Car Fatality Analysis on Interstates in United States

Karmanya Pathak

2020-09-06

Car Crash Analysis

Loading Data

Our project aims to analyze all the fatal crashes in US from 2016 -2018. The data is provided by FARS (Fatality Analysis Reporting System) and has extensive data regarding each fatal accident. The project aims to classify these fatal accidents to bring more light on potential causes of accidents that cannot be addressed by current safety standards

Data source: Accidental crash data from NHTSA(National Highway Traffic Safety Administration) Link: https://www.nhtsa.gov/node/97996/221

Jump to Top

Data Cleanup:

Per type 1 indicates the observation is driver, veh no 1 indicates that the car which caused the crash, and the rest of the filters on variables is removing the unkown/outlier data in the data set

Jump to Top

Splitting all the data into Train and Test data

We split the primary data into 75% training data and 25% test data where we would use the training data as a reference to train our model to predict the severity of fatal crashes and compare it with the real test data for accuracy of the model. All things remaining same, a more accurate model can be used to predict the chances and severity of such events happening in the future.

Jump to Top

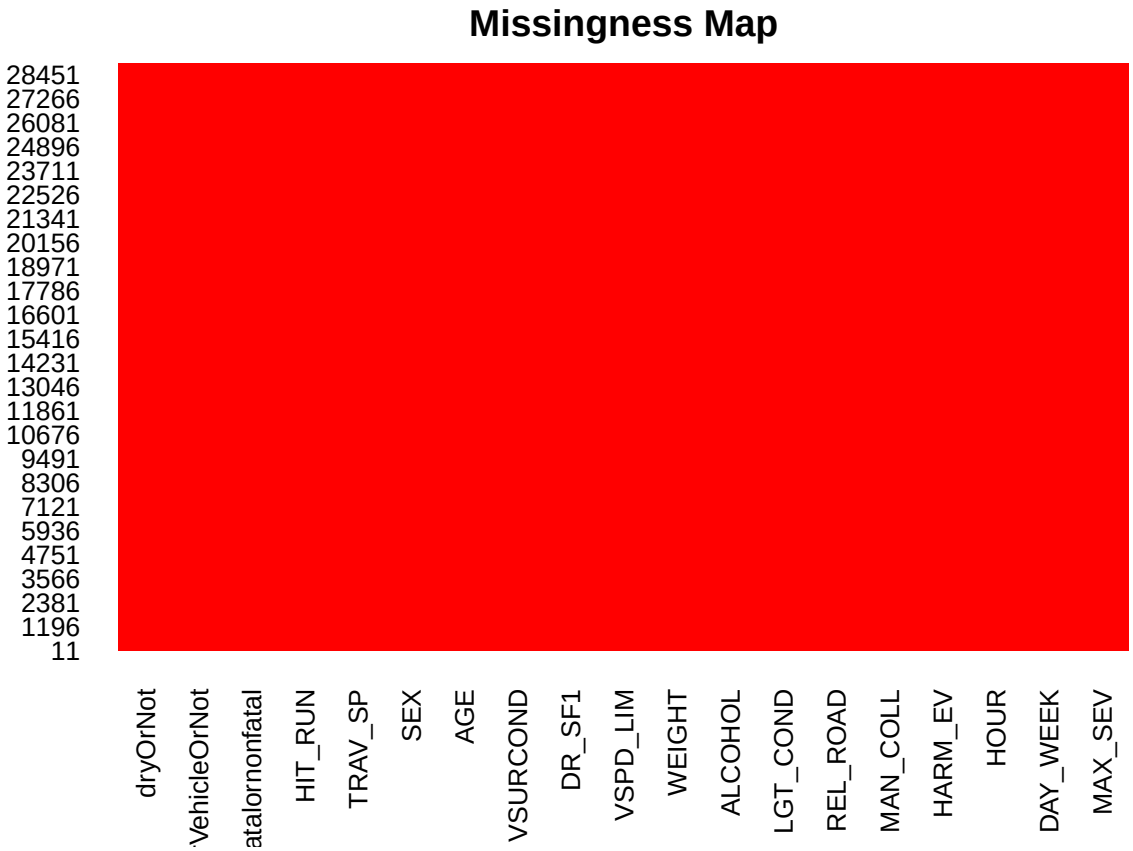
Exploratory Data Analysis

Our initial exploratory analysis aims to understand the structure of the data and the summary.

Checking for Missing data We can see that out of 19 variable we have 3 logical, 1 integer and the rest numerical.

From the chart you can see that there were 0 missing values.

```
## 'data.frame':    38695 obs. of  19 variables:
##  $ MAX_SEV          : int  0 3 2 0 3 0 0 0 0 2 ...
##  $ DAY_WEEK         : int  2 3 2 4 2 5 1 5 6 5 ...
##  $ HOUR             : int  12 6 8 15 17 18 12 12 7 20 ...
##  $ HARM_EV          : int  12 43 8 12 12 12 12 12 14 33 ...
##  $ MAN_COLL         : int  7 0 0 1 7 6 6 1 0 0 ...
##  $ REL_ROAD         : int  1 4 1 1 1 1 1 1 7 4 ...
##  $ LGT_COND         : int  1 3 1 1 1 3 1 1 1 3 ...
##  $ ALCOHOL          : int  2 2 2 2 2 2 2 9 1 1 ...
##  $ WEIGHT           : num  219.8 32.2 37.3 379.9 42.7 ...
##  $ VSPD_LIM         : int  30 50 45 40 55 35 35 35 30 30 ...
##  $ DR_SF1           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ VSURCOND         : int  1 1 1 1 1 1 2 1 1 2 ...
##  $ AGE              : int  72 32 61 63 54 73 23 40 50 20 ...
##  $ SEX              : int  1 2 1 1 1 2 1 2 2 1 ...
##  $ TRAV_SP          : int  20 50 3 5 55 5 10 0 40 40 ...
##  $ HIT_RUN          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ fatalornonfatal  : logi  FALSE TRUE TRUE FALSE TRUE FALSE ...
##  $ motorVehicleOrNot: logi  TRUE FALSE FALSE TRUE TRUE TRUE ...
##  $ dryOrNot         : num  1 1 1 1 1 1 0 1 1 0 ...
```



Furthermore, we explored the relationship between the fatalities of drivers with different factors recorded in the crash report.

From the figure titled 'Fatality by Travel Speed', it is evident from the boxplot that the mean travel speed of the vehicle is much higher in crashes that resulted in fatality.

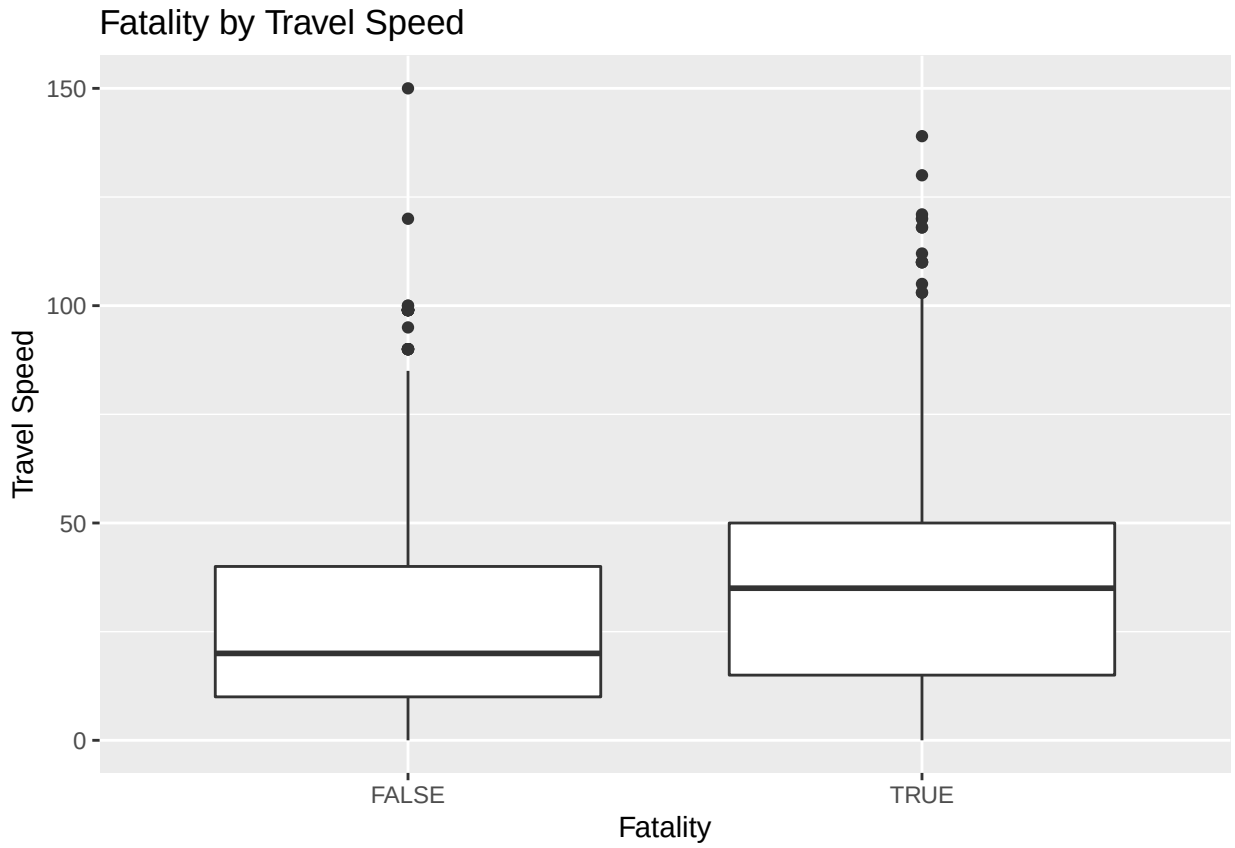
In figure 'Fatality by Days of the Week', we have tried to see how fatalities has ranged in terms of day of the week. The crashes involving fatalities are spread over the entire weekdays whereas non-fatal crashes are more concentrated between Tuesday to Friday.

There is no difference in the fatality of the traveler based on Age.

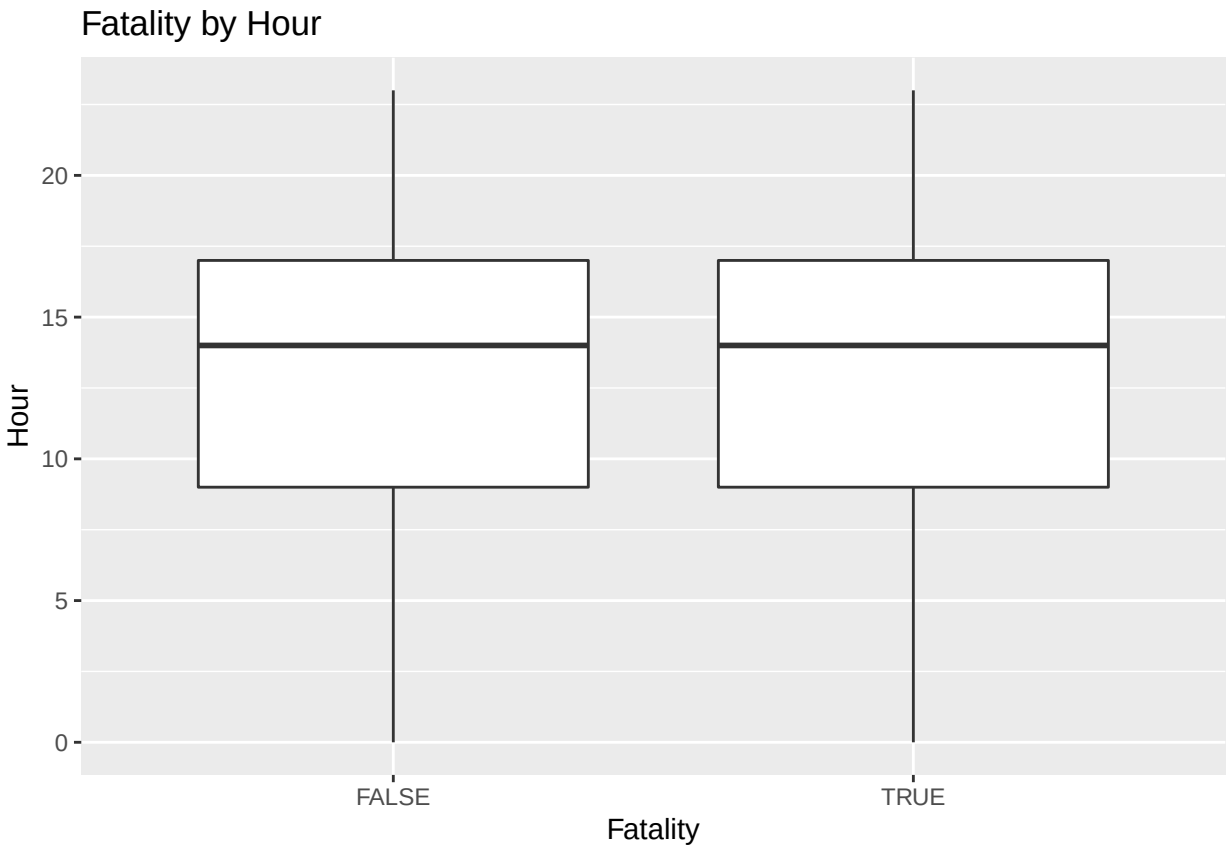
The travel speed for fatal accidents span a wider range as compared to non-fatal accidents.

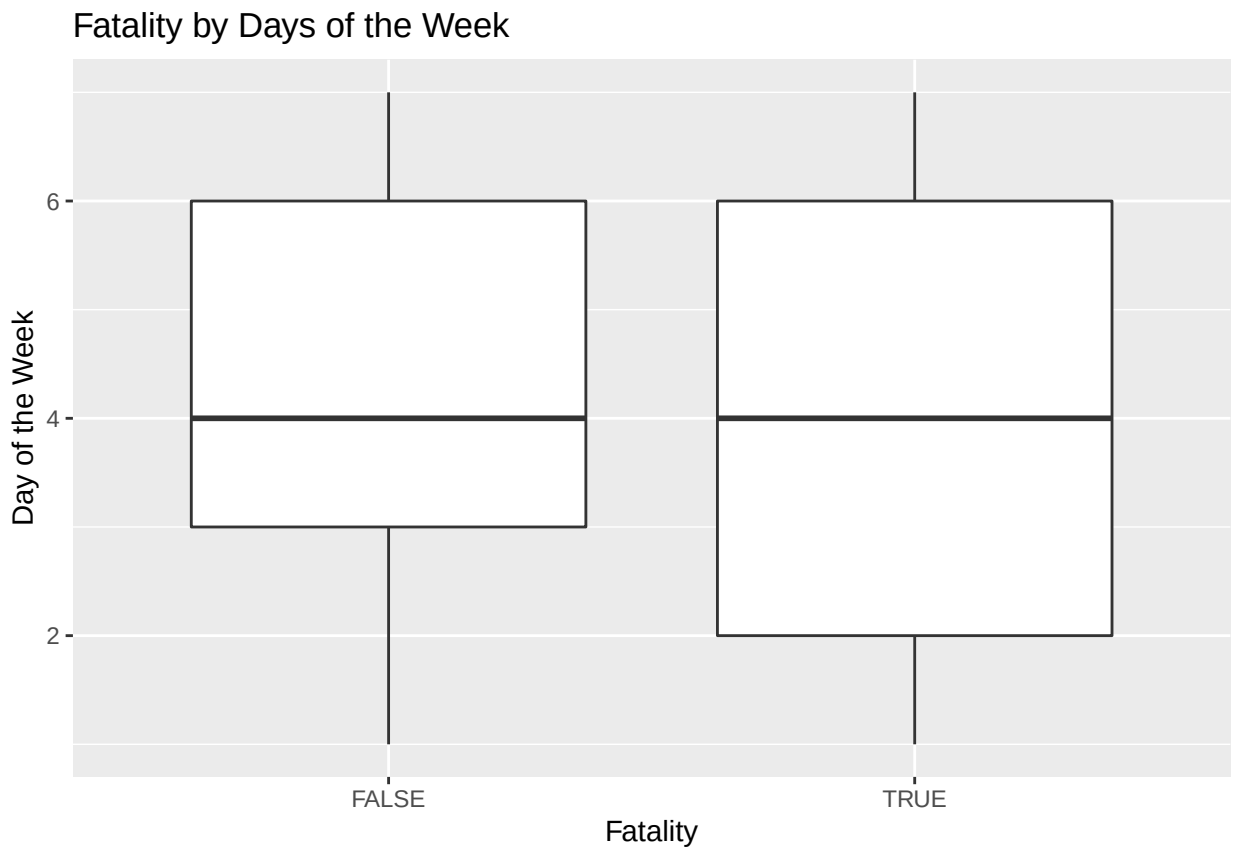
After comparing the road's speed limit with the traveling speed of the vehicle involved in an accident, we can see that there were greater fatal accidents had vehicles traveling a lot higher than the posted speed limits for the road as compared to non-fatal accidents.

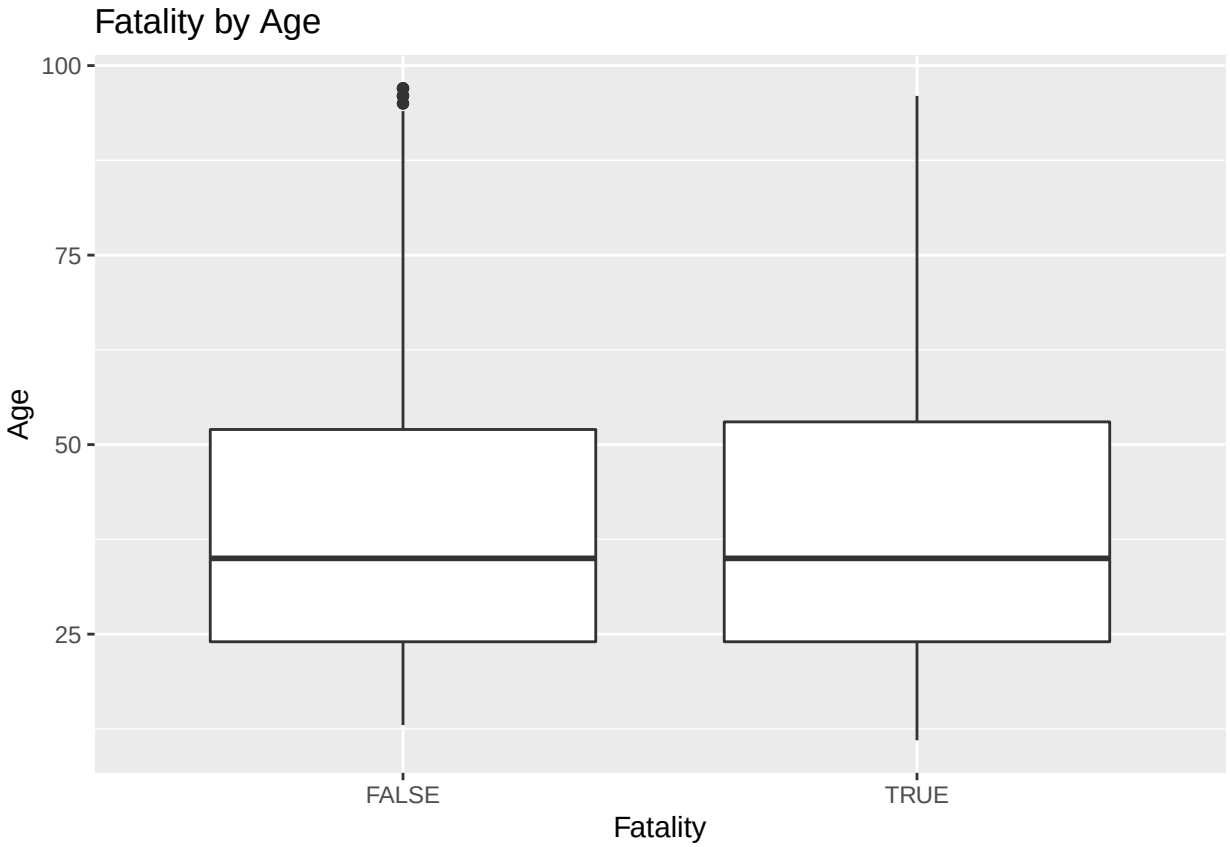
The last figure explores the spread of all crashes based on ages. We can see the major chunk of crashes being attributed to drivers between the age of 16 - 40.

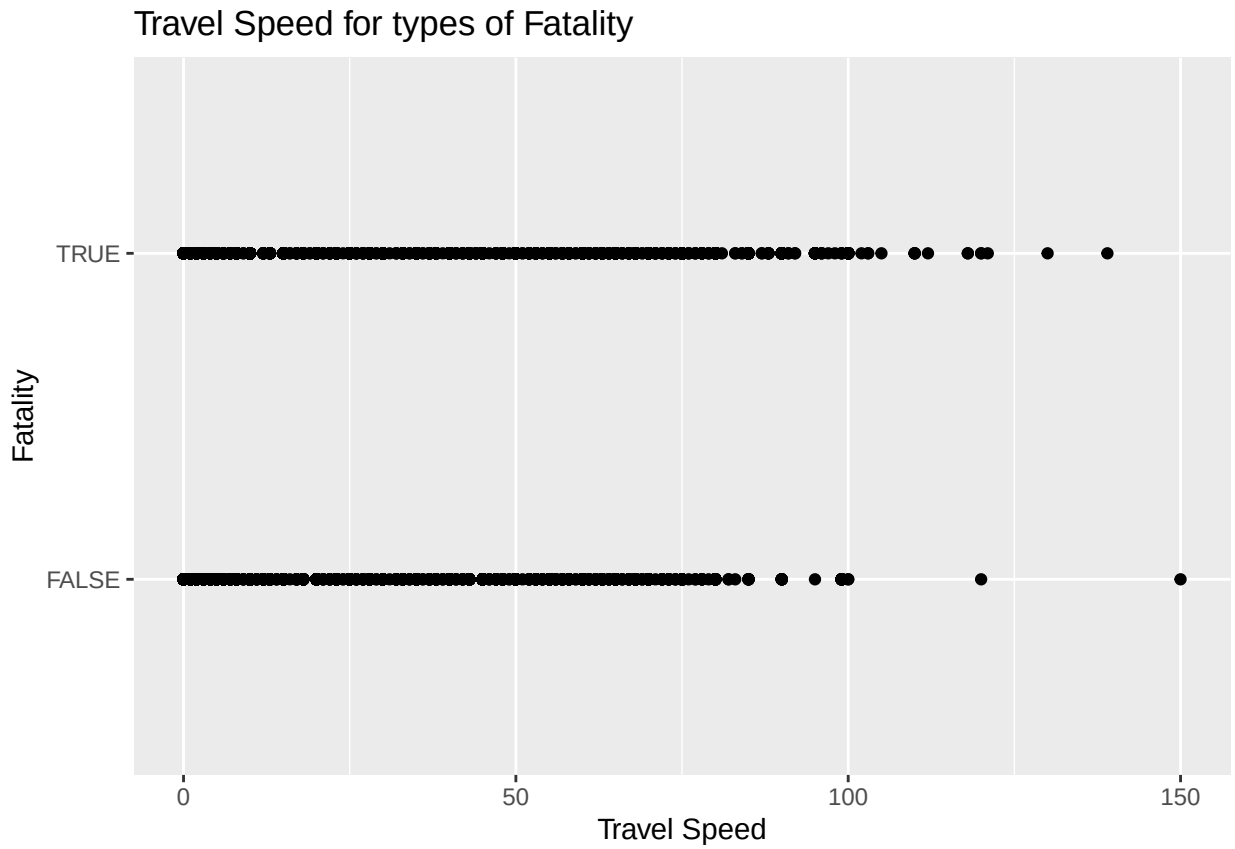


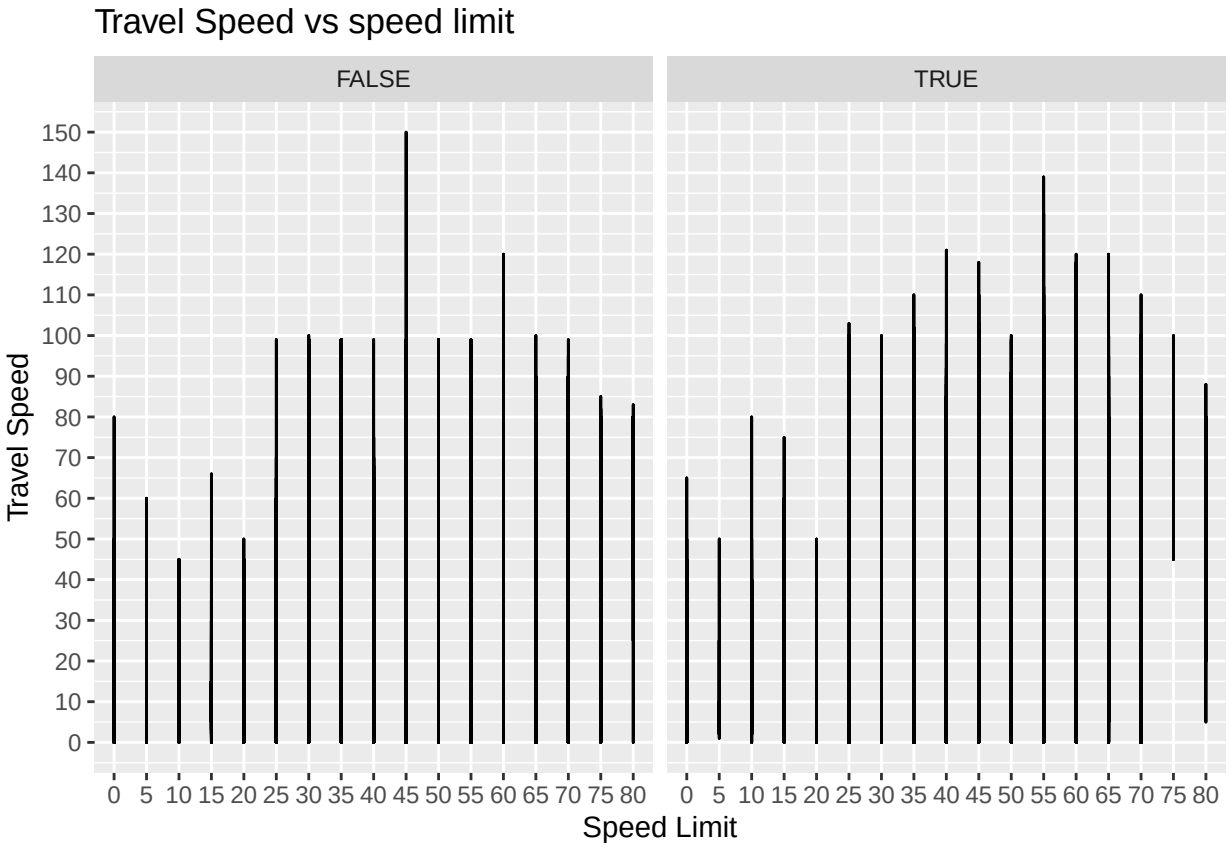

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   20.00   25.25   40.00  120.00

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   15.00   35.00   34.22   50.00  139.00
```

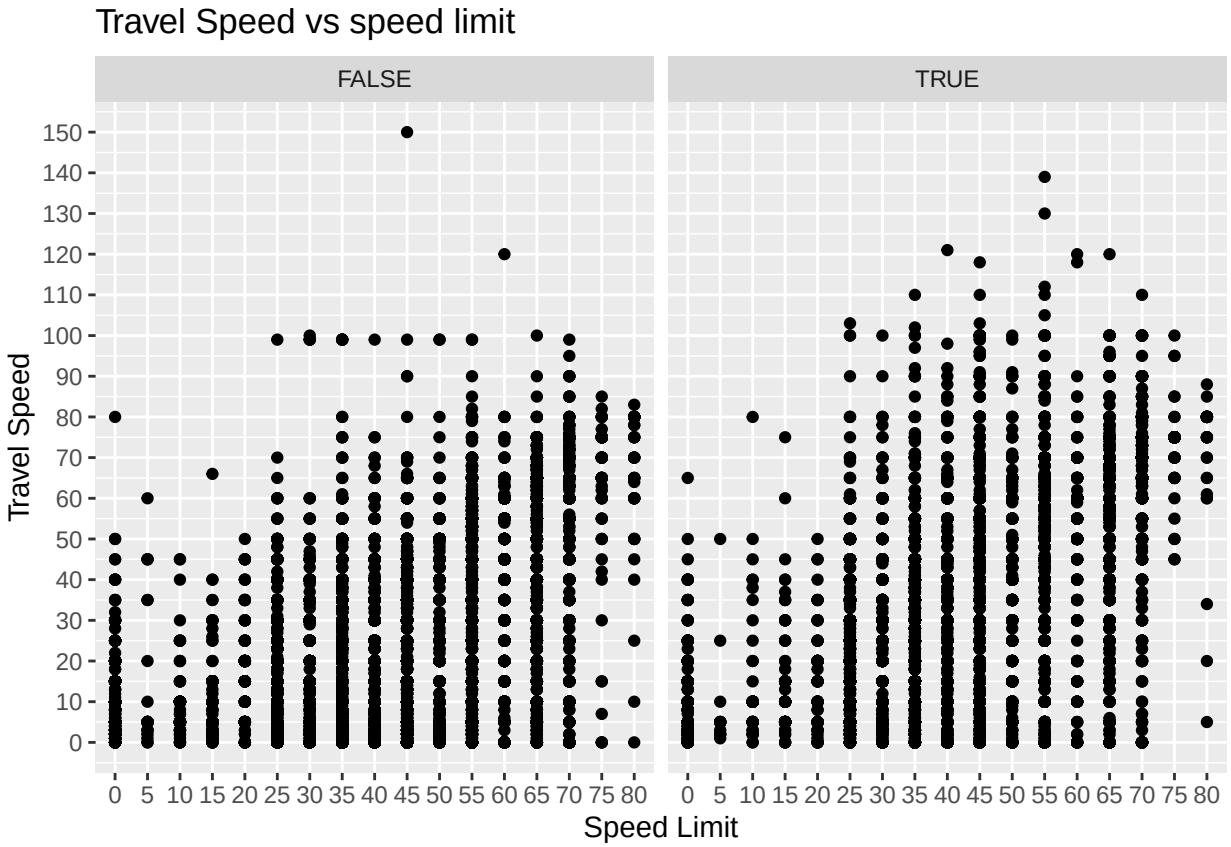

Fatality by Hour
Hour
20
15
10
5
0
FALSE
TRUE
Fatality


Fatality by Days of the Week
Day of the Week
6
4
2
FALSE
TRUE
Fatality


Fatality by Age
100
75
50
25
Age
FALSE
TRUE
Fatality


Travel Speed for types of Fatality
Fatality
TRUE
FALSE
0
50
100
150
Travel Speed

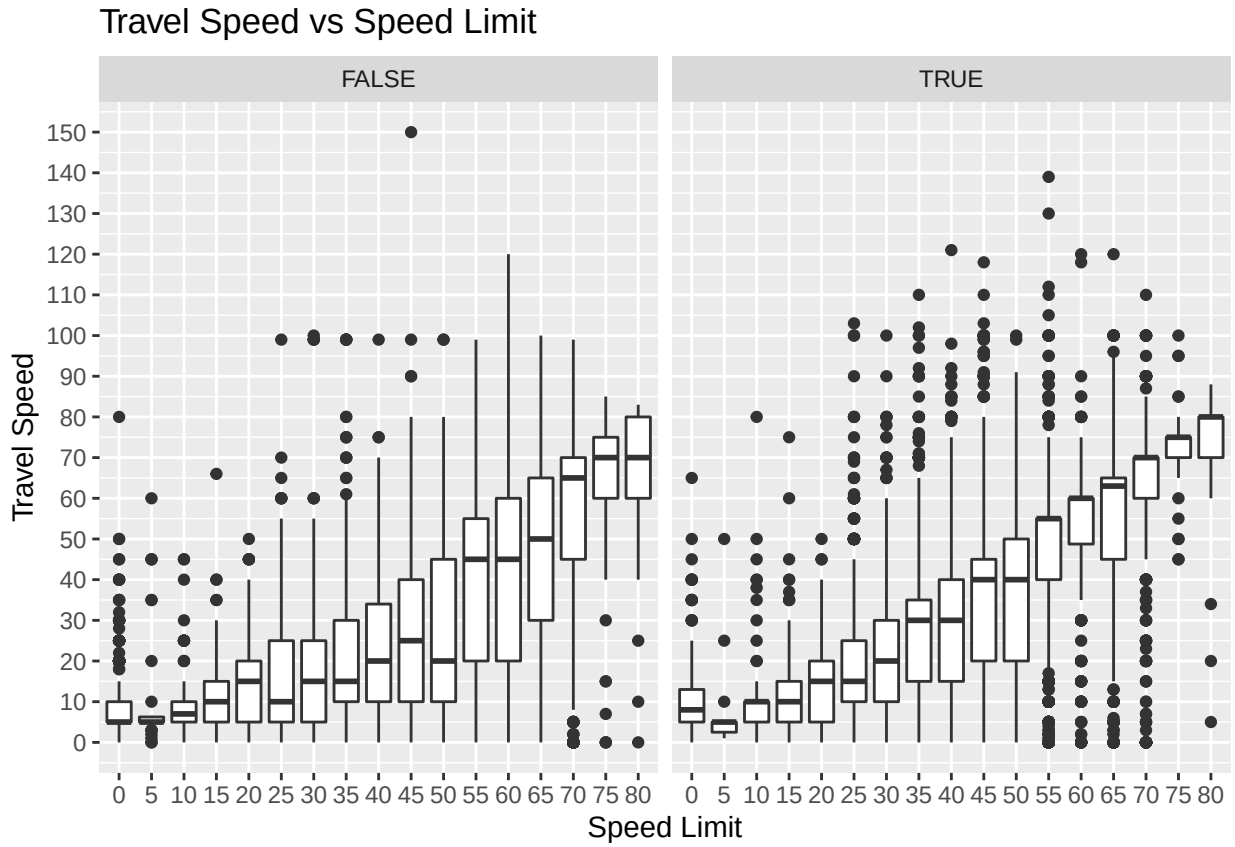

Travel Speed vs speed limit
FALSE
TRUE
Travel Speed
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80
Speed Limit


Travel Speed vs speed limit
FALSE
TRUE
Travel Speed
Speed Limit
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80

Jump to Top

Inference:

Our research involves understanding the factors affecting and leading to fatal and non-fatal crashes. Amongst all the factors in our dataset, we focused on analyzing the effects of the following factors: Travel Speed (TRAV_SP), Alcohol Involved in Crash(ALCOHOL), Speed Limit (VSPD_LIM), Roadway Surface Condition (dryOrNot), collision with moving vehicle(motorVehicleOrNot), Light Condition (LGT_COND), Sex (SEX), Age (AGE). Deep diving into analyzing the Travel Speed and the Speed limit variables, we can observe that for roads with a speed limit of 55 mph, fatal crashes had a mean speed of 47 mph (median value of 55 mph) with a maximum speed of 130 mph, whereas non-fatal crashes had a lower mean speed of over 37 mph (lower median speed of 45 mph) and a maximum speed of 99 mph. This indicates that higher traveling speeds for a road's speed limit have a significant impact on vehicle crashes leading to fatality.



Null Hypothesis : There's no impact on fatality of the driver based on accident attributes such as light condition, alcohol, road condition, Speed limit, hour of accident , day of week, sex of driver, age, collision vehicle type. Alternate Hypothesis: There is a significant impact of accident attributes on fatality of driver.

Jump to Top

Model Fit:

```
##
## Call:
## glm(formula = fatalornonfatal ~ TRAV_SP + ALCOHOL + VSPD_LIM +
##     dryOrNot + motorVehicleOrNot + LGT_COND + SEX + AGE, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8849  -0.9006  -0.7312   1.1691   2.0766
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.7903955  0.0799026  -9.892  < 2e-16 ***
## TRAV_SP                0.0195326  0.0008061  24.230  < 2e-16 ***
## ALCOHOL                0.0065941  0.0069462   0.949    0.342
## VSPD_LIM              -0.0099263  0.0010232  -9.701  < 2e-16 ***
## dryOrNot               0.3909748  0.0373128  10.478  < 2e-16 ***
## motorVehicleOrNotTRUE -0.8439493  0.0290149 -29.087  < 2e-16 ***
```

```
## LGT_COND               0.0770066  0.0160462   4.799 1.59e-06 ***
## SEX                    0.0069534  0.0261949   0.265    0.791
## AGE                    0.0049967  0.0007259   6.884 5.84e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38251  on 29020  degrees of freedom
## Residual deviance: 35739  on 29012  degrees of freedom
## AIC: 35757
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## glm(formula = fatalornonfatal ~ TRAV_SP + VSPD_LIM + dryOrNot +
##     motorVehicleOrNot + LGT_COND + AGE, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8871  -0.9008  -0.7313   1.1689   2.0771
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.7635576  0.0673090 -11.344  < 2e-16 ***
## TRAV_SP                0.0195230  0.0008044  24.270  < 2e-16 ***
## VSPD_LIM              -0.0099296  0.0010232  -9.705  < 2e-16 ***
## dryOrNot               0.3909002  0.0373113  10.477  < 2e-16 ***
## motorVehicleOrNotTRUE -0.8440172  0.0289717 -29.132  < 2e-16 ***
## LGT_COND               0.0768016  0.0160339   4.790 1.67e-06 ***
## AGE                    0.0049924  0.0007254   6.883 5.88e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38251  on 29020  degrees of freedom
## Residual deviance: 35740  on 29014  degrees of freedom
## AIC: 35754
##
## Number of Fisher Scoring iterations: 4
```

Model1. In this model we run regression model on Travel speed, alcohol, speed limit, road condition, collision with a moving vehicle, light condition, sex and age. Fatality ~ Travel speed+alcohol+ speed limit+road condition+collision with a moving vehicle+light condition+sex+age In this model we found the Alcohol and Sex is not a significant factor contributing to the fatality of the driver.(See code for actual results) Model 2: In this model we drop alcohol and sex and run the model again. Fatality ~ Travel speed+speedlimit+road condition+collision with a moving vehicle+light condition+age

Jump to Top

Testing model, CI, exponential coefficients:

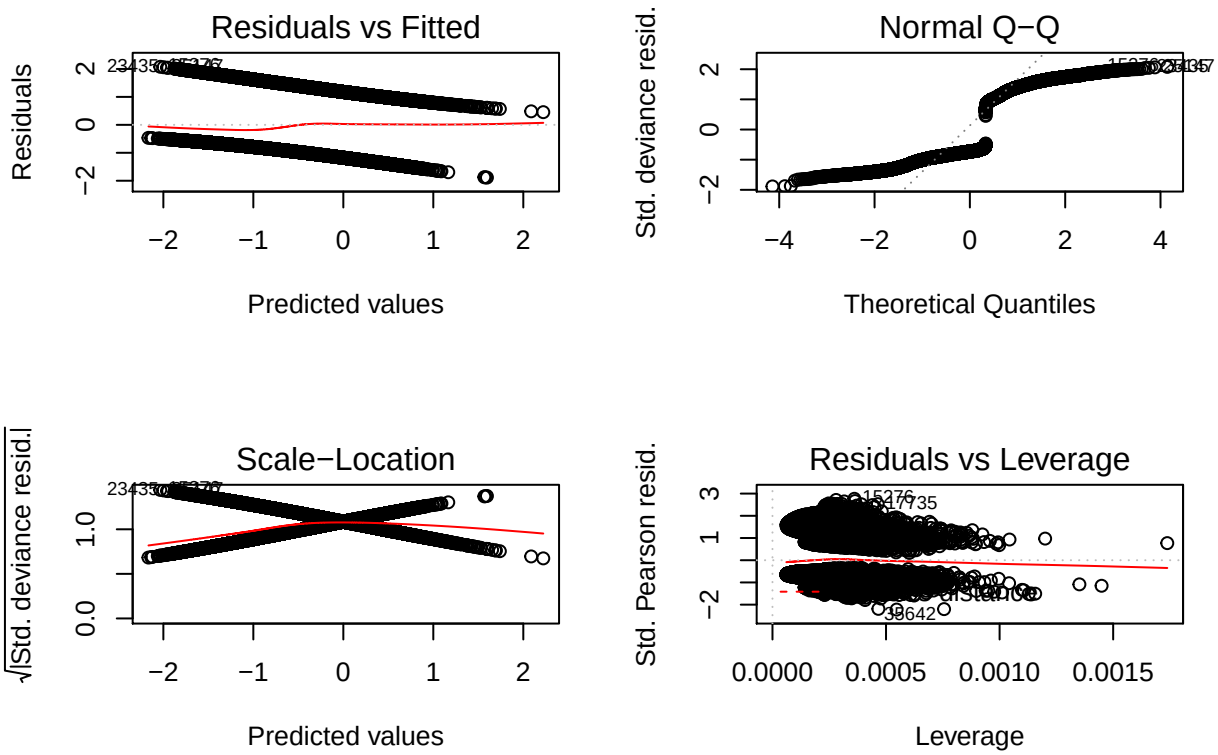
We use Model2 for our prediction. Fatality = -0.763 + 0.019(TRAV_SP) -0.099(VSPD_LIM) +0.390(dry-ornot) – 0.844(motorVehicleornot) + 0.07(LGT_COND) + (0.004)AGE

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: fatalornonfatal
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                29020      38251
## TRAV_SP            1  1231.22     29019      37019 < 2.2e-16 ***
## VSPD_LIM           1   237.44     29018      36782 < 2.2e-16 ***
## dryOrNot           1    60.22     29017      36722 8.470e-15 ***
## motorVehicleOrNot  1   916.72     29016      35805 < 2.2e-16 ***
## LGT_COND           1    17.80     29015      35787 2.456e-05 ***
## AGE                1    47.27     29014      35740 6.197e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: fatalornonfatal
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                29020      38251
## TRAV_SP            1  1231.22     29019      37019 < 2.2e-16 ***
## ALCOHOL            1     2.41     29018      37017    0.1205
## VSPD_LIM           1   237.23     29017      36780 < 2.2e-16 ***
## dryOrNot           1    60.26     29016      36719 8.314e-15 ***
## motorVehicleOrNot  1   915.42     29015      35804 < 2.2e-16 ***
## LGT_COND           1    17.82     29014      35786 2.424e-05 ***
## SEX                1     0.00     29013      35786    0.9910
## AGE                1    47.28     29012      35739 6.158e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## glm(formula = fatalornonfatal ~ TRAV_SP + VSPD_LIM + dryOrNot +
##     motorVehicleOrNot + LGT_COND + AGE, family = binomial, data = train)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.8871  -0.9008  -0.7313   1.1689   2.0771
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.7635576  0.0673090 -11.344  < 2e-16 ***
## TRAV_SP                0.0195230  0.0008044  24.270  < 2e-16 ***
## VSPD_LIM              -0.0099296  0.0010232  -9.705  < 2e-16 ***
## dryOrNot               0.3909002  0.0373113  10.477  < 2e-16 ***
## motorVehicleOrNotTRUE -0.8440172  0.0289717 -29.132  < 2e-16 ***
## LGT_COND               0.0768016  0.0160339   4.790 1.67e-06 ***
## AGE                    0.0049924  0.0007254   6.883 5.88e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38251  on 29020  degrees of freedom
## Residual deviance: 35740  on 29014  degrees of freedom
## AIC: 35754
##
## Number of Fisher Scoring iterations: 4
```




```
##                              2.5 %       97.5 %
## (Intercept)           -0.89560326 -0.631747001
```

```
## TRAV_SP                0.01794933  0.021102678
## VSPD_LIM              -0.01193617 -0.007925281
## dryOrNot               0.31798125  0.464247916
## motorVehicleOrNotTRUE -0.90082795 -0.787258131
## LGT_COND               0.04533648  0.108191251
## AGE                    0.00357038  0.006413860
```

```
##           (Intercept)               TRAV_SP              VSPD_LIM
##             0.4660056             1.0197148             0.9901195
##              dryOrNot motorVehicleOrNotTRUE              LGT_COND
##             1.4783109             0.4299797             1.0798278
##                   AGE
##             1.0050049
```

```
##                            2.5 %    97.5 %
## (Intercept)           0.4083612 0.5316622
## TRAV_SP               1.0181114 1.0213269
## VSPD_LIM              0.9881348 0.9921060
## dryOrNot              1.3743505 1.5908173
## motorVehicleOrNotTRUE 0.4062332 0.4550909
## LGT_COND              1.0463799 1.1142608
## AGE                   1.0035768 1.0064345
```

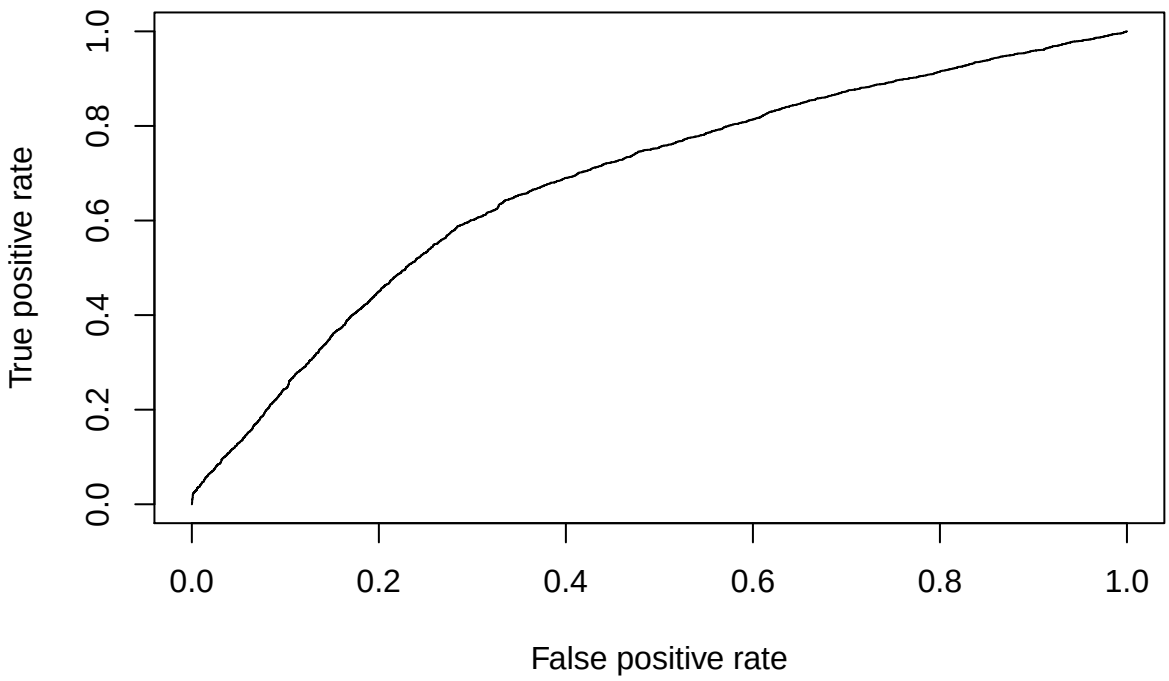
Jump to Top

Prediction

We applied the logistic regression model to the training dataset and tested it against our test dataset. The accuracy of the dataset came to approximately 66%. We further plotted a ROC curve to understand the predictability of the model. We also found the area under the ROC curve which came out to 0.68. An AUC of 1 is ideal and the model should strive to be closer to it.

```
##    MAX_SEV DAY_WEEK HOUR HARM_EV MAN_COLL REL_ROAD LGT_COND ALCOHOL    WEIGHT
## 14       0        6   12      12        1        1        1       2 198.91180
## 15       0        5   22      11        0        1        2       2 219.79548
## 23       2        5   20      12        1        1        2       2 130.17619
## 34       0        4   18      12        7        1        3       2 215.21812
## 36       3        3   13      32        0        3        1       2  31.45521
## 44       0        6   23      12        6        1        3       2 190.36350
##    VSPD_LIM DR_SF1 VSURCOND AGE SEX TRAV_SP HIT_RUN fatalornonfatal
## 14       35      0        3  66   1      10       0           FALSE
## 15       45      0        1  42   1      45       0           FALSE
## 23       55      0        2  20   1      55       0            TRUE
## 34       55      0        1  24   1      40       0           FALSE
## 36       60      0        1  46   1      70       0            TRUE
## 44       25      0        3  21   2      20       0           FALSE
##    motorVehicleOrNot dryOrNot probFatal
## 14              TRUE        0 0.2052857
## 15             FALSE        1 0.6040408
## 23              TRUE        0 0.3043887
## 34              TRUE        1 0.3471344
## 36             FALSE        1 0.6692134
## 44              TRUE        0 0.2441410
```

```
## [1] "Accuracy 0.664564812900558"
```




```
## [1] 0.6830216
```

Jump to Top

Conclusion:

In conclusion based on the Anova test, the p-vales of all the factors in the model are less than the significant level of 5%, we reject the null Hypothesis H0, and accept the alternate hypothesis HA that the factors based on model 2 significantly impact the fatality of the driver involved in the crash. We chose logistic regression to predict the fatality in a crash based on various factors. The predictability came to approximately 66%. This is not a very high accuracy rate but still significant. In cases of crash, it is preferable to have more false positives than false negatives. More factors or different modeling techniques are required to increase the accuracy of prediction. One of the surprising results of this analysis was that alcohol did not have any significant impact on the fatality of the crash. Although, the number of cases with alcohol were comparatively much lower.

Jump to Top

References

1. Preusser, D. F., Williams, A. F., & Ulmer, R. G. (2000, January 27). Analysis of fatal motorcycle crashes: crash typing. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/0001457595000275

2. Yasmin, S., Eluru, N., & Pinjari, A. R. (2015, November). Pooling data from fatality analysis reporting system (FARS) and generalized estimates system (GES) to explore the continuum of injury severity spectrum. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26342892