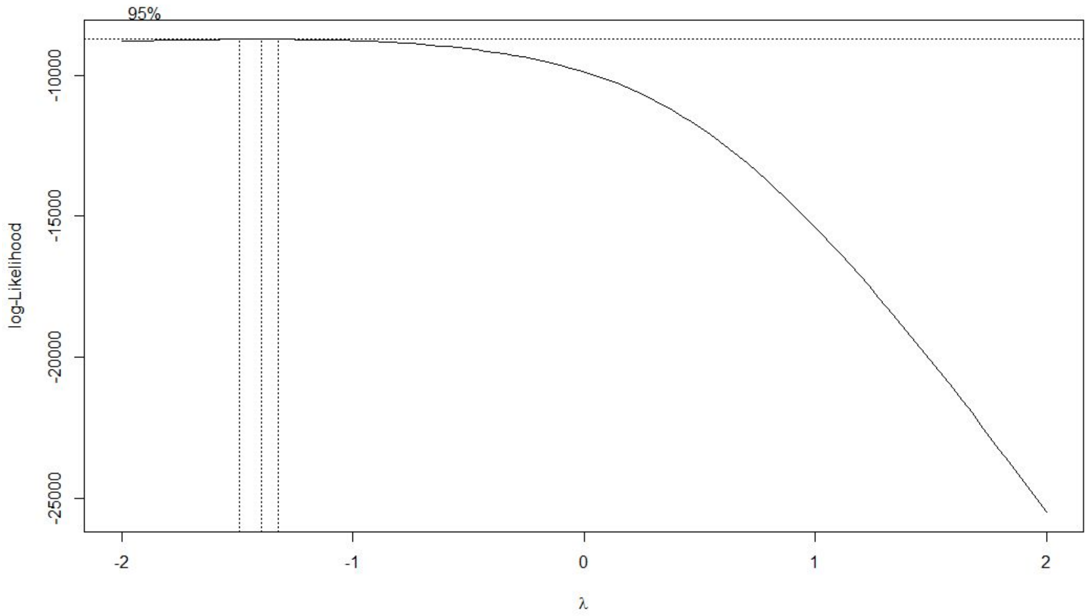

Figure 24: Boxcox plot of linear model 2 before transformation

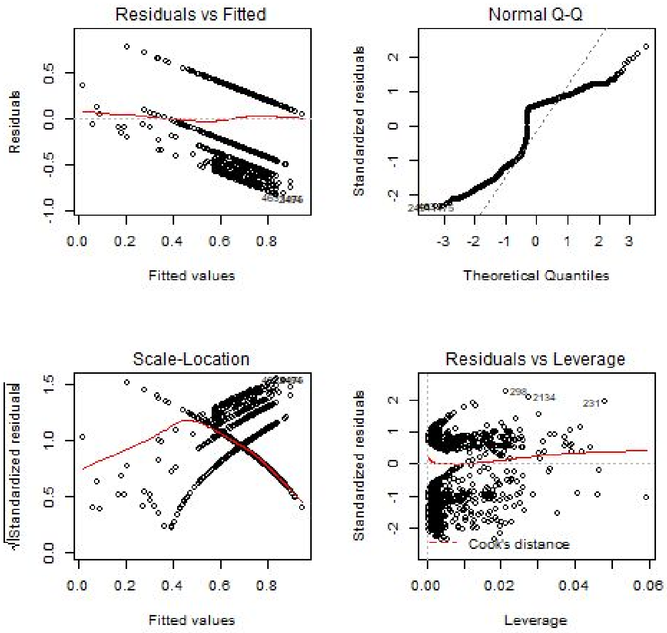

Figure 25: Diagnostic plots of Linear Model 2 - Transformed

Next, some model comparison techniques were carried out to further enforce a model decision. The creation of tests sets was one of the employed techniques. The test sets will be run on the stepwise regression of Linear Model 1, Linear Model 2, and Linear Model 3. In order to evaluate these test sets, the predicted mean square error (PMSE) values will be generated on a graph of 20 different test runs, to account for the values fluctuating with each run. This graph can be seen in **Figure 26**. In the graph, the blue line represents Linear Model 1 - Step, the red line represents Linear Model 2, and the green line represents Linear Model 3. While the three models follow the same general pattern, the blue line typically falls below the two others. This implies that the blue line, or Linear Model 1 - Step, typically produces a lower PMSE compared to the other models. The result of test sets implies that Linear Model 1 - Step is the strongest model, but cross-validation will also be employed to either enforce or refute these conclusions.

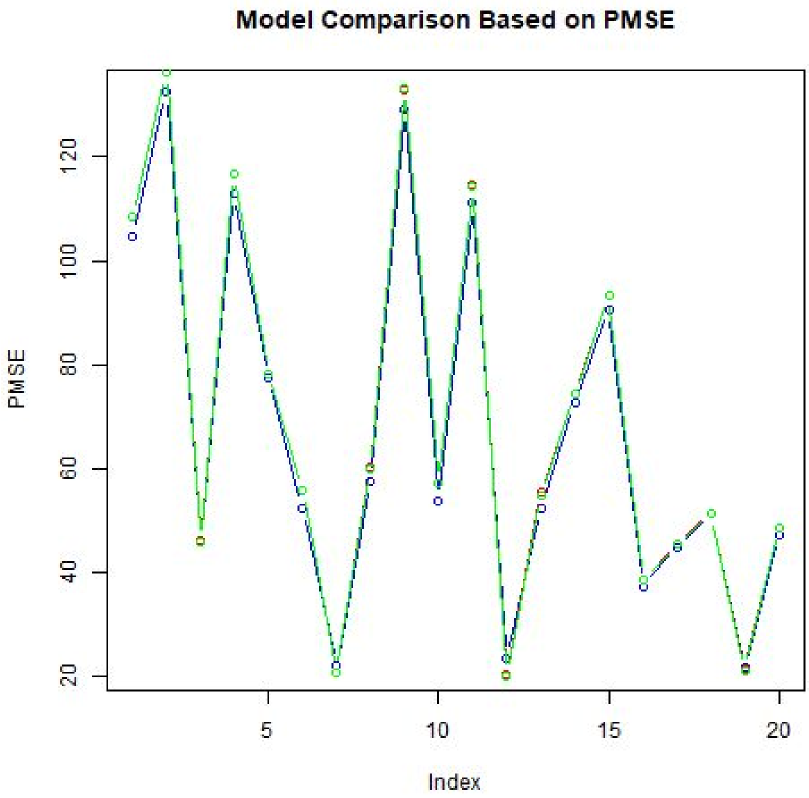

Figure 26: The PMSE over 20 test runs for the test sets of Hypothesis 2

Also, cross-validation can be run on the models to provide further evidence for the analytical decisions. After running cross validation on each model, the predicted mean square error (MSE) for each transformed model can be analyzed. After testing for a few times, it seems that the trends are about the same. **Table 8** below shows an example set of values. It can be seen that Linear Model 1 - Step, has the smallest predicted MSE, followed by Linear Model 3 and then Linear Model 2. Therefore, Linear Model 1 - Step exhibits the best predicted MSE compared to the other two models; based on cross-validation, that would make Linear Model 1 - Step the preferred model among all 3.

Linear Model 1 - Step	Linear Model 2	Linear Model 3
67.51567	69.48035	69.39995

Table 8: Predicted MSE from cross-validation for each transformed model

Therefore, these different analytical techniques provide evidence that Linear Model 1 - Step is superior to Linear Model 2 and 3. Recall Model 1 below from the Analysis section. As it pertains to the objective of factors that impact Casualty, significant variables include: LOADP1, TypeD, HIGHSPD:CauseH, and HIGHSPD:TypeD.

Model 1: Casualty = 1.610663 + 0.300130 LOADP1 + 0.048023*CauseH + 0.003757*HIGHSPD + -2.741850*TypeD + 0.074050*CauseH:HIGHSPD + 0.171774*HIGHSPD:TypeD+* ε

As it pertains to the second subgoal's first objective, it can be stated that some factors that impact the accumulated damage of extreme accidents are: LOADP1, CauseH, HIGHSPD, TYPED, CauseH:HIGHSPD, and HIGHSPD:TYPED. These are variables that appear significant in this model. Specifically shifting to the hypothesis for this section, this model demonstrates for every mile per hour that high speed increases, holding the fact that the accident is of cause H and all else constant, the number of casualties will increase by 0.074. Overall, we are confident in this model as it has a p-value of 2.2*10^-16, which is very close to 0. Therefore, we feel that this information can be used in the recommendation section to address the second objective of subgoal #2.

4 Recommendation

In terms of the ACCDMG investigation, we accepted the hypothesis that high speed derailments cause more monetary damage than high accidents of other types according to our linear regression model. For every 1 unit increase in HIGHSPD:TypeD, there is a -8.431*10^-6 change in ACCDMG^(-0.5). The following recommendations aim to reduce the number of high speed derailments.

- Provide better training for staff to avoid derailments
  - Certain areas of tracks, such as tracks that are curved or high in traffic so that they require frequent stops, requiring different types of conducting (ie. speed changes)
  - In situations that are more likely to result in derailments, such as on track areas where derailments have occurred in the past, train speeds should be reduced
- For the future, consider further research into derailment prevention techniques that can be implemented
  - Technology could be added to the rails and surrounding areas to monitor train speed
- Our final transformed model has a p-value of 2.2*10^-16, as well as an adjusted r-squared value of 0.266, meaning that 26.6% of the variance on the data is

explained by our model(**Table 4**). This is better than the other models we compared against, and the t-value of our hypothesis parameter (TypeD:HIGHSPD) is statistically significant with a p-value around 0. Therefore, we are confident in the accuracy of this mode and therefore our recommendations. More generally, our recommendation aims to change the situation from the top-right point to the bottom-right point in the interaction plot **(Figure 9)**.

In terms of the Casualty investigation, we accepted the hypothesis that trains traveling at high speeds that are of accident cause "H" (human factors) result in significant increases in human casualties as compared to accidents of other causes according to our linear regression model. The following recommendations aim to reduce the number of high speed accidents caused by human factors:

- Better train conductors and engineers, as reducing human error is pertinent to reducing accident severity
- Better monitor staff performance, speed tendencies, and behaviors in order to prevent employees with dangerous habits from causing accidents
    - For every 1 mile per hour increase, in accidents caused by human error, there is an increase in casualties of 0.074 casualties
- Again, enforce strict speed regulations on train staff, especially when there is a high number of passenger cars on a given trip
- In the future as technology advances, consider implementing train automation in order to limit the impact of humans and human error on train performance
- Our model has a p-value of 2.2*10^-16, meaning that our model is extremely significant, despite the adjusted r-squared value being 0.05. This means 5% of the variance in our data is explained by our model, which is far better than the other two models we used to predict Casualty. Therefore, we are confident in this model and the recommendations that we based on its results. Also, the t-value of our hypothesis parameter (HIGHSPD:CauseH) is statistically significant with a p-value near 0. More generally, our recommendation aims to change the situation from the top-right point to the bottom-right point in the interaction plot **(Figure 10)**.

5 References

[1] L. E. Barnes, “Project 1: Train accidents,” *Class project in SYS 4021*, 2017.

[2] L. E. Barnes, “Project 1 template,” *Class template in SYS 4021*, 2017.

[3] F. R. Administration. (2015) Federal railroad administration office of safety analysis. http://safetydata.fra.dot.gov/.

[4] Gardner, W., Mulvey, E.P., and Shaw, E.C (1995). “Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models”, Psychological Bulletin, 118, 392-404. http://www.theanalysisfactor.com/regression-models-for-count-data/

[5] Nobles, A. (2017, Oct. 10). Office Hours.

A Optional Appendices

It was decided not to include appendices for this project.