All Roads Lead to Data: Examining Factors in Traffic-Related Fatalities

Michael Pearson
Jessica Nguyen
Adrian Fletcher

Contents

```
knitr::opts_chunk$set(echo = TRUE, fig.width = 7, fig.height = 4)
if(!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(dplyr, ggplot2, magrittr, gridExtra, reshape, rmarkdown, leaps, glmnet, knitr, pROC, resh
```

Executive Summary

Traffic accidents are a leading cause of fatalities. The goal of our project is to understand which factors lead to fatalities in order to create recommendations for legislation that can reduce traffic-related deaths. The audience of this project is U.S. state legislators. We used data from the U.S. National Highway Traffic Safety Administration (NHTSA) to determine the leading predictors of fatalities. Our data is from 2016 and consists of a random sample of ~44,000 accidents. We include predictors such as drivers' and vehicles' characteristics, weather conditions, use of safety equipment, and types of roads.

We build two models: 1) Multiple Logistic Regression with cross-validation and 2) Random Forest. In our original data set, 1.94% of observations consisted of fatalities. To achieve a balanced dataset, we used bootstrapping to oversample the minority class and undersample the majority class. Our final data set has 50% fatalities

Both of our models mostly agreed on what the most important predictors are. Among our most important findings were that the level of lighting is an important predictor of fatalities. In general, lower levels of lighting are associated with a higher probability of fatalities. One potential recommendation is to increase the level of lighting by adding more traffic lights. Another one of our key findings were that high speeds were associated with an increase in the probability of fatalities. One recommendation would be to increase the use of techniques that enforce speed limits (e.g., automated cameras that monitor speed, more patrol officers on the road). Finally, our models also indicated that a fire increases the probability of fatalities. Therefore, we recommend that legislation require every car to carry a fire extinguisher or to manufacture cars that use more fire-retardant materials. However, both of our models performed very well out-of-sample. We expect that the NHTSA also has a model that performs very well. Yet, while we believe that we know very well what predictors contribute to traffic fatalities, fatalities still occur because many predictors are uncontrollable and cannot be acted upon. For example, from our model, precipitation is one of the most important predictors to fatalities but it is impossible for us to control the weather or enact legislation that would prevent people from traveling during bad weather.

While our model performs well, we still have many limitations. For the sake of parsimony, we excluded pedestrian data, but recognize that this limits our overall findings and suggestions in this research paper. Additionally, we would like to have more data from the NHTSA on cases with fatalities. This would allow us to forego bootstrapping and analyze non-artificial data. Finally, another limitation to our methodology is that by using LASSO, we assume that sparsity exists in the data (and that a few predictors are actually important)

Background

Traffic related deaths are the leading cause of fatality around the world for those aged 5-29. The World Health Organization estimates that 1.25 million people are killed each year from traffic related incidents with an additional 20-40 million people injured. Specifically for the U.S.[1], they estimate that there are "12.4 deaths per 100,000 people - or about 50% higher than similar nations (in Western Europe, Canada, Australia and Japan)." This is important for state legislators to consider when improving traffic safety laws.

Data

Our data comes from the (National Highway Traffic Safety Administration) NHTSA for the year 2016. Our data is publicly available online (https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system-crss). The data consists of 46,511 accidents that have been randomly sampled by NHTSA from 60 areas in the U.S. that are representative of the U.S. population (see map below). This original data set is called `ACCIDENT.csv`. We have merged this data set with select variables from other data sets that are part of the same series and are also available through the NHTSA. We have done extensive data pre-processing (described in more detail below).

[1] https://usa.streetsblog.org/2018/12/13/why-the-u-s-trails-the-developed-world-on-traffic-deaths/

Response Variable

We have coded our response variable, `fatal` as binary. where,

$$fatal = \begin{cases} 1 & \text{if at least one fatality in accident} \\ 0 & \text{if no fatalities} \end{cases}$$

`fatal` is derived from a variable from the original data set called `MAXSEV_IM`. This variable describes the maximum severity within an accident (ranging from minor injuries to deaths). We coded any accident that `MAXSEV_IM` = 4 (the level representing fatality) as `fatal` = 1.

We also removed from the original data variables that were too similar or derived from `MAXSEV_IM`. These variables do not add predictive power and can make the regressions we do too unstable. These variables are `NUM_INJ` (represents number of people injuried in an accident) AND `NO_INJ_IM` (the same as `NUM_INJ` but values have been imputed for observations missing this value).

As mentioned, the creators of this data set have imputed in values that were not reported. Therefore, we also remove `MAX_SEV` which is the unimputed version of `MAXSEV_IM`.

Data Pre-Processing and Predictor Variables

We have done the following things to the original dataset to process it:
1. Removed variables for which there was an imputed variable equivalent (`MINUTE`, `YEAR`, `DAY_WEEK`, `HOUR`, `RELJCT1`, `RELJCT2`, `HARM_EV`, `LGT_COND`, `WEATHER`, `ALCOHOL`, `MAN_COLL`, `WEATHER1`, `WEATHER2`) from the `ACCIDENT.csv` data set (our main data set).
2. Made the following variables from the `VEHICLE.csv`, `DISTRACT.csv`, `VISION.csv`, and `PERSON.csv` datasets.

The following are our predictor variables that we created:

Variable	Description
`hit_run_indicator`	Indicates whether there was a hit-and-run. It equals 1 if there was a hit-and-run and 0 otherwise. This was derived from the `HITRUN_IM` variable in the `VEHICLE.csv` data set. Since this was an imputed variable, there are no missing data for this variable.
`max_travel_speed`	Numeric variable representing the speed traveled by the fastest vehicle in the accident. Derived from the `TRAV_SP` variable in the original `VEHICLE.csv` data set representing the speed traveled by each car in the accident.
`rollover_indicator`	Indicates whether there was a car that rolled over. It equals '1' if there was a rollover and '0' otherwise. This was derived from the `ROLLOVER` variable in the `VEHICLE.csv` data set.
`max_deformity`	A categorical variable representing the maximum damage to any vehicle in the accident. It equals '0' if "No Damage", '2' for "Minor Damage", '4' for "Functional Damage", '6' for "Disabling Damage", '8' for "Unreported", and '9' for "Unknown." This variable is derived from the `DEFORMED` variable in the original `VEHICLE.csv` data set which had the same levels but for each individual car.
`speed_related_accident`	Indicates whether this accident was related to speed. It equals 1 if yes and 0 if no. This was derived from the `SPEEDREL` variable in the `VEHICLE.csv` data set. Missing values are indicated as *NA*.
`max_nlanes`	Numeric variable indicating the maximum number of lanes a vehicle in the accident was traveling on just prior to the accident, derived from `VNUM_LAN`in the original `VEHICLE.csv` data set.

Variable	Description
max_speed_limit	Numerical variable indicating the maximum speed limit of any road involved in the accident. Derived from the VSPD_LIM variable in the original VEHICLE.csv data set that is the maximum speed limit on a road for each vehicle in the data set. NOTE: Two vehicles might have been driving on roads with two different speed limits.
any_wet_road	Indicator variable of whether there were wet roads during the accident (1 = Yes, 0 = Otherwise), derived from VSURCONDin the originalVEHICLE.csv dataset.
any_gradient	indicator variable of whether any vehicle was traveling on a road with a gradient during the accident (1 = yes, 0 = otherwise), derived from VPROFILE variable in the original VEHICLE.csv dataset.
any_maneuver	Indicator variable of whether any vehicle performed a maneuver in the accident (1 = yes, 0 = Otherwise, NA = missing), derived from P_CRASH3 in VEHICLE.csv.
any_distracted	Indicator variable of whether any driver involved in the accident was distracted (1 = yes, 0 = no), derived from MDRDSTRD variable from DISTRACT.csv.
any_vision_obstructed	Indicator variable of whether any driver involved in the accident had their vision obstructed (1 = yes, 0 = no), derived form MVISOBSC variable from VISION.csv.
fire_indicator	Indicates whether there was a fire. It equals 1 if there was a fire and 0 otherwise. This was derived from the FIRE_EXP variable in the VEHICLE.csv data set.
rest_mis_indicator	Indicator variable of whether there was any misuse of restraint systems (1 = yes, 0 = no), derived from the REST_MIS variable from the PERSON.csv data set
air_bag_indicator	Indicator variable of whether an air bag failed to deploy (1 = failed, 0 = successfully deployed), derived from the AIR_BAG variable from the PERSON.csv data set.
ejection_indicator	Indicator variable of whether any person involved in the accident was ejected from their set (1 = yes, 0 = no), derived from the EJECTION variable from the PERSON.csv dataset.

Missing Data

Thankfully, our missing data has mostly been imputed by the creators of this data set. Check the CRSS handbook for this data set on more information about the imputation process. Whatever data was not imputed is a categorical variable. Therefore, for some variables with missing data, there is an additional category called "Other".

Dealing with Unbalanced Data Set with ROSE

In the original data set, we have 903 accidents that include at least one fatality. There are 45,608 observations that do not have fatalities. Having an inbalanced data set (1.94% cases with fatalities) prevent us from having accurate and stable results from our model. To solve this problem, we will bootstrap the observations in the minority class (accidents with fatalities) and add them to our data set. We will also randomly undersample from the majority class. One limitation to this method is that it will yield an inaccurate model if our original data set is not representative of the population. However, we have no reason to believe this as the NHTSA has randomly selected this sample. We will use the package ROSE. After bootstrapping, the minority class represents 50% of the data set.

Exploratory Data Analysis for Predictor Variables

Distributions of Predictor Variables

All variables are defaulted to "none" unless otherwise stated. We found a few variables of interest, namely

- `Hit and Runs`
- `Region`
- `Ejection`
- `Number of Lanes`
- `Month`

The rest of our EDA results are in the Appendix.

Hit-and-Runs

Hit and runs returned a surprising result. It appears that most traffic-related fatalities in our data set did not result from hit and runs. On second-thought, it makes sense that if fatalities are involved, people may be less likely to escape the accident and instead try to stay and help those who were severely injured

Region

Concerning region, our data is split into 4 regions:
1. Northeast *(PA, NJ, NY, NH, VT, RI, MA, ME, CT)*
2. Midwest *(OH, IN, IL, MI, WI, MN, ND, SD, NE, IA, MO, KS)*
3. South *(MD, DE, DC, WV, VA, KY, TN, NC, SC, GA, FL, AL, MS, LA, AR, OK, TX)*
4. West *(MT, ID, WA, OR, CA, NV, NM, AZ, UT, CO, WY, AK, HI)*

Region 1 and 4 have the largest proportion of traffic-related accidents. They have about 60-70% traffic related accidents, although we find that the majority of the data comes from region 3. This has to be taken into account with out future data.

Ejection

We find that most of the people that are ejected from their cars (`ejection indicator = 1`) result in fatalities. This makes sense and is unsurprising.

Number of Lanes

We originally hypothesized that a large number of lanes might result in higher fatalities because more lanes might mean more traffic and opportunities to get into crashes. Surprisingly, there is large proportion of accidents that occur in 2 lanes. However, the amount of 2-lane roads are likely most common.

Month

July is the highest month for fatalities. This could be attributed to the fact that there are more seasonal/outdoor events (e.g., Independence Day) occurring during this time that necessitate more cars on the road..

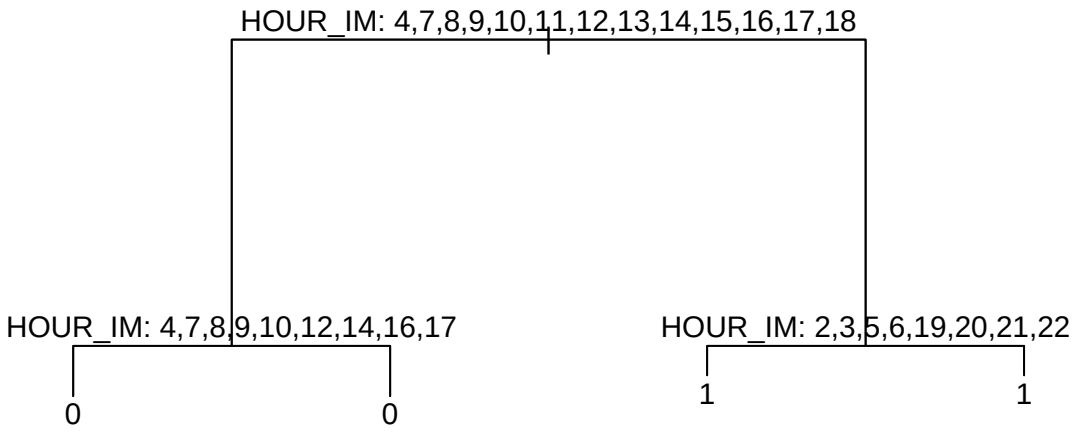
Collapsing Categorical Levels

Since we have many categorical variables with a total of more than 200 levels, we can quickly run into the problem of multiplicity (making every predictor appear significant despite should having about 5% false positives with alpha level 0.05). In order to determine which levels to collapse, we ran single decision trees

with just the categorical variable vs. the response variable (`fatal`). Here, we just show the tree for the `HOUR_IM` variable. However, we the rest of the trees are in Appendix 2.

HOUR_IM

For example, this is our decision tree for the `HOUR_IM` variable which represents what hour of the day the accident occurred. From this decision tree, we grouped observations where accidents occurred in the 4, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18 hours together into one new variable called `hour_night`. `hour_night` is equal to '0' for these observations because these hours occur during the day time. The other hours (0, 1, 2, 3, 5, 6, 19, 20, 21, 22, 23) occur during the night time and are coded as `hour_night` = 1.



max_deformity

We collapsed `max_deformity` according to the tree below. `max_deformity_group1` is the new variable which equals 1 if `max_deformity` is '2', '4', '8', or '9'. `max_deformity` is '0' if there is 'no damage', '2' if there is 'minor damage', '4' if ther eis 'functional dmaage', '6' if there is 'disabling damage', '8' if it is 'not reported' and '9' if it us 'unknown'.

MONTH

Following our tree, we collapsed the `MONTH` variable so that levels 1, 2, 3, 4, 5, 9, 11 were categorized as `month_group1` to 1. Note that these months are the winter months. The other months were set to `month_group1` = 0.

TYP_INT

`TYP_INT` is a variable representing the type of intersection that the traffic occurred at. Following our tree, we grouped observations with `TYP_INT`equal to 2, 4, 6, 10 into one category such that `typ_int_group1` = 0. All other observations were assigned `type_int_group1` = 0.

WRK_ZONE

WRK_ZONE is a variable that represents whether the accident occurred in a work zone or not. Additionally, it specifies what type of work zone the accident occurred in. WRK_ZONE is 0 for no work zone, 1 for construction, 2 for maintenance, 3 for utility, 4 for unknown workzone. However, our tree indicated that it is a singlenode tree. This means that there is not much variability in this variable. Therefore, we remove it from our data.

REL_ROAD

REL_ROAD is a categorical variable with the following levels (1 = On Roadway, 2 = On Shoulder, 3 = On Median, 4 = On Roadside, 5 = Outside Trafficway, 6 = Off Roadway - Location Unknown, 7 = In Parking Lane/Zone, 8 = Gore, 10 = Separator, 11 = Continuous Left Turn Lane). Following our tree, we set observations with REL_ROAD equal to 1, 2, 5, 6, 7, 8, 10 to have rel_road_group1 equal to 1. The other categories were set rel_road_group2 equal to 0.

CF1

CF1 is a categorical variable representing events that happened during the crash. It equals 0 if none of the following events occurred, 3 if road was experiencing maintenance or construction-created condition, 5 if surface under water, 7 if surface washed out (caved in, road slippage), 13 if aggressive driving/road rage by non-contact vehicle driver, 14 if motor vehicle struck by failing cargo or something that came loose from or something that was set in motion by a vehicle, 15 if non-occupant struck by falling cargo or something came loose from or something that was set in motion by a vehicle, 16 is non-occupant struck vehicle, 17 is vehicle set in motion by non-driver, 19 is recent previous crash scene nearby, 20 is police-pursuit-involved, 21 is within designated school zone, 23 is indication of a stalled/disabled vehicle, 24 is unstabilized situation began and all harmful events occurred off the roadway, 25 is toll booth/plaza related, 26 is backup due to prior non-recurring incident, 27 is backup due to prior crash, 28 is backup due to regulation congestion, and 99 is unknown.

From our tree, we set observations with CF1 equal to 0, 5 7, 15, 16, 21, 25, 26, 27, 28 as cf1_group1 as 1 and the other levels as 0.

CF2 and CF3

These variables have the same definition as for CF1. However, they are the second and third most important events that occur (the first is recorded by CF1). After doing trees, we found that these variables yield single-node trees. This gives us reason to believe that these predictors do not have variability and may not be significant. Therefore, we remove them from our analysis.

WKDY_IM

WKDY_IM is a categorical variable with levels 1 - 7. It represents the day of the week that the accident occurred. Sunday is coded as WKDY_IM = '1'. Monday is coded as '2' and so on. Our single tree has grouped levels '2', '3', '4', '5', and '6'. These are actually weekdays (Monday through Friday). We have coded this category as wkdy = '1'.

MANCOL_IM

MANCOL_IM represents the manner of collision. It is defined as '0' is "Not collision with Motor Vehicle in Transport", '1' is 'Front-to_Rear', 2 is 'Front-to-Front', '6' is "Angle", '7' is "Sideswipe, Same Direction", '8' is "Sideswipe, Opposite Direction", '9' is "Rear-to-Side", '10' is "Rear-to-Rear", '11' is "Other." Our tree has sorted levels '1', '7', '8', '9', '10', and '11' into one group. We will sent this group to have the variable mancol_group1 = '1'. The other levels will have this variable equal to '0'.

RELJCT2_IM

RELJCT2_IM is a variable that describes the type of junction the accident was located on. It is equal to '1' if "Non-Junction", '2' if 'Intersection', '3' if 'Intersection Related' '4' if 'Driveway Access', '5' if 'Entrance/Exit

Ramp Related', '6' if 'Railway Grade Crossing', '7' if 'Crossover Related', '8' if "Driveway Access Related", '16' if "Shared_uSe Path Crossing", '17' if "Acceleration/Deceleration Lane", '18' if "Through Roadway", '19' if 'Other Location Within Interchange Area', and '20' if "Entrance/Exit Ramp". Out tree grouped levels '3', '4', '5', '6', '7', '8', '17', '18', '19', and '20' together. We have given these groups `reljct2_group1` = '1'. The other levels had this variable set to '0'.

WEATHR_IM

`WEATHR_IM` describes the type of weather during the accident. It is equal to '0' if 'No Additional Atmospheric Conditions', '1' if "clear", '2' if 'Rain', '3' if 'Sleet or Hail', '4' if 'Snow', '5' if "Fog, Smog, Smoke", '6' if Severe Crosswinds', '7' if 'Blowing Sand, Soil, Dirt, '8' if Other', '10' if 'Cloudy', '11' if 'Blowing Snow', '12' if 'Freezing Rain or Drizzle'. Our tree has grouped levels '2', '3', '4', '5', '8', '11', and '12' together. These are all precipitations. The other level does not include precipitation. We have set the precipitations levels to have `precipitation` = '1'. The other levels have 'precipitation' = '0'.

LGTCON_IM

`LGTCON_IM` represents the type/level of lighting that was present during the time of the accident. Level '1' is 'Daylight, '2' is Dark-Not Lighted, '3' is 'Dark-Lighted', '4' is 'Dawn', '5' is 'Dusk', '6' is 'Dark-Unknown Lighting', '7' is 'Other', '8' is Not Reported', and '9' is 'Unknown.' Our tree has grouped levels '1', '4', and '6' together. It is unclear why these levels have been grouped together. We have set 'lgtcon_group1' to equal '1' for levels in this group and '0' for all other levels.

REGION

`REGION` is a four-level categorical variable. It equals '1' if the crash was in the Northeast U.S. (states of PA, NJ, NY, NJ, VT, RI, MA, ME, and CT). It equals '2' for the Midwest (OH, IN, IL, MI, VI, MN, ND, SD, NE, IA, MO, and KS). It equals '3' for the South (MD, DE, DC, WV, VA, KY, TN, NC, SC, GA, FL, AL, MS, LA, AR, OK, and TX). It equals '4' for the West (MT, ID, WA, OR, CA, NV, NM, AZ, UT, CO, WY, AK, and HI).

Our tree has grouped '1', '3' and '4' together. Therefore we set the variable `midwest` = '1' if `REGION` equals '2' and '0' otherwise.

After performing all of our EDA and data-processing, this is a summary of our data.

```
summary(data_test_removed2)
```

```
##  rest_mis_indicator air_bag_indicator ejection_indicator fire_indicator
##  0:45804            0: 8746           0:43332            0:44893
##  1:  707            1:37765           1: 3179            1: 1618
##
##
##
##
##  any_vision_obstructed any_distracted any_maneuver any_gradient
##  0:44100               0:38862        0:25564      0:39327
##  1: 2411               1: 7649        1:20947      1: 7184
##
##
##
##
##  any_wet_road max_speed_limit   max_nlanes    speed_related_accident
##  0:40580      Min.   : 5.00   Min.   :0.000   0:40551
##  1: 5931      1st Qu.:40.00   1st Qu.:2.000   1: 5960
##               Median :45.00   Median :2.000
##               Mean   :47.96   Mean   :2.924
```

```
##               3rd Qu.:55.00   3rd Qu.:4.000
##               Max.   :80.00   Max.   :7.000
##  rollover_indicator    VE_TOTAL        VE_FORMS        PVH_INVL
##  0:45595            Min.   :1.000   Min.   :1.000   Min.   :0.000000
##  1:  837            1st Qu.:2.000   1st Qu.:2.000   1st Qu.:0.000000
##  2:   79            Median :2.000   Median :2.000   Median :0.000000
##                     Mean   :2.083   Mean   :2.078   Mean   :0.004945
##                     3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.000000
##                     Max.   :8.000   Max.   :8.000   Max.   :3.000000
##       PEDS           PERMVIT         PERNOTMVIT     SCH_BUS   INT_HWY
##  Min.   :0.0000   Min.   : 1.000   Min.   :0.0000   0:46323   0:39205
##  1st Qu.:0.0000   1st Qu.: 2.000   1st Qu.:0.0000   1:  188   1: 7306
##  Median :0.0000   Median : 2.000   Median :0.0000
##  Mean   :0.1217   Mean   : 3.023   Mean   :0.1225
##  3rd Qu.:0.0000   3rd Qu.: 4.000   3rd Qu.:0.0000
##  Max.   :2.0000   Max.   :20.000   Max.   :3.0000
##  RELJCT1_IM ALCHL_IM  URBANICITY fatal     hit_run_indicator
##  0:45926    1: 8643   1:36079    0:23291   0:44934
##  1:  585    2:37868   2:10432    1:23220   1: 1577
##
##
##
##
##  max_travel_speed hour_night max_deformity_group1 month_group1
##  Min.   :  0.00   0:29127    0:31399              0:23710
##  1st Qu.: 30.00   1:17384    1:15112              1:22801
##  Median : 45.00
##  Mean   : 42.94
##  3rd Qu.: 60.00
##  Max.   :110.00
##  typ_int_group1 rel_road_group1 cf1_group1 wkdy      mancol_group1
##  0:36870        0: 6006         0: 2993    0:11942   0:24088
##  1: 9641        1:40505         1:43518    1:34569   1:22423
##
##
##
##
##  reljct2_group1 precipitation lgtcon_group1 midwest
##  0:35350        0:42936       0:18042       0:43057
##  1:11161        1: 3575       1:28469       1: 3454
##
##
##
##
```

Analysis

We chose to run two models: 1. Logistic Regression with LASSO 2. Random Forest

We chose to do logistic regression with LASSO because we have many predictor variables and levels of categorical variables. We chose to compare this to a random forest because of the random forest's well-known predictive abilities and computational efficiency.

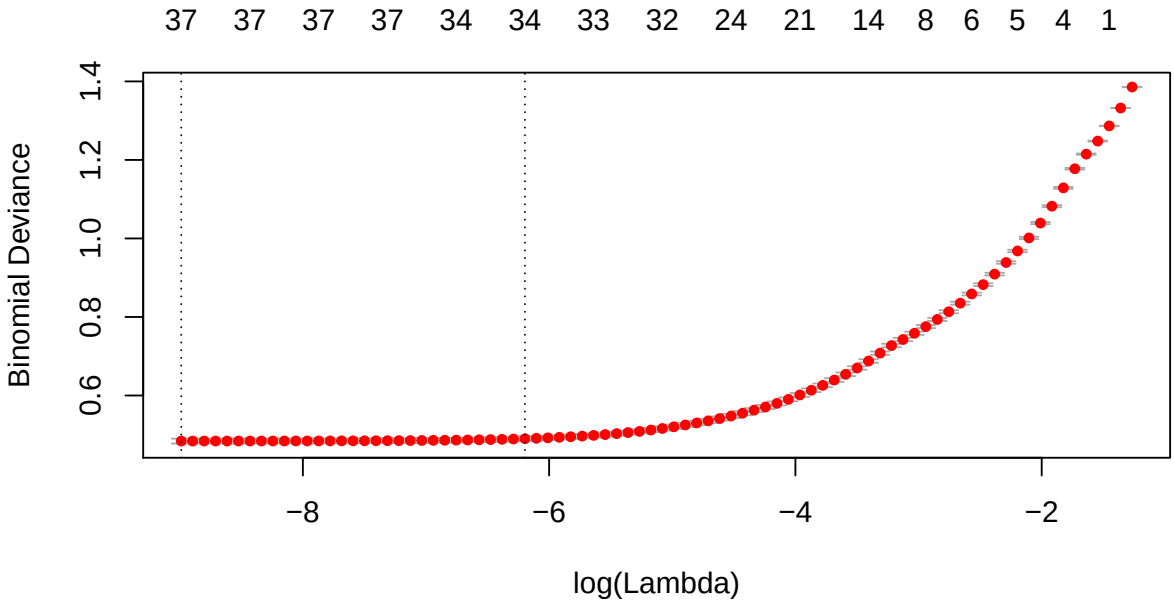
Training and Testing Data

We split our data of ~44,000 observations into a training and testing set. We put 37,208 observations (80%) in the training set. The remaining 9,303 observations (20%) are in the testing set. We train our multiple logistic regression on the training data set

Logistic Regression with LASSO

LASSO

We build the model using 10 fold cross-validation. We chose our model based on which model performed best in terms of minimizing deviance.



We will take lambda 1se. There will be about 34 non-zero predictors. The following are the variables that have been chosen by our LASSO.

```
##  [1] "(Intercept)"             "rest_mis_indicator1"
##  [3] "air_bag_indicator1"      "ejection_indicator1"
##  [5] "fire_indicator1"         "any_distracted1"
##  [7] "any_maneuver1"           "any_gradient1"
##  [9] "any_wet_road1"           "max_speed_limit"
## [11] "speed_related_accident1" "rollover_indicator1"
## [13] "rollover_indicator2"     "VE_FORMS"
## [15] "PVH_INVL"                "PEDS"
## [17] "SCH_BUS1"                "INT_HWY1"
## [19] "RELJCT1_IM1"             "ALCHL_IM2"
## [21] "URBANICITY2"             "hit_run_indicator1"
## [23] "max_travel_speed"        "hour_night1"
## [25] "max_deformity_group11"   "month_group11"
## [27] "typ_int_group11"         "rel_road_group11"
## [29] "cf1_group11"             "wkdy1"
## [31] "mancol_group11"          "reljct2_group11"
## [33] "lgtcon_group11"          "precipitation1"
```

```
## [35] "midwest1"
```

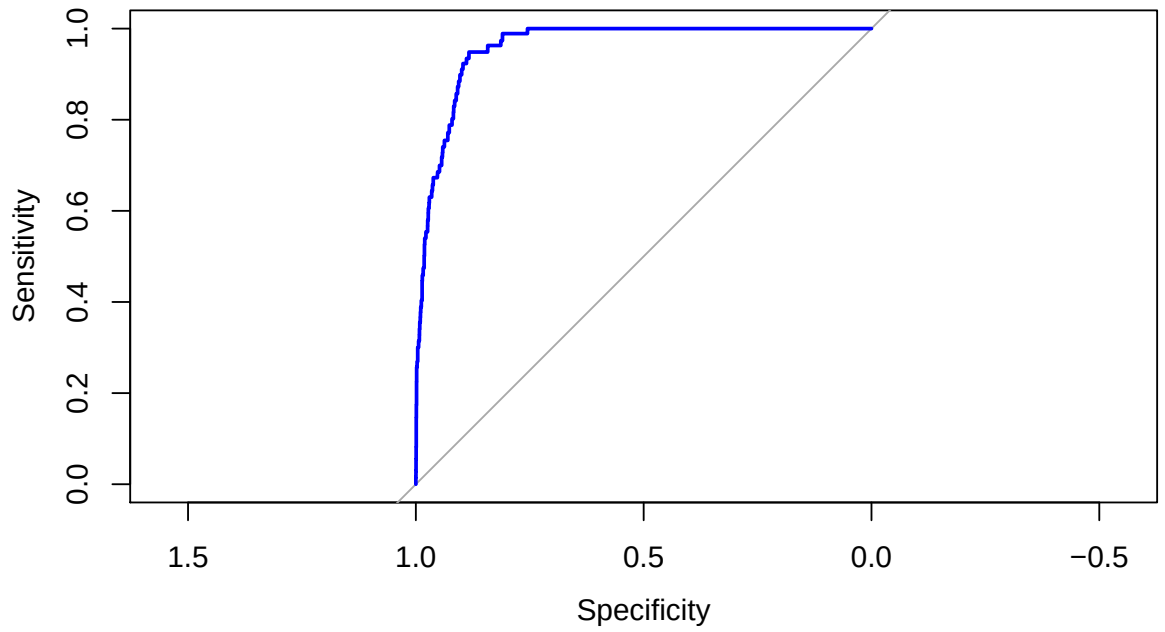
Logistic Regression (Relaxed LASSO)

To reduce bias, we take our inputs from LASSO and fit it into a logistic regression (a relaxed LASSO). We include a summary of our results in the Appendix. Here is our ROC curve.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
## Area under the curve: 0.9605
```

Results and Interpretation

Our model indicates that most predictors are statistically significant. The summary of our model is below. We will just discuss a few predictors that are actionable to legislators.

- **Fire**: Whether a car catches on fire or not is predictive of a fatality. A coefficient of 0.71 indicates that fires that result from crashes are more likely to result in fatalities
- **Ejection**: We found that whether or not a passenger was ejected from the vehicle was predictive of a fatality. Its positive estimate indicates that an ejection is more likely to result in a fatality.

- **Alcohol Impairment**: We find that accidents that do not involve alcohol impairment are statistically significant. The negative coefficient shows us that there will be a decrease in fatalities when a crash does not involve alcohol.

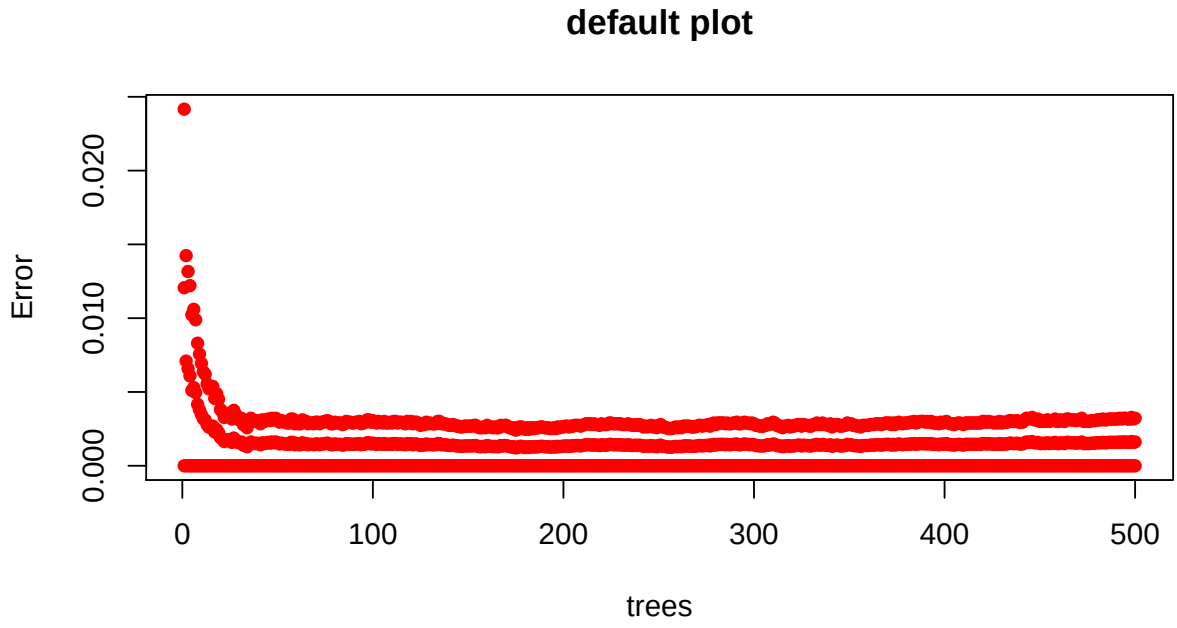
- **Maximum Speed Limit**: The maximum speed limit is a significant variable. Since it is positive, this shows that higher speed limits results in more fatalities

- **Lighting**: The amount of lighting has significant predictive power in our data set. Accidents in the dark are more lightly to result in fatalities.

- Additionally, we obtain coefficients that are not statistically different from 0. One of these coefficients is `School Bus`. Although the presence of a school bus is shown to be negative in our dataset, we do not have evidence to conclude that school buses are necessarily safer.

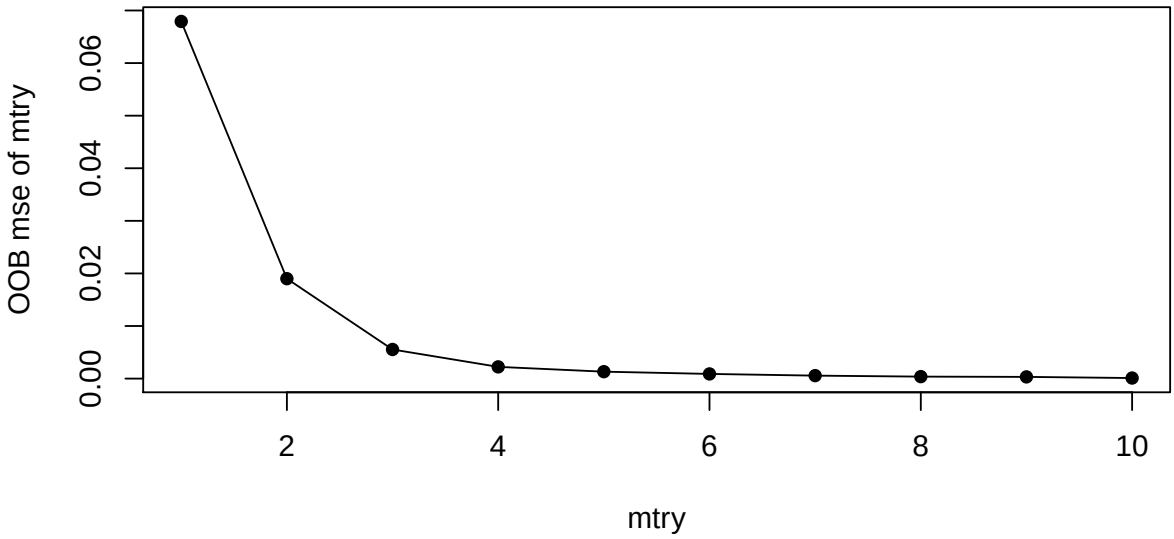
Random Forest

We will now run a random forest using `mtry` = 5 and `ntree` = 500. We will tune our tree and change these tuning parameters later. We run our random forest on the training data set. This is the plot of our initial random forest.



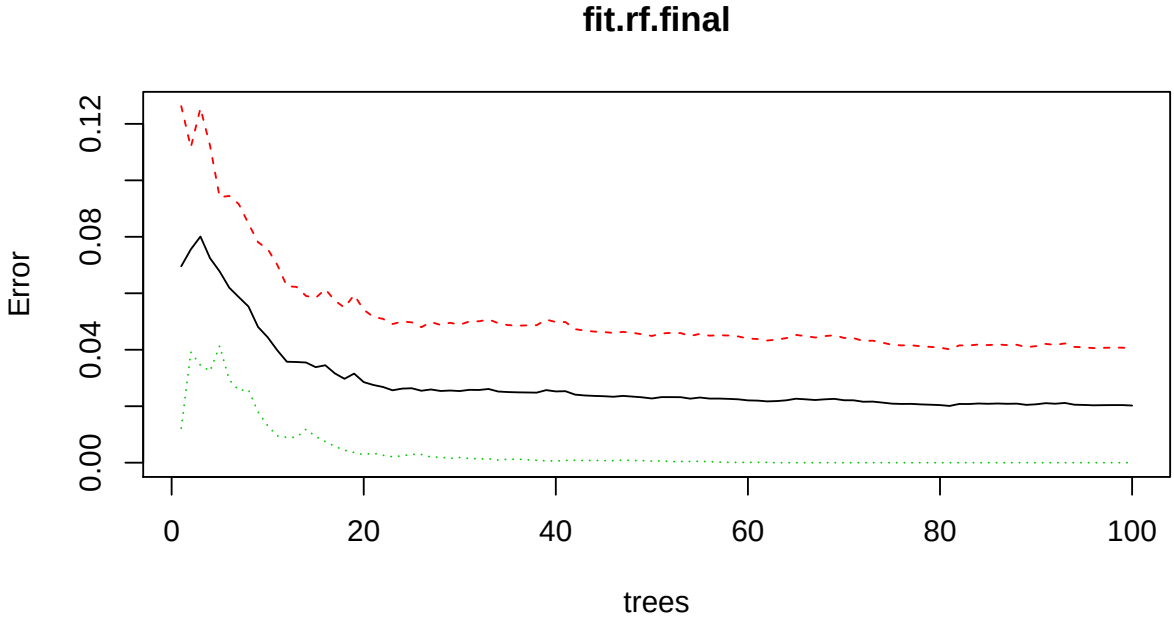
Tuning Model

For our random forest we first tuned `ntree`. To do this we created a random forest with the training dataset with an `mtry` of 5 and an `ntree` of 500 and identified at what level of `ntree` the error leveled off. We determined that the error leveled off where `ntree` equaled 100. Next we tuned `mtry` by creating random forests with the training dataset for each level of `mtry` from 1 to 10 with an `ntree` of 100 and comparing the OOB MSE. Using the elbow rule, we selected an `mtry` of 2. After this we were able to create our final random forest using the training dataset with an `mtry` of 2 and a `ntree` of 100.



Results and Interpretation

After tuning the model we get the following results for variable importances. The five most important factors in our random forest were `max_travel_speed`, `max_deformity_group1`, `mancol_group1`, `max_speed_limit`, and `hour_night`. This is in line with our relaxed LASSO results in which all five of these variables are statistically significant.




```
##  [1] "rest_mis_indicator"     "air_bag_indicator"
```

```
##  [3] "ejection_indicator"     "fire_indicator"
##  [5] "any_vision_obstructed"  "any_distracted"
##  [7] "any_maneuver"           "any_gradient"
##  [9] "any_wet_road"           "max_speed_limit"
## [11] "max_nlanes"             "speed_related_accident"
## [13] "rollover_indicator"     "VE_TOTAL"
## [15] "VE_FORMS"               "PVH_INVL"
## [17] "PEDS"                   "PERMVIT"
## [19] "PERNOTMVIT"             "SCH_BUS"
## [21] "INT_HWY"                "RELJCT1_IM"
## [23] "ALCHL_IM"               "URBANICITY"
## [25] "hit_run_indicator"      "max_travel_speed"
## [27] "hour_night"             "max_deformity_group1"
## [29] "month_group1"           "typ_int_group1"
## [31] "rel_road_group1"        "cf1_group1"
## [33] "wkdy"                   "mancol_group1"
## [35] "reljct2_group1"         "precipitation"
## [37] "lgtcon_group1"          "midwest"
```

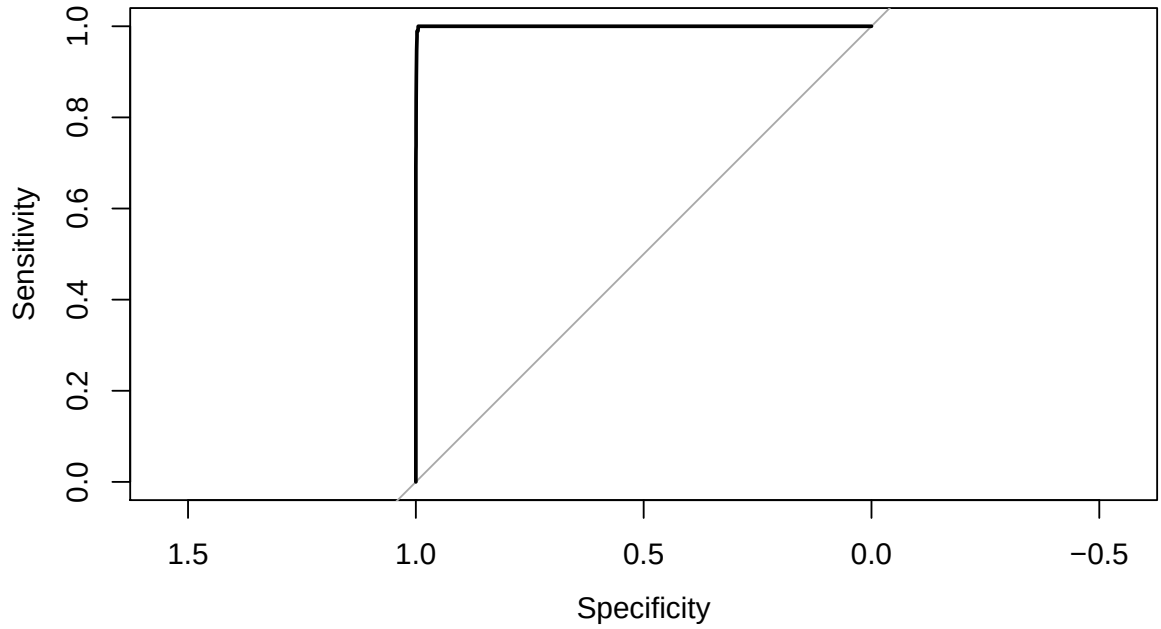
```
##                     names MeanDecreaseGini
## 26       max_travel_speed      1678.452786
## 34          mancol_group1      1138.459490
## 28   max_deformity_group1      1021.632586
## 10        max_speed_limit      1014.734419
## 23               ALCHL_IM       822.803130
## 27             hour_night       690.854811
## 37          lgtcon_group1       604.642191
## 29           month_group1       520.941444
## 18                PERMVIT       519.992823
## 14               VE_TOTAL       481.375443
## 15               VE_FORMS       442.845112
## 11             max_nlanes       416.756402
## 2       air_bag_indicator       363.156011
## 35         reljct2_group1       361.826529
## 19             PERNOTMVIT       344.612476
## 12 speed_related_accident       311.696703
## 7            any_maneuver       277.888111
## 17                   PEDS       257.161271
## 3      ejection_indicator       254.947557
## 30         typ_int_group1       249.322846
## 38                midwest       243.971856
## 32             cf1_group1       220.674966
## 21                INT_HWY       208.362392
## 33                   wkdy       188.711605
## 6          any_distracted       186.071654
## 36          precipitation       160.133083
## 31        rel_road_group1       137.631412
## 8            any_gradient       132.203183
## 24             URBANICITY       130.887263
## 9            any_wet_road       110.014281
## 4          fire_indicator        81.624985
## 5   any_vision_obstructed        74.295701
## 25     hit_run_indicator        69.598069
## 1      rest_mis_indicator        49.274365
```

```
## 22             RELJCT1_IM        43.085587
## 13     rollover_indicator        28.964384
## 16                PVH_INVL        13.783006
## 20                 SCH_BUS         6.579843
```

The following is the ROC curve for our random forest.

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = data.test$fatal, predictor = predict.rf[,     2], plot = TRUE)
##
## Data: predict.rf[, 2] in 4661 controls (data.test$fatal 0) < 4642 cases (data.test$fatal 1).
## Area under the curve: 0.9997
```

Model Selection and Final Model

To compare out LASSO logistic regression with our random forest, we calculated the AUC on the testing dataset. The AUC for LASSO was 0.9592 and the AUC for our random forest was 0.9995. Therefore, we determined that the random forest was the better model. Both of these models had very high AUCs. We believe this occurred because our predictors have very higher predictive power in predicting fatalities. We do not find this surprising and believe that NHTSA most likely knows what factors contribute to traffic fatalities. The real problem relates to legislation. It is significantly easier to determine the factors that predictive traffic fatalities than it is to create legislation to lower traffic fatalities. It is our goal to provide meaningful recommendations for legislation with the aim of reducing traffic fatalities.

Recommendations and Conclusions

We propose a few policy solutions in order to combat traffic-related incidents. Based on the models that we ran, we propose these solutions concerning:

- **Fire**: We can pressure car manufacturers to create more fire retardant coverings or protections on their vehicles. If not mandatory, it might be possible to offer a tax incentive to consumers to use fire-retardant protections
- **Ejection**: Similar to fire, we can add in protections and stricter seatbelt safety laws
- **Alcohol Impairment**: While we admit that there are already strict alcohol laws, additional protections can be added to cars such as breathalyzers in order to prevent drunk driving.
- **Maximum Travel Speed**: We can enforce the maximum speed limit that cars are traveling with harsher laws for speeders
- **Lighting**: Increase access to more lighting as dark areas tend to have more accidents
- **Buses**: Bus safety laws are very important to parents. However, our model shows that buses are not effective in predicting the amount of fatalities. In fact, data from the NHTSA has shown that buses are one of the safest vehicles on the road due to their size. Since buses are not found to improve our model, we suggest focusing energy on the other factors that we have described.

Limitations

Our model performs very well but still has many limitations. The first limitation that we excluded data about pedestrians. We did this for the sake of parsimony but we recognize that this limits the scope of our recommendations and suggestions. We excluded data about vehicle passengers for parsimony as well and recognize that this also limits our recommendations. Additionally, we would like to have more data from the NHTSA on cases with fatalities. This would allow us to forego bootstrapping and analyze non-artificial data. Another limitation of our model is that by using LASSO we assume that there is sparsity in the data, meaning that few of the predictors are actually important.

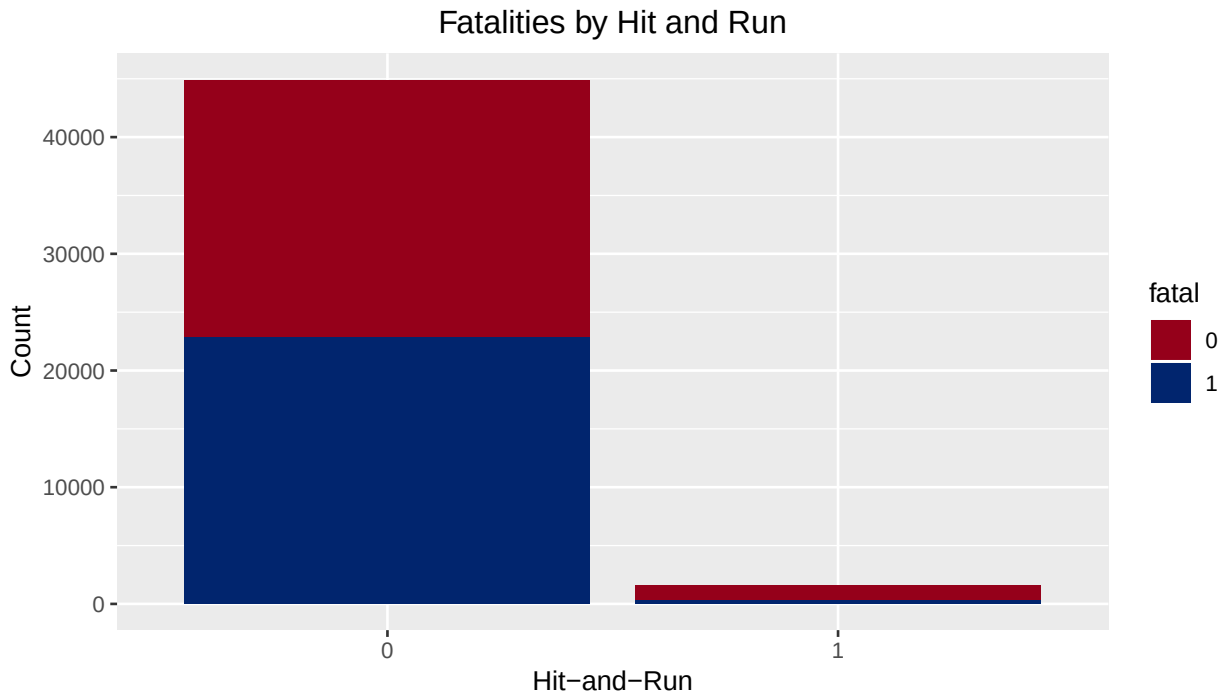
Appendix

EDA graphs

Graphs Mentioned in EDA Section

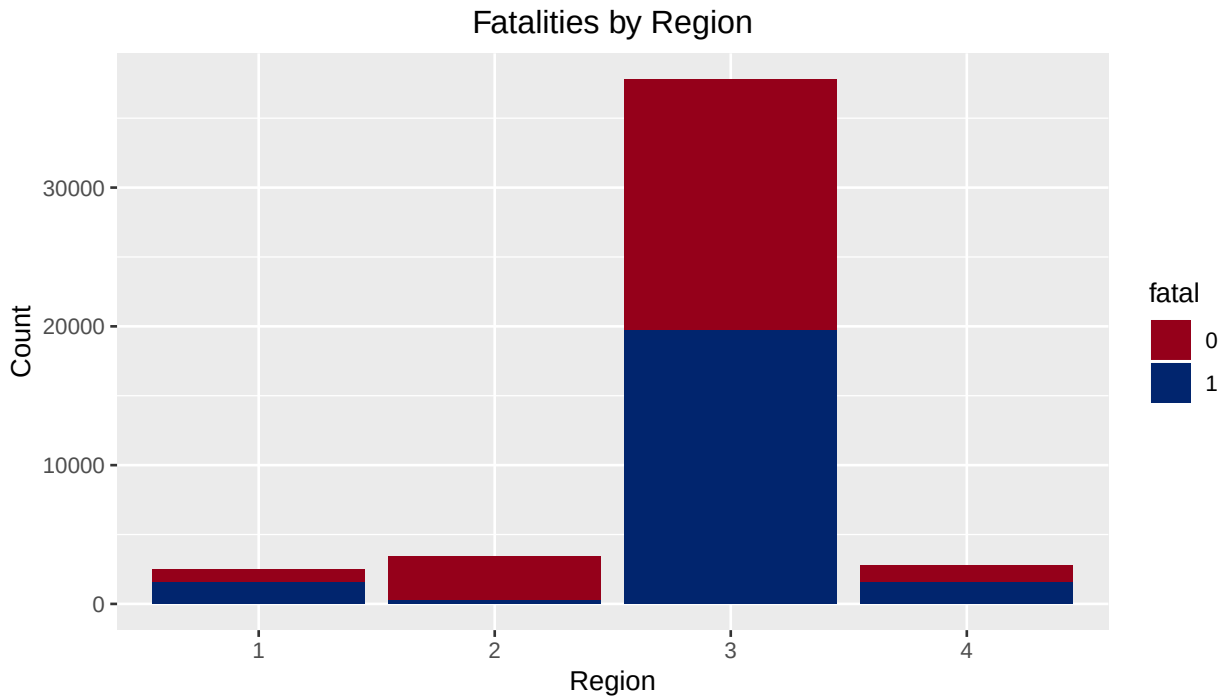
```
data %>%
  ggplot(aes(hit_run_indicator, fill=fatal)) +
  geom_histogram(stat="count") +
  ggtitle("Fatalities by Hit and Run") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Hit-and-Run", y = "Count") +
  scale_fill_manual(values=c(PennRed, PennBlue))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



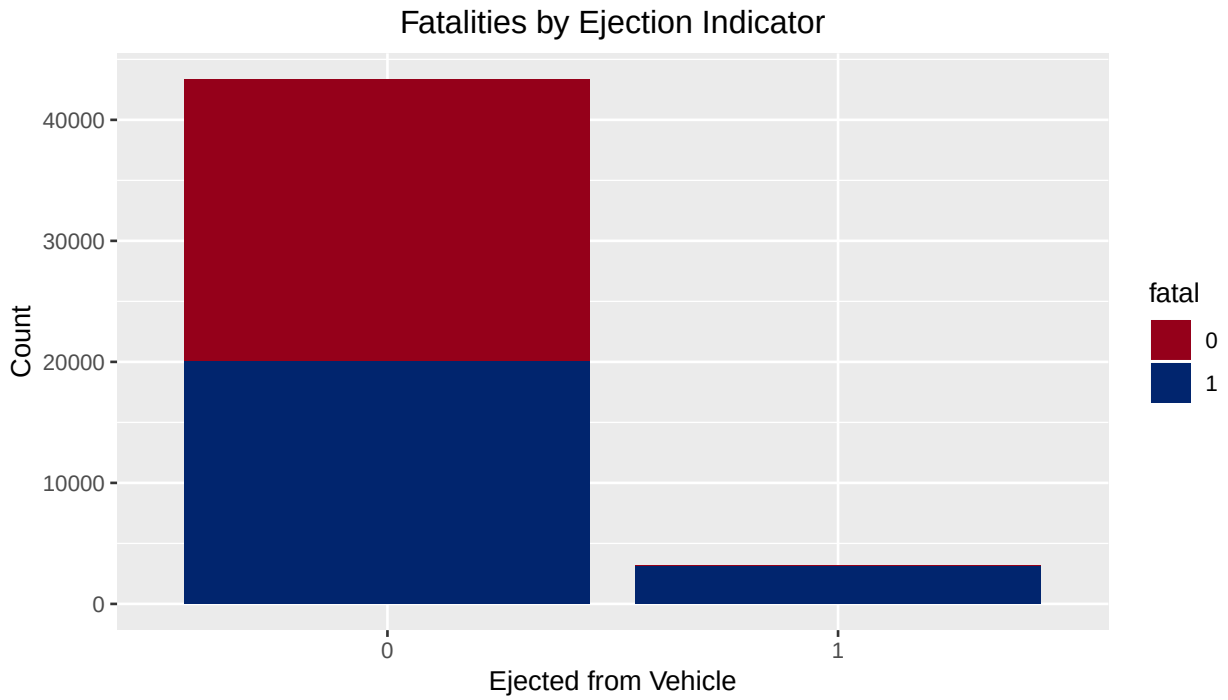

```
data %>%
  ggplot(aes(REGION, fill=fatal)) +
  geom_histogram(stat="count")+
  ggtitle("Fatalities by Region") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Region", y = "Count") +
  scale_fill_manual(values=c(PennRed, PennBlue))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



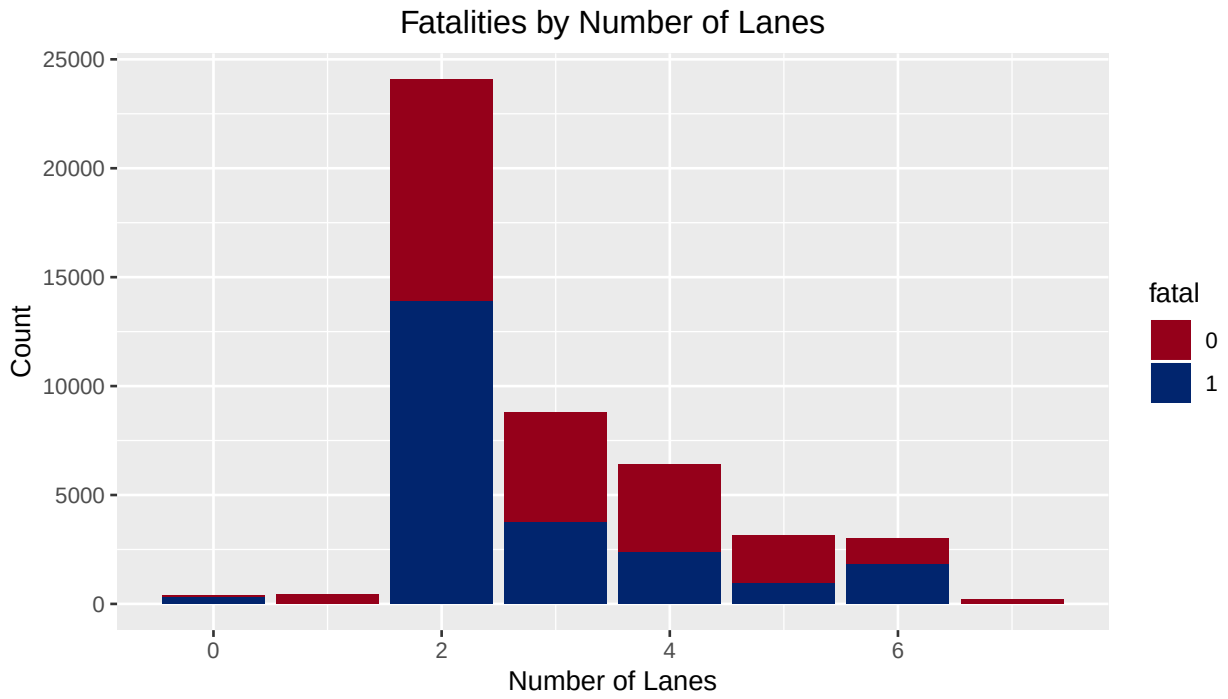

```
data %>%
  ggplot(aes(ejection_indicator, fill=fatal)) +
  geom_histogram(stat="count") +
  ggtitle("Fatalities by Ejection Indicator") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Ejected from Vehicle", y = "Count") +
  scale_fill_manual(values=c(PennRed, PennBlue))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



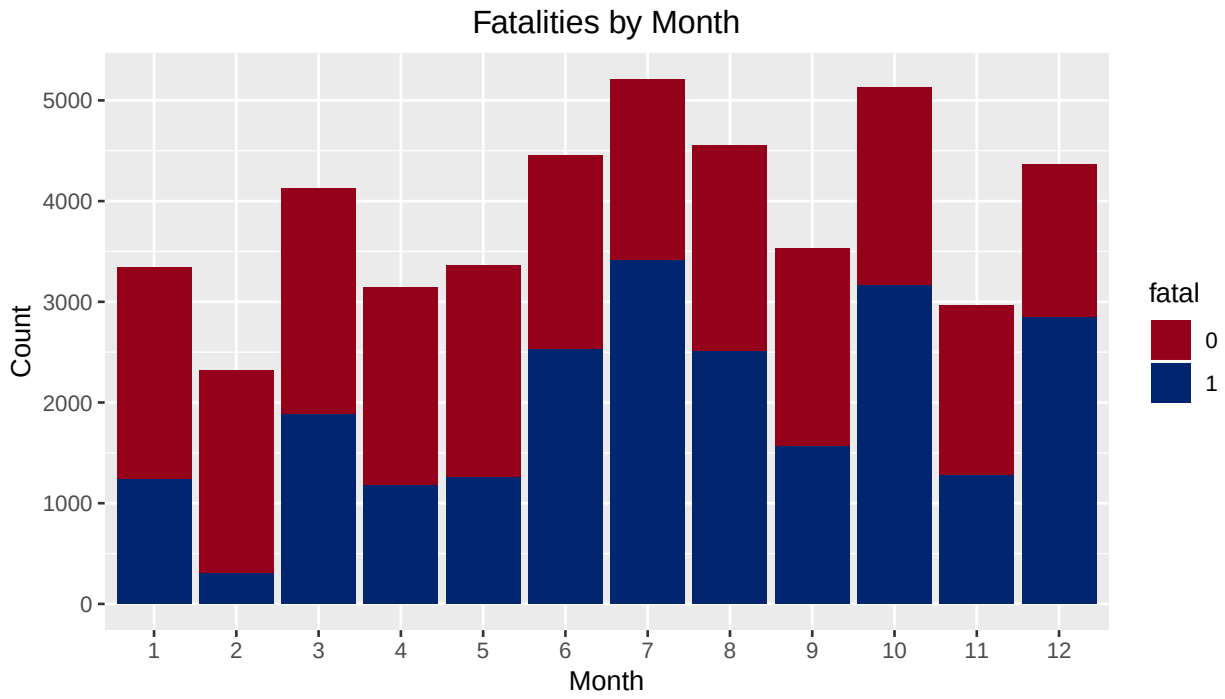

```
data %>%
  ggplot(aes(max_nlanes, fill=fatal)) +
  geom_histogram(stat="count") +
  ggtitle("Fatalities by Number of Lanes") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Number of Lanes", y = "Count") +
  scale_fill_manual(values=c(PennRed, PennBlue))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



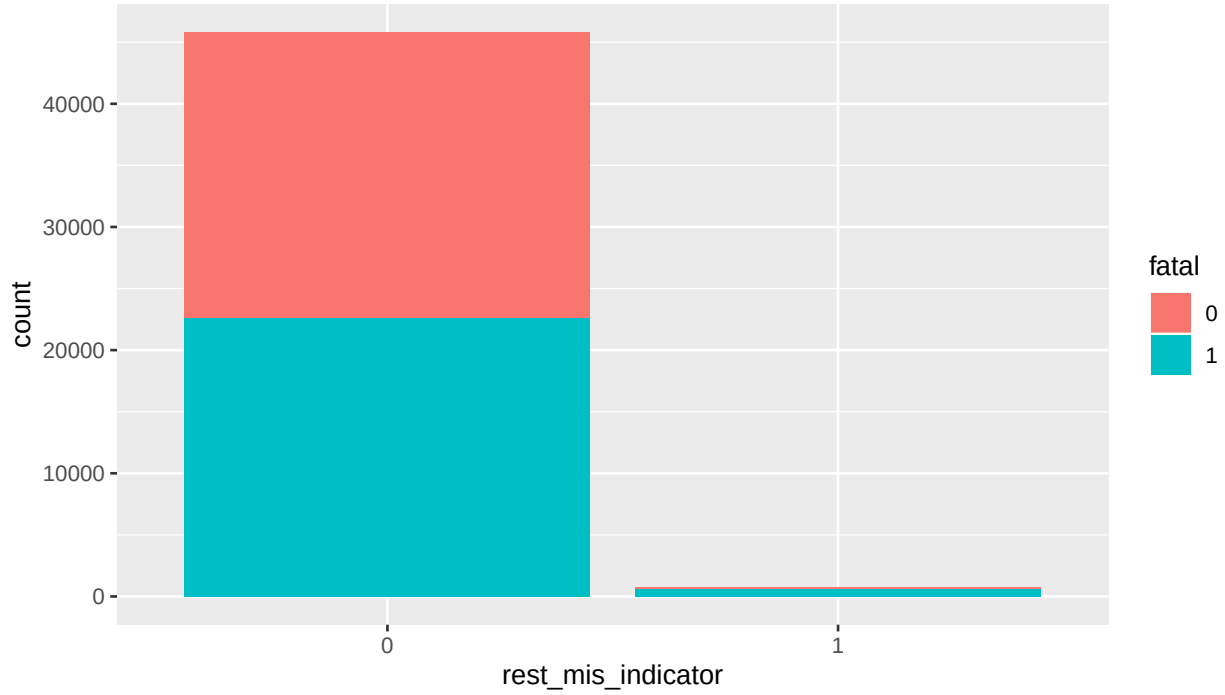

```
data %>%
  ggplot(aes(MONTH, fill=fatal)) +
  geom_histogram(stat="count") +
  ggtitle("Fatalities by Month") +
    theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Month", y = "Count") +
  scale_fill_manual(values=c(PennRed, PennBlue))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

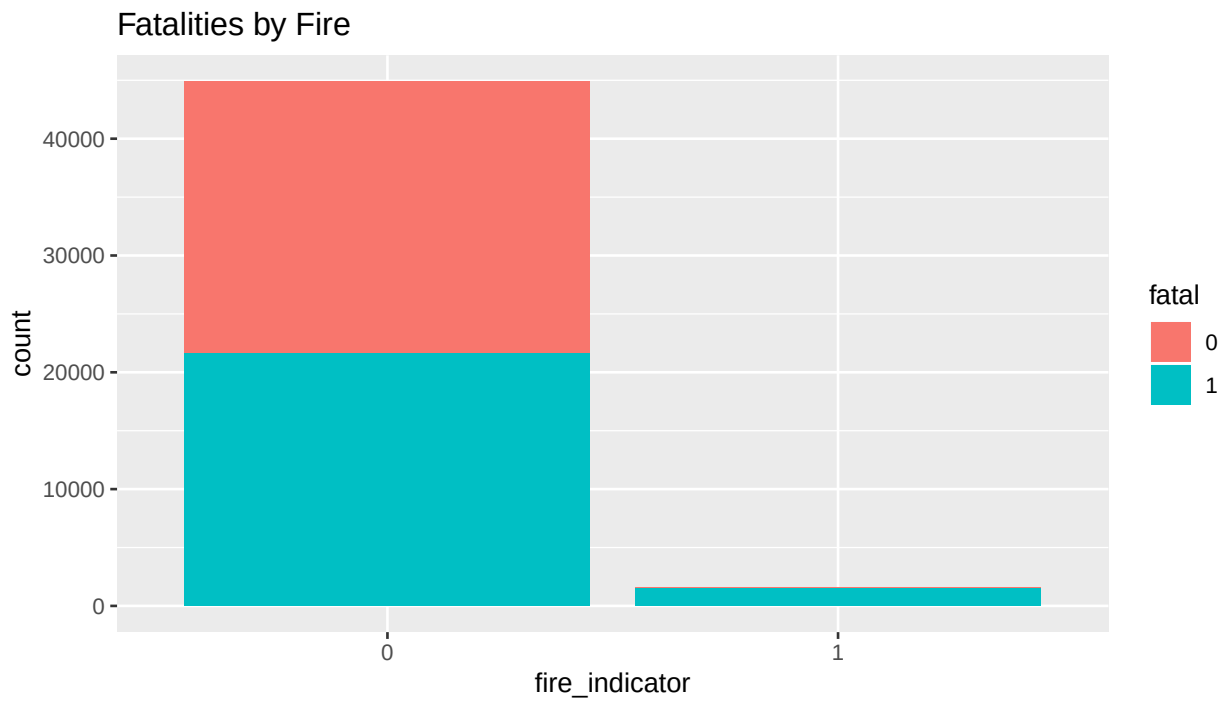


Other Graphs

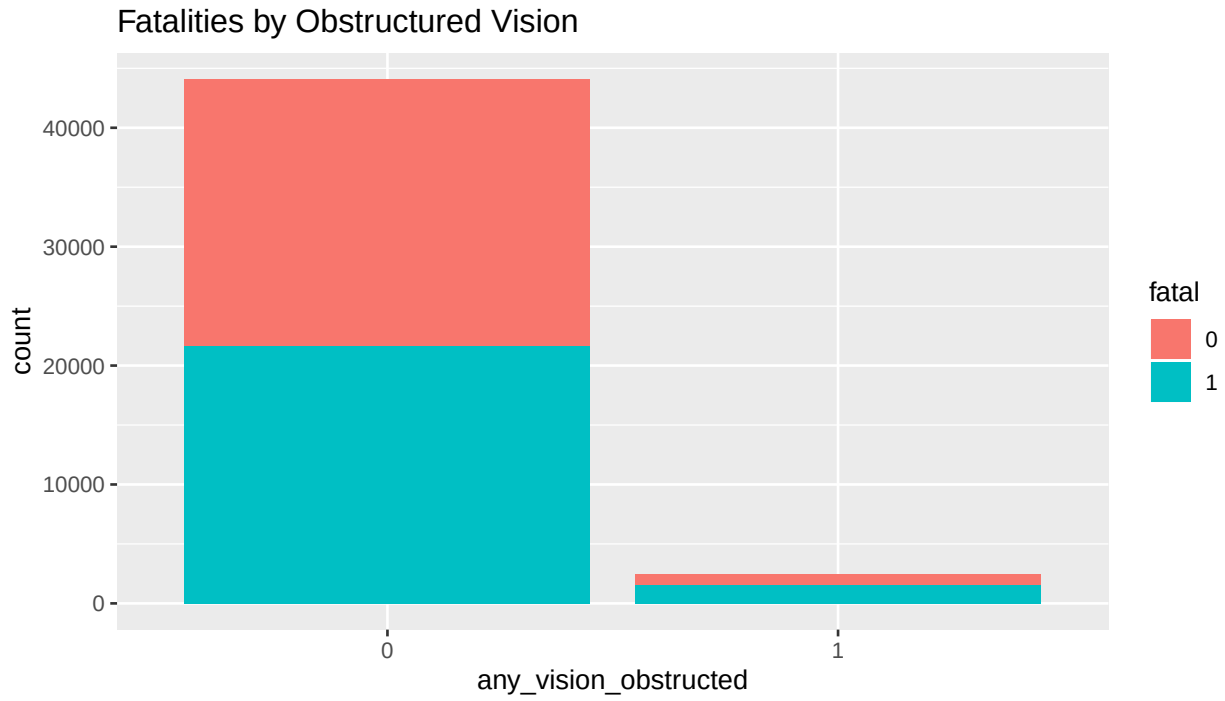
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



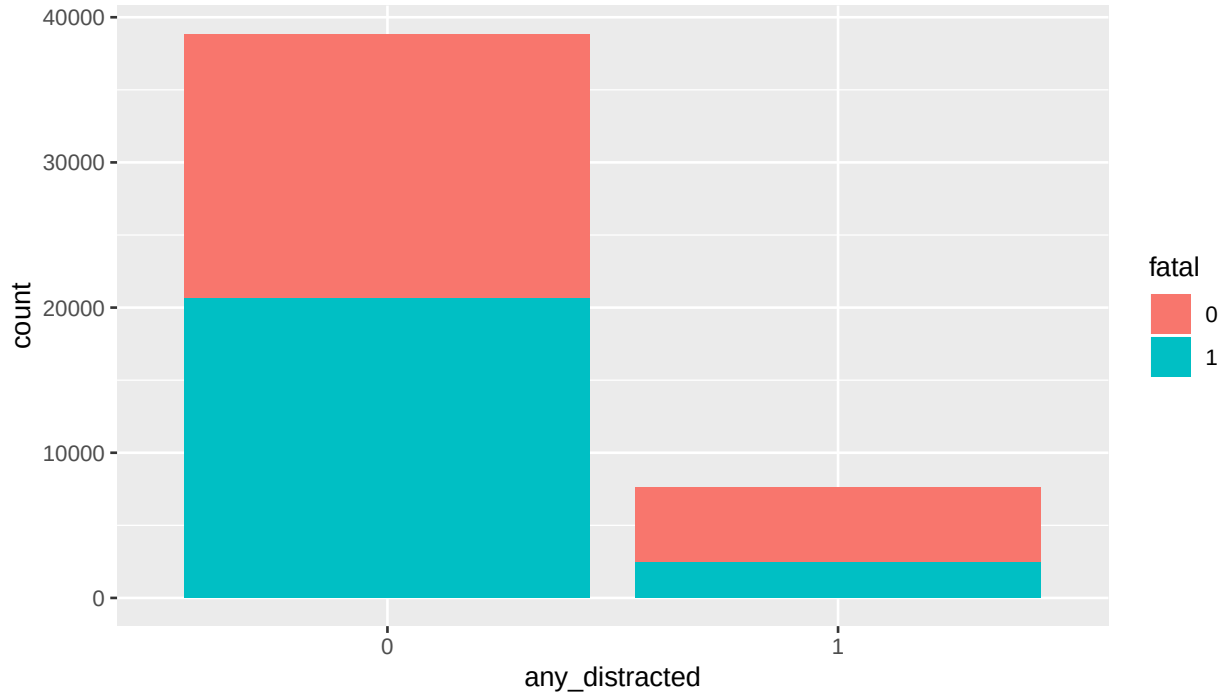

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



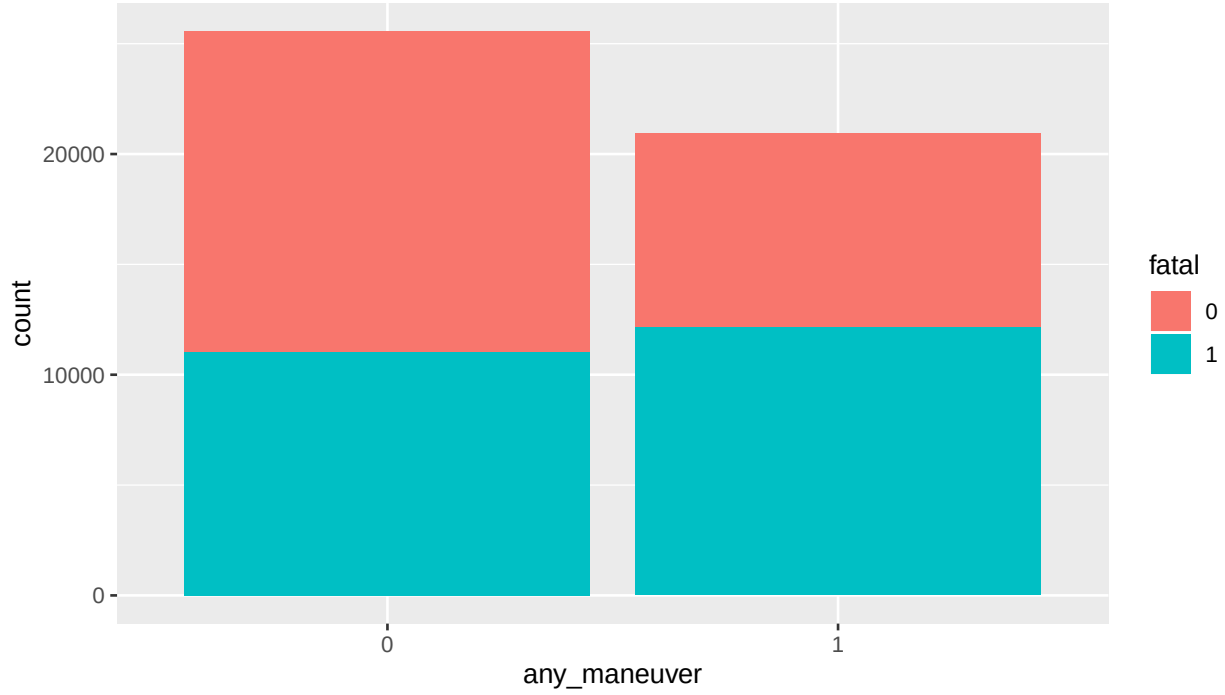

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



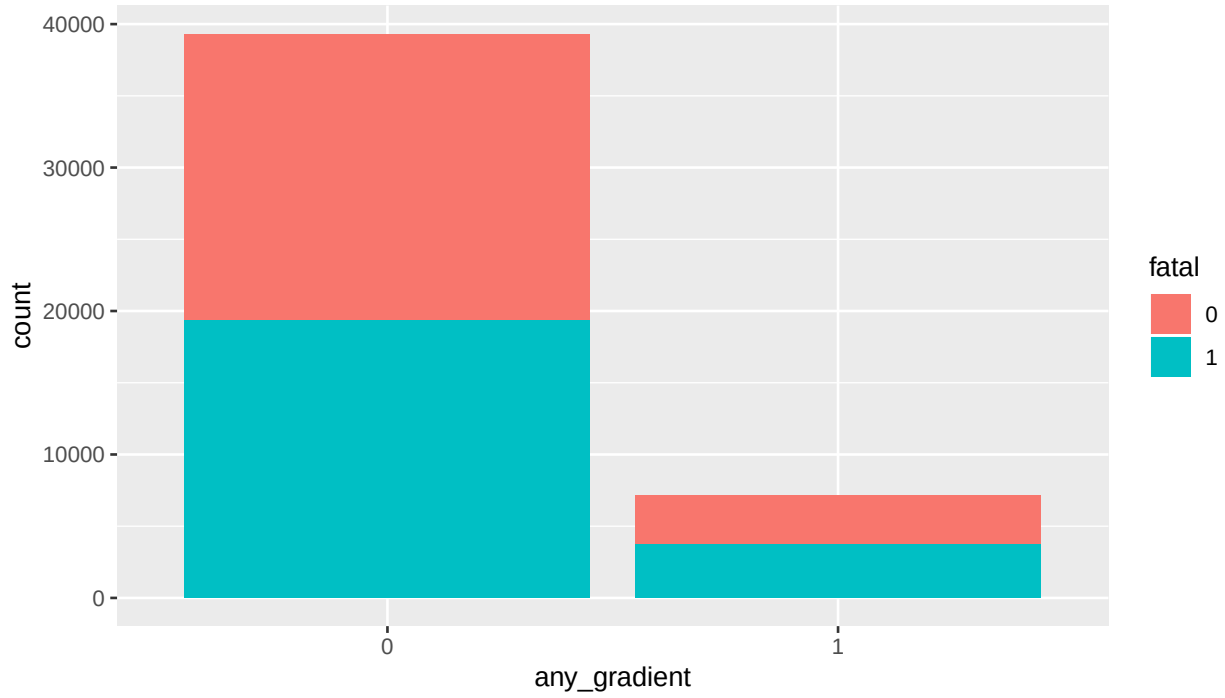

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



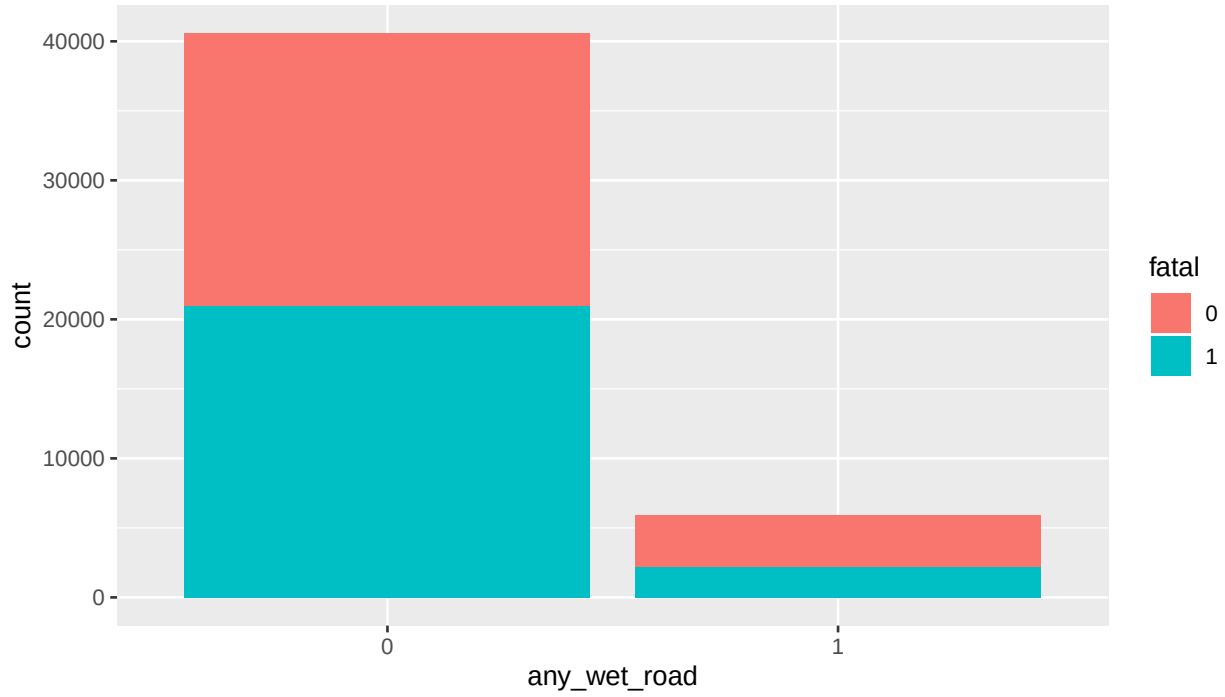

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



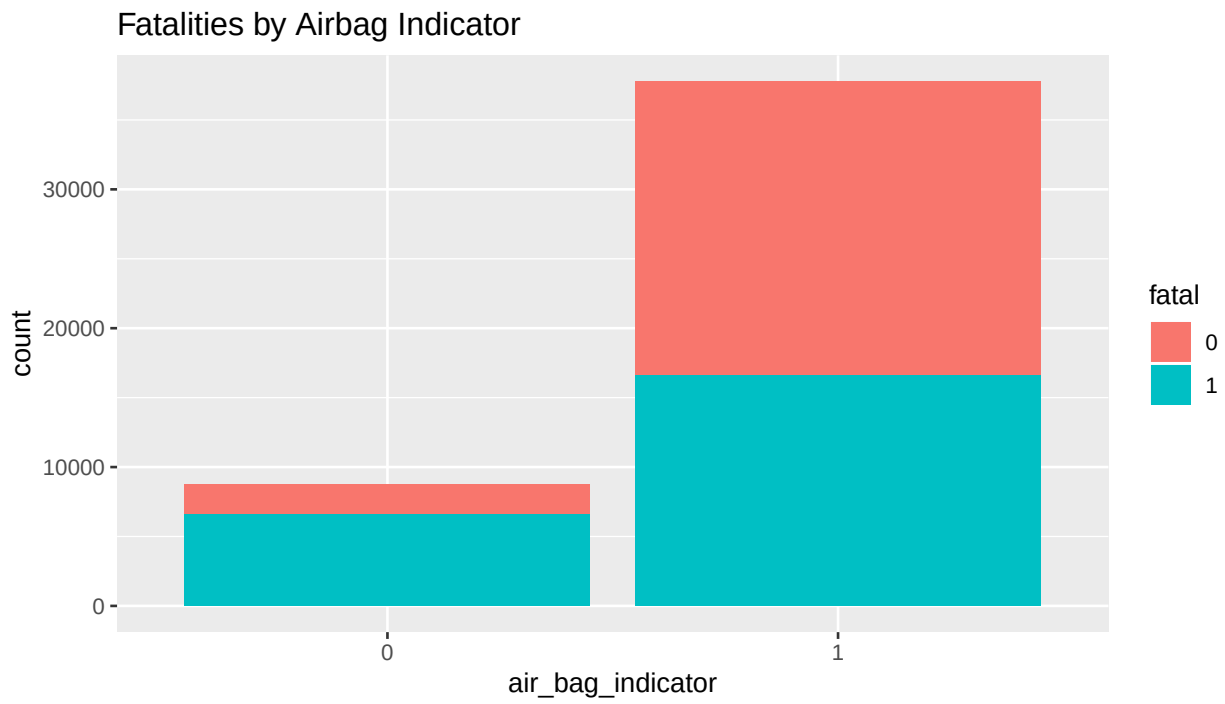

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



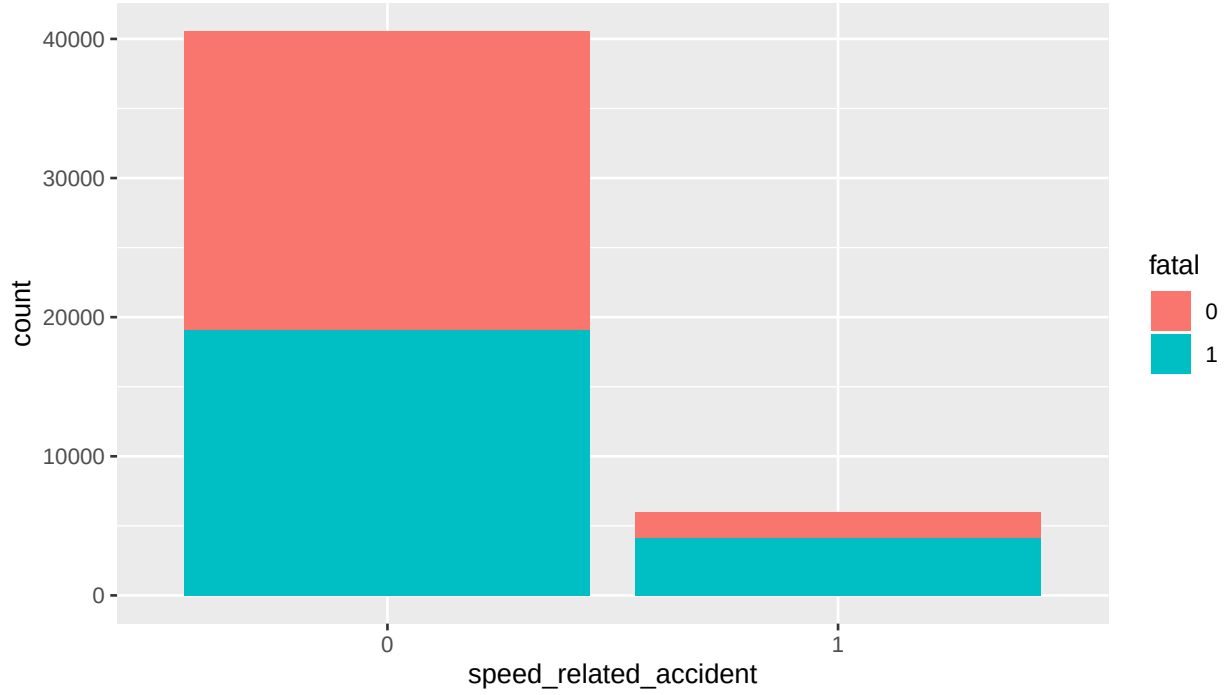

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



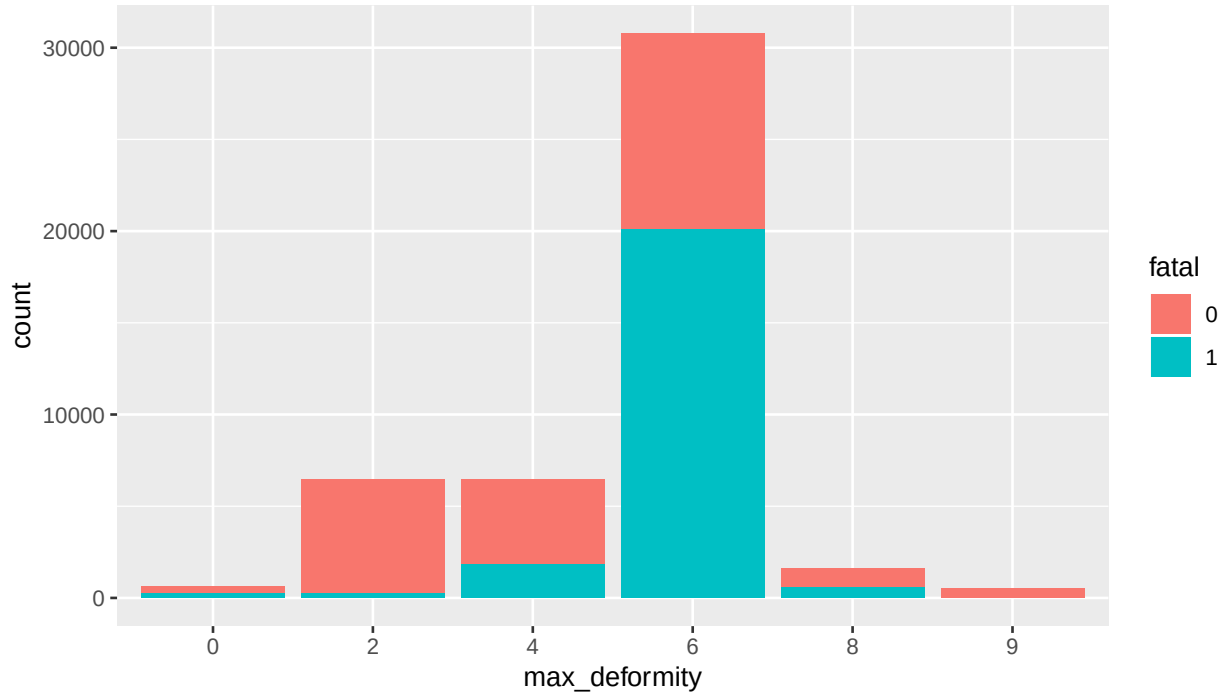

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



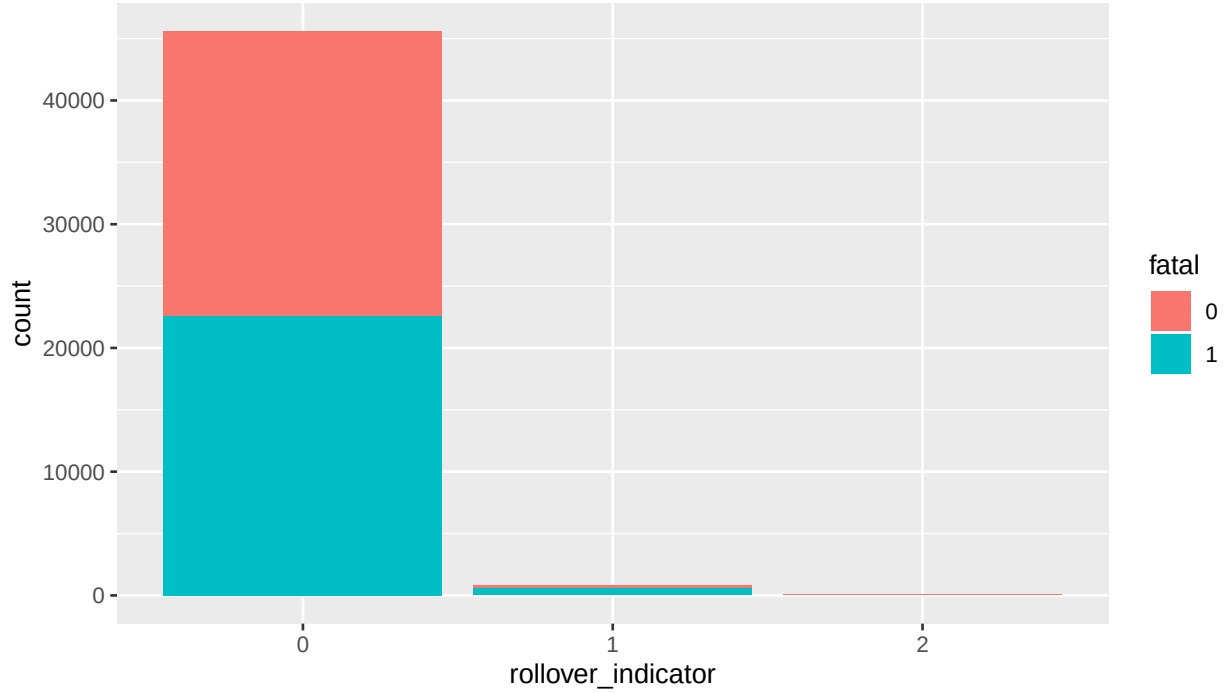

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



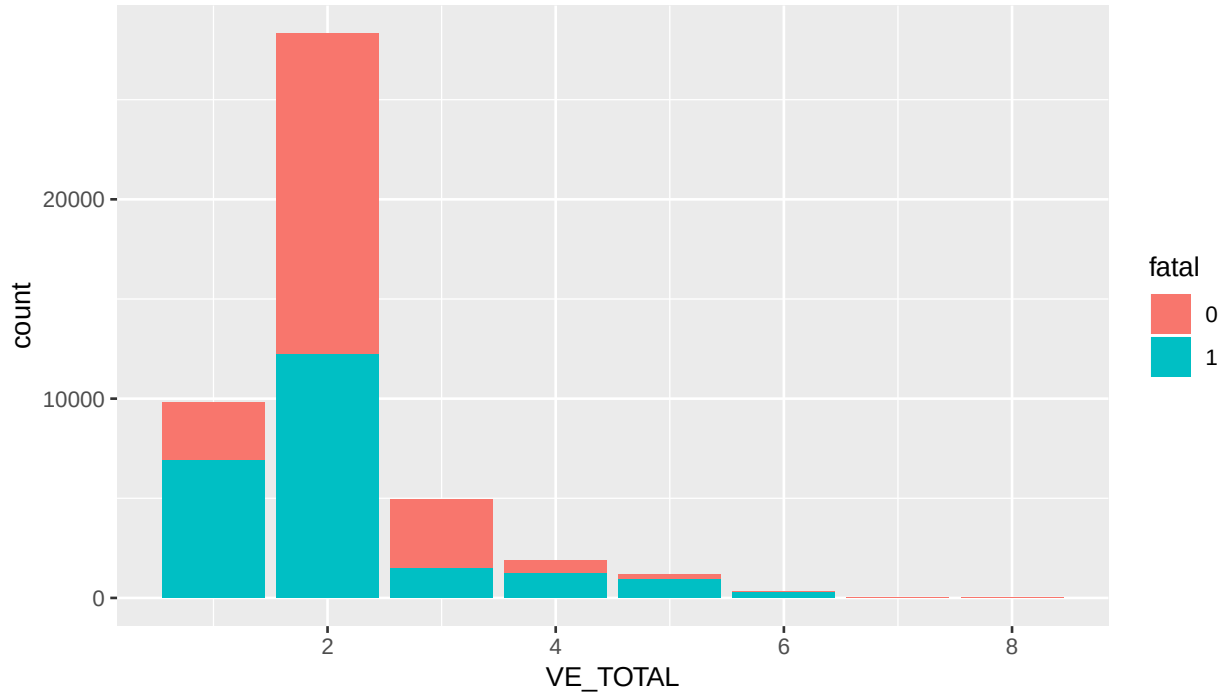

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



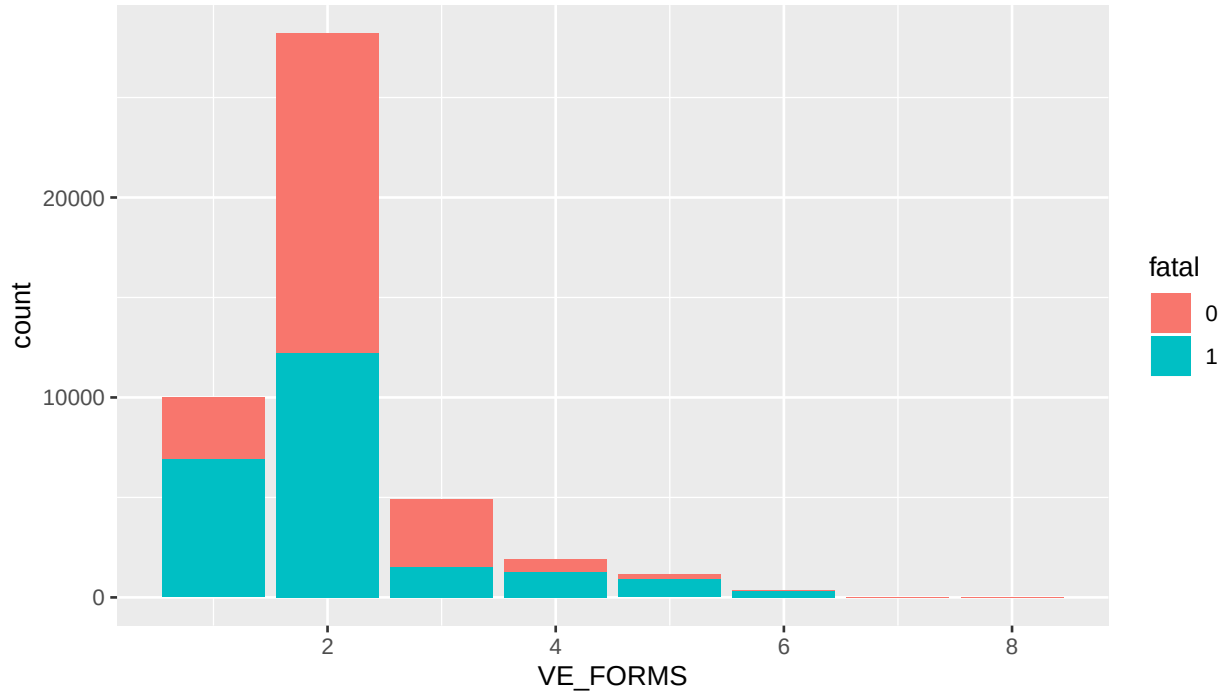

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



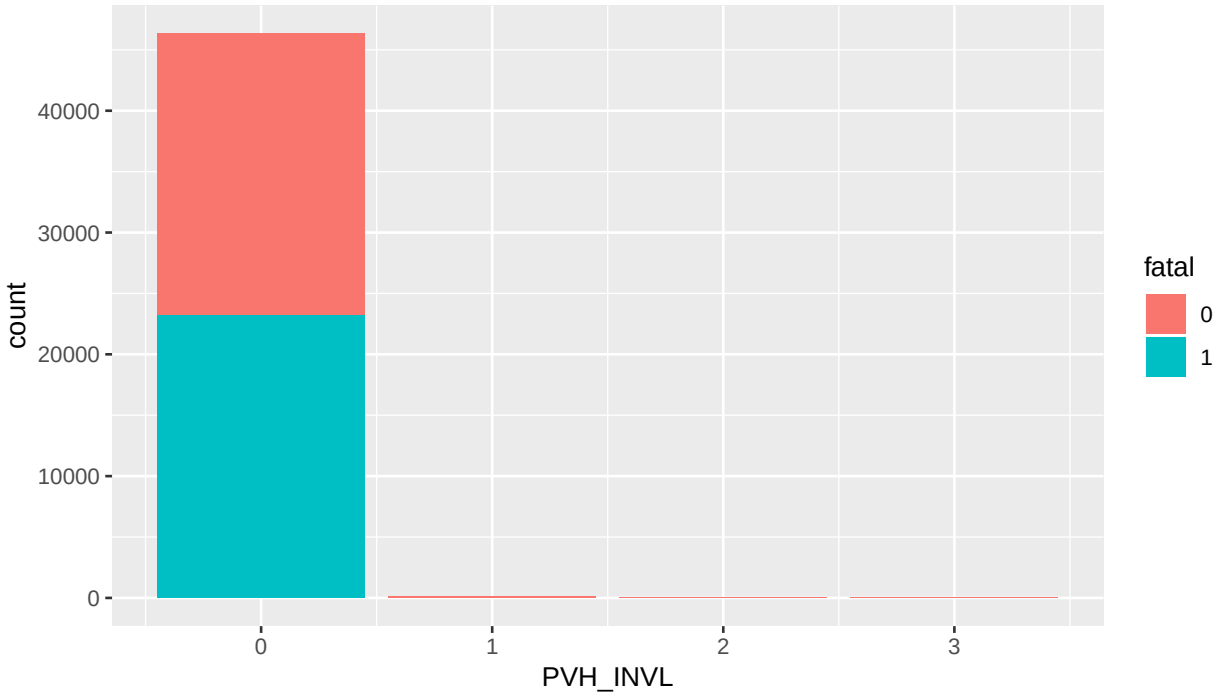

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



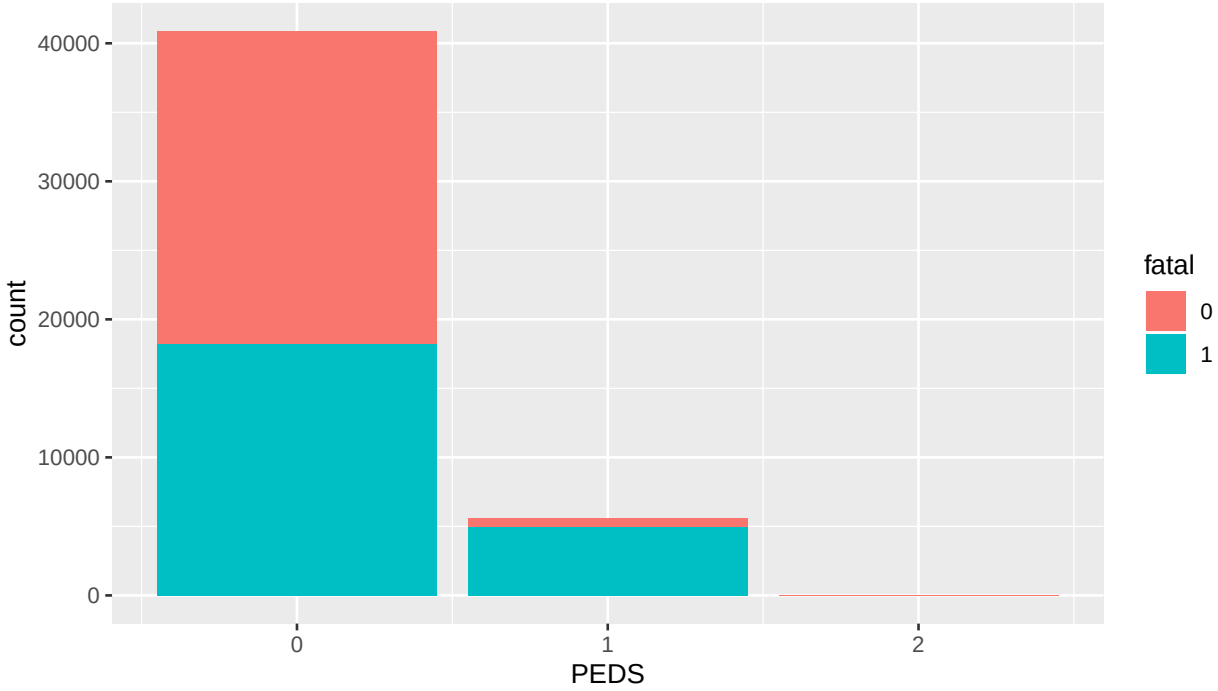

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



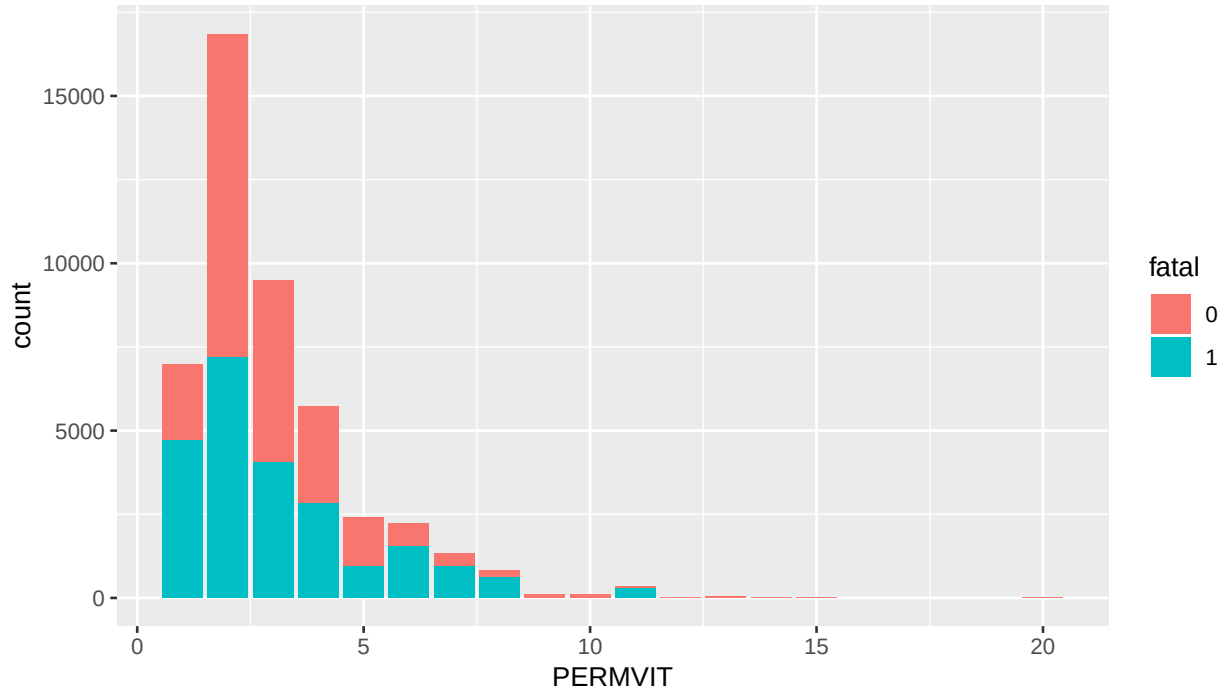

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



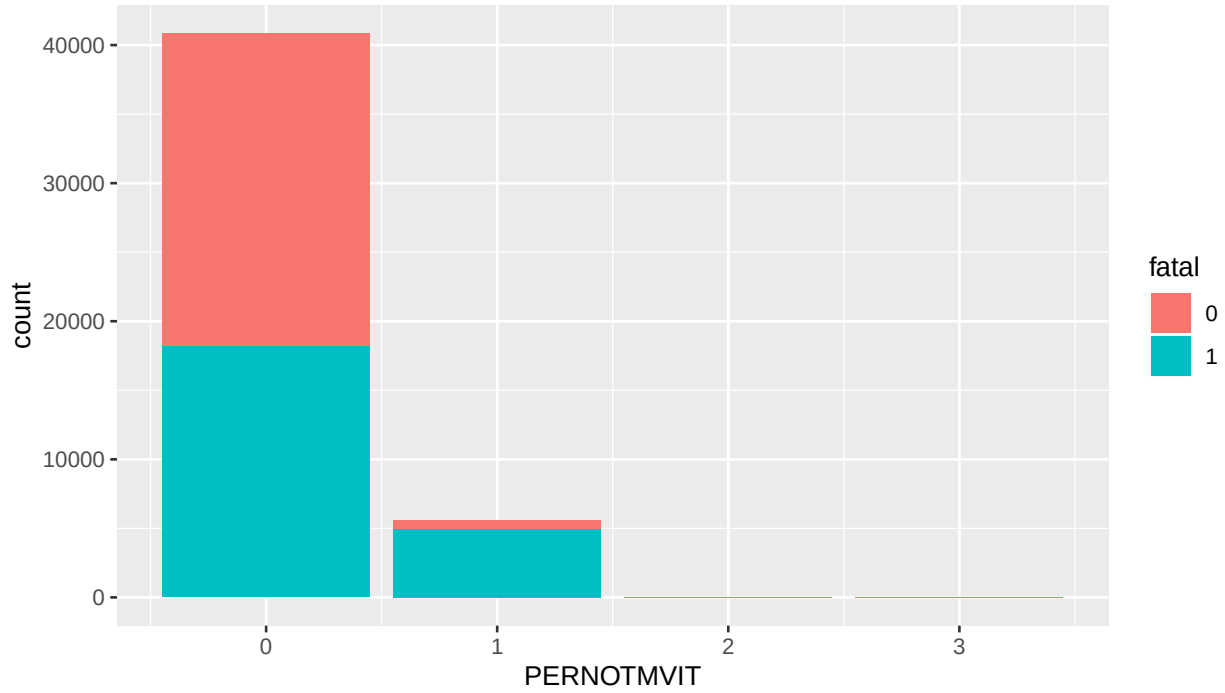

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



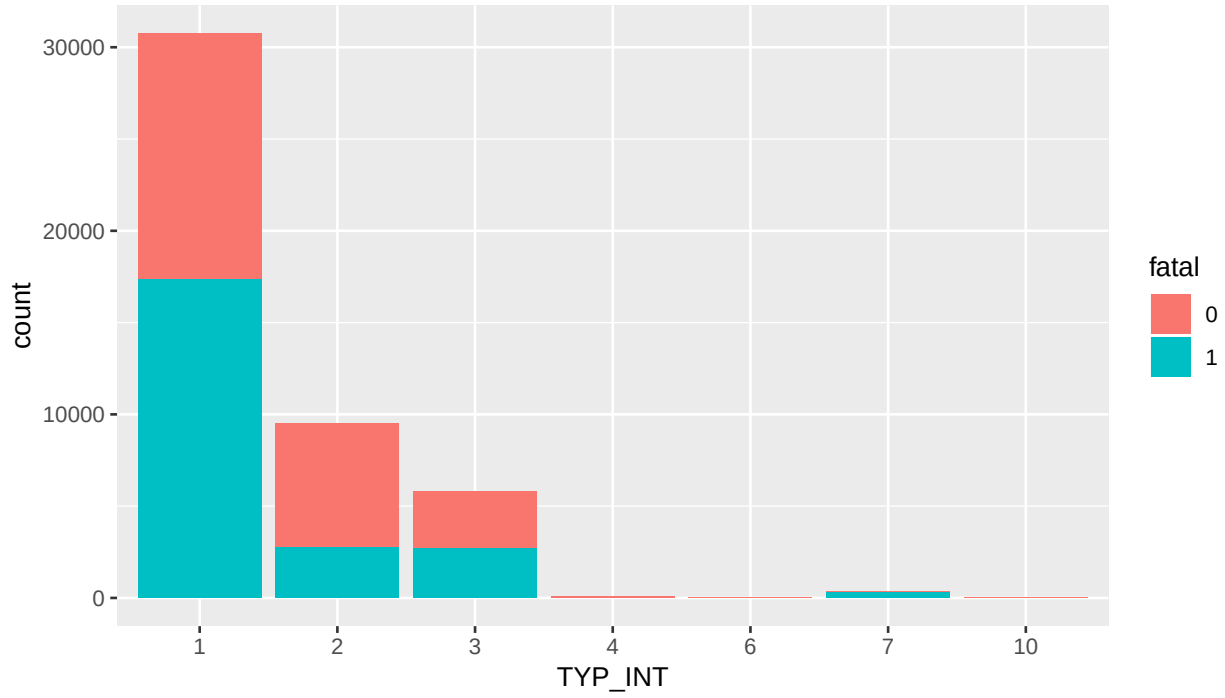

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



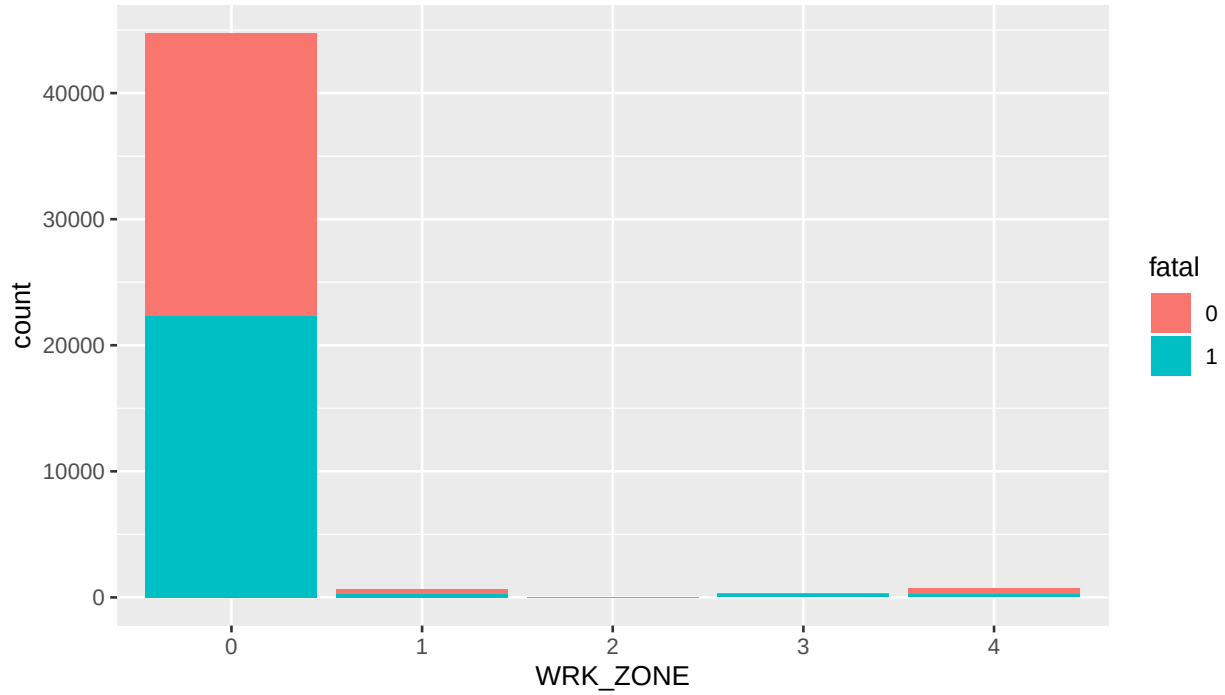

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



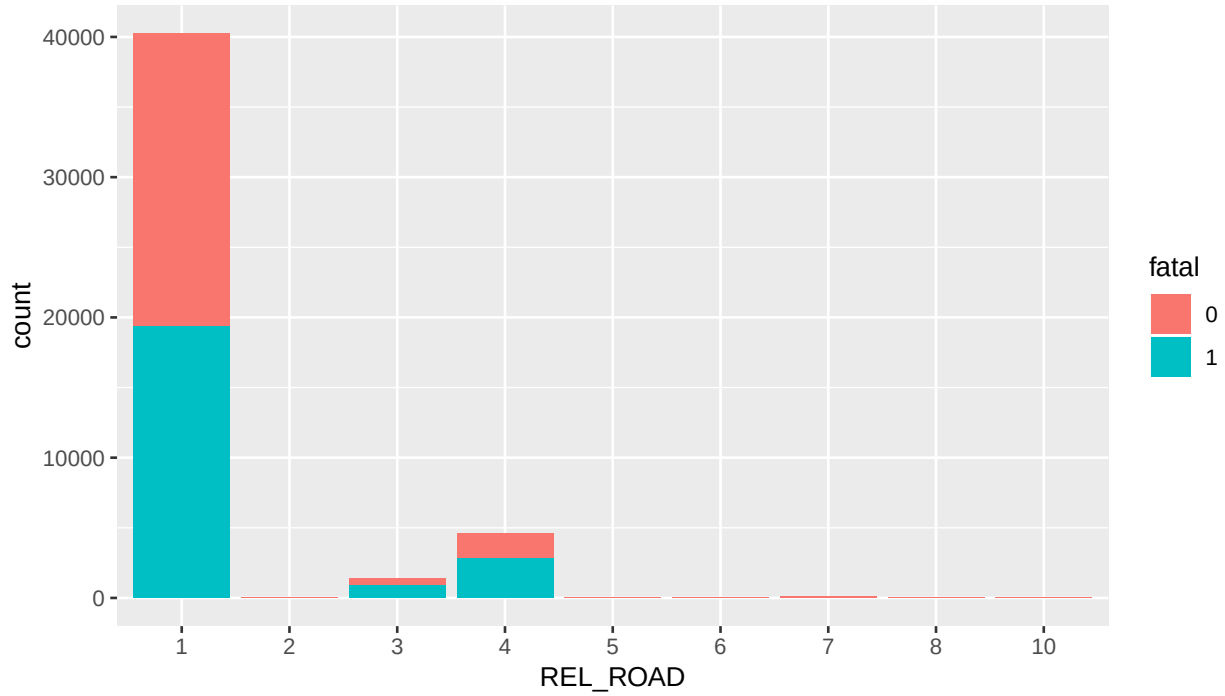

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



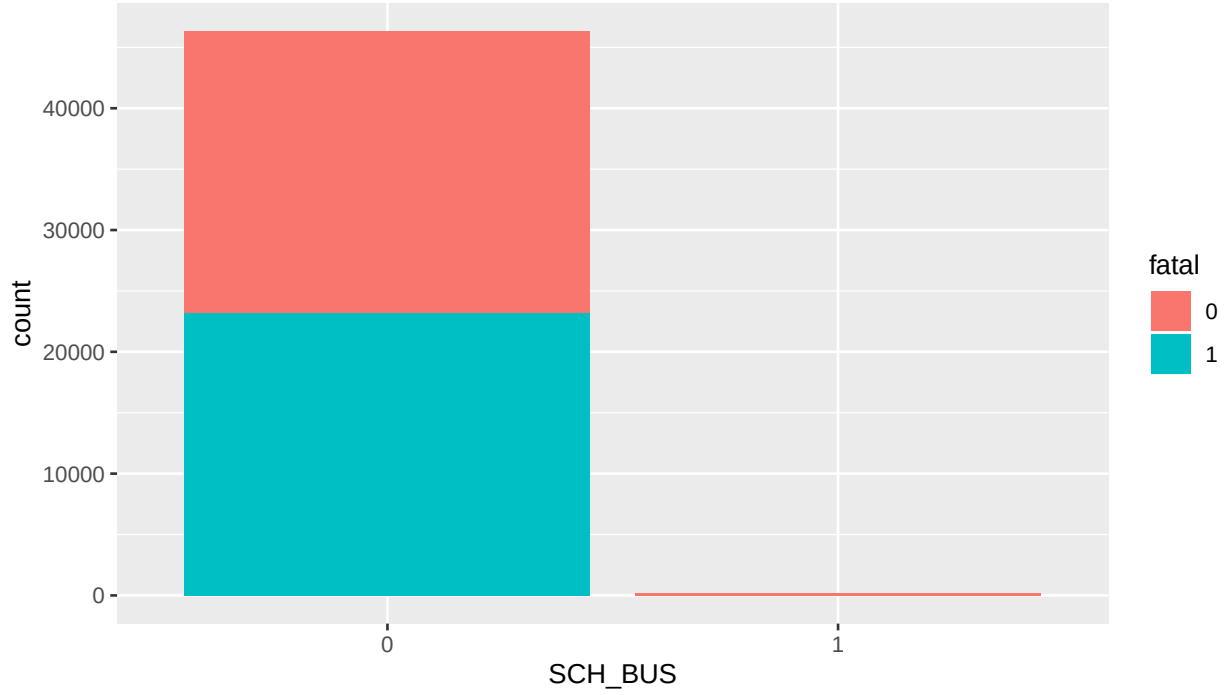

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



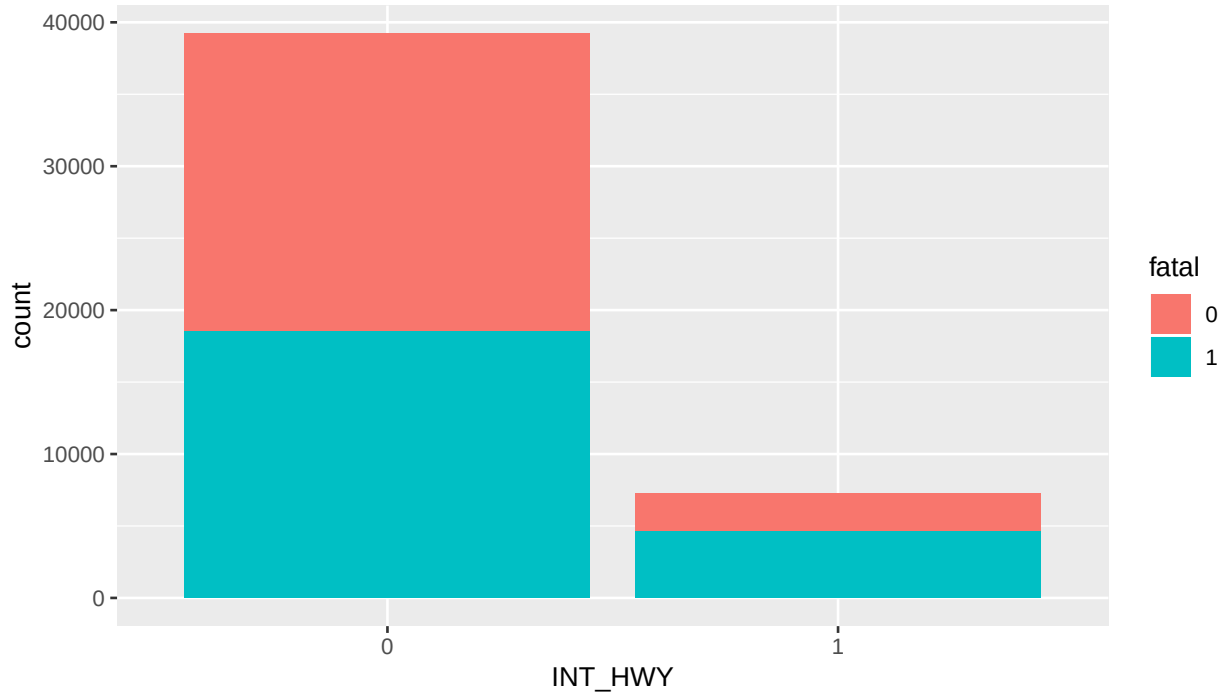

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



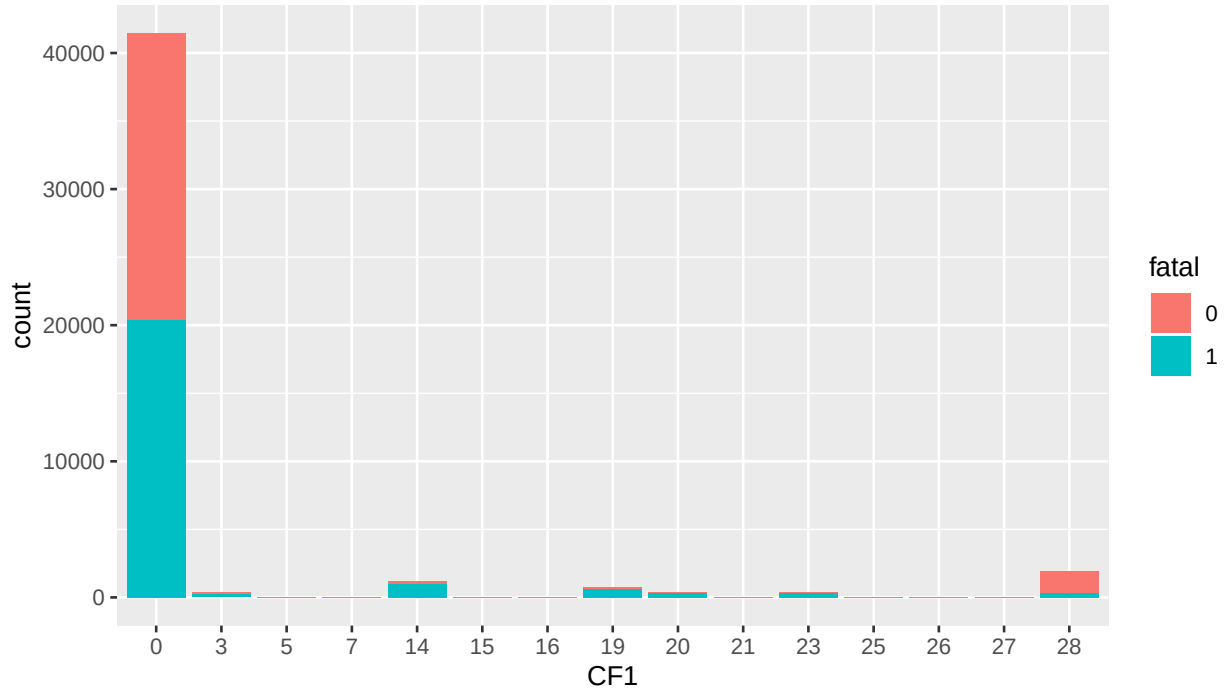

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



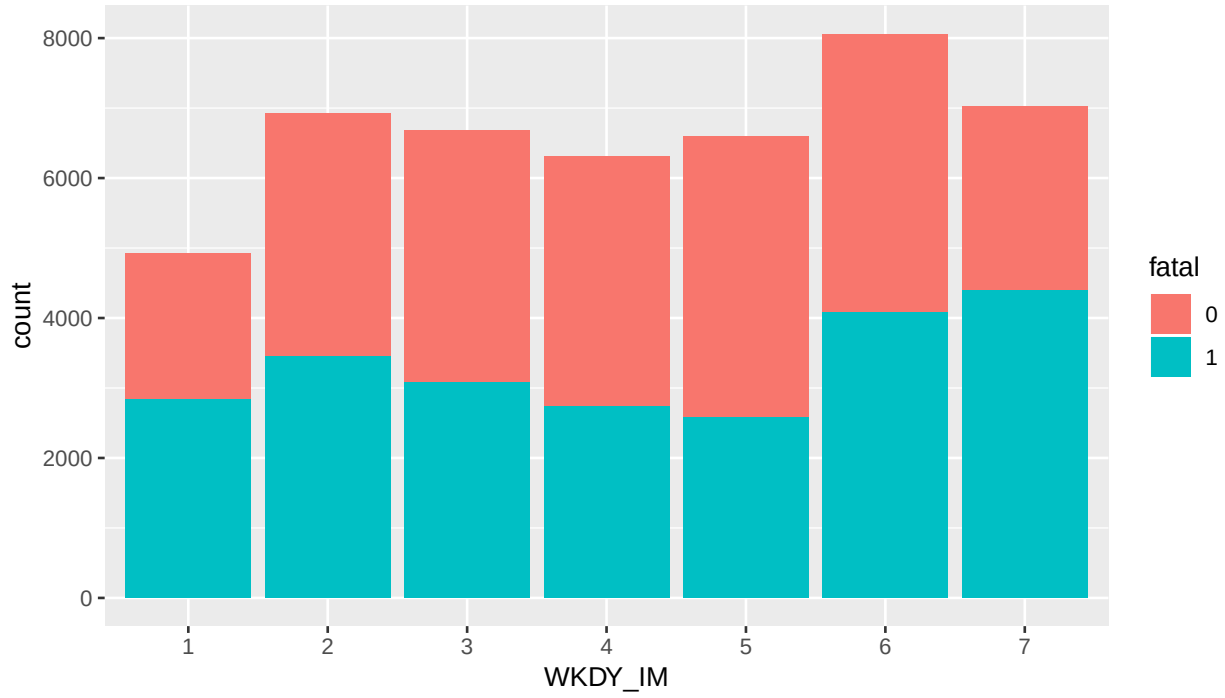

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



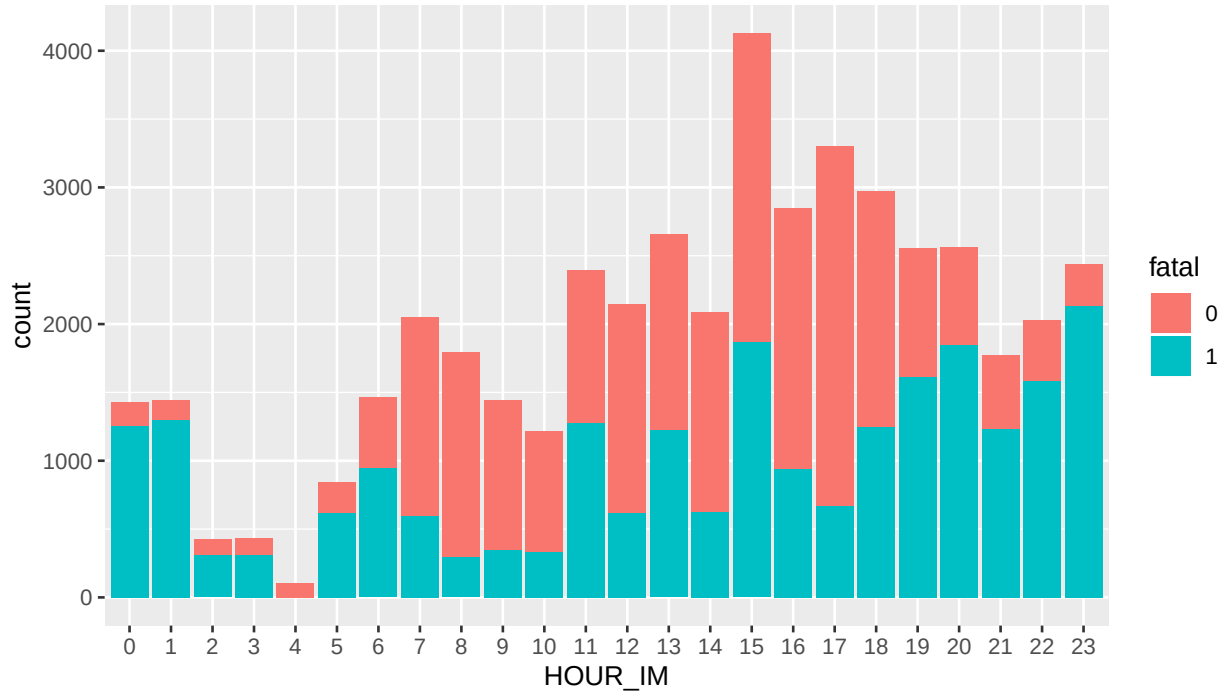

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



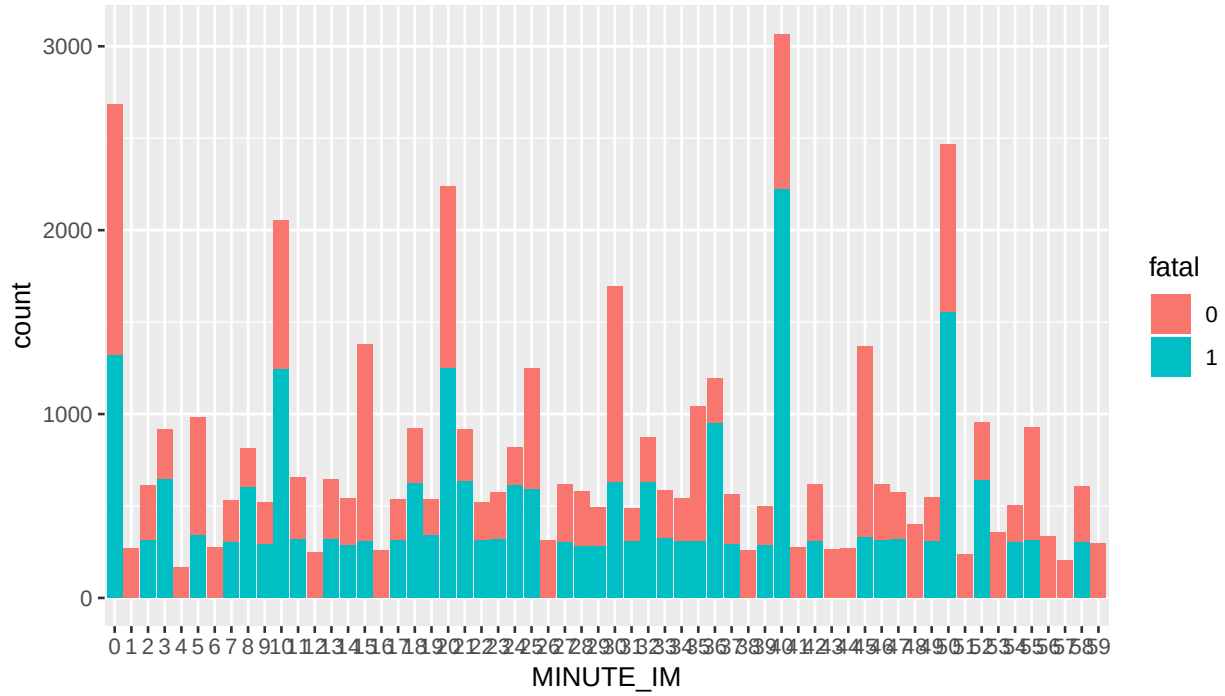

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



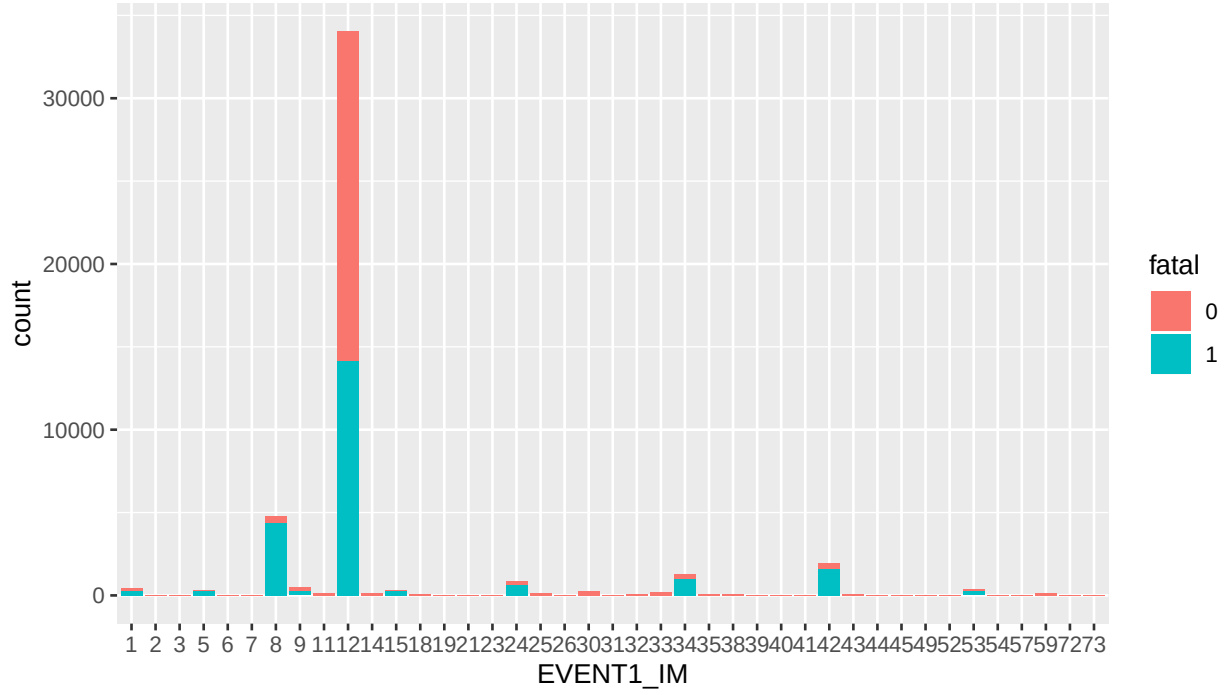

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



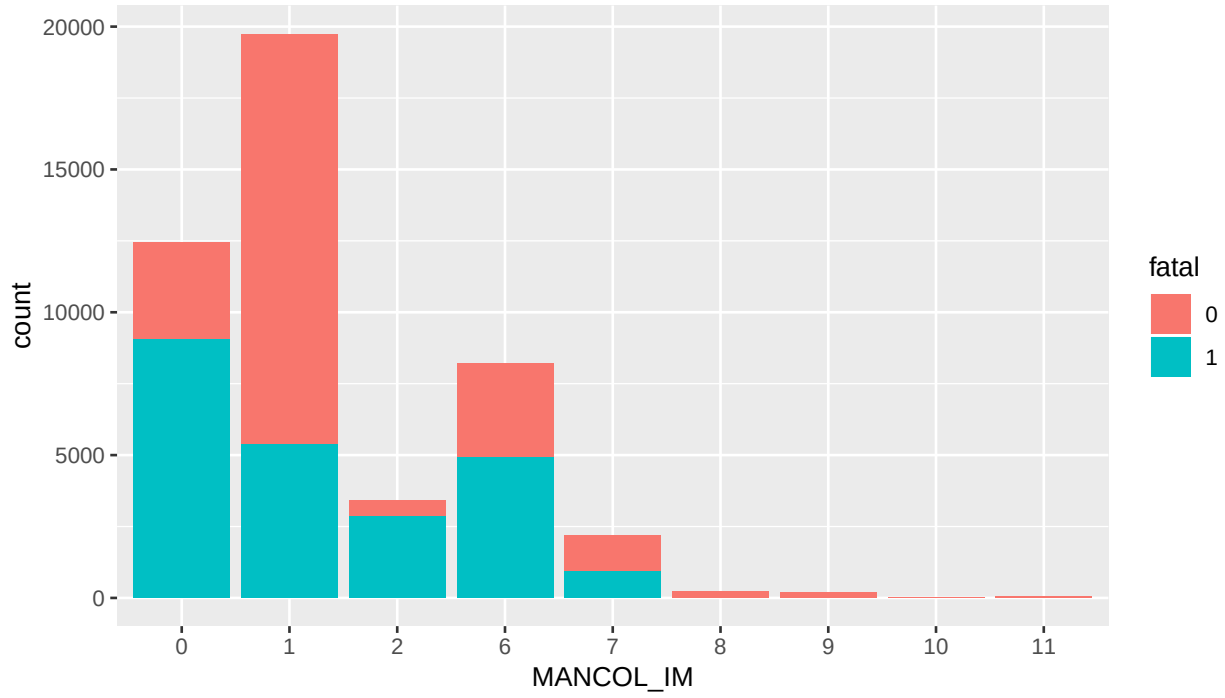

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



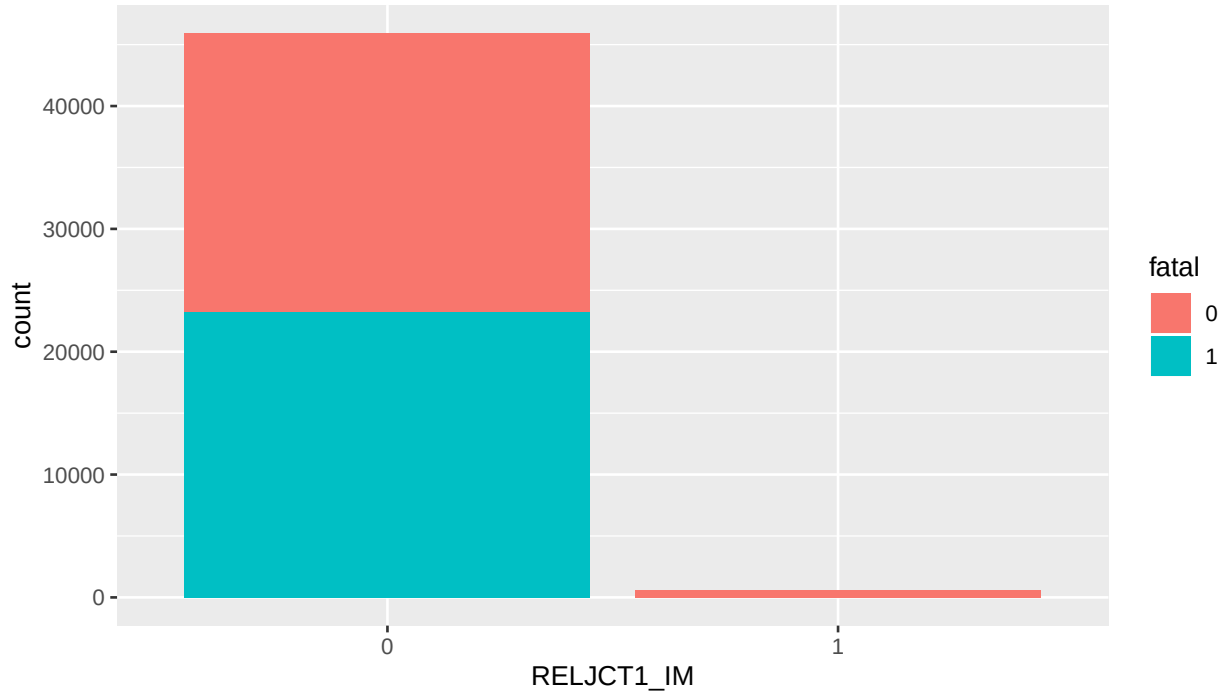

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



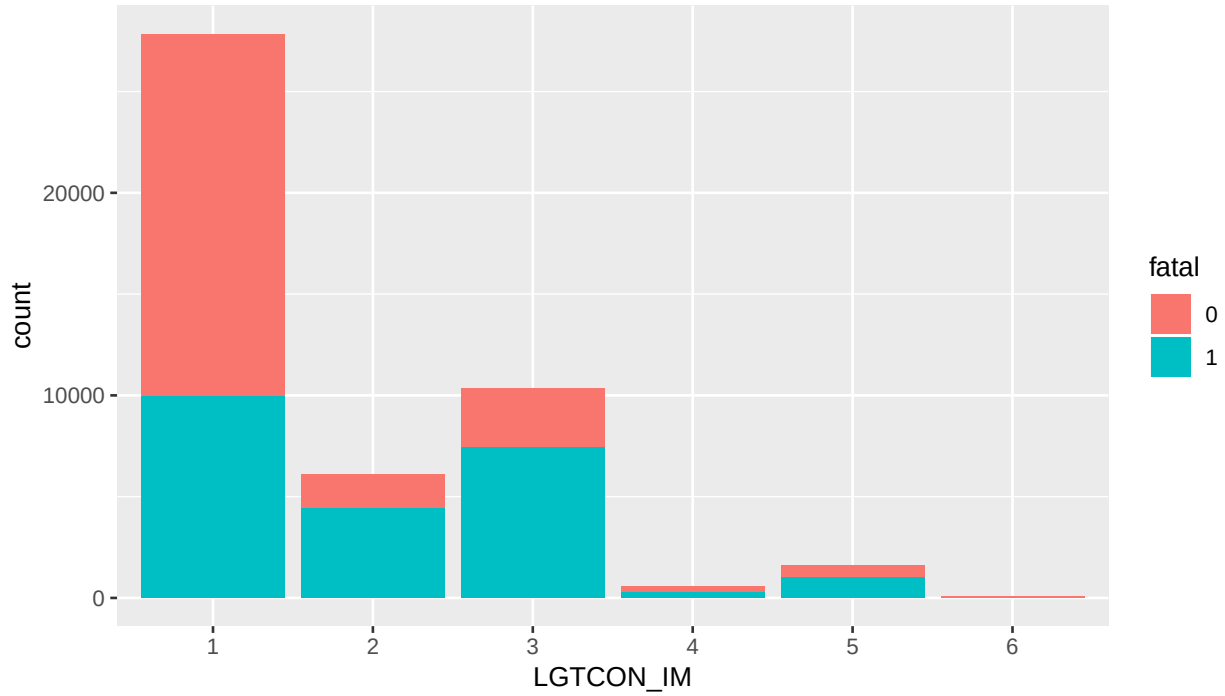

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



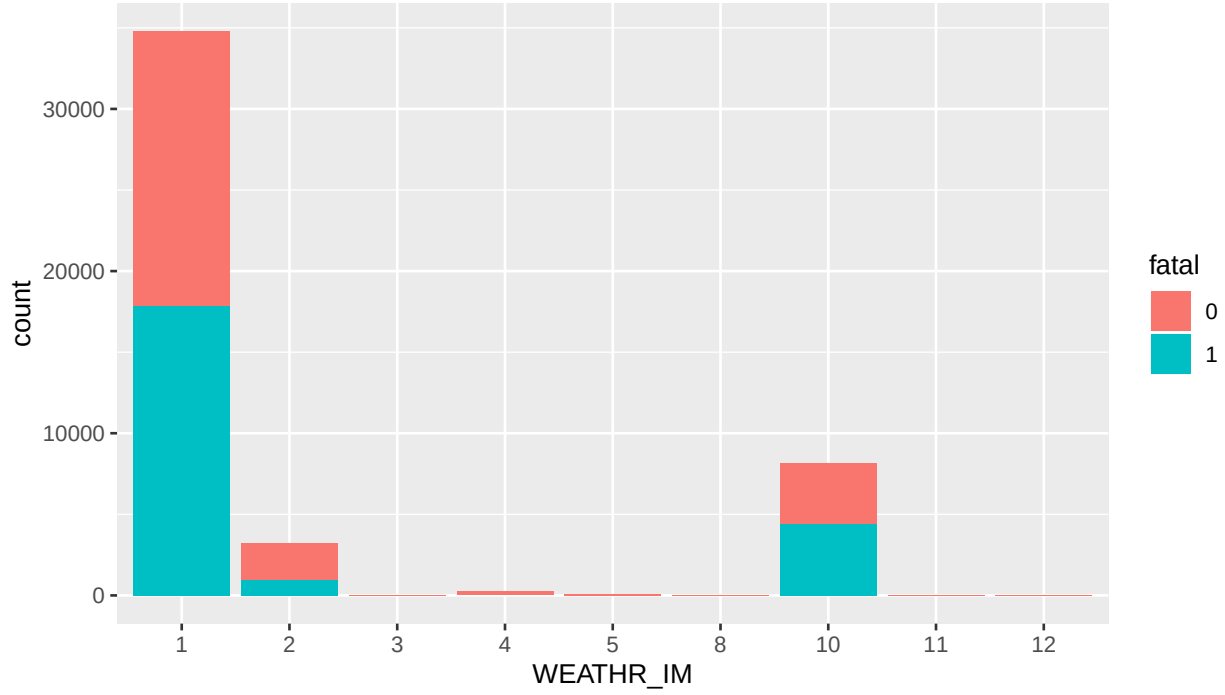

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



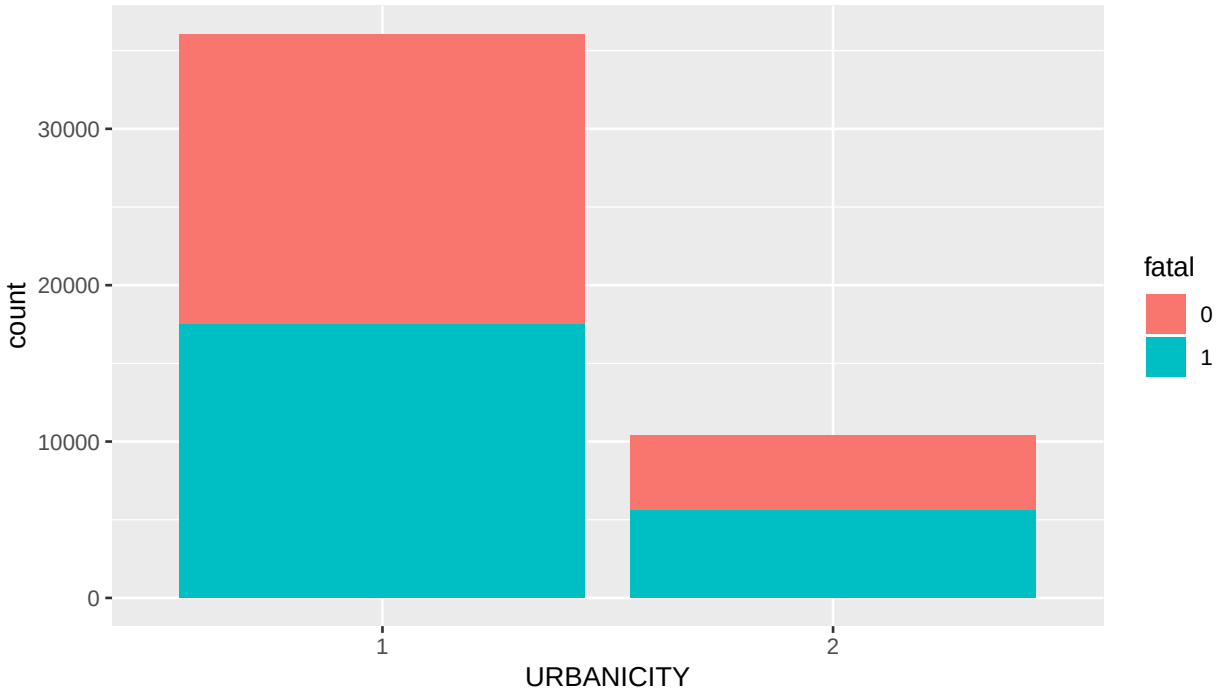

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

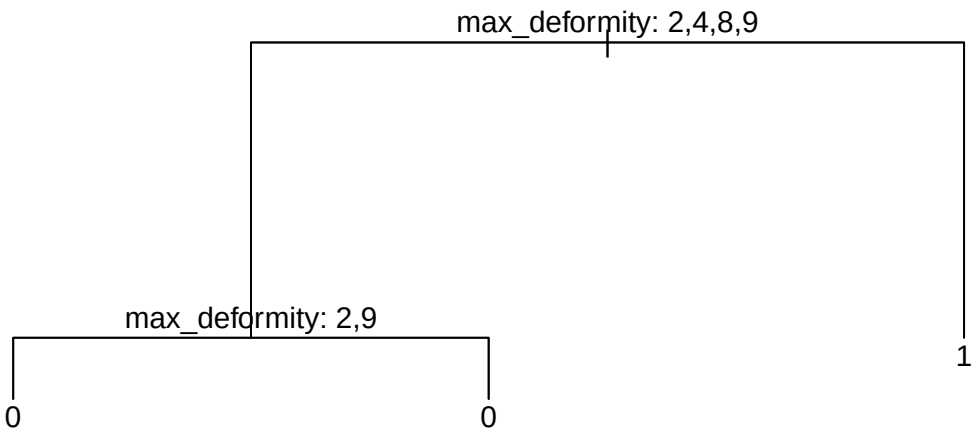



```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

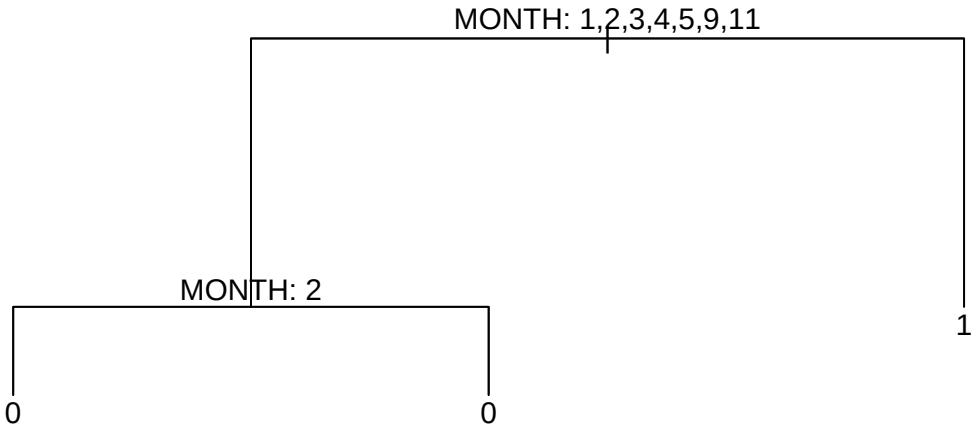


Single Trees for Collapsin Categorical Levels

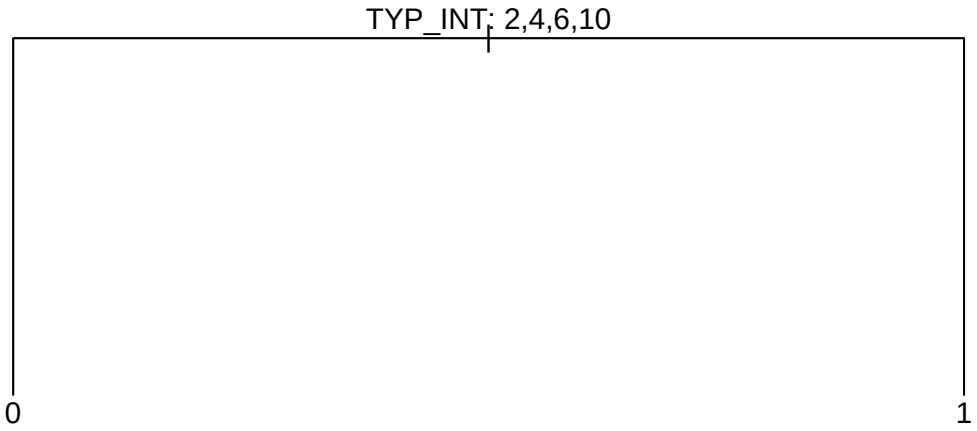
max_deformity



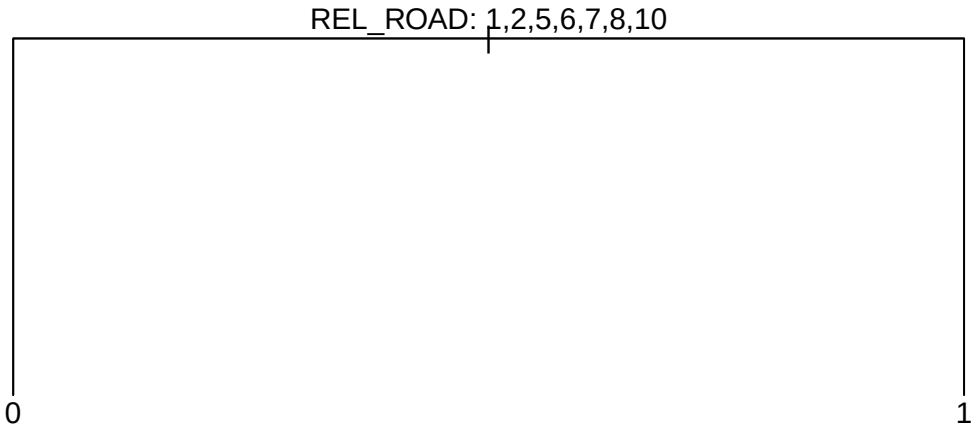
MONTH



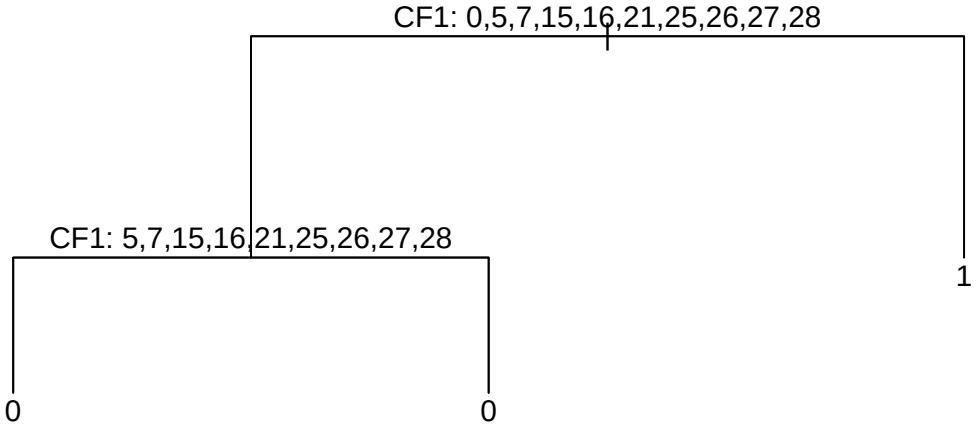
TYP_INT



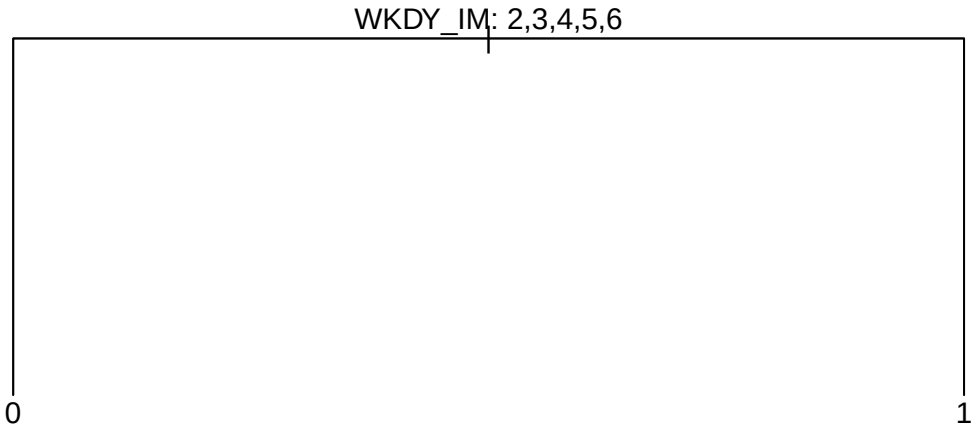
REL_ROAD


REL_ROAD: 1,2,5,6,7,8,10
0
1

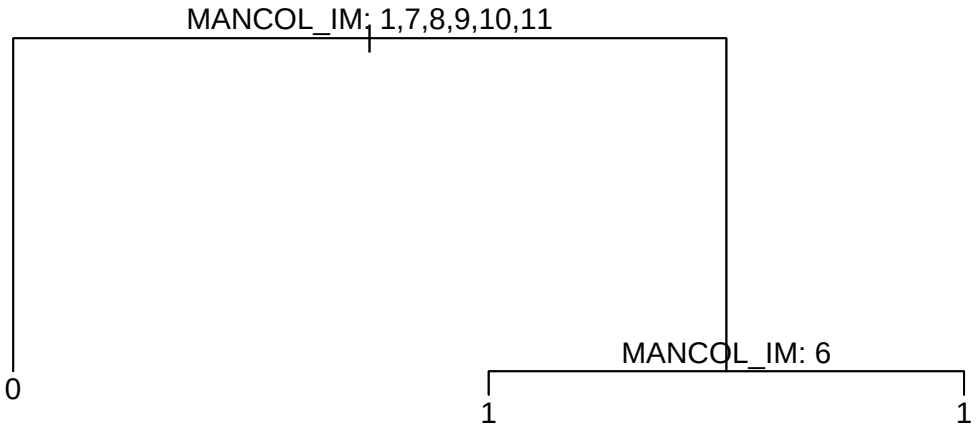
CF1


CF1: 0,5,7,15,16,21,25,26,27,28
CF1: 5,7,15,16,21,25,26,27,28
1
0
0

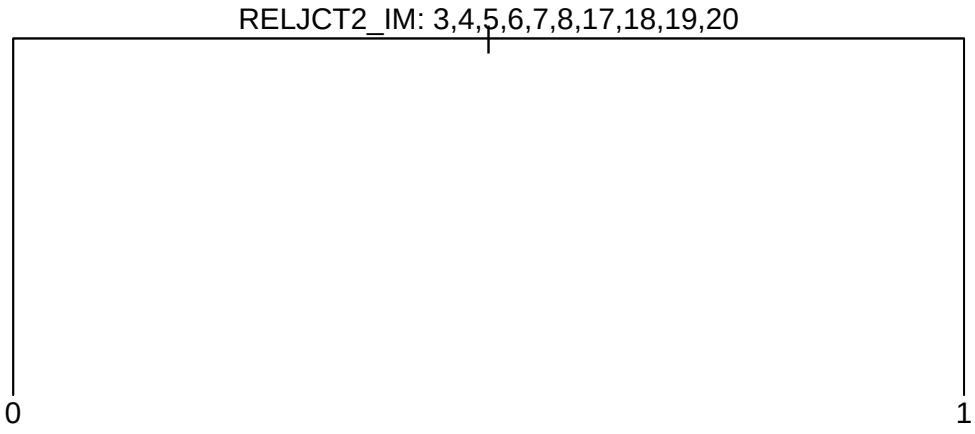
WKDY_IM



MANCOL_IM

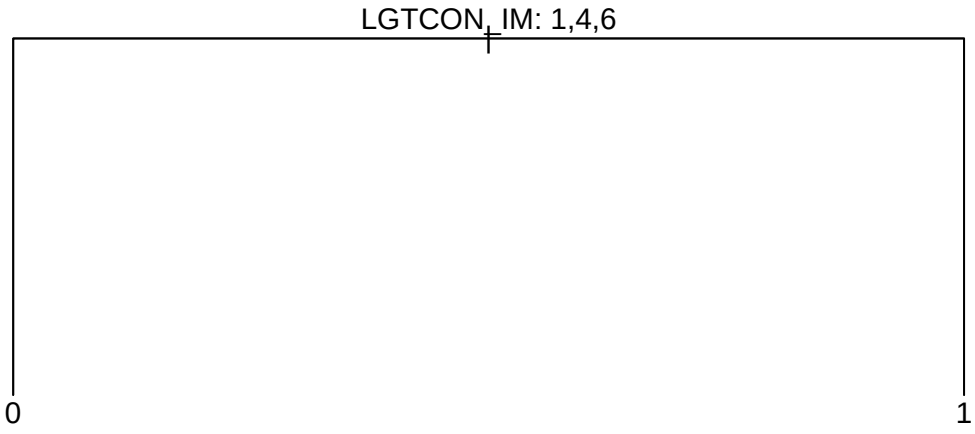


RELJCT2_IM

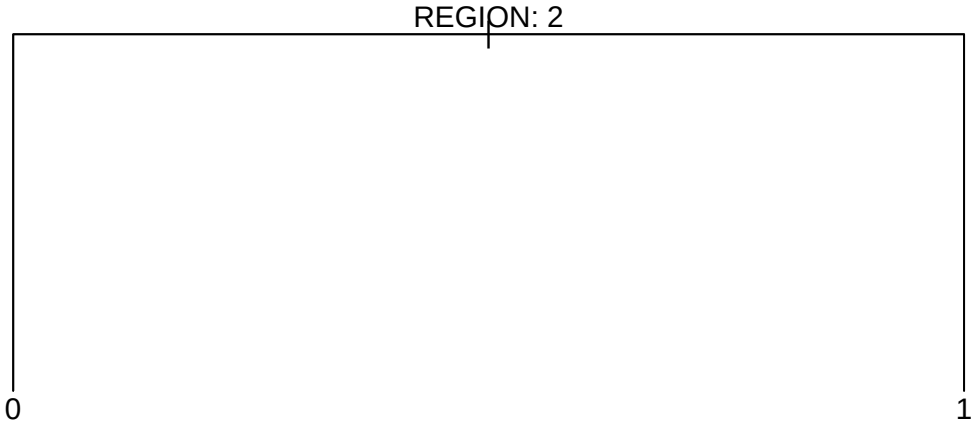


WEATHR_IM

LGTCON_IM



REGION



Relaxed LASSO Output

```
##
## Call:
## glm(formula = fatal ~ rest_mis_indicator + air_bag_indicator +
##     ejection_indicator + fire_indicator + any_distracted + any_maneuver +
```

```
##     any_gradient + any_wet_road + max_speed_limit + speed_related_accident +
##     rollover_indicator + VE_FORMS + PVH_INVL + PEDS + SCH_BUS +
##     INT_HWY + RELJCT1_IM + ALCHL_IM + URBANICITY + hit_run_indicator +
##     max_travel_speed + hour_night + max_deformity_group1 + month_group1 +
##     typ_int_group1 + rel_road_group1 + cf1_group1 + wkdy + mancol_group1 +
##     reljct2_group1 + lgtcon_group1 + precipitation + midwest,
##     family = binomial(logit), data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6223  -0.1575   0.0000   0.3577   1.8419
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.675112   0.194488  -3.471 0.000518 ***
## rest_mis_indicator1       1.448857   0.245159   5.910 3.42e-09 ***
## air_bag_indicator1       -0.743141   0.054964 -13.521  < 2e-16 ***
## ejection_indicator1       4.170564   0.234583  17.779  < 2e-16 ***
## fire_indicator1           0.868778   0.224592   3.868 0.000110 ***
## any_distracted1          -0.550084   0.058579  -9.391  < 2e-16 ***
## any_maneuver1            -0.294623   0.045409  -6.488 8.69e-11 ***
## any_gradient1            -0.414046   0.057868  -7.155 8.37e-13 ***
## any_wet_road1            -0.410000   0.085089  -4.818 1.45e-06 ***
## max_speed_limit           0.026012   0.003290   7.907 2.64e-15 ***
## speed_related_accident1   1.476183   0.060363  24.455  < 2e-16 ***
## rollover_indicator1      -3.405661   0.282677 -12.048  < 2e-16 ***
## rollover_indicator2     -18.854286 387.616696  -0.049 0.961205
## VE_FORMS                  0.832117   0.029794  27.929  < 2e-16 ***
## PVH_INVL                -18.878437 189.252375  -0.100 0.920541
## PEDS                      4.109470   0.107999  38.051  < 2e-16 ***
## SCH_BUS1                -16.048844 234.868693  -0.068 0.945522
## INT_HWY1                 -0.434127   0.077563  -5.597 2.18e-08 ***
## RELJCT1_IM1             -16.050053 136.994187  -0.117 0.906734
## ALCHL_IM2                -1.544856   0.073409 -21.045  < 2e-16 ***
## URBANICITY2              -0.277968   0.054864  -5.067 4.05e-07 ***
## hit_run_indicator1       -0.595062   0.132183  -4.502 6.74e-06 ***
## max_travel_speed          0.054639   0.002081  26.251  < 2e-16 ***
## hour_night1               0.575434   0.078283   7.351 1.97e-13 ***
## max_deformity_group11    -2.721089   0.078955 -34.464  < 2e-16 ***
## month_group11            -1.495163   0.045105 -33.149  < 2e-16 ***
## typ_int_group11          -0.778612   0.058459 -13.319  < 2e-16 ***
## rel_road_group11          0.958439   0.069734  13.744  < 2e-16 ***
## cf1_group11              -0.917338   0.108816  -8.430  < 2e-16 ***
## wkdy1                    -0.299719   0.048242  -6.213 5.20e-10 ***
## mancol_group11           -1.928963   0.061220 -31.508  < 2e-16 ***
## reljct2_group11          -0.299239   0.057383  -5.215 1.84e-07 ***
## lgtcon_group11           -0.969043   0.072487 -13.368  < 2e-16 ***
## precipitation1           -2.387811   0.135625 -17.606  < 2e-16 ***
## midwest1                 -1.941898   0.118705 -16.359  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 51581  on 37207  degrees of freedom
## Residual deviance: 18031  on 37173  degrees of freedom
## AIC: 18101
##
## Number of Fisher Scoring iterations: 16
```