Workbook to investigate Modeling

Team 3

25/09/2020

Introduciton

Workbook to investigate modeling of crash data usig logistic regression as the target.

Engineer Data

Steps taken:

- Created Crash_Severity_Fac which is an ordinal factor version of Crash_Severity
- Created Crash_Severity_Num which is a numeric version of Crash_Severity_Fac
- created dummy attributes for Crash_Nature
- all columnm names renamed to replace spaces with "_".
- created logical fatal_accident column, equals 1 for a fatal accident, 0 for a non-fatal accident
- create factor fatal_accident_fac based on fatal_accident
- created Crash_Nature_Num as a numeric version of Crash_Nature

NOTE: from this point on analysis should use the CrashDF dataframe

```
# will check to see if we previously saved away the engineered data set and load it
# rather than re-engineering the main crash data again.
# NOTE: this means that if additional engineering code is added the saved copy of q1CrashDF.Rds
# should be deleted OR the forceDfRebuild variable should be set to TRUE.
forceDfRebuild <- TRUE
dfFileName <-"CrashDF.Rds"

if( (!forceDfRebuild) & file.exists(dfFileName)) {
  # if already engineered use the saved copy
  crashDF <- read_rds(dfFileName)
} else {
  # grab the original data set created by William, Rhoin and Luci
  origData <- read_rds("../datasets/main.Rds")

  # create extra dummy columns for the Crash_Nature - should we drop the first one?
  #     remove_first_dummy = TRUE
  crashDF <- fastDummies::dummy_cols(origData, select_columns= "Crash_Nature")
  names(crashDF)<-str_replace_all(names(crashDF), c(" " = "_" , "," = "", "-"="_" ))
  names(crashDF)<-str_replace_all(names(crashDF), c("___" = "_"  ))

  # replace spaces with underscore in Crash_Severity
  crashDF$Crash_Severity <- str_replace_all(crashDF$Crash_Severity, " ", "_")


  # may not need this as we have Count_Fatal, Count_Hospitalisation etc
```

```
  crashDF$Crash_Severity_Fac <- factor(crashDF$Crash_Severity, c("Property_damage_only",
        "Minor_injury",  "Medical_treatment", "Hospitalisation", "Fatal"), ordered=TRUE)

  # numeric version of Crash_Severity_Fac
  crashDF$Crash_Severity_Num <- as.numeric(crashDF$Crash_Severity_Fac)

  # create a logical - 1 if fatal, 0 if non-fatal
  crashDF %<>% mutate(fatal_accident = Count_Casualty_Fatality > 0)
  crashDF$fatal_accident_fac  <- factor(crashDF$fatal_accident, levels=c("FALSE", "TRUE"))

  crashDF$Crash_Nature_Fac <- factor(crashDF$Crash_Nature)

  crashDF$Crash_Nature_Num <- as.numeric(crashDF$Crash_Nature_Fac)

  # remove spaces from Crash_Speed_Limit
  crashDF$Crash_Speed_Limit <- str_replace_all(crashDF$Crash_Speed_Limit, c(" "= "", "-"="to", "/"=""))

  # create an ordinal version of speed limit
  crashDF$Crash_Speed_Limit_Fac <- factor(crashDF$Crash_Speed_Limit,
                                          c("0to50kmh", "60kmh", "70kmh", "80to90kmh", "100to110kmh")
                                          ordered=TRUE)
  # create a numeric version of speed limit
  crashDF$Crash_Speed_Limit_Num <- as.numeric(crashDF$Crash_Speed_Limit_Fac)

   # create dummy variables for Speed limit
  crashDF <- fastDummies::dummy_cols(crashDF, select_columns= "Crash_Speed_Limit")


  print(names(crashDF))

  write_rds(crashDF, dfFileName)

}
```

```
##  [1] "X1"
##  [2] "crash_id"
##  [3] "Crash_Severity"
##  [4] "Crash_Year"
##  [5] "Crash_Month"
##  [6] "Crash_Day_Of_Week"
##  [7] "Crash_Hour"
##  [8] "Crash_Nature"
##  [9] "Crash_Type"
## [10] "Crash_Longitude_GDA94"
## [11] "Crash_Latitude_GDA94"
## [12] "Crash_Street"
## [13] "Loc_Suburb"
## [14] "Loc_Local_Government_Area"
## [15] "Loc_Post_Code"
## [16] "Loc_Main_Roads_Region"
## [17] "Loc_ABS_Remoteness"
## [18] "Crash_Controlling_Authority"
## [19] "Crash_Roadway_Feature"
## [20] "Crash_Traffic_Control"
```

```
## [21] "Crash_Speed_Limit"
## [22] "Crash_Road_Surface_Condition"
## [23] "Crash_Atmospheric_Condition"
## [24] "Crash_Lighting_Condition"
## [25] "Crash_Road_Horiz_Align"
## [26] "Crash_Road_Vert_Align"
## [27] "Crash_DCA_Code"
## [28] "Crash_DCA_Description"
## [29] "Crash_DCA_Group_Description"
## [30] "DCA_Key_Approach_Dir"
## [31] "Count_Casualty_Fatality"
## [32] "Count_Casualty_Hospitalised"
## [33] "Count_Casualty_MedicallyTreated"
## [34] "Count_Casualty_MinorInjury"
## [35] "Count_Casualty_Total"
## [36] "Count_Unit_Motorcycle_Moped"
## [37] "site_id_1"
## [38] "site_name_1"
## [39] "distance_1"
## [40] "site_id_list_2"
## [41] "site_name_2"
## [42] "distance_2"
## [43] "site_id_list_3"
## [44] "site_name_3"
## [45] "distance_3"
## [46] "Lat"
## [47] "Lon"
## [48] "rainfall"
## [49] "Crash_Nature_Angle"
## [50] "Crash_Nature_Collision_miscellaneous"
## [51] "Crash_Nature_Fall_from_vehicle"
## [52] "Crash_Nature_Head_on"
## [53] "Crash_Nature_Hit_animal"
## [54] "Crash_Nature_Hit_object"
## [55] "Crash_Nature_Hit_parked_vehicle"
## [56] "Crash_Nature_Hit_pedestrian"
## [57] "Crash_Nature_Non_collision_miscellaneous"
## [58] "Crash_Nature_Overturned"
## [59] "Crash_Nature_Rear_end"
## [60] "Crash_Nature_Sideswipe"
## [61] "Crash_Nature_Struck_by_external_load"
## [62] "Crash_Severity_Fac"
## [63] "Crash_Severity_Num"
## [64] "fatal_accident"
## [65] "fatal_accident_fac"
## [66] "Crash_Nature_Fac"
## [67] "Crash_Nature_Num"
## [68] "Crash_Speed_Limit_Fac"
## [69] "Crash_Speed_Limit_Num"
## [70] "Crash_Speed_Limit_0to50kmh"
## [71] "Crash_Speed_Limit_100to110kmh"
## [72] "Crash_Speed_Limit_60kmh"
## [73] "Crash_Speed_Limit_70kmh"
## [74] "Crash_Speed_Limit_80to90kmh"
```

Additional exploration

Accidents per speed zone

```
crashDF %>%  count(Crash_Speed_Limit_Fac) %>% rename(total_accidents = n) %>%
  arrange( -total_accidents)
```

```
## # A tibble: 5 x 2
##   Crash_Speed_Limit_Fac total_accidents
##   <ord>                           <int>
## 1 60kmh                           11994
## 2 0to50kmh                         4046
## 3 100to110kmh                      3196
## 4 80to90kmh                        2291
## 5 70kmh                            1230
```

Do some breakdown of fatalities

```
accidentTableToDF <-function(colName, accidentTable) {
  colnames(accidentTable) <- c("non_fatal", "fatal")

 # tmpDF <- cbind(crash_nature = row.names(accidentTable),
#                 as.data.frame.matrix(accidentTable))
  tmpDF <- mutate(as.data.frame.matrix(accidentTable), !!colName := row.names(accidentTable))

  total_accidents = sum(tmpDF$fatal + tmpDF$non_fatal)
  tmpDF %<>% mutate( percent_accidents = ((fatal+non_fatal)/(total_accidents))*100 )
  tmpDF %<>% mutate( percent_fatal = (fatal/(fatal+non_fatal))*100 ) %>%
    arrange(-percent_fatal, -percent_accidents)


  return(tmpDF)

}
```

Total accidents by speed zone

```
accidentTable =  table(crashDF$Crash_Speed_Limit_Fac, crashDF$fatal_accident)

tmpDF <- accidentTableToDF("Crash_Speed_limit", accidentTable)

tmpDF
```

```
##   non_fatal fatal Crash_Speed_limit percent_accidents percent_fatal
## 1      2958   238       100to110kmh          14.04403      7.446809
## 2      2146   145         80to90kmh          10.06723      6.329114
## 3      1186    44             70kmh           5.40493      3.577236
## 4     11737   257             60kmh          52.70466      2.142738
## 5      3969    77          0to50kmh          17.77914      1.903114
```

Total accidents by crash nature

```
accidentTable =  table(crashDF$Crash_Nature, crashDF$fatal_accident)
colnames(accidentTable) <- c("non_fatal", "fatal")

tmpDF <- accidentTableToDF("Crash_Nature", accidentTable)

tmpDF
```

```
##    non_fatal fatal                  Crash_Nature percent_accidents
## 1        455    93                     Head-on        2.40805027
## 2       3198   252                   Hit object       15.16017050
## 3         15     1 Non-collision - miscellaneous        0.07030804
## 4         55     3    Collision - miscellaneous        0.25486663
## 5         32     1      Struck by external load        0.14501033
## 6        207     6               Hit pedestrian        0.93597574
## 7       6986   184                        Angle       31.50678912
## 8        229     6           Hit parked vehicle        1.03264929
## 9        580    14                   Hit animal        2.61018588
## 10      5617   125            Fall from vehicle       25.23179681
## 11      1778    30                    Sideswipe        7.94480819
## 12      2796    46                     Rear-end       12.48846509
## 13        48     0                   Overturned        0.21092411
##    percent_fatal
## 1      16.970803
## 2       7.304348
## 3       6.250000
## 4       5.172414
## 5       3.030303
## 6       2.816901
## 7       2.566248
## 8       2.553191
## 9       2.356902
## 10      2.176942
## 11      1.659292
## 12      1.618578
## 13      0.000000
```

Accidents by nature for each speed zone.

```
crashDF %>% filter(Crash_Speed_Limit_0to50kmh == 1) -> tmpDF
accidentTable =  table( tmpDF$Crash_Nature, tmpDF$fatal_accident)
print("Speed Zone 0-50kmh")
```

```
## [1] "Speed Zone 0-50kmh"
```

```
print(accidentTableToDF("Crash_Nature", accidentTable))
```

```
##    non_fatal fatal                  Crash_Nature percent_accidents
## 1        520    33                   Hit object       13.66782007
## 2         92     5                      Head-on        2.39742956
## 3        117     5           Hit parked vehicle        3.01532378
## 4       1524    20                        Angle       38.16114681
## 5        965    11            Fall from vehicle       24.12259021
## 6        326     2                     Rear-end        8.10677212
## 7        260     1                    Sideswipe        6.45081562
## 8         77     0               Hit pedestrian        1.90311419
## 9         72     0                   Hit animal        1.77953534
## 10         6     0    Collision - miscellaneous        0.14829461
## 11         5     0                   Overturned        0.12357884
## 12         4     0 Non-collision - miscellaneous        0.09886307
## 13         1     0      Struck by external load        0.02471577
##    percent_fatal
## 1      5.9674503
## 2      5.1546392
```

```
## 3      4.0983607
## 4      1.2953368
## 5      1.1270492
## 6      0.6097561
## 7      0.3831418
## 8      0.0000000
## 9      0.0000000
## 10     0.0000000
## 11     0.0000000
## 12     0.0000000
## 13     0.0000000
```

```
crashDF %>% filter(Crash_Speed_Limit_60kmh == 1) -> tmpDF
accidentTable =  table( tmpDF$Crash_Nature, tmpDF$fatal_accident)
print("Speed Zone 60kmh")
```

```
## [1] "Speed Zone 60kmh"
```

```
print(accidentTableToDF("Crash_Nature",accidentTable))
```

```
##    non_fatal fatal                  Crash_Nature percent_accidents
## 1        196    20                      Head-on        1.80090045
## 2       1429    88                   Hit object       12.64799066
## 3        103     6                Hit pedestrian        0.90878773
## 4       4537    78                         Angle       38.47757212
## 5       2607    41             Fall from vehicle       22.07770552
## 6       1017    13                     Sideswipe        8.58762715
## 7       1600    11                      Rear-end       13.43171586
## 8        122     0                   Hit animal        1.01717525
## 9         81     0            Hit parked vehicle        0.67533767
## 10        20     0     Collision - miscellaneous        0.16675004
## 11        19     0                    Overturned        0.15841254
## 12         3     0 Non-collision - miscellaneous        0.02501251
## 13         3     0       Struck by external load        0.02501251
##    percent_fatal
## 1      9.2592593
## 2      5.8009229
## 3      5.5045872
## 4      1.6901408
## 5      1.5483384
## 6      1.2621359
## 7      0.6828057
## 8      0.0000000
## 9      0.0000000
## 10     0.0000000
## 11     0.0000000
## 12     0.0000000
## 13     0.0000000
```

```
crashDF %>% filter(Crash_Speed_Limit_70kmh == 1) -> tmpDF
accidentTable =  table( tmpDF$Crash_Nature, tmpDF$fatal_accident)
print("Speed Zone 70kmh")
```

```
## [1] "Speed Zone 70kmh"
```

```r
print(accidentTableToDF("Crash_Nature",accidentTable))
```

```
##    non_fatal fatal                  Crash_Nature percent_accidents
## 1         22     2                       Head-on        1.95121951
## 2        166    15                    Hit object       14.71544715
## 3        324    19                         Angle       27.88617886
## 4        274     6             Fall from vehicle       22.76422764
## 5        233     2                      Rear-end       19.10569106
## 6        133     0                     Sideswipe       10.81300813
## 7         12     0                    Hit animal        0.97560976
## 8         11     0                Hit pedestrian        0.89430894
## 9          6     0            Hit parked vehicle        0.48780488
## 10         2     0     Collision - miscellaneous        0.16260163
## 11         1     0 Non-collision - miscellaneous        0.08130081
## 12         1     0                    Overturned        0.08130081
## 13         1     0       Struck by external load        0.08130081
##    percent_fatal
## 1      8.3333333
## 2      8.2872928
## 3      5.5393586
## 4      2.1428571
## 5      0.8510638
## 6      0.0000000
## 7      0.0000000
## 8      0.0000000
## 9      0.0000000
## 10     0.0000000
## 11     0.0000000
## 12     0.0000000
## 13     0.0000000
```

```r
crashDF %>% filter(Crash_Speed_Limit_80to90kmh == 1) -> tmpDF
accidentTable =  table( tmpDF$Crash_Nature, tmpDF$fatal_accident)
print("Speed Zone 80-90kmh")
```

```
## [1] "Speed Zone 80-90kmh"
```

```r
print(accidentTableToDF("Crash_Nature",accidentTable))
```

```
##    non_fatal fatal                  Crash_Nature percent_accidents
## 1         66    27                       Head-on        4.05936272
## 2         10     1     Collision - miscellaneous        0.48013968
## 3        433    41                    Hit object       20.68965517
## 4        368    32                         Angle       17.45962462
## 5        260    11                      Rear-end       11.82889568
## 6        727    27             Fall from vehicle       32.91139241
## 7        153     5                     Sideswipe        6.89655172
## 8         90     1                    Hit animal        3.97206460
## 9         12     0            Hit parked vehicle        0.52378874
## 10        10     0                Hit pedestrian        0.43649062
## 11         8     0                    Overturned        0.34919249
## 12         8     0       Struck by external load        0.34919249
## 13         1     0 Non-collision - miscellaneous        0.04364906
##    percent_fatal
## 1      29.032258
```

```
## 2       9.090909
## 3       8.649789
## 4       8.000000
## 5       4.059041
## 6       3.580902
## 7       3.164557
## 8       1.098901
## 9       0.000000
## 10      0.000000
## 11      0.000000
## 12      0.000000
## 13      0.000000
```

```
crashDF %>% filter(Crash_Speed_Limit_100to110kmh == 1) -> tmpDF
accidentTable =  table( tmpDF$Crash_Nature, tmpDF$fatal_accident)
print("Speed Zone 100-110kmh")
```

```
## [1] "Speed Zone 100-110kmh"
```

```
print(accidentTableToDF("Crash_Nature",accidentTable))
```

```
##    non_fatal fatal                  Crash_Nature percent_accidents
## 1         79    39                      Head-on         3.6921151
## 2          6     1 Non-collision - miscellaneous         0.2190238
## 3        233    35                         Angle         8.3854819
## 4         17     2     Collision - miscellaneous         0.5944931
## 5        650    75                    Hit object        22.6846058
## 6         13     1            Hit parked vehicle         0.4380476
## 7        377    20                      Rear-end        12.4217772
## 8         19     1       Struck by external load         0.6257822
## 9        215    11                     Sideswipe         7.0713392
## 10       284    13                    Hit animal         9.2928661
## 11      1044    40             Fall from vehicle        33.9173967
## 12        15     0                    Overturned         0.4693367
## 13         6     0                Hit pedestrian         0.1877347
##    percent_fatal
## 1      33.050847
## 2      14.285714
## 3      13.059701
## 4      10.526316
## 5      10.344828
## 6       7.142857
## 7       5.037783
## 8       5.000000
## 9       4.867257
## 10      4.377104
## 11      3.690037
## 12      0.000000
## 13      0.000000
```

Pre-modeling EDA

Probably need to show we looked into some statistical measures before we jump right into the modeling. This needs to be padded out a bit

```
corrData <- select(crashDF,  Count_Casualty_Fatality, Count_Casualty_Hospitalised,
                   Count_Casualty_MedicallyTreated, Count_Casualty_MinorInjury,
                   Crash_Nature_Angle, Crash_Nature_Collision_miscellaneous,
                   Crash_Nature_Fall_from_vehicle, Crash_Nature_Head_on,
                   Crash_Nature_Hit_animal, Crash_Nature_Hit_object,
                   Crash_Nature_Hit_parked_vehicle, Crash_Nature_Hit_pedestrian,
                   Crash_Nature_Non_collision_miscellaneous, Crash_Nature_Overturned,
                   Crash_Nature_Rear_end, Crash_Nature_Sideswipe,
                   Crash_Nature_Struck_by_external_load
                   )
 result <- corrData %>% correlate()
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
 result %>% focus(names(result)[2])  # Can't seem to be able to use the name Count_Casulaty_Fatality di
```

```
## # A tibble: 16 x 2
##    rowname                                Count_Casualty_Fatality
##    <chr>                                                    <dbl>
##  1 Count_Casualty_Hospitalised                            -0.143
##  2 Count_Casualty_MedicallyTreated                        -0.0770
##  3 Count_Casualty_MinorInjury                             -0.0465
##  4 Crash_Nature_Angle                                     -0.0269
##  5 Crash_Nature_Collision_miscellaneous                    0.00471
##  6 Crash_Nature_Fall_from_vehicle                         -0.0390
##  7 Crash_Nature_Head_on                                    0.125
##  8 Crash_Nature_Hit_animal                                -0.00942
##  9 Crash_Nature_Hit_object                                 0.0889
## 10 Crash_Nature_Hit_parked_vehicle                        -0.00480
## 11 Crash_Nature_Hit_pedestrian                            -0.00320
## 12 Crash_Nature_Non_collision_miscellaneous                0.00400
## 13 Crash_Nature_Overturned                                -0.00845
## 14 Crash_Nature_Rear_end                                  -0.0360
## 15 Crash_Nature_Sideswipe                                 -0.0270
## 16 Crash_Nature_Struck_by_external_load                   -0.000819
```

```
result %>% focus(names(result)[3])
```

```
## # A tibble: 16 x 2
##    rowname                                Count_Casualty_Hospitalised
##    <chr>                                                        <dbl>
##  1 Count_Casualty_Fatality                                   -0.143
##  2 Count_Casualty_MedicallyTreated                           -0.594
##  3 Count_Casualty_MinorInjury                                -0.320
##  4 Crash_Nature_Angle                                        -0.0162
##  5 Crash_Nature_Collision_miscellaneous                       0.0108
##  6 Crash_Nature_Fall_from_vehicle                             0.0117
##  7 Crash_Nature_Head_on                                       0.0578
##  8 Crash_Nature_Hit_animal                                    0.0312
##  9 Crash_Nature_Hit_object                                    0.0632
## 10 Crash_Nature_Hit_parked_vehicle                           -0.000883
## 11 Crash_Nature_Hit_pedestrian                                0.0311
## 12 Crash_Nature_Non_collision_miscellaneous                  -0.00121
```

```
## 13 Crash_Nature_Overturned                                0.0186
## 14 Crash_Nature_Rear_end                                 -0.0698
## 15 Crash_Nature_Sideswipe                                -0.0547
## 16 Crash_Nature_Struck_by_external_load                  -0.0112
```

```
result %>% focus(names(result)[4])
```

```
## # A tibble: 16 x 2
##    rowname                                Count_Casualty_MedicallyTreated
##    <chr>                                                            <dbl>
##  1 Count_Casualty_Fatality                                        -0.0770
##  2 Count_Casualty_Hospitalised                                    -0.594
##  3 Count_Casualty_MinorInjury                                     -0.187
##  4 Crash_Nature_Angle                                              0.0250
##  5 Crash_Nature_Collision_miscellaneous                           -0.00697
##  6 Crash_Nature_Fall_from_vehicle                                 -0.0193
##  7 Crash_Nature_Head_on                                           -0.0176
##  8 Crash_Nature_Hit_animal                                        -0.0121
##  9 Crash_Nature_Hit_object                                        -0.0773
## 10 Crash_Nature_Hit_parked_vehicle                                -0.0293
## 11 Crash_Nature_Hit_pedestrian                                    0.0304
## 12 Crash_Nature_Non_collision_miscellaneous                       0.00892
## 13 Crash_Nature_Overturned                                        0.00793
## 14 Crash_Nature_Rear_end                                          0.0620
## 15 Crash_Nature_Sideswipe                                         0.0289
## 16 Crash_Nature_Struck_by_external_load                           0.0162
```

```
result %>% focus(names(result)[5])
```

```
## # A tibble: 16 x 2
##    rowname                                Count_Casualty_MinorInjury
##    <chr>                                                       <dbl>
##  1 Count_Casualty_Fatality                                   -0.0465
##  2 Count_Casualty_Hospitalised                               -0.320
##  3 Count_Casualty_MedicallyTreated                           -0.187
##  4 Crash_Nature_Angle                                         0.0288
##  5 Crash_Nature_Collision_miscellaneous                      -0.00593
##  6 Crash_Nature_Fall_from_vehicle                            -0.0380
##  7 Crash_Nature_Head_on                                      -0.00751
##  8 Crash_Nature_Hit_animal                                   -0.0306
##  9 Crash_Nature_Hit_object                                   -0.0711
## 10 Crash_Nature_Hit_parked_vehicle                           -0.0170
## 11 Crash_Nature_Hit_pedestrian                                0.0360
## 12 Crash_Nature_Non_collision_miscellaneous                  -0.0104
## 13 Crash_Nature_Overturned                                   -0.0104
## 14 Crash_Nature_Rear_end                                      0.0654
## 15 Crash_Nature_Sideswipe                                     0.0469
## 16 Crash_Nature_Struck_by_external_load                      -0.00882
```

Correlation analysis doesn't reveal any correlation of any strength between severity of accident and the nature of the accident. The Highest positive correlation is 0.12 between Head On accidents and Fatalities, however this i not a significant correlation. All remaining correlations are above -0.01 and below 0.01 which are no correlation at all.

```
corrData <- select(crashDF,  fatal_accident,
                   Crash_Nature_Angle, Crash_Nature_Collision_miscellaneous,
```

```
                  Crash_Nature_Fall_from_vehicle, Crash_Nature_Head_on,
                  Crash_Nature_Hit_animal, Crash_Nature_Hit_object,
                  Crash_Nature_Hit_parked_vehicle, Crash_Nature_Hit_pedestrian,
                  Crash_Nature_Non_collision_miscellaneous, Crash_Nature_Overturned,
                  Crash_Nature_Rear_end, Crash_Nature_Sideswipe,
                  Crash_Nature_Struck_by_external_load
                  )
 result <- corrData %>% correlate()
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
 result %>% focus(names(result)[2])
```

```
## # A tibble: 13 x 2
##    rowname                                fatal_accident
##    <chr>                                           <dbl>
##  1 Crash_Nature_Angle                           -0.0293
##  2 Crash_Nature_Collision_miscellaneous          0.00514
##  3 Crash_Nature_Fall_from_vehicle               -0.0377
##  4 Crash_Nature_Head_on                          0.119
##  5 Crash_Nature_Hit_animal                      -0.00899
##  6 Crash_Nature_Hit_object                       0.0931
##  7 Crash_Nature_Hit_parked_vehicle              -0.00449
##  8 Crash_Nature_Hit_pedestrian                  -0.00285
##  9 Crash_Nature_Non_collision_miscellaneous      0.00429
## 10 Crash_Nature_Overturned                      -0.00855
## 11 Crash_Nature_Rear_end                        -0.0363
## 12 Crash_Nature_Sideswipe                       -0.0275
## 13 Crash_Nature_Struck_by_external_load         -0.000665
```

Correlation of fatal accidents and speed

Not sure if this is jutifiable to encode speed zone as numeric. There is a low correlation between it and fatal accident.

```
corrData <- select(crashDF,  fatal_accident,
                   Crash_Speed_Limit_Num
                  )
 result <- corrData %>% correlate()
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
 result %>% focus(names(result)[2])
```

```
## # A tibble: 1 x 2
##   rowname               fatal_accident
##   <chr>                          <dbl>
## 1 Crash_Speed_Limit_Num          0.113
```

Looking at speed zone slightly differently encoded as dummy variables (do we need to leave one out). Doesn't seem to be a correlation between the dummy variables and fatal accidents

```
corrData <- select(crashDF,  fatal_accident,
                   Crash_Speed_Limit_0to50kmh,
                   Crash_Speed_Limit_100to110kmh,
                   Crash_Speed_Limit_60kmh,
                   Crash_Speed_Limit_70kmh,
                   Crash_Speed_Limit_80to90kmh
                  )
 result <- corrData %>% correlate()
```

```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
 result %>% focus(names(result)[2])
```

```
## # A tibble: 5 x 2
##   rowname                       fatal_accident
##   <chr>                                  <dbl>
## 1 Crash_Speed_Limit_0to50kmh          -0.0373
## 2 Crash_Speed_Limit_100to110kmh        0.0922
## 3 Crash_Speed_Limit_60kmh             -0.0705
## 4 Crash_Speed_Limit_70kmh              0.00310
## 5 Crash_Speed_Limit_80to90kmh          0.0556
```

Correlation still low. It would seem that the nature of an accident by itself has no correlation to the crash severity. However as can seen by the histograms at the start of the document a large number of accidents are the result of specific types of accidents, namely a collision at an angle and riders falling off. It is not clear if riders fall off as a result of some other accident type.

Just in case I'm doing things wrong with the above correlation let's look at a chi-squared test

```
chisq.test(crashDF$Crash_Severity_Fac, crashDF$Crash_Nature, correct=FALSE)
```

```
## Warning in chisq.test(crashDF$Crash_Severity_Fac, crashDF$Crash_Nature, : Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  crashDF$Crash_Severity_Fac and crashDF$Crash_Nature
## X-squared = 1549.7, df = 48, p-value < 2.2e-16
```

Given the low p value the chi-squared test does imply that the Nature and Severity are dependent.

Traing and test set

The crash data set will be divided into a training data set (70%) and test data set (30%). However, this won't be a straight split. Because of the small number of fatalities to non-fatalities to ensure a similar proportion in the training and test sets the crash data will first be divided into fatal and non-fatal and each of these subsets will be sampled for their contribution to the training and test sets.

```
genTrainTestSets <- function( df ){
  # - will return a list element 1 is the train set, element two is the test set
  # - will split the data ~ 70% to 30% - TODO: should make this a parameter
  # - the input dataset will be split into fatal and non-fatal observations and
  #   the these subsets will divided (as described above) into the train and test sets.
  #   TODO: make the query for the split a parameter
```

```
  set.seed(42)

  fatalDF <- filter(df, fatal_accident == 1)
  nonFatalDF <- filter(df, fatal_accident == 0)

  # work out how much of the non-fatal data goes into the train and test set
  trainSize <- floor(0.7*nrow(nonFatalDF))
  trainIndexes <- sample(seq_len(nrow(nonFatalDF)), size = trainSize)

  # assign nofatal data to test and train data sets
  trainDF  <- nonFatalDF[trainIndexes, ]
  testDF   <- nonFatalDF[-trainIndexes, ]

  # work out how much of the fatal data goes into the train and test set
  trainSize <- floor(0.7*nrow(fatalDF))
  trainIndexes <- sample(seq_len(nrow(fatalDF)), size = trainSize)

  # add fatal data to test and train data set
  trainDF <- rbind(trainDF, fatalDF[trainIndexes,])
  testDF  <- rbind(testDF, fatalDF[-trainIndexes,])

  retList <- list( trainDF, testDF)
  names(retList) = c("TrainSet", "TestSet")

  return(retList)
}
```

```
#  split data into 70% train 30% test.

trainTestSets <- genTrainTestSets(crashDF)

trainDF  <- trainTestSets$TrainSet
print(paste("nrow Train set=", nrow(trainDF)))
```

```
## [1] "nrow Train set= 15929"
```

```
testDF   <- trainTestSets$TestSet
print(paste("nrow Test set=", nrow(testDF)))
```

```
## [1] "nrow Test set= 6828"
```

Modeling all years

This is using the entire data set. Following sections will look at data from specific time periods. upto 2008 ? 2008 to 2017, 2017 to now.

Basic logistic model - Fatal_accident as the target, start with speed zone as the predictor. Try with speed limit as a factor and then coded as a dummy variable to see if there are differences.

Will probably have to add additional predictors to get anything meaningful.

NOTE: need to check this article out in detail, can make our analysis sound fancy ;-), https://stats.idre.ucla.edu/r/dae/logit-regression/

Model 1 fatal_accident ~ Speed limit encoded as a numeric

Again not sure if this makes sense to encode speed zone as numeirc. Should repeat with ordinal regression.

Both intercept and speed seem to be signficiant based on p-values.

Prediction is way off. seems to just predict FALSE. - which gives us an accuracy of ~95% because of low numbers of fatalities

```
model1 <- glm( fatal_accident ~ Crash_Speed_Limit_Num, data=trainDF, family="binomial")
summary(model1)
```

```
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_Num, family = "binomial",
##     data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4038  -0.2681  -0.2178  -0.2178   2.8866
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.57175    0.10904  -41.93   <2e-16 ***
## Crash_Speed_Limit_Num  0.42121    0.03063   13.75   <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: 4662.8  on 15928  degrees of freedom
## Residual deviance: 4479.8  on 15927  degrees of freedom
## AIC: 4483.8
##
## Number of Fisher Scoring iterations: 6
```

```
model1.predictions <- predict(model1, newdata = testDF, type="response") >= 0.5


cm = confusionMatrix(factor(model1.predictions, levels=c("FALSE","TRUE")), factor(testDF$fatal_accident
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  6599  229
##      TRUE      0    0
##
##                Accuracy : 0.9665
##                  95% CI : (0.9619, 0.9706)
##     No Information Rate : 0.9665
##     P-Value [Acc > NIR] : 0.5176
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
```

```
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96646
##              Prevalence : 0.03354
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
```

Model 2 - ordinal linear regression

Try the same thing but with ordinal linear regression. Similar result as above

```
model2<- clm(data=trainDF, fatal_accident_fac ~ Crash_Speed_Limit_Fac,
    link="logit")

summary(model2)
```

```
## formula: fatal_accident_fac ~ Crash_Speed_Limit_Fac
## data:    trainDF
##
##  link  threshold nobs  logLik   AIC     niter max.grad cond.H
##  logit flexible  15929 -2235.49 4480.99 7(0)  1.28e-10 1.6e+01
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## Crash_Speed_Limit_Fac.L  1.2740542  0.1090511  11.683  < 2e-16 ***
## Crash_Speed_Limit_Fac.Q  0.0082054  0.1332952   0.062  0.95091
## Crash_Speed_Limit_Fac.C -0.2853677  0.0950291  -3.003  0.00267 **
## Crash_Speed_Limit_Fac^4  0.0001645  0.1440521   0.001  0.99909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##            Estimate Std. Error z value
## FALSE|TRUE  3.25022    0.05452   59.62
```

```
model2.predictions <- predict(model2, newdata=testDF, type="class")

cm = confusionMatrix(model2.predictions$fit, testDF$fatal_accident_fac, positive="TRUE")
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  6599  229
##      TRUE      0    0
##
##                Accuracy : 0.9665
##                  95% CI : (0.9619, 0.9706)
```

```
##     No Information Rate : 0.9665
##     P-Value [Acc > NIR] : 0.5176
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96646
##              Prevalence : 0.03354
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
```

Model 3 - logistic regression, dummy speed variables

Back to logistic regression but with dummy variables for speed. Model still producing all FALSE values.

```
model3 <- glm( fatal_accident ~
               Crash_Speed_Limit_0to50kmh + Crash_Speed_Limit_100to110kmh +
               Crash_Speed_Limit_60kmh + Crash_Speed_Limit_70kmh +
               Crash_Speed_Limit_80to90kmh,
               data=trainDF, family="binomial")
summary(model3)
```

```
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_0to50kmh + Crash_Speed_Limit_100to110kmh +
##     Crash_Speed_Limit_60kmh + Crash_Speed_Limit_70kmh + Crash_Speed_Limit_80to90kmh,
##     family = "binomial", data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3915  -0.2752  -0.2067  -0.2067   2.8214
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.6691     0.1019 -26.196  < 2e-16 ***
## Crash_Speed_Limit_0to50kmh     -1.2922     0.1721  -7.510  5.9e-14 ***
## Crash_Speed_Limit_100to110kmh   0.1388     0.1298   1.070  0.28474
## Crash_Speed_Limit_60kmh        -1.1667     0.1271  -9.180  < 2e-16 ***
## Crash_Speed_Limit_70kmh        -0.5854     0.2070  -2.828  0.00468 **
## Crash_Speed_Limit_80to90kmh         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4662.8  on 15928  degrees of freedom
```

```
## Residual deviance: 4471.0  on 15924  degrees of freedom
## AIC: 4481
##
## Number of Fisher Scoring iterations: 6
```

```
model3.predictions <- predict(model3, newdata = testDF, type="response") >= 0.5
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
cm = confusionMatrix(factor(model3.predictions, levels=c("FALSE","TRUE")), factor(testDF$fatal_accident
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  6599  229
##      TRUE      0    0
##
##                Accuracy : 0.9665
##                  95% CI : (0.9619, 0.9706)
##     No Information Rate : 0.9665
##     P-Value [Acc > NIR] : 0.5176
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96646
##              Prevalence : 0.03354
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
```

Model 4 - logistic regression with refined dummy speed variables

As model 3 but drop the 110Km zone.

```
model4 <- glm( fatal_accident ~
               Crash_Speed_Limit_0to50kmh +
               Crash_Speed_Limit_60kmh + Crash_Speed_Limit_70kmh +
               Crash_Speed_Limit_80to90kmh,
               data=trainDF, family="binomial")
summary(model4)
```

```
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_0to50kmh + Crash_Speed_Limit_60kmh +
##     Crash_Speed_Limit_70kmh + Crash_Speed_Limit_80to90kmh, family = "binomial",
```

```
##     data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3915  -0.2752  -0.2067  -0.2067   2.8214
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.5303     0.0804 -31.469  < 2e-16 ***
## Crash_Speed_Limit_0to50kmh   -1.4311     0.1603  -8.929  < 2e-16 ***
## Crash_Speed_Limit_60kmh      -1.3056     0.1106 -11.803  < 2e-16 ***
## Crash_Speed_Limit_70kmh      -0.7242     0.1973  -3.671 0.000242 ***
## Crash_Speed_Limit_80to90kmh  -0.1389     0.1298  -1.070 0.284741
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4662.8  on 15928  degrees of freedom
## Residual deviance: 4471.0  on 15924  degrees of freedom
## AIC: 4481
##
## Number of Fisher Scoring iterations: 6
```

```
model4.predictions <- predict(model4, newdata = testDF, type="response") >= 0.5


cm = confusionMatrix(factor(model4.predictions, levels=c("FALSE","TRUE")), factor(testDF$fatal_accident
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  6599  229
##      TRUE      0    0
##
##                Accuracy : 0.9665
##                  95% CI : (0.9619, 0.9706)
##     No Information Rate : 0.9665
##     P-Value [Acc > NIR] : 0.5176
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96646
##              Prevalence : 0.03354
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
```

```
##        'Positive' Class : TRUE
##
```

Year ranges

Going to run models against data at interesting date ranges:

- 2017 to pressent
- 2018 to 2016 inclusive
- prior to 2018

```
modelYearlyData <- function( df, year)
{
  #  split data into 70% train 30% test.
  trainTestSets <- genTrainTestSets(df)

  trainDF <- trainTestSets$TrainSet
  print(paste("Year ", year, " Train Set size=", nrow(trainDF)))

  testDF <- trainTestSets$TestSet
  print(paste("Year ", year, " Test Set size=", nrow(testDF)))

  #logistic regression model
  print(paste("********   YearModel1 for ", year))
  yearModel1 <- glm( fatal_accident ~ Crash_Speed_Limit_Num, data=trainDF, family="binomial")
  print(summary(yearModel1))

  yearModel1.predictions <- predict(yearModel1, newdata = testDF, type="response") >= 0.5

  cm = confusionMatrix(factor(yearModel1.predictions, levels=c("FALSE","TRUE")),
                       factor(testDF$fatal_accident, levels=c("FALSE", "TRUE")), positive="TRUE" )
  print(cm)

  #ordinal regression model
  print(paste("******  YearModel2 for ", year))
  yearModel2<- clm(data=trainDF, fatal_accident_fac ~ Crash_Speed_Limit_Fac,
    link="logit")
  print(summary(yearModel2))

  yearModel2.predictions <- predict(yearModel2, newdata=testDF, type="class")

  cm = confusionMatrix(yearModel2.predictions$fit, testDF$fatal_accident_fac, positive="TRUE")
  print(cm)

  # logistic regression with speed dummy values
  print(paste("******* YearModel3 for ", year))
  yearModel3 <- glm( fatal_accident ~
              Crash_Speed_Limit_0to50kmh + Crash_Speed_Limit_100to110kmh +
              Crash_Speed_Limit_60kmh + Crash_Speed_Limit_70kmh +
              Crash_Speed_Limit_80to90kmh,
              data=trainDF, family="binomial")
   print(summary(yearModel3))

   yearModel3.predictions <- predict(yearModel3, newdata = testDF, type="response") >= 0.5
```

```
   cm = confusionMatrix(factor(yearModel3.predictions, levels=c("FALSE","TRUE")),
                        factor(testDF$fatal_accident, levels=c("FALSE", "TRUE")),
                        positive="TRUE" )
   print(cm)

   # logistic regression dummy speed varialbes but with one removed (100-110)
   # NOTE: it was decided to remvoe the 100-110 speed limit because it didn't have a good
   # p-value when modeling for all the years. However, this could chnage for year subsets which
   # we are now processing
   # so this is not optimal way of doing things
   print(paste("*********** YearModel4 for ", year))
   yearModel4 <- glm( fatal_accident ~
               Crash_Speed_Limit_0to50kmh +
               Crash_Speed_Limit_60kmh + Crash_Speed_Limit_70kmh +
               Crash_Speed_Limit_80to90kmh,
               data=trainDF, family="binomial")
    print(summary(yearModel4))

    yearModel4.predictions <- predict(yearModel4, newdata = testDF, type="response") >= 0.5
    cm = confusionMatrix(factor(yearModel4.predictions, levels=c("FALSE","TRUE")),
                         factor(testDF$fatal_accident, levels=c("FALSE", "TRUE")), positive="TRUE" )
    print(cm)

}
```

2017 to present

```
crash2017DF <- filter(crashDF, Crash_Year >= 2017)
```

There are 2256 observations. With 67 fatalities

```
summary( crash2017DF)
```

```
##        X1           crash_id      Crash_Severity       Crash_Year
##  Min.   : 2626   Min.   : 32184   Length:2256        Min.   :2017
##  1st Qu.: 9664   1st Qu.:114228   Class :character   1st Qu.:2017
##  Median :16818   Median :196542   Mode  :character   Median :2018
##  Mean   :15767   Mean   :183953                      Mean   :2018
##  3rd Qu.:23617   3rd Qu.:277619                      3rd Qu.:2018
##  Max.   :28332   Max.   :328244                      Max.   :2018
##
##  Crash_Month        Crash_Day_Of_Week    Crash_Hour   Crash_Nature
##  Length:2256        Length:2256        Min.   : 0.0   Length:2256
##  Class :character   Class :character   1st Qu.: 9.0   Class :character
##  Mode  :character   Mode  :character   Median :13.0   Mode  :character
##                                        Mean   :12.9
##                                        3rd Qu.:17.0
##                                        Max.   :23.0
##
##   Crash_Type        Crash_Longitude_GDA94 Crash_Latitude_GDA94
##  Length:2256        Min.   :138.4         Min.   :-28.654
##  Class :character   1st Qu.:152.7         1st Qu.:-27.662
##  Mode  :character   Median :153.0         Median :-27.466
```

```
##                    Mean   :152.2         Mean   :-26.376
##                    3rd Qu.:153.1         3rd Qu.:-26.704
##                    Max.   :153.5         Max.   : -9.751
##
##  Crash_Street        Loc_Suburb        Loc_Local_Government_Area
##  Length:2256        Length:2256        Length:2256
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Loc_Post_Code      Loc_Main_Roads_Region Loc_ABS_Remoteness
##  Length:2256        Length:2256           Length:2256
##  Class :character   Class :character      Class :character
##  Mode  :character   Mode  :character      Mode  :character
##
##
##
##
##  Crash_Controlling_Authority Crash_Roadway_Feature Crash_Traffic_Control
##  Length:2256                 Length:2256           Length:2256
##  Class :character            Class :character      Class :character
##  Mode  :character            Mode  :character      Mode  :character
##
##
##
##
##  Crash_Speed_Limit  Crash_Road_Surface_Condition Crash_Atmospheric_Condition
##  Length:2256        Length:2256                  Length:2256
##  Class :character   Class :character             Class :character
##  Mode  :character   Mode  :character             Mode  :character
##
##
##
##
##  Crash_Lighting_Condition Crash_Road_Horiz_Align Crash_Road_Vert_Align
##  Length:2256              Length:2256           Length:2256
##  Class :character         Class :character      Class :character
##  Mode  :character         Mode  :character      Mode  :character
##
##
##
##
##  Crash_DCA_Code  Crash_DCA_Description Crash_DCA_Group_Description
##  Min.   :  1.0   Length:2256           Length:2256
##  1st Qu.:202.8   Class :character      Class :character
##  Median :309.0   Mode  :character      Mode  :character
##  Mean   :452.4
##  3rd Qu.:705.0
##  Max.   :905.0
##
##  DCA_Key_Approach_Dir Count_Casualty_Fatality Count_Casualty_Hospitalised
##  Length:2256          Min.   :0.00000         Min.   :0.0000
```

```
##  Class :character     1st Qu.:0.00000         1st Qu.:0.0000
##  Mode  :character     Median :0.00000         Median :1.0000
##                       Mean   :0.03059         Mean   :0.6848
##                       3rd Qu.:0.00000         3rd Qu.:1.0000
##                       Max.   :2.00000         Max.   :6.0000
##
##  Count_Casualty_MedicallyTreated Count_Casualty_MinorInjury
##  Min.   :0.0000                  Min.   :0.00000
##  1st Qu.:0.0000                  1st Qu.:0.00000
##  Median :0.0000                  Median :0.00000
##  Mean   :0.2837                  Mean   :0.09929
##  3rd Qu.:1.0000                  3rd Qu.:0.00000
##  Max.   :4.0000                  Max.   :3.00000
##
##  Count_Casualty_Total Count_Unit_Motorcycle_Moped   site_id_1
##  Min.   : 1.000       Min.   :1.000               Min.   : 27042
##  1st Qu.: 1.000       1st Qu.:1.000               1st Qu.: 40237
##  Median : 1.000       Median :1.000               Median : 40782
##  Mean   : 1.098       Mean   :1.024               Mean   : 40495
##  3rd Qu.: 1.000       3rd Qu.:1.000               3rd Qu.: 40913
##  Max.   :13.000       Max.   :4.000               Max.   :140009
##
##  site_name_1          distance_1       site_id_list_2   site_name_2
##  Length:2256        Min.   :   46.61   Min.   : 27006   Length:2256
##  Class :character   1st Qu.: 2034.22   1st Qu.: 40212   Class :character
##  Mode  :character   Median : 3364.64   Median : 40460   Mode  :character
##                     Mean   : 4123.24   Mean   : 40894
##                     3rd Qu.: 5514.78   3rd Qu.: 40861
##                     Max.   :90222.14   Max.   :140009
##
##    distance_2       site_id_list_3   site_name_3          distance_3
##  Min.   :   496.2   Min.   : 27005   Length:2256        Min.   :  1724
##  1st Qu.:  3792.7   1st Qu.: 40212   Class :character   1st Qu.:  5414
##  Median :  5372.3   Median : 40609   Mode  :character   Median :  7453
##  Mean   :  6506.6   Mean   : 40796                      Mean   :  8913
##  3rd Qu.:  7884.2   3rd Qu.: 40878                      3rd Qu.: 10649
##  Max.   :140228.7   Max.   :140009                      Max.   :149404
##
##       Lat              Lon           rainfall       Crash_Nature_Angle
##  Min.   :-28.66   Min.   :138.8   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.:-27.67   1st Qu.:152.7   1st Qu.: 0.6387   1st Qu.:0.0000
##  Median :-27.48   Median :153.0   Median : 1.3377   Median :0.0000
##  Mean   :-26.38   Mean   :152.2   Mean   : 2.9680   Mean   :0.2979
##  3rd Qu.:-26.73   3rd Qu.:153.1   3rd Qu.: 4.0196   3rd Qu.:1.0000
##  Max.   :-10.05   Max.   :153.5   Max.   :56.0714   Max.   :1.0000
##
##  Crash_Nature_Collision_miscellaneous Crash_Nature_Fall_from_vehicle
##  Min.   :0                            Min.   :0.0000
##  1st Qu.:0                            1st Qu.:0.0000
##  Median :0                            Median :0.0000
##  Mean   :0                            Mean   :0.2748
##  3rd Qu.:0                            3rd Qu.:1.0000
##  Max.   :0                            Max.   :1.0000
##
```

```
##  Crash_Nature_Head_on Crash_Nature_Hit_animal Crash_Nature_Hit_object
##  Min.   :0.00000      Min.   :0.00000         Min.   :0.0000
##  1st Qu.:0.00000      1st Qu.:0.00000         1st Qu.:0.0000
##  Median :0.00000      Median :0.00000         Median :0.0000
##  Mean   :0.01817      Mean   :0.02305         Mean   :0.1268
##  3rd Qu.:0.00000      3rd Qu.:0.00000         3rd Qu.:0.0000
##  Max.   :1.00000      Max.   :1.00000         Max.   :1.0000
##
##  Crash_Nature_Hit_parked_vehicle Crash_Nature_Hit_pedestrian
##  Min.   :0.00000                 Min.   :0.000000
##  1st Qu.:0.00000                 1st Qu.:0.000000
##  Median :0.00000                 Median :0.000000
##  Mean   :0.01152                 Mean   :0.007092
##  3rd Qu.:0.00000                 3rd Qu.:0.000000
##  Max.   :1.00000                 Max.   :1.000000
##
##  Crash_Nature_Non_collision_miscellaneous Crash_Nature_Overturned
##  Min.   :0                                Min.   :0.000000
##  1st Qu.:0                                1st Qu.:0.000000
##  Median :0                                Median :0.000000
##  Mean   :0                                Mean   :0.004876
##  3rd Qu.:0                                3rd Qu.:0.000000
##  Max.   :0                                Max.   :1.000000
##
##  Crash_Nature_Rear_end Crash_Nature_Sideswipe
##  Min.   :0.0000        Min.   :0.00000
##  1st Qu.:0.0000        1st Qu.:0.00000
##  Median :0.0000        Median :0.00000
##  Mean   :0.1348        Mean   :0.09973
##  3rd Qu.:0.0000        3rd Qu.:0.00000
##  Max.   :1.0000        Max.   :1.00000
##
##  Crash_Nature_Struck_by_external_load           Crash_Severity_Fac
##  Min.   :0.00000                     Property_damage_only:   0
##  1st Qu.:0.00000                     Minor_injury        : 153
##  Median :0.00000                     Medical_treatment   : 569
##  Mean   :0.00133                     Hospitalisation     :1467
##  3rd Qu.:0.00000                     Fatal               :  67
##  Max.   :1.00000
##
##  Crash_Severity_Num fatal_accident  fatal_accident_fac
##  Min.   :2.000      Mode :logical   FALSE:2189
##  1st Qu.:3.000      FALSE:2189      TRUE :  67
##  Median :4.000      TRUE :67
##  Mean   :3.642
##  3rd Qu.:4.000
##  Max.   :5.000
##
##            Crash_Nature_Fac Crash_Nature_Num Crash_Speed_Limit_Fac
##  Angle             :672     Min.   : 1.000   0to50kmh   : 506
##  Fall from vehicle:620      1st Qu.: 1.000   60kmh      :1095
##  Rear-end          :304     Median : 3.000   70kmh      : 147
##  Hit object        :286     Mean   : 4.953   80to90kmh  : 224
##  Sideswipe         :225     3rd Qu.: 7.000   100to110kmh: 284
```

```
##  Hit animal       : 52     Max.   :13.000
##  (Other)          : 97
##  Crash_Speed_Limit_Num Crash_Speed_Limit_0to50kmh Crash_Speed_Limit_100to110kmh
##  Min.   :1.000         Min.   :0.0000            Min.   :0.0000
##  1st Qu.:2.000         1st Qu.:0.0000            1st Qu.:0.0000
##  Median :2.000         Median :0.0000            Median :0.0000
##  Mean   :2.417         Mean   :0.2243            Mean   :0.1259
##  3rd Qu.:3.000         3rd Qu.:0.0000            3rd Qu.:0.0000
##  Max.   :5.000         Max.   :1.0000            Max.   :1.0000
##
##  Crash_Speed_Limit_60kmh Crash_Speed_Limit_70kmh Crash_Speed_Limit_80to90kmh
##  Min.   :0.0000          Min.   :0.00000         Min.   :0.00000
##  1st Qu.:0.0000          1st Qu.:0.00000         1st Qu.:0.00000
##  Median :0.0000          Median :0.00000         Median :0.00000
##  Mean   :0.4854          Mean   :0.06516         Mean   :0.09929
##  3rd Qu.:1.0000          3rd Qu.:0.00000         3rd Qu.:0.00000
##  Max.   :1.0000          Max.   :1.00000         Max.   :1.00000
##
```

```
modelYearlyData(crash2017DF, "2017to2019")
```

```
## [1] "Year  2017to2019  Train Set size= 1578"
## [1] "Year  2017to2019  Test Set size= 678"
## [1] "********   YearModel1 for  2017to2019"
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_Num, family = "binomial",
##     data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3650  -0.2564  -0.2145  -0.1793   2.8771
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -4.4845     0.3485 -12.868  < 2e-16 ***
## Crash_Speed_Limit_Num   0.3618     0.1035   3.495 0.000474 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 415.89  on 1577  degrees of freedom
## Residual deviance: 404.31  on 1576  degrees of freedom
## AIC: 408.31
##
## Number of Fisher Scoring iterations: 6
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE   657   21
##      TRUE      0    0
##
```

```
##                Accuracy : 0.969
##                  95% CI : (0.953, 0.9807)
##     No Information Rate : 0.969
##     P-Value [Acc > NIR] : 0.5577
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 1.275e-05
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96903
##              Prevalence : 0.03097
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
## [1] "******  YearModel2 for  2017to2019"
## formula: fatal_accident_fac ~ Crash_Speed_Limit_Fac
## data:    trainDF
##
##  link  threshold nobs logLik  AIC    niter max.grad cond.H
##  logit flexible  1578 -201.36 412.72 7(0)  5.44e-12 2.1e+01
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## Crash_Speed_Limit_Fac.L   1.1023     0.3347   3.294 0.000988 ***
## Crash_Speed_Limit_Fac.Q   0.2927     0.4752   0.616 0.537898
## Crash_Speed_Limit_Fac.C  -0.3181     0.3102  -1.025 0.305217
## Crash_Speed_Limit_Fac^4  -0.4141     0.5544  -0.747 0.455088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##            Estimate Std. Error z value
## FALSE|TRUE   3.4281     0.1925    17.8
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE   657   21
##      TRUE      0    0
##
##                Accuracy : 0.969
##                  95% CI : (0.953, 0.9807)
##     No Information Rate : 0.969
##     P-Value [Acc > NIR] : 0.5577
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 1.275e-05
```

```
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96903
##              Prevalence : 0.03097
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
## [1] "******* YearModel3 for  2017to2019"
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_0to50kmh + Crash_Speed_Limit_100to110kmh +
##     Crash_Speed_Limit_60kmh + Crash_Speed_Limit_70kmh + Crash_Speed_Limit_80to90kmh,
##     family = "binomial", data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3564  -0.2044  -0.2044  -0.1984   2.8062
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.75860    0.34373  -8.025 1.01e-15 ***
## Crash_Speed_Limit_0to50kmh    -1.15912    0.51364  -2.257  0.02403 *
## Crash_Speed_Limit_100to110kmh  0.03402    0.45492   0.075  0.94038
## Crash_Speed_Limit_60kmh       -1.09949    0.42658  -2.577  0.00995 **
## Crash_Speed_Limit_70kmh       -1.12296    0.79272  -1.417  0.15660
## Crash_Speed_Limit_80to90kmh         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 415.89  on 1577  degrees of freedom
## Residual deviance: 402.72  on 1573  degrees of freedom
## AIC: 412.72
##
## Number of Fisher Scoring iterations: 6

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE   657   21
##      TRUE      0    0
##
##                Accuracy : 0.969
##                  95% CI : (0.953, 0.9807)
##     No Information Rate : 0.969
```

```
##     P-Value [Acc > NIR] : 0.5577
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 1.275e-05
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96903
##              Prevalence : 0.03097
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
## [1] "*********** YearModel4 for  2017to2019"
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_0to50kmh + Crash_Speed_Limit_60kmh +
##     Crash_Speed_Limit_70kmh + Crash_Speed_Limit_80to90kmh, family = "binomial",
##     data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3564  -0.2044  -0.2044  -0.1984   2.8062
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.72458    0.29799  -9.143  < 2e-16 ***
## Crash_Speed_Limit_0to50kmh  -1.19314    0.48423  -2.464  0.01374 *
## Crash_Speed_Limit_60kmh     -1.13352    0.39065  -2.902  0.00371 **
## Crash_Speed_Limit_70kmh     -1.15698    0.77398  -1.495  0.13495
## Crash_Speed_Limit_80to90kmh -0.03402    0.45492  -0.075  0.94038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 415.89  on 1577  degrees of freedom
## Residual deviance: 402.72  on 1573  degrees of freedom
## AIC: 412.72
##
## Number of Fisher Scoring iterations: 6
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE   657   21
##      TRUE      0    0
##
##                Accuracy : 0.969
```

```
##                  95% CI : (0.953, 0.9807)
##     No Information Rate : 0.969
##     P-Value [Acc > NIR] : 0.5577
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 1.275e-05
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96903
##              Prevalence : 0.03097
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
```

2008 to 2016 Inclusive

```
crash2008DF <- filter(crashDF, Crash_Year >= 2008 & Crash_Year < 2017)
```

There are 11754 observations. With 404 fatalities

```
summary( crash2008DF)
```

```
##        X1           crash_id      Crash_Severity       Crash_Year
##  Min.   : 1122   Min.   : 16909   Length:11754       Min.   :2008
##  1st Qu.: 6832   1st Qu.: 80104   Class :character   1st Qu.:2009
##  Median :15298   Median :181045   Mode  :character   Median :2012
##  Mean   :14410   Mean   :170013                      Mean   :2012
##  3rd Qu.:21321   3rd Qu.:247118                      3rd Qu.:2014
##  Max.   :28322   Max.   :328148                      Max.   :2016
##
##  Crash_Month        Crash_Day_Of_Week    Crash_Hour    Crash_Nature
##  Length:11754       Length:11754       Min.   : 0.00   Length:11754
##  Class :character   Class :character   1st Qu.: 9.00   Class :character
##  Mode  :character   Mode  :character   Median :13.00   Mode  :character
##                                        Mean   :12.84
##                                        3rd Qu.:16.00
##                                        Max.   :23.00
##
##   Crash_Type        Crash_Longitude_GDA94 Crash_Latitude_GDA94
##  Length:11754       Min.   :138.0         Min.   :-29.00
##  Class :character   1st Qu.:152.0         1st Qu.:-27.59
##  Mode  :character   Median :153.0         Median :-27.41
##                     Mean   :151.8         Mean   :-25.90
##                     3rd Qu.:153.1         3rd Qu.:-26.30
##                     Max.   :153.5         Max.   :-10.79
##
##  Crash_Street        Loc_Suburb        Loc_Local_Government_Area
##  Length:11754       Length:11754       Length:11754
##  Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Loc_Post_Code      Loc_Main_Roads_Region Loc_ABS_Remoteness
##  Length:11754       Length:11754          Length:11754
##  Class :character   Class :character      Class :character
##  Mode  :character   Mode  :character      Mode  :character
##
##
##
##
##  Crash_Controlling_Authority Crash_Roadway_Feature Crash_Traffic_Control
##  Length:11754                Length:11754          Length:11754
##  Class :character            Class :character      Class :character
##  Mode  :character            Mode  :character      Mode  :character
##
##
##
##
##  Crash_Speed_Limit  Crash_Road_Surface_Condition Crash_Atmospheric_Condition
##  Length:11754       Length:11754                 Length:11754
##  Class :character   Class :character             Class :character
##  Mode  :character   Mode  :character             Mode  :character
##
##
##
##
##  Crash_Lighting_Condition Crash_Road_Horiz_Align Crash_Road_Vert_Align
##  Length:11754             Length:11754          Length:11754
##  Class :character         Class :character      Class :character
##  Mode  :character         Mode  :character      Mode  :character
##
##
##
##
##  Crash_DCA_Code  Crash_DCA_Description Crash_DCA_Group_Description
##  Min.   :  0.0   Length:11754          Length:11754
##  1st Qu.:202.0   Class :character      Class :character
##  Median :400.0   Mode  :character      Mode  :character
##  Mean   :455.1
##  3rd Qu.:705.0
##  Max.   :907.0
##
##  DCA_Key_Approach_Dir Count_Casualty_Fatality Count_Casualty_Hospitalised
##  Length:11754         Min.   :0.00000         Min.   :0.000
##  Class :character     1st Qu.:0.00000         1st Qu.:0.000
##  Mode  :character     Median :0.00000         Median :1.000
##                       Mean   :0.03531         Mean   :0.614
##                       3rd Qu.:0.00000         3rd Qu.:1.000
##                       Max.   :2.00000         Max.   :4.000
##
##  Count_Casualty_MedicallyTreated Count_Casualty_MinorInjury
```

```
##  Min.   :0.0000                  Min.   :0.0000
##  1st Qu.:0.0000                  1st Qu.:0.0000
##  Median :0.0000                  Median :0.0000
##  Mean   :0.3295                  Mean   :0.1228
##  3rd Qu.:1.0000                  3rd Qu.:0.0000
##  Max.   :7.0000                  Max.   :3.0000
##
##  Count_Casualty_Total Count_Unit_Motorcycle_Moped   site_id_1
##  Min.   :0.000        Min.   :1.000               Min.   : 27005
##  1st Qu.:1.000        1st Qu.:1.000               1st Qu.: 40145
##  Median :1.000        Median :1.000               Median : 40460
##  Mean   :1.102        Mean   :1.021               Mean   : 39214
##  3rd Qu.:1.000        3rd Qu.:1.000               3rd Qu.: 40908
##  Max.   :8.000        Max.   :4.000               Max.   :140007
##
##  site_name_1          distance_1        site_id_list_2   site_name_2
##  Length:11754       Min.   :    33.14   Min.   : 27006   Length:11754
##  Class :character   1st Qu.:  1952.22   1st Qu.: 40157   Class :character
##  Mode  :character   Median :  3267.34   Median : 40412   Mode  :character
##                     Mean   :  4098.91   Mean   : 40266
##                     3rd Qu.:  5461.06   3rd Qu.: 40875
##                     Max.   :135185.90   Max.   :140010
##
##    distance_2       site_id_list_3   site_name_3          distance_3
##  Min.   :   174.3   Min.   : 27006   Length:11754       Min.   :   509.5
##  1st Qu.:  3807.4   1st Qu.: 40151   Class :character   1st Qu.:  5354.3
##  Median :  5423.0   Median : 40476   Mode  :character   Median :  7556.4
##  Mean   :  6607.1   Mean   : 39962                      Mean   :  9122.2
##  3rd Qu.:  8039.7   3rd Qu.: 40874                      3rd Qu.: 11279.8
##  Max.   :135299.1   Max.   :140010                      Max.   :160293.0
##
##       Lat              Lon           rainfall       Crash_Nature_Angle
##  Min.   :-29.00   Min.   :138.7   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.:-27.58   1st Qu.:152.0   1st Qu.: 0.7735   1st Qu.:0.0000
##  Median :-27.42   Median :153.0   Median : 2.0917   Median :0.0000
##  Mean   :-25.90   Mean   :151.8   Mean   : 3.6295   Mean   :0.2967
##  3rd Qu.:-26.32   3rd Qu.:153.1   3rd Qu.: 4.7186   3rd Qu.:1.0000
##  Max.   :-10.72   Max.   :153.5   Max.   :68.5400   Max.   :1.0000
##
##  Crash_Nature_Collision_miscellaneous Crash_Nature_Fall_from_vehicle
##  Min.   :0.000000                     Min.   :0.0000
##  1st Qu.:0.000000                     1st Qu.:0.0000
##  Median :0.000000                     Median :0.0000
##  Mean   :0.002042                     Mean   :0.2665
##  3rd Qu.:0.000000                     3rd Qu.:1.0000
##  Max.   :1.000000                     Max.   :1.0000
##
##  Crash_Nature_Head_on Crash_Nature_Hit_animal Crash_Nature_Hit_object
##  Min.   :0.00000      Min.   :0.00000        Min.   :0.0000
##  1st Qu.:0.00000      1st Qu.:0.00000        1st Qu.:0.0000
##  Median :0.00000      Median :0.00000        Median :0.0000
##  Mean   :0.02535      Mean   :0.02739        Mean   :0.1588
##  3rd Qu.:0.00000      3rd Qu.:0.00000        3rd Qu.:0.0000
##  Max.   :1.00000      Max.   :1.00000        Max.   :1.0000
```

```
##
##  Crash_Nature_Hit_parked_vehicle Crash_Nature_Hit_pedestrian
##  Min.   :0.00000                 Min.   :0.000000
##  1st Qu.:0.00000                 1st Qu.:0.000000
##  Median :0.00000                 Median :0.000000
##  Mean   :0.01021                 Mean   :0.009018
##  3rd Qu.:0.00000                 3rd Qu.:0.000000
##  Max.   :1.00000                 Max.   :1.000000
##
##  Crash_Nature_Non_collision_miscellaneous Crash_Nature_Overturned
##  Min.   :0.0000000                        Min.   :0.000000
##  1st Qu.:0.0000000                        1st Qu.:0.000000
##  Median :0.0000000                        Median :0.000000
##  Mean   :0.0003403                        Mean   :0.002893
##  3rd Qu.:0.0000000                        3rd Qu.:0.000000
##  Max.   :1.0000000                        Max.   :1.000000
##
##  Crash_Nature_Rear_end Crash_Nature_Sideswipe
##  Min.   :0.0000        Min.   :0.00000
##  1st Qu.:0.0000        1st Qu.:0.00000
##  Median :0.0000        Median :0.00000
##  Mean   :0.1262        Mean   :0.07342
##  3rd Qu.:0.0000        3rd Qu.:0.00000
##  Max.   :1.0000        Max.   :1.00000
##
##  Crash_Nature_Struck_by_external_load            Crash_Severity_Fac
##  Min.   :0.000000                     Property_damage_only:  99
##  1st Qu.:0.000000                     Minor_injury        :1048
##  Median :0.000000                     Medical_treatment   :3399
##  Mean   :0.001191                     Hospitalisation     :6804
##  3rd Qu.:0.000000                     Fatal               : 404
##  Max.   :1.000000
##
##  Crash_Severity_Num fatal_accident  fatal_accident_fac
##  Min.   :1.000      Mode :logical   FALSE:11350
##  1st Qu.:3.000      FALSE:11350     TRUE :  404
##  Median :4.000      TRUE :404
##  Mean   :3.542
##  3rd Qu.:4.000
##  Max.   :5.000
##
##           Crash_Nature_Fac Crash_Nature_Num Crash_Speed_Limit_Fac
##  Angle            :3487    Min.   : 1.000   0to50kmh   :2244
##  Fall from vehicle:3132    1st Qu.: 1.000   60kmh      :6064
##  Hit object       :1867    Median : 3.000   70kmh      : 616
##  Rear-end         :1483    Mean   : 4.752   80to90kmh  :1195
##  Sideswipe        : 863    3rd Qu.: 6.000   100to110kmh:1635
##  Hit animal       : 322    Max.   :13.000
##  (Other)          : 600
##  Crash_Speed_Limit_Num Crash_Speed_Limit_0to50kmh Crash_Speed_Limit_100to110kmh
##  Min.   :1.000         Min.   :0.0000            Min.   :0.0000
##  1st Qu.:2.000         1st Qu.:0.0000            1st Qu.:0.0000
##  Median :2.000         Median :0.0000            Median :0.0000
##  Mean   :2.482         Mean   :0.1909            Mean   :0.1391
```

```
##  3rd Qu.:3.000         3rd Qu.:0.0000             3rd Qu.:0.0000
##  Max.   :5.000         Max.   :1.0000             Max.   :1.0000
##
##  Crash_Speed_Limit_60kmh Crash_Speed_Limit_70kmh Crash_Speed_Limit_80to90kmh
##  Min.   :0.0000          Min.   :0.00000         Min.   :0.0000
##  1st Qu.:0.0000          1st Qu.:0.00000         1st Qu.:0.0000
##  Median :1.0000          Median :0.00000         Median :0.0000
##  Mean   :0.5159          Mean   :0.05241         Mean   :0.1017
##  3rd Qu.:1.0000          3rd Qu.:0.00000         3rd Qu.:0.0000
##  Max.   :1.0000          Max.   :1.00000         Max.   :1.0000
##
```

```
modelYearlyData(crash2008DF, "2008to2016")
```

```
## [1] "Year  2008to2016  Train Set size= 8226"
## [1] "Year  2008to2016  Test Set size= 3528"
## [1] "********   YearModel1 for  2008to2016"
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_Num, family = "binomial",
##     data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4250  -0.2708  -0.2155  -0.2155   2.9086
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.67938    0.15256  -30.67   <2e-16 ***
## Crash_Speed_Limit_Num  0.46404    0.04208   11.03   <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: 2456.7  on 8225  degrees of freedom
## Residual deviance: 2337.5  on 8224  degrees of freedom
## AIC: 2341.5
##
## Number of Fisher Scoring iterations: 6
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3406  122
##      TRUE      0    0
##
##                Accuracy : 0.9654
##                  95% CI : (0.9589, 0.9712)
##     No Information Rate : 0.9654
##     P-Value [Acc > NIR] : 0.5241
##
##                   Kappa : 0
##
```

```
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96542
##              Prevalence : 0.03458
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
## [1] "******  YearModel2 for  2008to2016"
## formula: fatal_accident_fac ~ Crash_Speed_Limit_Fac
## data:    trainDF
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  8226 -1166.65 2343.30 7(0)  5.18e-11 1.9e+01
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## Crash_Speed_Limit_Fac.L  1.38460    0.14725   9.403   <2e-16 ***
## Crash_Speed_Limit_Fac.Q  0.13934    0.19130   0.728   0.4664
## Crash_Speed_Limit_Fac.C -0.26118    0.13034  -2.004   0.0451 *
## Crash_Speed_Limit_Fac^4 -0.06599    0.21379  -0.309   0.7576
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##            Estimate Std. Error z value
## FALSE|TRUE  3.25285    0.07777   41.83
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3406  122
##      TRUE      0    0
##
##                Accuracy : 0.9654
##                  95% CI : (0.9589, 0.9712)
##     No Information Rate : 0.9654
##     P-Value [Acc > NIR] : 0.5241
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96542
##              Prevalence : 0.03458
##          Detection Rate : 0.00000
```

```
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
## [1] "******* YearModel3 for  2008to2016"
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_0to50kmh + Crash_Speed_Limit_100to110kmh +
##     Crash_Speed_Limit_60kmh + Crash_Speed_Limit_70kmh + Crash_Speed_Limit_80to90kmh,
##     family = "binomial", data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4181  -0.2595  -0.2040  -0.2040   2.8277
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.6555     0.1382 -19.209  < 2e-16 ***
## Crash_Speed_Limit_0to50kmh     -1.3239     0.2329  -5.685 1.31e-08 ***
## Crash_Speed_Limit_100to110kmh   0.2624     0.1746   1.503   0.1329
## Crash_Speed_Limit_60kmh        -1.2061     0.1756  -6.867 6.56e-12 ***
## Crash_Speed_Limit_70kmh        -0.7192     0.3049  -2.358   0.0184 *
## Crash_Speed_Limit_80to90kmh         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2456.7  on 8225  degrees of freedom
## Residual deviance: 2333.3  on 8221  degrees of freedom
## AIC: 2343.3
##
## Number of Fisher Scoring iterations: 6

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading

## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3406  122
##      TRUE      0    0
##
##                Accuracy : 0.9654
##                  95% CI : (0.9589, 0.9712)
##     No Information Rate : 0.9654
##     P-Value [Acc > NIR] : 0.5241
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
```

```
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96542
##              Prevalence : 0.03458
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
## [1] "*********** YearModel4 for  2008to2016"
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_0to50kmh + Crash_Speed_Limit_60kmh +
##     Crash_Speed_Limit_70kmh + Crash_Speed_Limit_80to90kmh, family = "binomial",
##     data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4181  -0.2595  -0.2040  -0.2040   2.8277
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.3931     0.1066 -22.445  < 2e-16 ***
## Crash_Speed_Limit_0to50kmh   -1.5862     0.2156  -7.357 1.88e-13 ***
## Crash_Speed_Limit_60kmh      -1.4684     0.1520  -9.661  < 2e-16 ***
## Crash_Speed_Limit_70kmh      -0.9815     0.2920  -3.362 0.000774 ***
## Crash_Speed_Limit_80to90kmh  -0.2624     0.1746  -1.503 0.132910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2456.7  on 8225  degrees of freedom
## Residual deviance: 2333.3  on 8221  degrees of freedom
## AIC: 2343.3
##
## Number of Fisher Scoring iterations: 6
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3406  122
##      TRUE      0    0
##
##                Accuracy : 0.9654
##                  95% CI : (0.9589, 0.9712)
##     No Information Rate : 0.9654
##     P-Value [Acc > NIR] : 0.5241
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
```

```
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96542
##              Prevalence : 0.03458
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
```

Before 2008

```
crash2001DF <- filter(crashDF, Crash_Year < 2008)
```

There are 8747 observations. With 290 fatalities

```
summary( crash2001DF)
```

```
##        X1           crash_id      Crash_Severity       Crash_Year
##  Min.   :    0   Min.   :    26   Length:8747        Min.   :2001
##  1st Qu.: 7086   1st Qu.: 82248   Class :character   1st Qu.:2003
##  Median :13616   Median :159311   Mode  :character   Median :2005
##  Mean   :13429   Mean   :158014                      Mean   :2004
##  3rd Qu.:21198   3rd Qu.:246003                      3rd Qu.:2006
##  Max.   :27110   Max.   :318341                      Max.   :2007
##
##  Crash_Month        Crash_Day_Of_Week    Crash_Hour    Crash_Nature
##  Length:8747        Length:8747        Min.   : 0.00   Length:8747
##  Class :character   Class :character   1st Qu.:10.00   Class :character
##  Mode  :character   Mode  :character   Median :14.00   Mode  :character
##                                        Mean   :13.24
##                                        3rd Qu.:17.00
##                                        Max.   :23.00
##
##   Crash_Type        Crash_Longitude_GDA94 Crash_Latitude_GDA94
##  Length:8747        Min.   :  0.0         Min.   :-28.90
##  Class :character   1st Qu.:151.9         1st Qu.:-27.58
##  Mode  :character   Median :153.0         Median :-27.44
##                     Mean   :151.6         Mean   :-25.70
##                     3rd Qu.:153.1         3rd Qu.:-25.55
##                     Max.   :153.5         Max.   :  0.00
##
##  Crash_Street        Loc_Suburb        Loc_Local_Government_Area
##  Length:8747        Length:8747        Length:8747
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Loc_Post_Code      Loc_Main_Roads_Region Loc_ABS_Remoteness
##  Length:8747        Length:8747           Length:8747
```

```
##  Class :character   Class :character      Class :character
##  Mode  :character   Mode  :character      Mode  :character
##
##
##
##
##  Crash_Controlling_Authority Crash_Roadway_Feature Crash_Traffic_Control
##  Length:8747                 Length:8747           Length:8747
##  Class :character            Class :character      Class :character
##  Mode  :character            Mode  :character      Mode  :character
##
##
##
##
##  Crash_Speed_Limit  Crash_Road_Surface_Condition Crash_Atmospheric_Condition
##  Length:8747        Length:8747                  Length:8747
##  Class :character   Class :character             Class :character
##  Mode  :character   Mode  :character             Mode  :character
##
##
##
##
##  Crash_Lighting_Condition Crash_Road_Horiz_Align Crash_Road_Vert_Align
##  Length:8747              Length:8747           Length:8747
##  Class :character         Class :character      Class :character
##  Mode  :character         Mode  :character      Mode  :character
##
##
##
##
##  Crash_DCA_Code Crash_DCA_Description Crash_DCA_Group_Description
##  Min.   :  0    Length:8747           Length:8747
##  1st Qu.:202    Class :character      Class :character
##  Median :308    Mode  :character      Mode  :character
##  Mean   :428
##  3rd Qu.:705
##  Max.   :904
##
##  DCA_Key_Approach_Dir Count_Casualty_Fatality Count_Casualty_Hospitalised
##  Length:8747          Min.   :0.00000         Min.   :0.0000
##  Class :character     1st Qu.:0.00000         1st Qu.:0.0000
##  Mode  :character     Median :0.00000         Median :0.0000
##                       Mean   :0.03395         Mean   :0.5275
##                       3rd Qu.:0.00000         3rd Qu.:1.0000
##                       Max.   :2.00000         Max.   :5.0000
##
##  Count_Casualty_MedicallyTreated Count_Casualty_MinorInjury
##  Min.   :0.0000                  Min.   :0.0000
##  1st Qu.:0.0000                  1st Qu.:0.0000
##  Median :0.0000                  Median :0.0000
##  Mean   :0.3395                  Mean   :0.1932
##  3rd Qu.:1.0000                  3rd Qu.:0.0000
##  Max.   :5.0000                  Max.   :4.0000
##
```

```
##  Count_Casualty_Total Count_Unit_Motorcycle_Moped   site_id_1
##  Min.   :0.000        Min.   :1.000               Min.   :27005
##  1st Qu.:1.000        1st Qu.:1.000               1st Qu.:40115
##  Median :1.000        Median :1.000               Median :40416
##  Mean   :1.094        Mean   :1.018               Mean   :38912
##  3rd Qu.:1.000        3rd Qu.:1.000               3rd Qu.:40861
##  Max.   :7.000        Max.   :5.000               Max.   :45063
##
##  site_name_1          distance_1       site_id_list_2   site_name_2
##  Length:8747        Min.   :      22   Min.   : 27006   Length:8747
##  Class :character   1st Qu.:    1854   1st Qu.: 40141   Class :character
##  Mode  :character   Median :    3108   Median : 40383   Mode  :character
##                     Mean   :    7262   Mean   : 40065
##                     3rd Qu.:    5188   3rd Qu.: 40848
##                     Max.   :14769184   Max.   :140009
##
##    distance_2       site_id_list_3   site_name_3          distance_3
##  Min.   :     308   Min.   : 27034   Length:8747        Min.   :     521
##  1st Qu.:    3671   1st Qu.: 40115   Class :character   1st Qu.:    5213
##  Median :    5147   Median : 40417   Mode  :character   Median :    7218
##  Mean   :    9615   Mean   : 39754                      Mean   :   12010
##  3rd Qu.:    7427   3rd Qu.: 40849                      3rd Qu.:   10894
##  Max.   :14793133   Max.   :140010                      Max.   :14793149
##
##       Lat              Lon           rainfall       Crash_Nature_Angle
##  Min.   :-29.00   Min.   :138.1   Min.   :  0.000   Min.   :0.0000
##  1st Qu.:-27.58   1st Qu.:151.9   1st Qu.:  0.669   1st Qu.:0.0000
##  Median :-27.43   Median :153.0   Median :  1.735   Median :0.0000
##  Mean   :-25.70   Mean   :151.7   Mean   :  2.777   Mean   :0.3442
##  3rd Qu.:-25.51   3rd Qu.:153.1   3rd Qu.:  3.540   3rd Qu.:1.0000
##  Max.   :-10.90   Max.   :153.5   Max.   :232.000   Max.   :1.0000
##
##  Crash_Nature_Collision_miscellaneous Crash_Nature_Fall_from_vehicle
##  Min.   :0.000000                     Min.   :0.0000
##  1st Qu.:0.000000                     1st Qu.:0.0000
##  Median :0.000000                     Median :0.0000
##  Mean   :0.003887                     Mean   :0.2275
##  3rd Qu.:0.000000                     3rd Qu.:0.0000
##  Max.   :1.000000                     Max.   :1.0000
##
##  Crash_Nature_Head_on Crash_Nature_Hit_animal Crash_Nature_Hit_object
##  Min.   :0.00000      Min.   :0.00000        Min.   :0.0000
##  1st Qu.:0.00000      1st Qu.:0.00000        1st Qu.:0.0000
##  Median :0.00000      Median :0.00000        Median :0.0000
##  Mean   :0.02389      Mean   :0.02515        Mean   :0.1483
##  3rd Qu.:0.00000      3rd Qu.:0.00000        3rd Qu.:0.0000
##  Max.   :1.00000      Max.   :1.00000        Max.   :1.0000
##
##  Crash_Nature_Hit_parked_vehicle Crash_Nature_Hit_pedestrian
##  Min.   :0.00000                 Min.   :0.0000
##  1st Qu.:0.00000                 1st Qu.:0.0000
##  Median :0.00000                 Median :0.0000
##  Mean   :0.01017                 Mean   :0.0104
##  3rd Qu.:0.00000                 3rd Qu.:0.0000
```

```
##  Max.   :1.00000                Max.   :1.0000
##
##  Crash_Nature_Non_collision_miscellaneous Crash_Nature_Overturned
##  Min.   :0.000000                         Min.   :0.000000
##  1st Qu.:0.000000                         1st Qu.:0.000000
##  Median :0.000000                         Median :0.000000
##  Mean   :0.001372                         Mean   :0.000343
##  3rd Qu.:0.000000                         3rd Qu.:0.000000
##  Max.   :1.000000                         Max.   :1.000000
##
##  Crash_Nature_Rear_end Crash_Nature_Sideswipe
##  Min.   :0.0000        Min.   :0.00000
##  1st Qu.:0.0000        1st Qu.:0.00000
##  Median :0.0000        Median :0.00000
##  Mean   :0.1206        Mean   :0.08231
##  3rd Qu.:0.0000        3rd Qu.:0.00000
##  Max.   :1.0000        Max.   :1.00000
##
##  Crash_Nature_Struck_by_external_load           Crash_Severity_Fac
##  Min.   :0.000000                    Property_damage_only: 155
##  1st Qu.:0.000000                    Minor_injury        :1362
##  Median :0.000000                    Medical_treatment   :2617
##  Mean   :0.001829                    Hospitalisation     :4323
##  3rd Qu.:0.000000                    Fatal               : 290
##  Max.   :1.000000
##
##  Crash_Severity_Num fatal_accident  fatal_accident_fac
##  Min.   :1.000      Mode :logical   FALSE:8457
##  1st Qu.:3.000      FALSE:8457      TRUE : 290
##  Median :4.000      TRUE :290
##  Mean   :3.369
##  3rd Qu.:4.000
##  Max.   :5.000
##
##           Crash_Nature_Fac Crash_Nature_Num Crash_Speed_Limit_Fac
##  Angle            :3011    Min.   : 1.000   0to50kmh   :1296
##  Fall from vehicle:1990    1st Qu.: 1.000   60kmh      :4835
##  Hit object       :1297    Median : 3.000   70kmh      : 467
##  Rear-end         :1055    Mean   : 4.654   80to90kmh  : 872
##  Sideswipe        : 720    3rd Qu.: 6.000   100to110kmh:1277
##  Hit animal       : 220    Max.   :13.000
##  (Other)          : 454
##  Crash_Speed_Limit_Num Crash_Speed_Limit_0to50kmh Crash_Speed_Limit_100to110kmh
##  Min.   :1.000         Min.   :0.0000             Min.   :0.000
##  1st Qu.:2.000         1st Qu.:0.0000             1st Qu.:0.000
##  Median :2.000         Median :0.0000             Median :0.000
##  Mean   :2.543         Mean   :0.1482             Mean   :0.146
##  3rd Qu.:3.000         3rd Qu.:0.0000             3rd Qu.:0.000
##  Max.   :5.000         Max.   :1.0000             Max.   :1.000
##
##  Crash_Speed_Limit_60kmh Crash_Speed_Limit_70kmh Crash_Speed_Limit_80to90kmh
##  Min.   :0.0000          Min.   :0.00000         Min.   :0.00000
##  1st Qu.:0.0000          1st Qu.:0.00000         1st Qu.:0.00000
##  Median :1.0000          Median :0.00000         Median :0.00000
```

```
##  Mean   :0.5528          Mean   :0.05339         Mean   :0.09969
##  3rd Qu.:1.0000          3rd Qu.:0.00000         3rd Qu.:0.00000
##  Max.   :1.0000          Max.   :1.00000         Max.   :1.00000
##
```

```
modelYearlyData(crash2008DF, "Prior2008")
```

```
## [1] "Year  Prior2008  Train Set size= 8226"
## [1] "Year  Prior2008  Test Set size= 3528"
## [1] "********   YearModel1 for  Prior2008"
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_Num, family = "binomial",
##     data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4250  -0.2708  -0.2155  -0.2155   2.9086
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.67938    0.15256  -30.67   <2e-16 ***
## Crash_Speed_Limit_Num  0.46404    0.04208   11.03   <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: 2456.7  on 8225  degrees of freedom
## Residual deviance: 2337.5  on 8224  degrees of freedom
## AIC: 2341.5
##
## Number of Fisher Scoring iterations: 6
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3406  122
##      TRUE      0    0
##
##                Accuracy : 0.9654
##                  95% CI : (0.9589, 0.9712)
##     No Information Rate : 0.9654
##     P-Value [Acc > NIR] : 0.5241
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96542
##              Prevalence : 0.03458
```

```
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
## [1] "******  YearModel2 for  Prior2008"
## formula: fatal_accident_fac ~ Crash_Speed_Limit_Fac
## data:    trainDF
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  8226 -1166.65 2343.30 7(0)  5.18e-11 1.9e+01
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## Crash_Speed_Limit_Fac.L  1.38460    0.14725   9.403   <2e-16 ***
## Crash_Speed_Limit_Fac.Q  0.13934    0.19130   0.728   0.4664
## Crash_Speed_Limit_Fac.C -0.26118    0.13034  -2.004   0.0451 *
## Crash_Speed_Limit_Fac^4 -0.06599    0.21379  -0.309   0.7576
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##            Estimate Std. Error z value
## FALSE|TRUE  3.25285    0.07777   41.83
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3406  122
##      TRUE      0    0
##
##                Accuracy : 0.9654
##                  95% CI : (0.9589, 0.9712)
##     No Information Rate : 0.9654
##     P-Value [Acc > NIR] : 0.5241
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96542
##              Prevalence : 0.03458
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
## [1] "******* YearModel3 for  Prior2008"
##
```

```
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_0to50kmh + Crash_Speed_Limit_100to110kmh +
##     Crash_Speed_Limit_60kmh + Crash_Speed_Limit_70kmh + Crash_Speed_Limit_80to90kmh,
##     family = "binomial", data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4181  -0.2595  -0.2040  -0.2040   2.8277
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.6555     0.1382 -19.209  < 2e-16 ***
## Crash_Speed_Limit_0to50kmh     -1.3239     0.2329  -5.685 1.31e-08 ***
## Crash_Speed_Limit_100to110kmh   0.2624     0.1746   1.503   0.1329
## Crash_Speed_Limit_60kmh        -1.2061     0.1756  -6.867 6.56e-12 ***
## Crash_Speed_Limit_70kmh        -0.7192     0.3049  -2.358   0.0184 *
## Crash_Speed_Limit_80to90kmh         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2456.7  on 8225  degrees of freedom
## Residual deviance: 2333.3  on 8221  degrees of freedom
## AIC: 2343.3
##
## Number of Fisher Scoring iterations: 6
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3406  122
##      TRUE      0    0
##
##                Accuracy : 0.9654
##                  95% CI : (0.9589, 0.9712)
##     No Information Rate : 0.9654
##     P-Value [Acc > NIR] : 0.5241
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96542
##              Prevalence : 0.03458
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
```

```
##
##        'Positive' Class : TRUE
##
## [1] "*********** YearModel4 for  Prior2008"
##
## Call:
## glm(formula = fatal_accident ~ Crash_Speed_Limit_0to50kmh + Crash_Speed_Limit_60kmh +
##     Crash_Speed_Limit_70kmh + Crash_Speed_Limit_80to90kmh, family = "binomial",
##     data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4181  -0.2595  -0.2040  -0.2040   2.8277
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.3931     0.1066 -22.445  < 2e-16 ***
## Crash_Speed_Limit_0to50kmh   -1.5862     0.2156  -7.357 1.88e-13 ***
## Crash_Speed_Limit_60kmh      -1.4684     0.1520  -9.661  < 2e-16 ***
## Crash_Speed_Limit_70kmh      -0.9815     0.2920  -3.362 0.000774 ***
## Crash_Speed_Limit_80to90kmh  -0.2624     0.1746  -1.503 0.132910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2456.7  on 8225  degrees of freedom
## Residual deviance: 2333.3  on 8221  degrees of freedom
## AIC: 2343.3
##
## Number of Fisher Scoring iterations: 6
##
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  3406  122
##      TRUE      0    0
##
##                Accuracy : 0.9654
##                  95% CI : (0.9589, 0.9712)
##     No Information Rate : 0.9654
##     P-Value [Acc > NIR] : 0.5241
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96542
##              Prevalence : 0.03458
##          Detection Rate : 0.00000
```

```
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
```

Speed zone modeling

Select data by speed zone and then run models on that.

100 to 100Km zone

model 1

Select all observations in the 100 to 110km zone. Run some linear regression models using grash nature.

```
natureModel1 <- glm( fatal_accident ~
                Crash_Nature_Angle+ Crash_Nature_Collision_miscellaneous+
                Crash_Nature_Fall_from_vehicle+ Crash_Nature_Head_on+
                Crash_Nature_Hit_animal+ Crash_Nature_Hit_object+
                Crash_Nature_Hit_parked_vehicle+ Crash_Nature_Hit_pedestrian+
                Crash_Nature_Non_collision_miscellaneous+ Crash_Nature_Overturned+
                Crash_Nature_Rear_end+ Crash_Nature_Sideswipe+
                Crash_Nature_Struck_by_external_load,
              data=trainDF, family="binomial")

print(summary(natureModel1))
```

```
##
## Call:
## glm(formula = fatal_accident ~ Crash_Nature_Angle + Crash_Nature_Collision_miscellaneous +
##     Crash_Nature_Fall_from_vehicle + Crash_Nature_Head_on + Crash_Nature_Hit_animal +
##     Crash_Nature_Hit_object + Crash_Nature_Hit_parked_vehicle +
##     Crash_Nature_Hit_pedestrian + Crash_Nature_Non_collision_miscellaneous +
##     Crash_Nature_Overturned + Crash_Nature_Rear_end + Crash_Nature_Sideswipe +
##     Crash_Nature_Struck_by_external_load, family = "binomial",
##     data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6030  -0.2334  -0.2271  -0.2200   2.9131
##
## Coefficients: (1 not defined because of singularities)
##                                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -1.557e+01  3.103e+02  -0.050    0.960
## Crash_Nature_Angle                       1.198e+01  3.103e+02   0.039    0.969
## Crash_Nature_Collision_miscellaneous     1.268e+01  3.103e+02   0.041    0.967
## Crash_Nature_Fall_from_vehicle           1.186e+01  3.103e+02   0.038    0.970
## Crash_Nature_Head_on                     1.395e+01  3.103e+02   0.045    0.964
## Crash_Nature_Hit_animal                  1.192e+01  3.103e+02   0.038    0.969
## Crash_Nature_Hit_object                  1.296e+01  3.103e+02   0.042    0.967
## Crash_Nature_Hit_parked_vehicle          1.211e+01  3.103e+02   0.039    0.969
## Crash_Nature_Hit_pedestrian              1.174e+01  3.103e+02   0.038    0.970
## Crash_Nature_Non_collision_miscellaneous -9.214e-11  6.316e+02   0.000    1.000
```

```
## Crash_Nature_Overturned                  -9.123e-11  3.960e+02   0.000    1.000
## Crash_Nature_Rear_end                     1.134e+01  3.103e+02   0.037    0.971
## Crash_Nature_Sideswipe                    1.156e+01  3.103e+02   0.037    0.970
## Crash_Nature_Struck_by_external_load             NA         NA      NA       NA
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4662.8  on 15928  degrees of freedom
## Residual deviance: 4416.2  on 15916  degrees of freedom
## AIC: 4442.2
##
## Number of Fisher Scoring iterations: 14
```

```
natureModel1.predictions <- predict(natureModel1, newdata = testDF, type="response") >= 0.5
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
    cm = confusionMatrix(factor(natureModel1.predictions, levels=c("FALSE","TRUE")),
                         factor(testDF$fatal_accident, levels=c("FALSE", "TRUE")), positive="TRUE" )
    print(cm)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  6599  229
##      TRUE      0    0
##
##                Accuracy : 0.9665
##                  95% CI : (0.9619, 0.9706)
##     No Information Rate : 0.9665
##     P-Value [Acc > NIR] : 0.5176
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96646
##              Prevalence : 0.03354
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
```

model 2

Remove some nature of crash predictors that make up low percentage of accidents.

```r
natureModel1 <- glm( fatal_accident ~
                Crash_Nature_Angle+
                Crash_Nature_Fall_from_vehicle+ Crash_Nature_Head_on+
                Crash_Nature_Hit_animal+ Crash_Nature_Hit_object+
                Crash_Nature_Rear_end+ Crash_Nature_Sideswipe,
              data=trainDF, family="binomial")

print(summary(natureModel1))
```

```
##
## Call:
## glm(formula = fatal_accident ~ Crash_Nature_Angle + Crash_Nature_Fall_from_vehicle +
##     Crash_Nature_Head_on + Crash_Nature_Hit_animal + Crash_Nature_Hit_object +
##     Crash_Nature_Rear_end + Crash_Nature_Sideswipe, family = "binomial",
##     data = trainDF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6030  -0.2334  -0.2233  -0.2200   2.9131
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.67883    0.32019 -11.489  < 2e-16 ***
## Crash_Nature_Angle              0.08973    0.33195   0.270  0.78692
## Crash_Nature_Fall_from_vehicle -0.03071    0.33628  -0.091  0.92723
## Crash_Nature_Head_on            2.06622    0.34866   5.926  3.1e-09 ***
## Crash_Nature_Hit_animal         0.03409    0.44250   0.077  0.93859
## Crash_Nature_Hit_object         1.07258    0.33014   3.249  0.00116 **
## Crash_Nature_Rear_end          -0.54976    0.37077  -1.483  0.13814
## Crash_Nature_Sideswipe         -0.33098    0.38573  -0.858  0.39086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4662.8  on 15928  degrees of freedom
## Residual deviance: 4420.6  on 15921  degrees of freedom
## AIC: 4436.6
##
## Number of Fisher Scoring iterations: 6
```

```r
natureModel1.predictions <- predict(natureModel1, newdata = testDF, type="response") >= 0.5
   cm = confusionMatrix(factor(natureModel1.predictions, levels=c("FALSE","TRUE")),
                        factor(testDF$fatal_accident, levels=c("FALSE", "TRUE")), positive="TRUE" )
   print(cm)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE TRUE
##      FALSE  6599  229
##      TRUE      0    0
##
##                Accuracy : 0.9665
```

```
##                  95% CI : (0.9619, 0.9706)
##     No Information Rate : 0.9665
##     P-Value [Acc > NIR] : 0.5176
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.96646
##              Prevalence : 0.03354
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : TRUE
##
```