City of Austin Car Crashes & Fatality/Injury Rates

Jakob Long, 2023/12/1

Dataset & Project Information -

Abstract:

Driving in Texas is always an interesting journey, from the jokes about the speed limit being a suggestion, to drivers being incapable of driving during inclement weather. The purpose of this investigation was to assess the frequency of which car crashes occur & if they were fatal or not. This is accompanied by visualizations regarding speed of crash, street types they occured on & the severity of the crash. Additonallly, a logistical model was developed, and cross validated, to predict whether or not a crash was fatal. This dataset was pulled from the public data portal managed by the city of Austin: https://data.austintexas.gov/Transportation-and-Mobility/Austin-Crash-Report-Data-Crash-Level-Records/y2wy-tgr5 (https://data.austintexas.gov/Transportation-and-Mobility/Austin-Crash-Report-Data-Crash-Level-Records/y2wy-tgr5).

This document: https://safety.fhwa.dot.gov/speedmgt/ref_mats/fhwasa1304/Resources3/08%20-%20The%20Relation%20Between%20Speed%20and%20Crashes.pdf (https://safety.fhwa.dot.gov/speedmgt/ref_mats/fhwasa1304/Resources3/08%20-%20The%20Relation%20Between%20Speed%20and%20Crashes.pdf) is a publication from SWOV, the Institute for Road Safety Research located in the Netherlands. Within this publication the research between the fatality of a crash & the speed of the vehicle causing the crash was investigated. This document helped in the choosing of this dataset as it provided a scientifically generalized idea that speed kills. I do however, wish that this dataset contained the weight the passenger cars so I could more closely follow the investigation regarding a vehicles mass when looking at fatal crashes at specific speeds.

Additional Dataset Information:

The dataset is populated by the TXDOT's, (Texas Department of Transportation) Crash Reporting Information System (CRIS), which is populated by reports documented by Police officers throughout the state. This data is holds data going back ten years only within the Austin Area & is managed by the Austin Transportation & Public Works Department.

Additionally, a link is provided within the sources section that aides in understanding the differences between the street types & their naming conventions! Digesting that information is helpful in fully understanding the implications of the results gathered here.

Uploading Dataset:

```
# Uploading datset
CarCrash <- read.csv("Austin_Crash_Report_Data_-_Crash_Level_Records_20231119.csv")
head(CarCrash)
```

```
##   crash_id crash_fatal_fl             crash_date crash_time   case_id
## 1 13961098              N 07/17/2014 07:44:00 PM   19:44:00 141981607
## 2 14110876              N 11/08/2014 01:27:00 AM   01:27:00 143120149
## 3 13650115              N 01/19/2014 08:50:00 PM   20:50:00 140191423
## 4 13635682              N 12/07/2013 03:35:00 AM   03:35:00 133410351
## 5 13596020              N 11/22/2013 09:24:00 PM   21:24:00 133261868
## 6 13563150              N 11/22/2013 11:40:00 AM   11:40:00 133260828
##   rpt_latitude rpt_longitude rpt_block_num rpt_street_pfx rpt_street_name
## 1           NA            NA                            W          HOWARD
## 2           NA            NA                              8011 CAMERON RD
## 3           NA            NA          2400                   NOT REPORTED
## 4           NA            NA                            S         MANCHACA
## 5           NA            NA                              GREAT HILLS TRL
## 6           NA            NA         13700                   NOT REPORTED
##   rpt_street_sfx crash_speed_limit road_constr_zone_fl latitude longitude
## 1             LN               -1                   N       NA        NA
## 2                              -1                   N       NA        NA
## 3           BLVD               60                   N 30.23265 -97.79379
## 4             RD               55                   N 30.15872 -97.83342
## 5            TRL               35                   N 30.39411 -97.74648
## 6            HWY               75                   N 30.47397 -97.77927
##       street_name street_nbr   street_name_2 street_nbr_2 crash_sev_id
## 1    W HOWARD LN             MC CALLEN PASS           NA            2
## 2 8011 CAMERON RD                       N/A           NA            5
## 3          US0290                    SL0343           NA            5
## 4          FM2304      11123       KAISER DR           NA            1
## 5          US0183            GREAT HILLS TRL           NA            5
## 6          SH0045                       N/A           NA            2
##   sus_serious_injry_cnt nonincap_injry_cnt poss_injry_cnt non_injry_cnt
## 1                     0                  1              0             2
## 2                     0                  0              0             1
## 3                     0                  0              0             1
## 4                     1                  0              1             0
## 5                     0                  0              0             1
## 6                     0                  1              0             0
##   unkn_injry_cnt tot_injry_cnt death_cnt contrib_factr_p1_id
## 1              0             1         0                  NA
## 2              0             0         0                  NA
## 3              0             0         0                  NA
## 4              0             2         0                  NA
## 5              0             0         0                  NA
## 6              0             1         0                  NA
##   contrib_factr_p2_id          units_involved
## 1                  NA           Passenger car
## 2                  NA           Passenger car
## 3                  NA           Passenger car
## 4                  NA           Passenger car
## 5                  NA           Passenger car
## 6                  NA Large passenger vehicle
##
atd_mode_category_metadata
## 1           [{"mode_id": 1, "mode_desc": "Passenger car", "unit_id": 2278307, "death_cnt": 0,
```

```
"sus_serious_injry_cnt": 0, "nonincap_injry_cnt": 1, "poss_injry_cnt": 0, "non_injry_cnt": 2, "u
nkn_injry_cnt": 0, "tot_injry_cnt": 1}]
## 2           [{"mode_id": 1, "mode_desc": "Passenger car", "unit_id": 2292812, "death_cnt": 0,
"sus_serious_injry_cnt": 0, "nonincap_injry_cnt": 0, "poss_injry_cnt": 0, "non_injry_cnt": 1, "u
nkn_injry_cnt": 0, "tot_injry_cnt": 0}]
## 3           [{"mode_id": 1, "mode_desc": "Passenger car", "unit_id": 2249542, "death_cnt": 0,
"sus_serious_injry_cnt": 0, "nonincap_injry_cnt": 0, "poss_injry_cnt": 0, "non_injry_cnt": 1, "u
nkn_injry_cnt": 0, "tot_injry_cnt": 0}]
## 4           [{"mode_id": 1, "mode_desc": "Passenger car", "unit_id": 2244002, "death_cnt": 0,
"sus_serious_injry_cnt": 1, "nonincap_injry_cnt": 0, "poss_injry_cnt": 1, "non_injry_cnt": 0, "u
nkn_injry_cnt": 0, "tot_injry_cnt": 2}]
## 5           [{"mode_id": 1, "mode_desc": "Passenger car", "unit_id": 2241305, "death_cnt": 0,
"sus_serious_injry_cnt": 0, "nonincap_injry_cnt": 0, "poss_injry_cnt": 0, "non_injry_cnt": 1, "u
nkn_injry_cnt": 0, "tot_injry_cnt": 0}]
## 6 [{"mode_id": 2, "mode_desc": "Large passenger vehicle", "unit_id": 2239077, "death_cnt": 0,
"sus_serious_injry_cnt": 0, "nonincap_injry_cnt": 1, "poss_injry_cnt": 0, "non_injry_cnt": 0, "u
nkn_injry_cnt": 0, "tot_injry_cnt": 1}]
##   pedestrian_fl motor_vehicle_fl motorcycle_fl bicycle_fl other_fl
## 1                             Y
## 2                             Y
## 3                             Y
## 4                             Y
## 5                             Y
## 6                             Y
##                            point apd_confirmed_fatality
## 1                                                     N
## 2                                                     N
## 3 POINT (-97.79379396 30.23265268)                    N
## 4 POINT (-97.83342355 30.15872477)                    N
## 5 POINT (-97.74647893 30.39410945)                    N
## 6 POINT (-97.77926848 30.47397305)                    N
##   apd_confirmed_death_count motor_vehicle_death_count
## 1                         0                         0
## 2                         0                         0
## 3                         0                         0
## 4                         0                         0
## 5                         0                         0
## 6                         0                         0
##   motor_vehicle_serious_injury_count bicycle_death_count
## 1                                  0                   0
## 2                                  0                   0
## 3                                  0                   0
## 4                                  1                   0
## 5                                  0                   0
## 6                                  0                   0
##   bicycle_serious_injury_count pedestrian_death_count
## 1                            0                      0
## 2                            0                      0
## 3                            0                      0
## 4                            0                      0
## 5                            0                      0
## 6                            0                      0
```

```
##   pedestrian_serious_injury_count motorcycle_death_count
## 1                               0                      0
## 2                               0                      0
## 3                               0                      0
## 4                               0                      0
## 5                               0                      0
## 6                               0                      0
##   motorcycle_serious_injury_count other_death_count other_serious_injury_count
## 1                               0                 0                         0
## 2                               0                 0                         0
## 3                               0                 0                         0
## 4                               0                 0                         0
## 5                               0                 0                         0
## 6                               0                 0                         0
##   onsys_fl private_dr_fl micromobility_serious_injury_count
## 1        N             N                                  0
## 2        N             N                                  0
## 3        Y             N                                  0
## 4        Y             N                                  0
## 5        Y             N                                  0
## 6        Y             N                                  0
##   micromobility_death_count micromobility_fl
## 1                         0
## 2                         0
## 3                         0
## 4                         0
## 5                         0
## 6                         0
```

```
dim(CarCrash)
```

```
## [1] 148482     54
```

I neglected original structure & summary of the dataset, to avoid clutter. But from the table above & the dimensions we can see that the original dataset contained 148,482 observations with 54 different variables for each observation. This is a lot of data to tidy!

Cleaning Data -

Tidying our Dataset:

```
# Reduces columns in dataset to those that are relevant for our investigation.
CarCrash <- CarCrash |>
  dplyr::select(crash_fatal_fl,rpt_street_sfx,crash_speed_limit,crash_sev_id,
         tot_injry_cnt,units_involved,motor_vehicle_fl,motorcycle_fl,road_constr_zone_fl) |>
  # removes rows that have a nonsense speed limit value & street suffix
  # removes rows that dont involve a Passenger Car
  filter(!is.na(crash_speed_limit),
         crash_speed_limit > -1,
         rpt_street_sfx != '',
         motor_vehicle_fl == 'Y')
```

```
CarCrash <- CarCrash |>
  # renames variables for easier/quicker understanding
  rename(Fatal_Crash = crash_fatal_fl, Street_Type = rpt_street_sfx,
         Speed_Limit = crash_speed_limit, Crash_Severity = crash_sev_id,
         Total_Injured = tot_injry_cnt, Construction_Zone = road_constr_zone_fl) |>
  # Factorizing street types & renaming Crash_Severity to a meaningful description & factorizes
it
  mutate(Street_Type = as.factor(Street_Type))
```

A significant amount of data tidying was not necessary, so I simply removed unnecessary columns, and reduced the number of observations as shown above. I additionally added levels to the Street_Type variable so that the logistical model would account for the differences between streets & their frequency in which crashes occur. I would have done the same for the units_involved variable; however, since each observation is a unique car crash the types of vehicles involved is too diverse to be able to factorize the variable for meaningful & effective information.

Printing Post-Cleaned Data:

```
summary(CarCrash)
```

```
##  Fatal_Crash         Street_Type     Speed_Limit    Crash_Severity
##  Length:74385       BLVD   :13461   Min.   : 0.00   Min.   : 0.000
##  Class :character   RD     :11170   1st Qu.:35.00   1st Qu.: 2.000
##  Mode  :character   ST     :10932   Median :40.00   Median : 5.000
##                     LN     :10356   Mean   :43.21   Mean   : 3.716
##                     HWY    :10187   3rd Qu.:50.00   3rd Qu.: 5.000
##                     DR     : 8298   Max.   :80.00   Max.   :99.000
##                     (Other): 9981
##  Total_Injured     units_involved     motor_vehicle_fl   motorcycle_fl
##  Min.   : 0.0000   Length:74385       Length:74385       Length:74385
##  1st Qu.: 0.0000   Class :character   Class :character   Class :character
##  Median : 0.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 0.6475
##  3rd Qu.: 1.0000
##  Max.   :18.0000
##
##  Construction_Zone
##  Length:74385
##  Class :character
##  Mode  :character
##
##
##
##
```

```
str(CarCrash)
```

```
## 'data.frame':    74385 obs. of  9 variables:
##  $ Fatal_Crash      : chr  "N" "N" "N" "N" ...
##  $ Street_Type      : Factor w/ 18 levels "AVE","BLVD","CIR",..: 2 15 17 9 7 10 15 7 6 2 ...
##  $ Speed_Limit      : int  60 55 35 75 65 30 35 65 35 30 ...
##  $ Crash_Severity   : int  5 1 5 2 5 5 5 5 3 5 ...
##  $ Total_Injured    : int  0 2 0 1 0 0 0 0 2 0 ...
##  $ units_involved   : chr  "Passenger car" "Passenger car" "Passenger car" "Large passenger v
ehicle" ...
##  $ motor_vehicle_fl : chr  "Y" "Y" "Y" "Y" ...
##  $ motorcycle_fl    : chr  "" "" "" "" ...
##  $ Construction_Zone: chr  "N" "N" "N" "N" ...
```

```
dim(CarCrash)
```

```
## [1] 74385     9
```

After tidying our data we reduced our number of observations from 148,482 -> 74,385, and down from 54 variables to 9! This allows for faster computation & hopefully an increase in our stability of our model and interpret-ability!

Data Analysis:

Investigating a predictive model for fatal crashes:

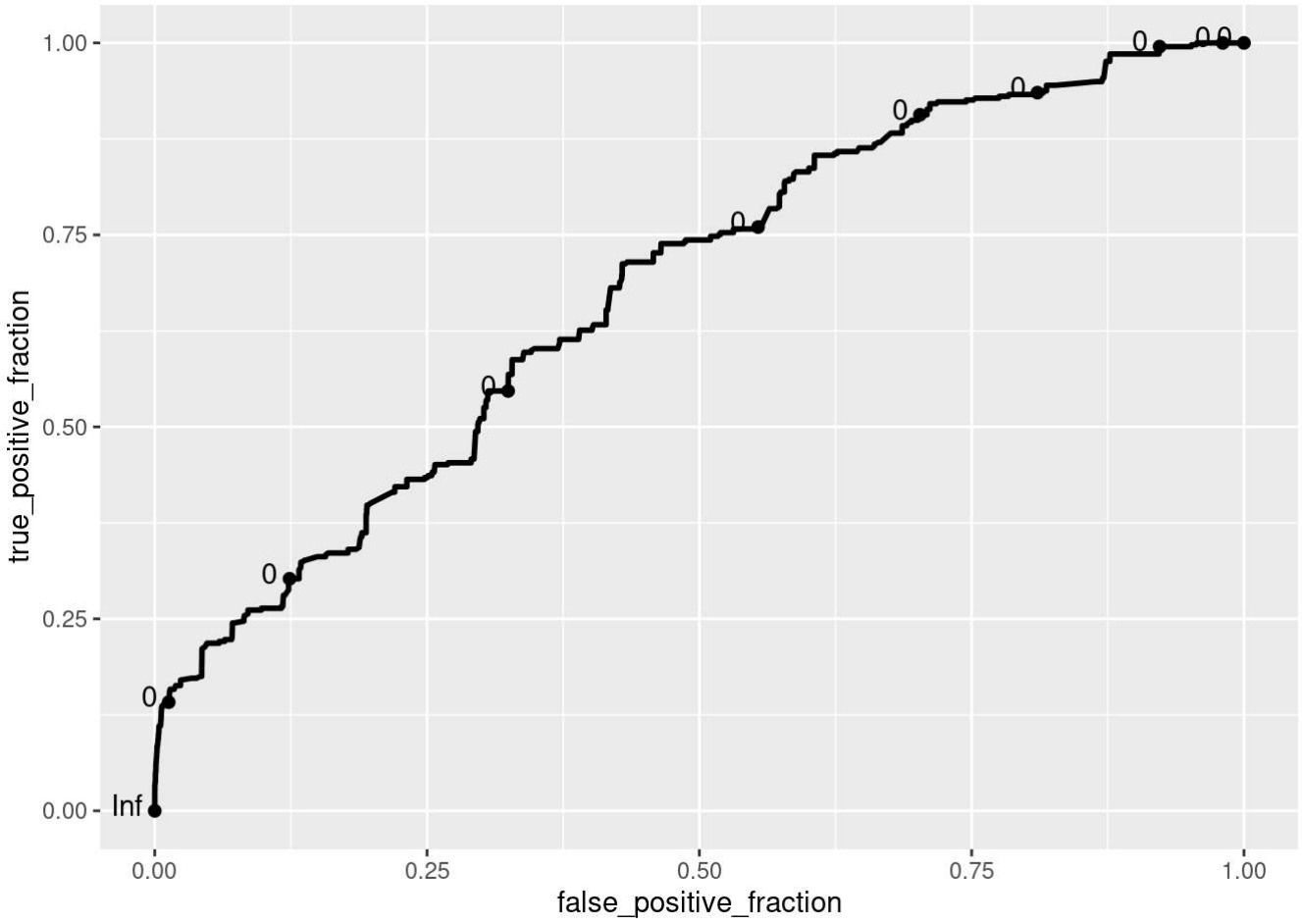
```
CarCrash <- CarCrash |>
  mutate(Fatal_Crash = ifelse(Fatal_Crash == 'Y',1,0),
         motorcycle_fl = ifelse(motorcycle_fl == 'Y',1,0),
         Construction_Zone = ifelse(Construction_Zone == 'Y',1,0))
crash_fatal_reg <- glm(Fatal_Crash ~ Street_Type + Speed_Limit +
                   Crash_Severity + Total_Injured + motorcycle_fl + Construction_Zone,
                   data = CarCrash,
                   family = 'binomial')
summary(crash_fatal_reg)
```

```
##
## Call:
## glm(formula = Fatal_Crash ~ Street_Type + Speed_Limit + Crash_Severity +
##     Total_Injured + motorcycle_fl + Construction_Zone, family = "binomial",
##     data = CarCrash)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -6.597e+00  3.163e-01 -20.858  < 2e-16 ***
## Street_TypeBLVD   -1.037e-01  2.675e-01  -0.388   0.6982
## Street_TypeCIR    -1.312e+01  6.003e+02  -0.022   0.9826
## Street_TypeCT     -1.315e+01  9.540e+02  -0.014   0.9890
## Street_TypeCV     -1.320e+01  8.597e+02  -0.015   0.9877
## Street_TypeDR     -1.151e-01  2.852e-01  -0.404   0.6865
## Street_TypeEXPY   -8.269e-01  3.545e-01  -2.332   0.0197 *
## Street_TypeFWY    -1.393e+01  3.004e+02  -0.046   0.9630
## Street_TypeHWY    -3.999e-01  2.951e-01  -1.355   0.1753
## Street_TypeLN     -2.121e-02  2.713e-01  -0.078   0.9377
## Street_TypeLOOP   -1.381e+01  3.645e+02  -0.038   0.9698
## Street_TypePARK   -1.320e+01  1.110e+03  -0.012   0.9905
## Street_TypePKWY    1.478e-01  4.140e-01   0.357   0.7212
## Street_TypePL     -1.327e+01  6.488e+02  -0.020   0.9837
## Street_TypeRD     -1.626e-01  2.713e-01  -0.599   0.5489
## Street_TypeST     -6.899e-01  2.964e-01  -2.328   0.0199 *
## Street_TypeTRL    -1.325e+01  3.304e+02  -0.040   0.9680
## Street_TypeWAY     2.970e-01  1.033e+00   0.287   0.7737
## Speed_Limit        2.719e-02  5.402e-03   5.034 4.80e-07 ***
## Crash_Severity     8.307e-02  1.082e-02   7.679 1.61e-14 ***
## Total_Injured     -5.166e-02  6.071e-02  -0.851   0.3948
## motorcycle_fl      2.426e+00  1.517e-01  15.987  < 2e-16 ***
## Construction_Zone  1.309e-01  2.335e-01   0.561   0.5751
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5155.1  on 74384  degrees of freedom
## Residual deviance: 4818.7  on 74362  degrees of freedom
## AIC: 4864.7
##
## Number of Fisher Scoring iterations: 17
```

Here I've constructed a logistic regression model to predict whether a crash is fatal or not. From the summary output above we're able to see that only 3 of our variables are statistically significant towards the model: MotorCycle_fl, Crash_Severity, Speed_Limit, while two others are statistically significant to the base level which for this model was Street_TypeAVE: Street_TypeST & Street_TypeEXPY.

Investigating Model Performance:

```
ROC_crash_fatal <- CarCrash |>
  # Make predictions
  mutate(probability = predict(crash_fatal_reg, type = "response")) |>
  ggplot() +
  geom_roc(aes(d = Fatal_Crash, m = probability), n.cuts = 10)
ROC_crash_fatal
```




```
calc_auc(ROC_crash_fatal)$AUC
```

```
## [1] 0.6774029
```

```
# Make this example reproducible by setting a seed
set.seed(322)

# Choose number of folds
k = 5

# Randomly order rows in the dataset
data <- CarCrash[sample(nrow(CarCrash)), ]

# Create k folds from the dataset
folds <- cut(seq(1:nrow(data)), breaks = k, labels = FALSE)

# Initialize a vector to keep track of the performance
perf_k <- NULL

# Use a for loop to get diagnostics for each test dataset
for(i in 1:k){
  # Create train and test datasets
  train_not_i <- data[folds != i, ] # all observations except in fold i
  test_i <- data[folds == i, ]  # all observations in fold i

  # Train model on train data (all but fold i)
  train_model_reg <- glm(Fatal_Crash ~ Street_Type + Speed_Limit +
                         Crash_Severity + Total_Injured + motorcycle_fl,
                         data = train_not_i,
                         family = 'binomial')

  # Performance listed for each test data (fold i)
  perf_k[i] <- sqrt(mean((
    test_i$Fatal_Crash - predict(train_model_reg, newdata = test_i))^2,
    na.rm = TRUE))

}

# Average performance over all k folds and variation
round(mean(perf_k),digits = 2)
```

```
## [1] 5.99
```

```
round(sd(perf_k),digits = 2)
```

```
## [1] 0.07
```

From above we can see that our model is not the best performing model, as it had an ROC AUC score of 0.677, or about 0.68. Meaning that the model has a 68% chance of accurately predicting if a crash will result in a fatality or not. This is not exactly ideal, as theres a significant margin of error when using this model.

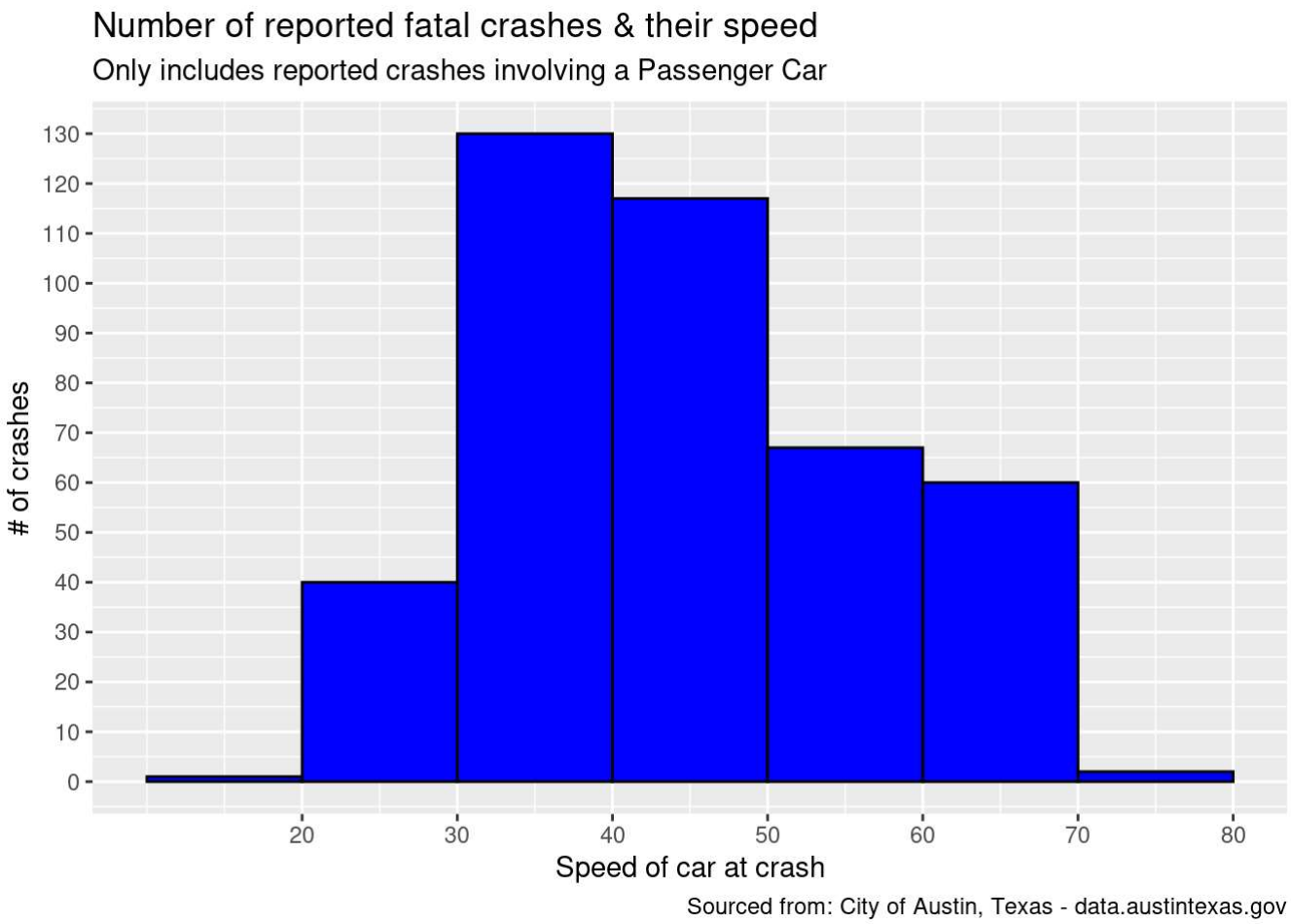
However, when the k-folds cross validation was performed the model performed fairly well across multiple test sets! As our standard deviation value was close to zero, meaning that across the 5 different folds our model had approximately the exact same performance! Meaning that the model is okay for testing new observations.

Investigating speed of crashes & their fatality:

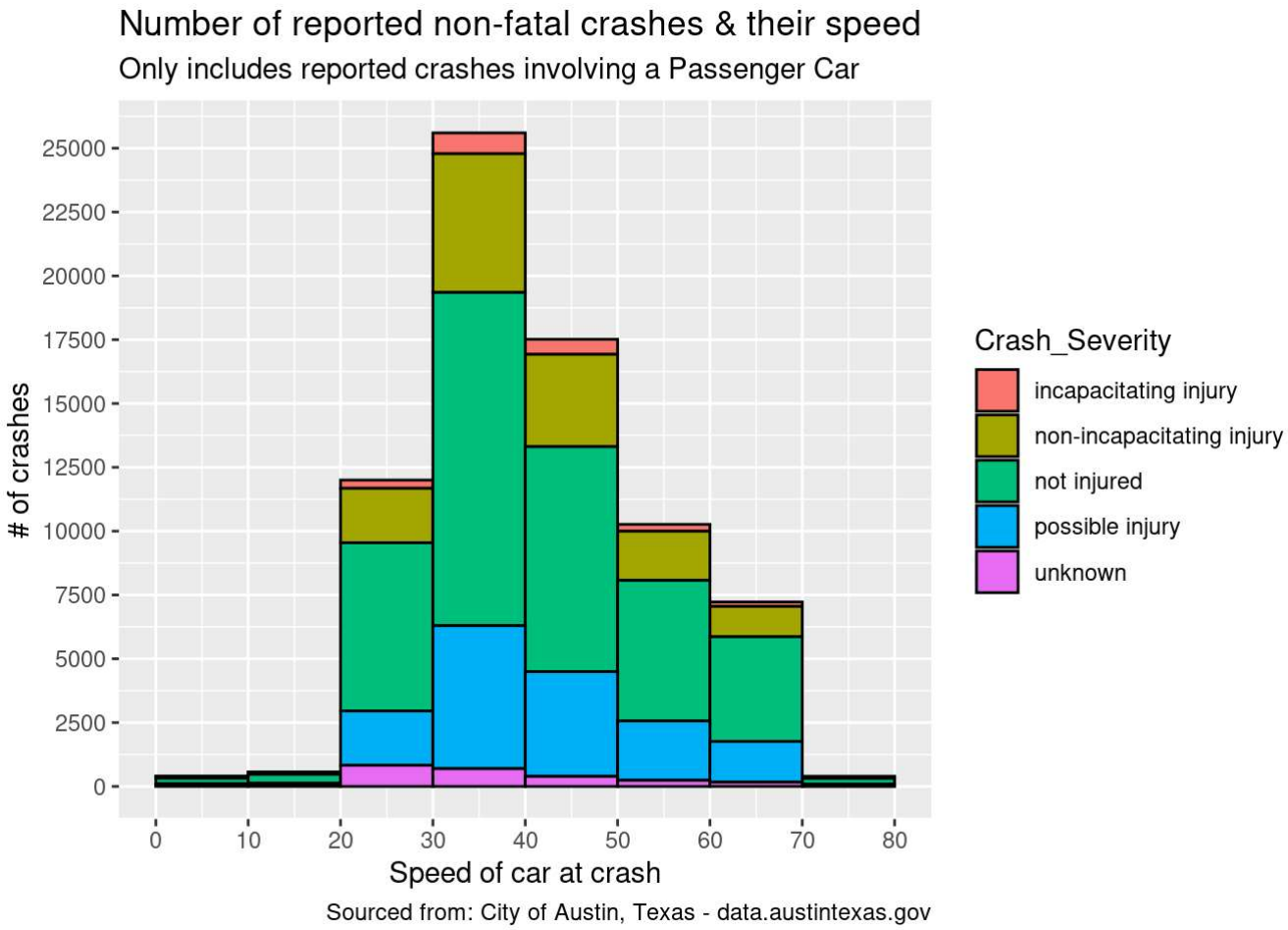
```
# Reinitializing Datframe from 1/0 -> Y/N & Renaming Crash_Severity for meaningful descriptions
CarCrash <- CarCrash |>
  mutate(Crash_Severity = recode(Crash_Severity,
                                 `0` = "unknown",
                                 `1` = "incapacitating injury",
                                 `2` = "non-incapacitating injury",
                                 `3` = "possible injury",
                                 `4` = "killed",
                                 `5` = "not injured",
                                 .default = 'unknown'),
         Fatal_Crash = ifelse(Fatal_Crash == '1','Y','N'))

# Getting percentages of fatal Crashes
Crash_Fatality <- CarCrash |>
  group_by(Fatal_Crash) |>
  summarize(Percentage = round((n()/nrow(CarCrash))*100,digits = 4))
Crash_Fatality
```

```
## # A tibble: 2 × 2
##   Fatal_Crash Percentage
##   <chr>            <dbl>
## 1 N               99.4
## 2 Y                0.561
```

```
# Plots distribution frequency of crash speed for crashes that were fatal
CarCrash |>
  filter(Fatal_Crash == 'Y') |>
  ggplot() +
  geom_histogram(aes(x = Speed_Limit),
                 binwidth = 10,
                 center = 5,
                 color = 'black',
                 fill = 'blue') +
  scale_x_continuous(breaks = seq(20,80,10)) +
  scale_y_continuous(breaks = seq(0,140,10)) +
  labs(x = 'Speed of car at crash',
       y = '# of crashes',
       title = 'Number of reported fatal crashes & their speed',
       subtitle = 'Only includes reported crashes involving a Passenger Car',
       caption = 'Sourced from: City of Austin, Texas - data.austintexas.gov')
```




```
# Plots distribution frequency of crash speed for crashes that were non-fatal
CarCrash |>
  filter(Fatal_Crash == 'N',
         !Crash_Severity == 'killed') |>
  ggplot() +
  geom_histogram(aes(x = Speed_Limit,fill = Crash_Severity),
                 binwidth = 10,
                 center = 5,
                 color = 'black') +
  scale_x_continuous(breaks = seq(0,80,10)) +
  scale_y_continuous(breaks = seq(0,25000,2500)) +
  labs(x = 'Speed of car at crash',
       y = '# of crashes',
       title = 'Number of reported non-fatal crashes & their speed',
       subtitle = 'Only includes reported crashes involving a Passenger Car',
       caption = 'Sourced from: City of Austin, Texas - data.austintexas.gov')
```



From the two above plots we're able to see that a majority of all crashes happen within the range of 30 Mph, and 50 Mph. Which makes sense, as drivers are typically more alert and aware when driving at both slower and faster speeds. As driving at a faster speed requires more caution and control over the car, while slower speeds have a larger tolerance for reaction timing for braking or swerving to avoid a crash.

Interestingly enough we also see that both the fatal and non-fatal crashes share a similarly shaped histogram, that appears to have a right skewed distribution.

Additionally, we can see from tibble above, less than 1% of the reported crashes we're observing resulted in a reported fatality.

Investigating street types & crash frequency:

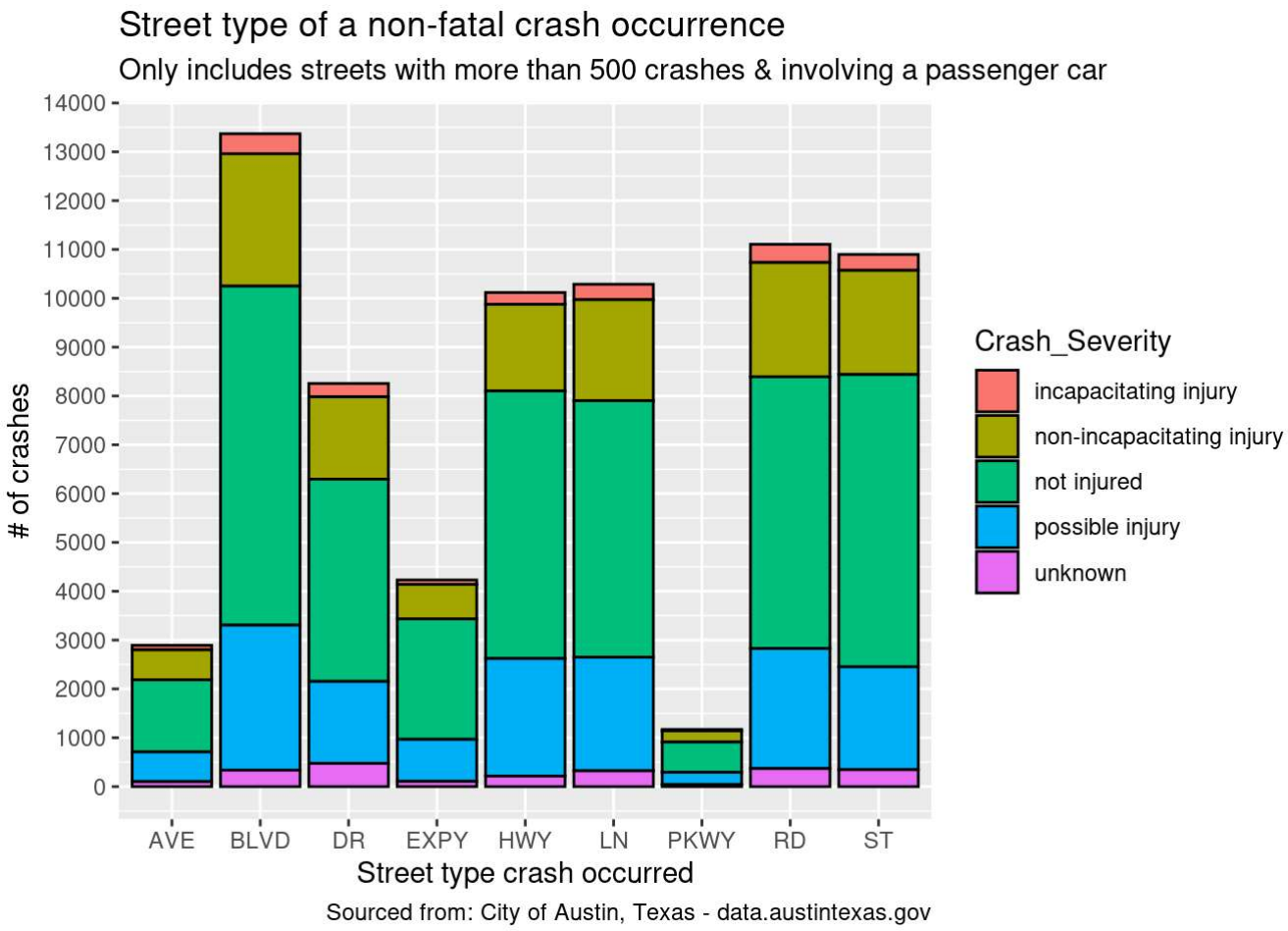
```
Crash_Fatality <- CarCrash |>
  filter(Fatal_Crash == 'Y') |>
  group_by(Street_Type) |>
  summarize(Percentage = round((n()/nrow(CarCrash))*100,digits = 4)) |>
  arrange(Percentage)
Crash_Fatality
```

```
## # A tibble: 10 × 2
##    Street_Type Percentage
##    <fct>            <dbl>
##  1 WAY             0.0013
##  2 PKWY            0.0121
##  3 AVE             0.0255
##  4 EXPY            0.0255
##  5 ST              0.0471
##  6 DR              0.0565
##  7 RD              0.0874
##  8 LN              0.0914
##  9 HWY             0.0928
## 10 BLVD            0.121
```

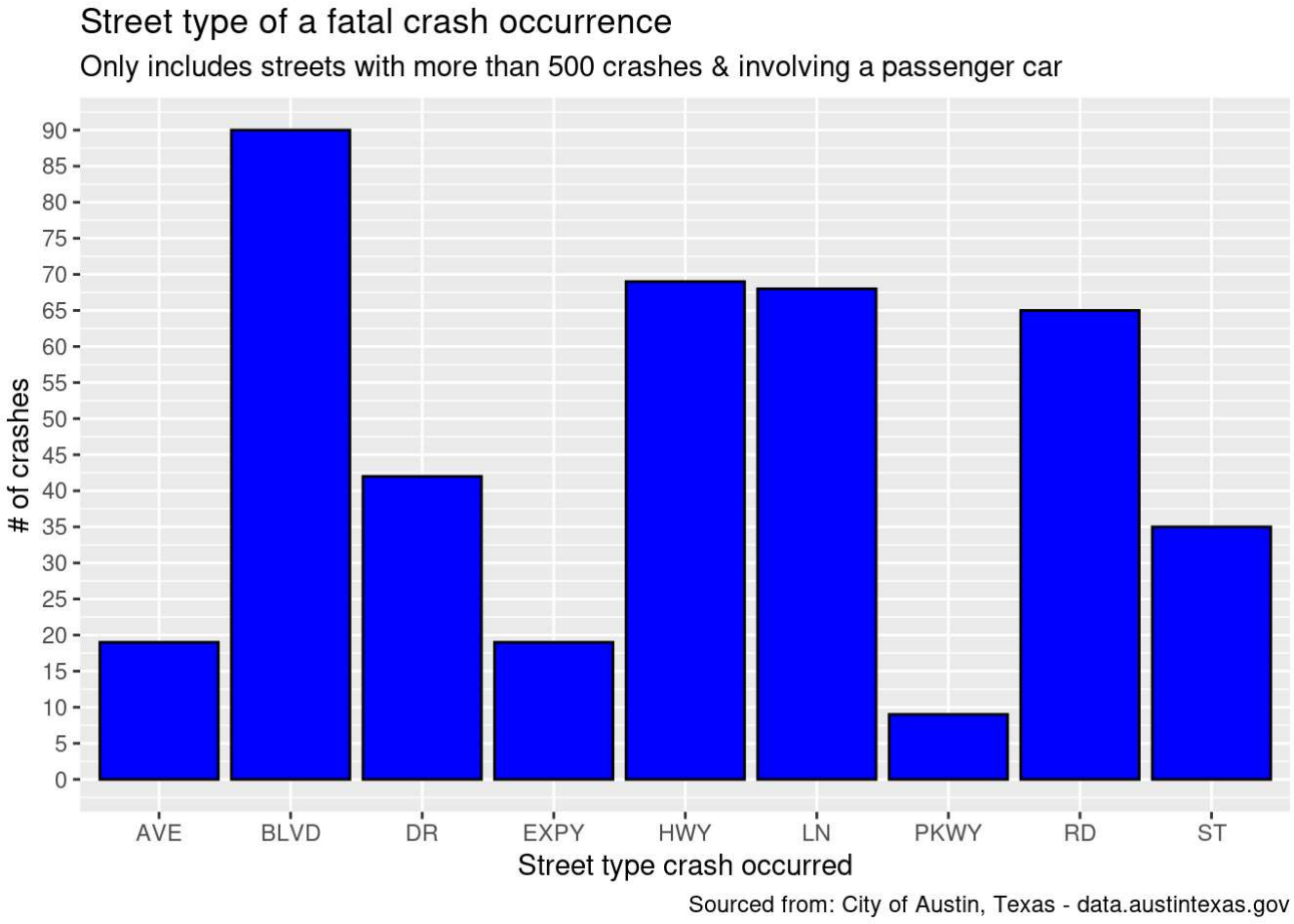
```
CarCrash |>
  group_by(Street_Type)
```

```
## # A tibble: 74,385 × 9
## # Groups:   Street_Type [18]
##    Fatal_Crash Street_Type Speed_Limit Crash_Severity            Total_Injured
##    <chr>       <fct>             <int> <chr>                           <int>
##  1 N           BLVD                 60 not injured                         0
##  2 N           RD                   55 incapacitating injury               2
##  3 N           TRL                  35 not injured                         0
##  4 N           HWY                  75 non-incapacitating injury           1
##  5 N           EXPY                 65 not injured                         0
##  6 N           LN                   30 not injured                         0
##  7 N           RD                   35 not injured                         0
##  8 N           EXPY                 65 not injured                         0
##  9 N           DR                   35 possible injury                     2
## 10 N           BLVD                 30 not injured                         0
## # ℹ 74,375 more rows
## # ℹ 4 more variables: units_involved <chr>, motor_vehicle_fl <chr>,
## #   motorcycle_fl <dbl>, Construction_Zone <dbl>
```

```
# Plots frequency distribution for the street type a crash occured
CarCrash |>
  group_by(Street_Type) |>
  # filters out street types that have less than 500 reported crashes, to not clog the plot
  filter(n() > 500,
         Fatal_Crash == 'N',
         !Crash_Severity == 'killed') |>
  ggplot() +
  geom_bar(aes(x = Street_Type, fill = Crash_Severity),
           color = 'black') +
  labs(x = 'Street type crash occurred',
       y = '# of crashes',
       title = 'Street type of a non-fatal crash occurrence',
       subtitle = 'Only includes streets with more than 500 crashes & involving a passenger ca
r',
       caption = 'Sourced from: City of Austin, Texas - data.austintexas.gov') +
  scale_y_continuous(breaks = seq(0,15000,1000))
```



```
# Plots frequency distribution for the street type a crash occurred
CarCrash |>
  group_by(Street_Type) |>
  # filters out street types that have less than 500 reported crashes, to not clog the data
  filter(n() > 500,
         # plots only fata crashes
         Fatal_Crash == 'Y') |>
  ggplot() +
  geom_bar(aes(x = Street_Type),
           color = 'black',
           fill = 'blue') +
  labs(x = 'Street type crash occurred',
       y = '# of crashes',
       title = 'Street type of a fatal crash occurrence',
       subtitle = 'Only includes streets with more than 500 crashes & involving a passenger ca
r',
       caption = 'Sourced from: City of Austin, Texas - data.austintexas.gov') +
  scale_y_continuous(breaks = seq(0,95,5))
```



Similar to the distribution of reported speeds of crashes, the street types where crashes occcurred share a similar trend between both the fatal and non-fatal crash plots.

Additionally, we're able to see that most fatal crashes occur on Boulevards which is interesting that it's not on a major roadway such as a Highway, Parkway, or Expressway.x`

Conclusion:

Reflection:

Despite the cross-validation performance being exceptional, the overall models performance is still lacking, thus the model is not quite a good fit for what we set out to achieve. However, if we were to possibly add more variables, such as vehicle weight as stated in the SWOV study, the performance of the model would likely increase! This could be done by adding the year, make & models of the vehicles involved in the accident so that way the weights of each vehicle could be added retroactively by referencing the manufacturer's spec sheets for the vehicle.

Ethical Concerns:

Some concerns with the analysis of this data could be an assumption that certain road types are more unsafe than others due to having a larger number of accidents. As the plots above displayed Boulevards having the highest frequency of accidents, but that doesn't necessarily mean that streets named as a boulevard are more dangerous to drive on. The data tested does not include weather conditions, time of day or year all of which are influential factors.

Construction of the Project:

While doing this project, the most difficult part, unsurprisingly, was the 'tidying', or reorganization of the data. As I had ran into multiple issues, such as having to factorize my Street_Type variable so that each different type of street is accounted for. I then attempted to do this with my Crash_Severity, but that resulted in a model that had a ROC AUC value of 0.999, which makes sense since it includes a level where a fatality was guaranteed to have occurred. So statistically using that model didn't make a lot of sense.

Additionally, if possible I would share with the city of Austin, that we should try and report the makes and models of vehicles so that way the weights of vehicles involved in the accidents can be added & accounted for. However, this would also need the speeds of all vehicles involved, and having a unique column for each in order to follow the scientific methods that were applied in the SWOV document below.

Sources/Acknowledgements:

https://www.kickassfacts.com/whats-the-difference-between-an-ave-rd-st-ln-dr-way-pl-blvd-etc/ (https://www.kickassfacts.com/whats-the-difference-between-an-ave-rd-st-ln-dr-way-pl-blvd-etc/)

https://data.austintexas.gov/Transportation-and-Mobility/Austin-Crash-Report-Data-Crash-Level-Records/y2wy-tgr5 (https://data.austintexas.gov/Transportation-and-Mobility/Austin-Crash-Report-Data-Crash-Level-Records/y2wy-tgr5)

https://safety.fhwa.dot.gov/speedmgt/ref_mats/fhwasa1304/Resources3/08%20-%20The%20Relation%20Between%20Speed%20and%20Crashes.pdf (https://safety.fhwa.dot.gov/speedmgt/ref_mats/fhwasa1304/Resources3/08%20-%20The%20Relation%20Between%20Speed%20and%20Crashes.pdf)