Car Crash Analysis

```
library(plyr)
library(dplyr)
library(ggplot2)
library(MASS)
library(caret)
library(glmnet)
library(rpart)
```

```
# load dataset
data <- read.csv("dataset.csv")
```

EDA

```
# structure
dim(data)
```

```
## [1] 23137    14
```

```
# as we can observe, it is a dataset with 14 dimensions and 23137 observations
```

```
# missing value
sum(is.na(data))
```

```
## [1] 0
```

```
# as we can observe, there is no missing value
```

```
# head
head(data)
```

```
##   Crash_Score year Month Time_of_Day Rd_Feature   Rd_Character  Rd_Class
## 1        6.56 2016     6           2       NONE STRAIGHT-LEVEL STATE HWY
## 2        6.53 2016     6           3       NONE STRAIGHT-LEVEL     OTHER
## 3        1.58 2016     6           5       NONE STRAIGHT-LEVEL STATE HWY
## 4        7.15 2016     6           3       NONE STRAIGHT-LEVEL     OTHER
## 5        9.57 2016     6           6       NONE STRAIGHT-LEVEL     OTHER
## 6        8.14 2016     6           3       NONE STRAIGHT-LEVEL     OTHER
##           Rd_Configuration     Rd_Surface Rd_Conditions        Light
## 1 TWO-WAY-PROTECTED-MEDIAN SMOOTH ASPHALT           DRY     DAYLIGHT
## 2        TWO-WAY-NO-MEDIAN COARSE ASPHALT           DRY     DAYLIGHT
## 3        TWO-WAY-NO-MEDIAN SMOOTH ASPHALT           DRY DARK-NOT-LIT
## 4        TWO-WAY-NO-MEDIAN SMOOTH ASPHALT           DRY     DAYLIGHT
## 5        TWO-WAY-NO-MEDIAN COARSE ASPHALT           DRY     DARK-LIT
## 6        TWO-WAY-NO-MEDIAN SMOOTH ASPHALT           DRY     DAYLIGHT
##   Weather Traffic_Control Work_Area
## 1   CLEAR            NONE        NO
## 2   CLEAR            NONE        NO
## 3   CLEAR            NONE        NO
## 4   CLEAR            NONE        NO
## 5   CLEAR            NONE        NO
## 6   CLEAR            NONE        NO
```

```
# as we can observe, the dependent variable Crash_Score is a numeric variable, and most of the independent variables are c
ategorical variables
```

```
colnames(data)
```

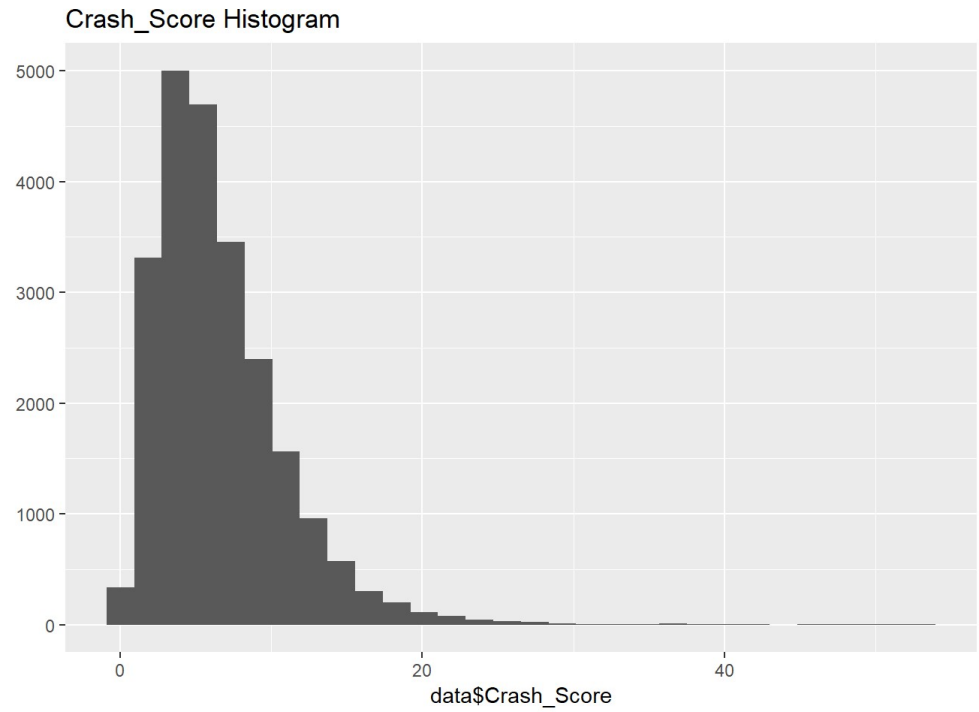
```
##  [1] "Crash_Score"      "year"             "Month"
##  [4] "Time_of_Day"      "Rd_Feature"       "Rd_Character"
##  [7] "Rd_Class"         "Rd_Configuration" "Rd_Surface"
## [10] "Rd_Conditions"    "Light"            "Weather"
## [13] "Traffic_Control"  "Work_Area"
```

```
# the dependent variable is the "Crash_Score", which measures the extent of the crash using factors such as number of inju
ries and fatalities, the number of vehicles involved, and other factors
# the independent variables could be categorized into
# time variables: "Year", "Month", "Time_of_Day"
# road variables: "Rd_Feature", "Rd_Character", "Rd_Class", "Rd_Configuration", "Rd_Surface", "Rd_Conditions"
# other variables: "Light", "Weather", "Work_Area", "Traffic_Control"
```

dependent variable

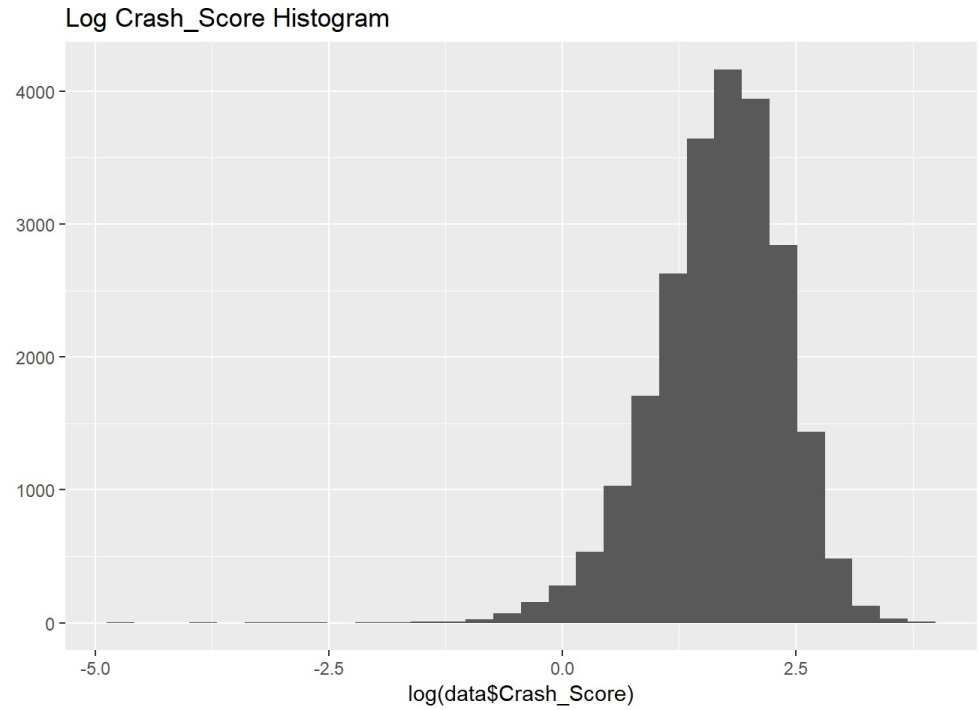
```
qplot(data$Crash_Score, geom="histogram", main="Crash_Score Histogram")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
qplot(log(data$Crash_Score), geom="histogram", main="Log Crash_Score Histogram")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
summary(data$Crash_Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.010   3.540   5.660   6.567   8.600  53.070
```

```
# as we can observe, the distribution of the Crash_Score is right skewed, with a median of 5.660, and a max of 53.070
# which indicate that most car crashes are slight car crashes and a small proportion are severe car crashes
```

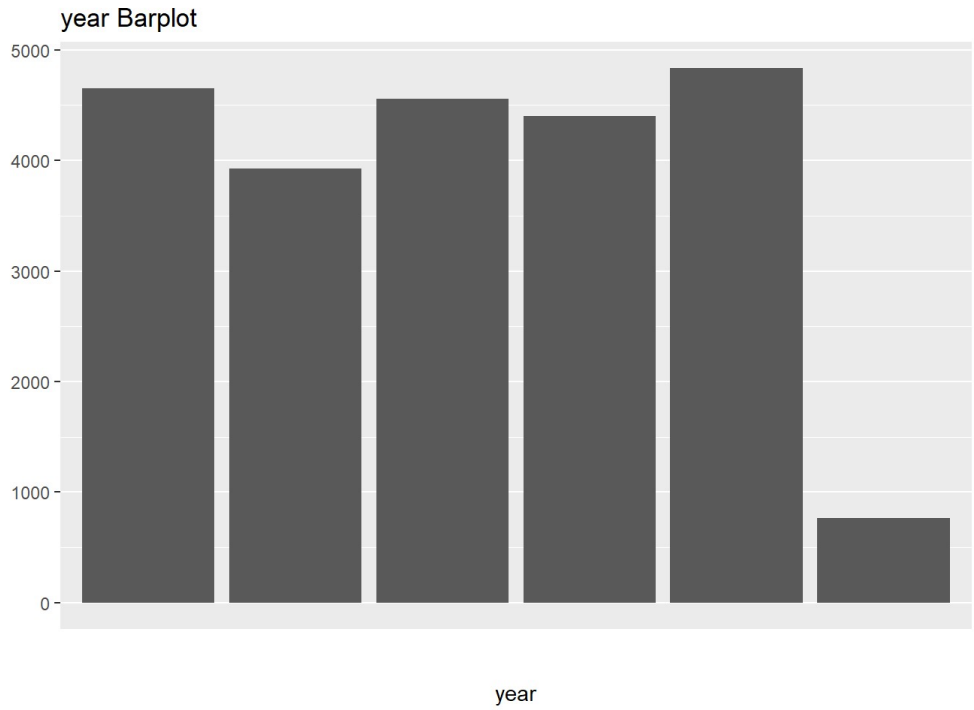
independent variable

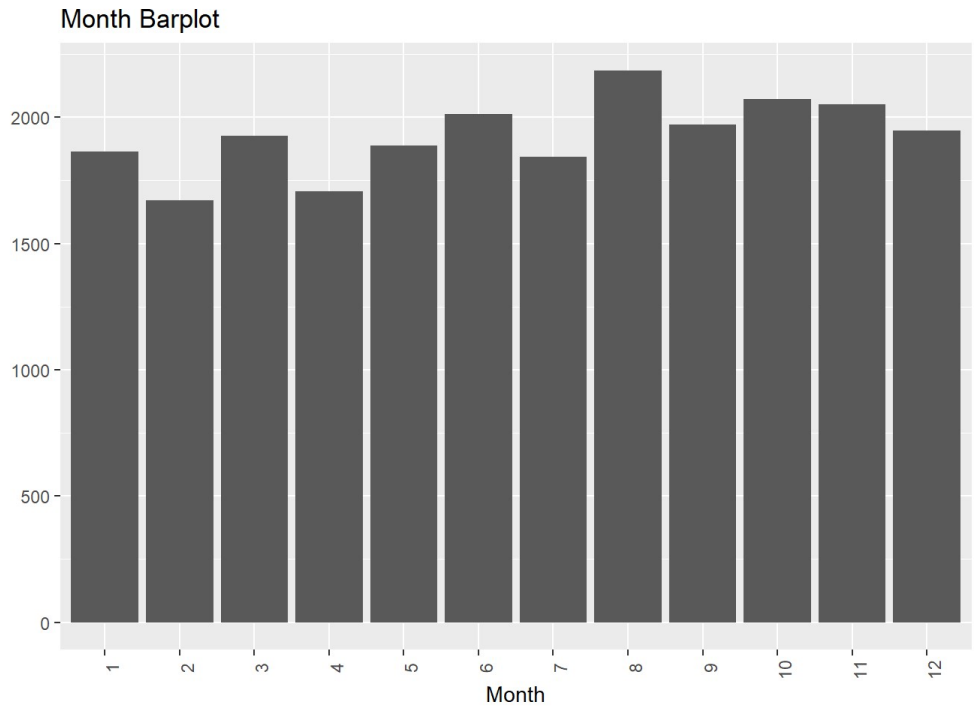
```
# relevel
vars <- colnames(data)[-(1:4)]
for (i in vars){
  table <- as.data.frame(table(data[,i]))
  # table counts the number of observations for each level of the categorical variable
  max <- which.max(table[,2])
  level.name <- as.character(table[max,1])
  data[,i] <- relevel(data[,i], ref=level.name)
}
# we relevel all the categorical variables, assign the base level to the level with the most observations
summary(data)
```

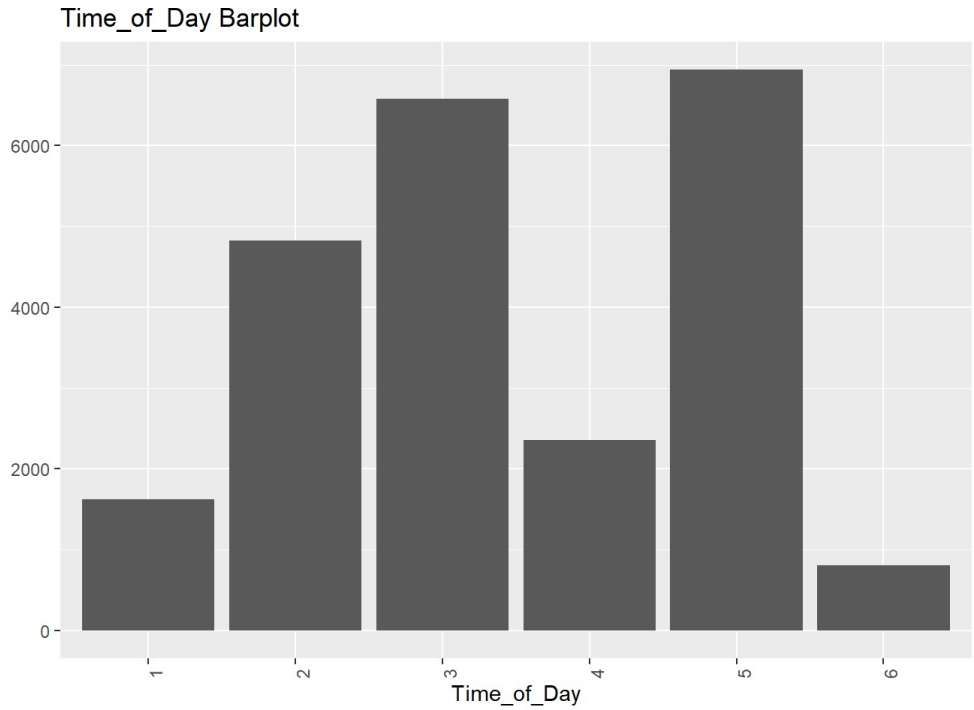
```
##   Crash_Score          year          Month        Time_of_Day
##  Min.   : 0.010   Min.   :2014   Min.   : 1.00   Min.   :1.000
##  1st Qu.: 3.540   1st Qu.:2015   1st Qu.: 3.00   1st Qu.:3.000
##  Median : 5.660   Median :2016   Median : 7.00   Median :4.000
##  Mean   : 6.567   Mean   :2016   Mean   : 6.56   Mean   :4.034
##  3rd Qu.: 8.600   3rd Qu.:2017   3rd Qu.:10.00   3rd Qu.:5.000
##  Max.   :53.070   Max.   :2019   Max.   :12.00   Max.   :6.000
##
##         Rd_Feature            Rd_Character        Rd_Class
##  NONE        :13025   STRAIGHT-LEVEL:18215   STATE HWY:10603
##  DRIVEWAY    : 2373   CURVE-GRADE   :  643   OTHER    : 9960
##  INTERSECTION: 6702   CURVE-LEVEL   :  725   US HWY   : 2574
##  OTHER       :  259   CURVE-OTHER   :  239
##  RAMP        :  778   OTHER         :   13
##                       STRAIGHT-GRADE: 2622
##                       STRAIGHT-OTHER:  680
##                   Rd_Configuration            Rd_Surface
##  TWO-WAY-NO-MEDIAN         :12076   SMOOTH ASPHALT  :20007
##  ONE-WAY                   : 1496   COARSE ASPHALT  : 1997
##  TWO-WAY-PROTECTED-MEDIAN  : 2627   CONCRETE        :  692
##  TWO-WAY-UNPROTECTED-MEDIAN: 6882   GROOVED CONCRETE:  371
##  UNKNOWN                   :   56   OTHER           :   70
##
##
##         Rd_Conditions            Light         Weather
##  DRY           :19262   DAYLIGHT    :18262   CLEAR :17393
##  ICE-SNOW-SLUSH:  322   DARK-LIT    : 3219   CLOUDY: 3234
##  OTHER         :  134   DARK-NOT-LIT:  708   OTHER :   85
##  WET           : 3419   DAWN        :  140   RAIN  : 2230
##                         DUSK        :  602   SNOW  :  195
##                         OTHER       :  206
##
##   Traffic_Control  Work_Area
##  NONE     :14028   NO :22823
##  OTHER    :  228   YES:  314
##  SIGNAL   : 6352
##  STOP-SIGN: 2269
##  YIELD    :  260
##
##
```

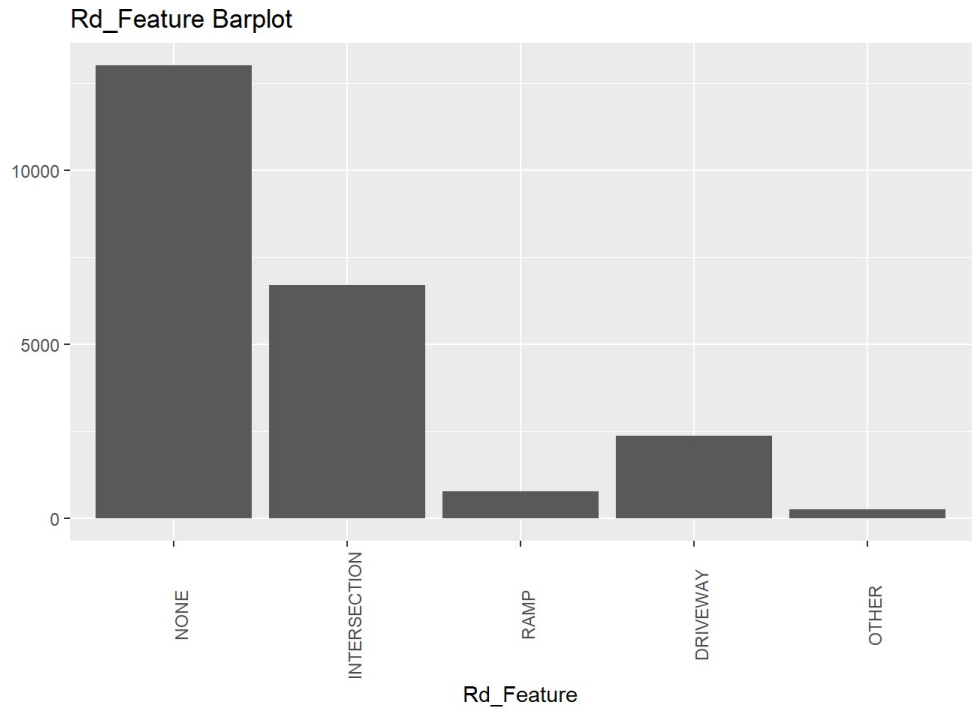
```
# here is a summary of all the variables
```

```
# barplot for count
vars <- colnames(data)[-1]
for (i in vars) {
  print(
    qplot(as.factor(data[,i]), geom="bar", main = paste(i, "Barplot")) +
    theme(axis.text.x=element_text(angle=90)) +
    scale_x_discrete(name=i, limits=unique(data[, i]))
    )
}
```

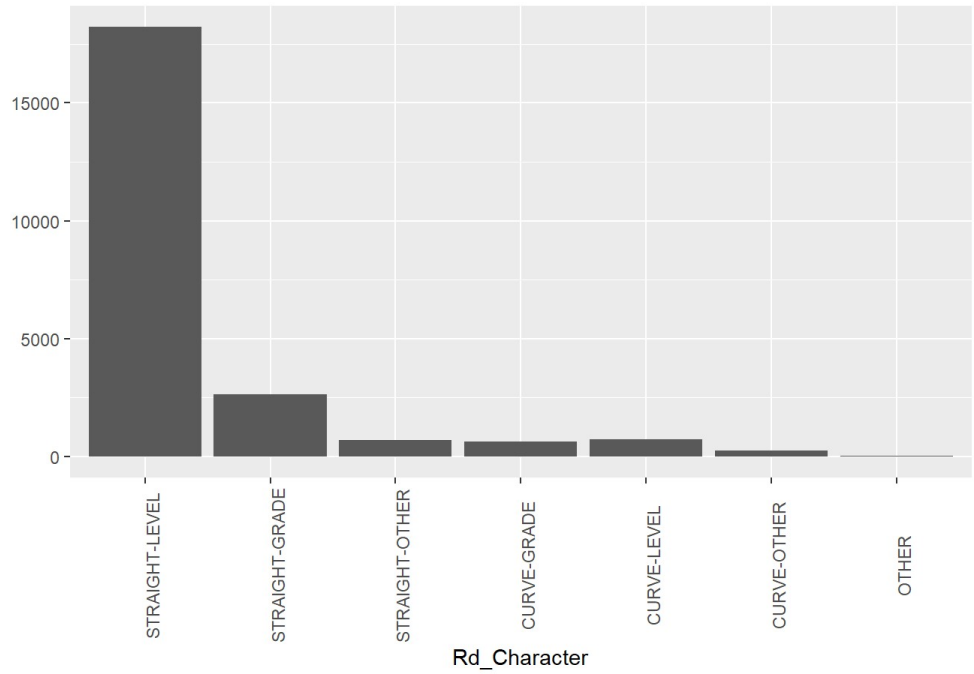

year Barplot
5000
4000
3000
2000
1000
0
year


Month Barplot
2000
1500
1000
500
0
1
2
3
4
5
6
7
8
9
10
11
12
Month

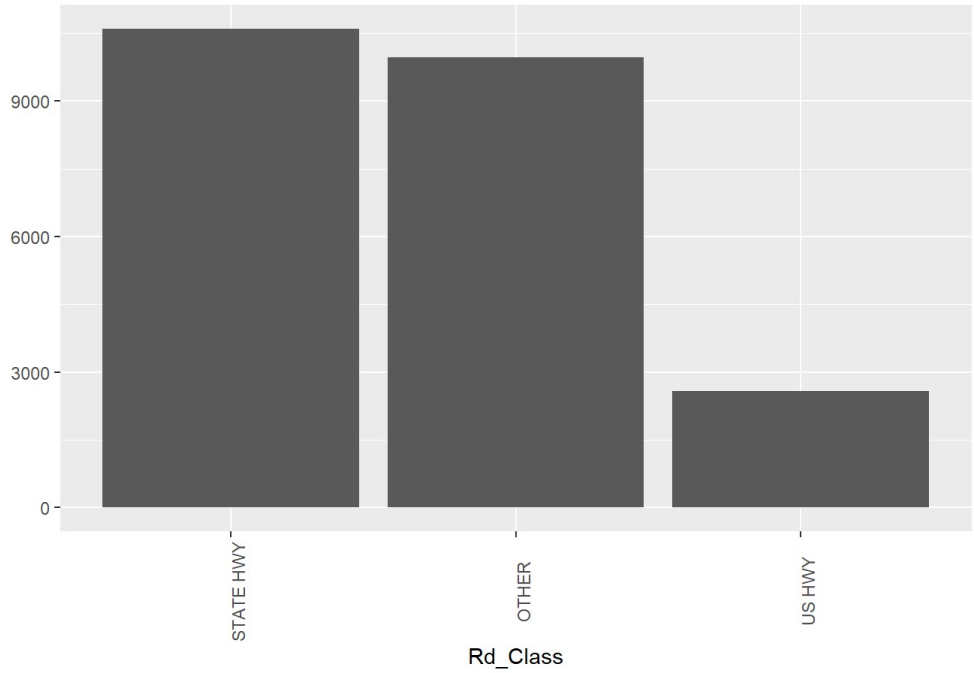

Time_of_Day Barplot
6000
4000
2000
0
1
2
3
4
5
6
Time_of_Day


Rd_Feature Barplot
10000
5000
0
NONE
INTERSECTION
RAMP
DRIVEWAY
OTHER
Rd_Feature

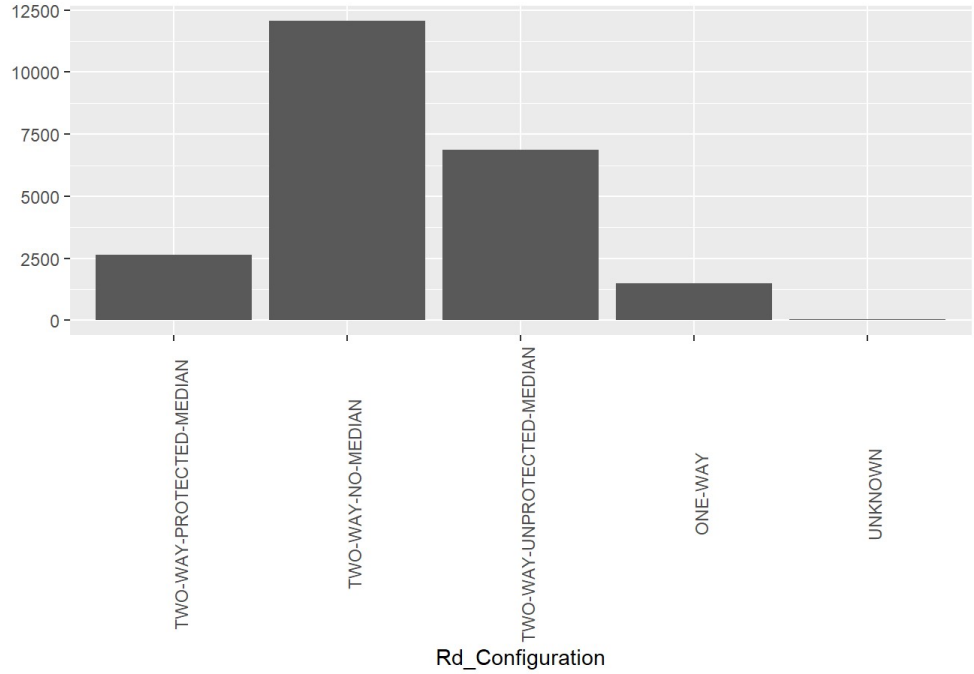
Rd_Character Barplot



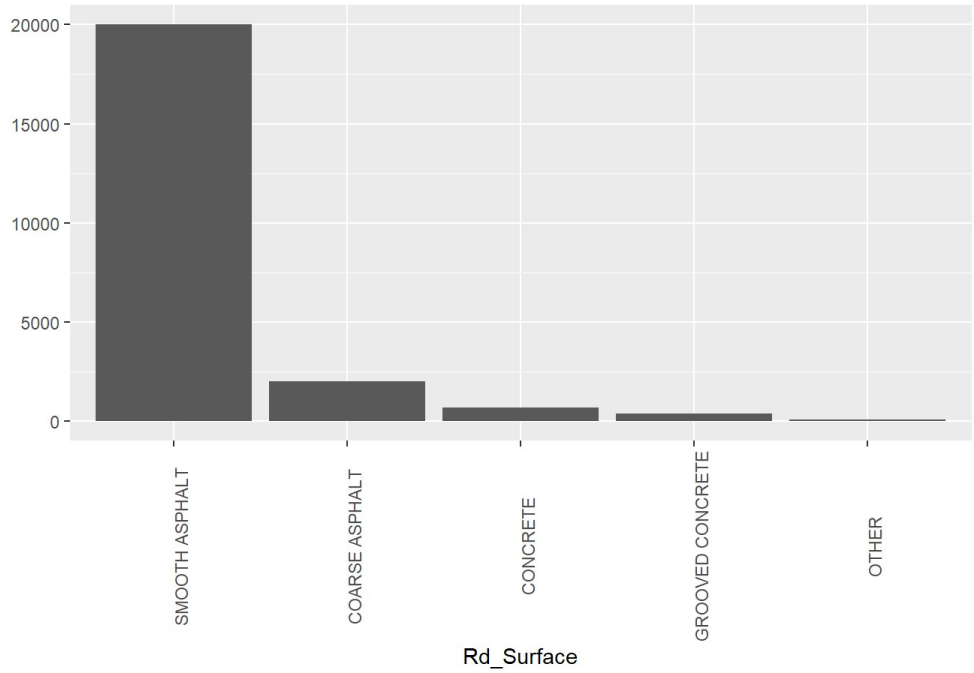
Rd_Class Barplot



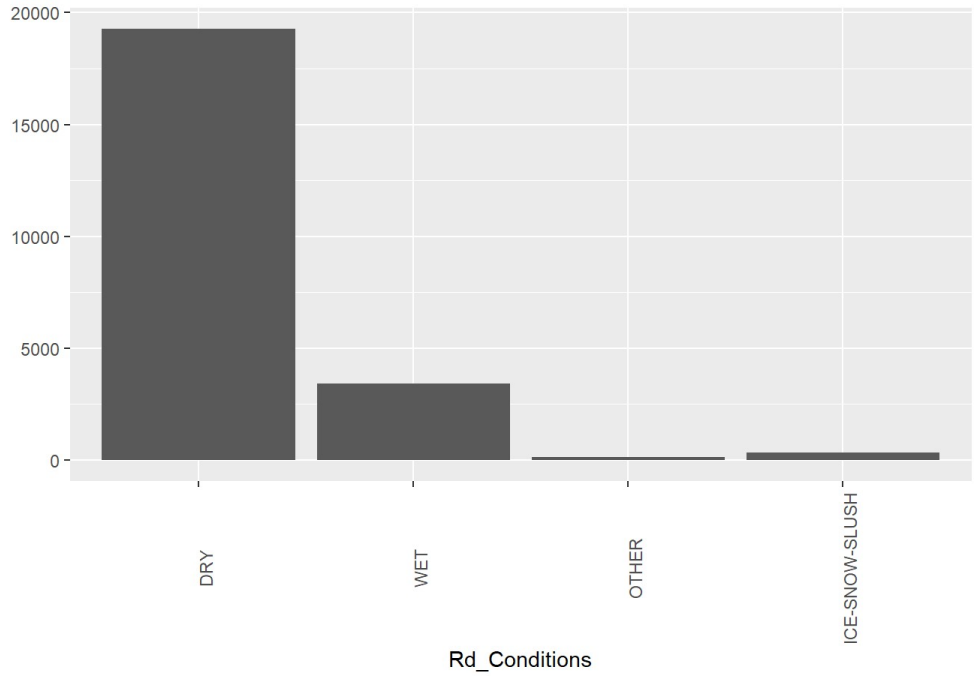
Rd_Configuration Barplot



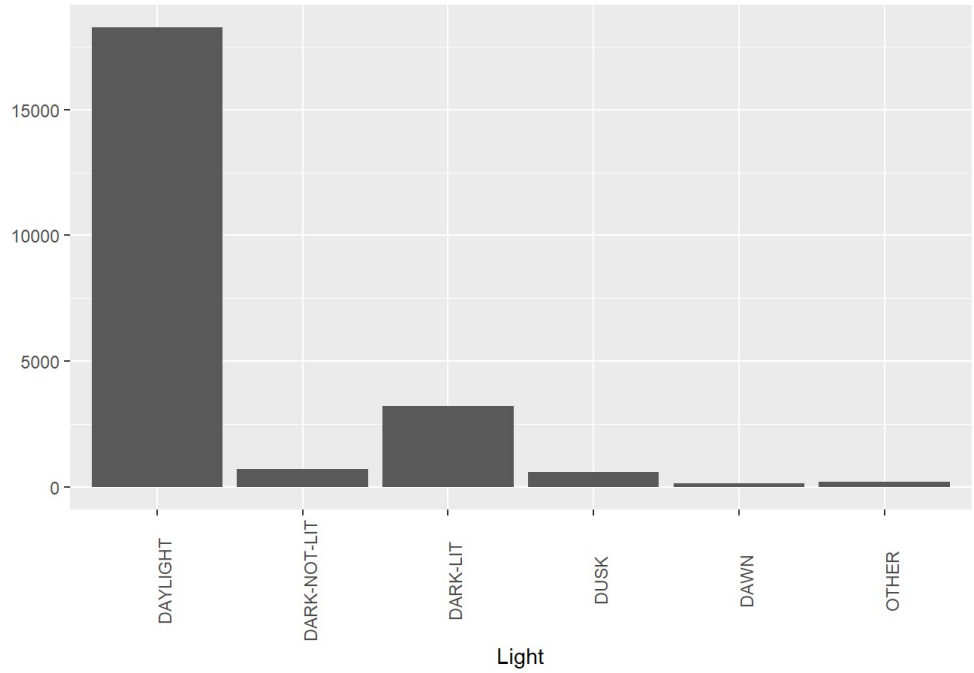
Rd_Surface Barplot



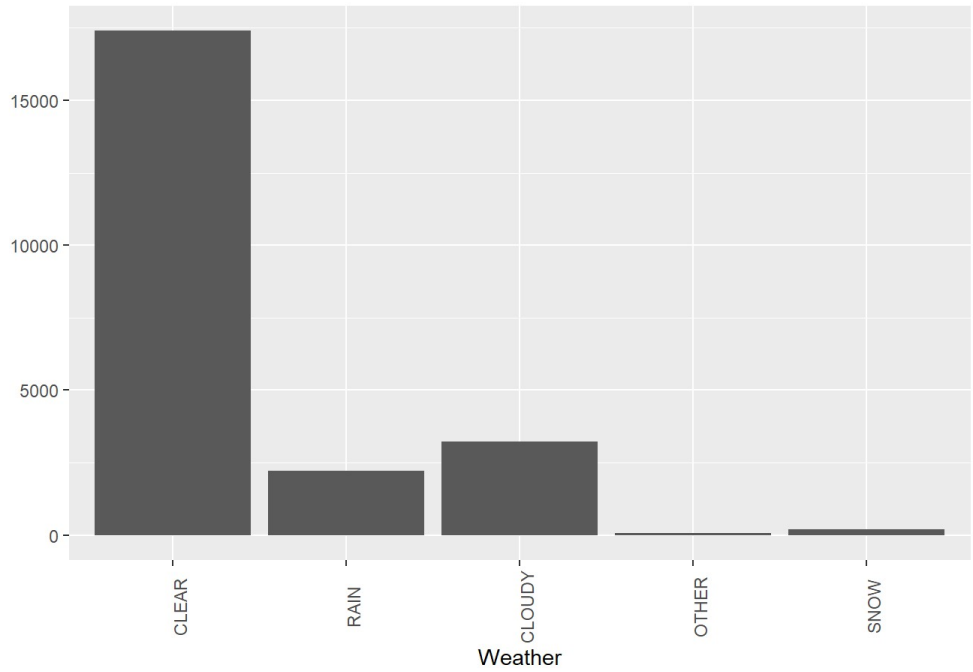
Rd_Conditions Barplot



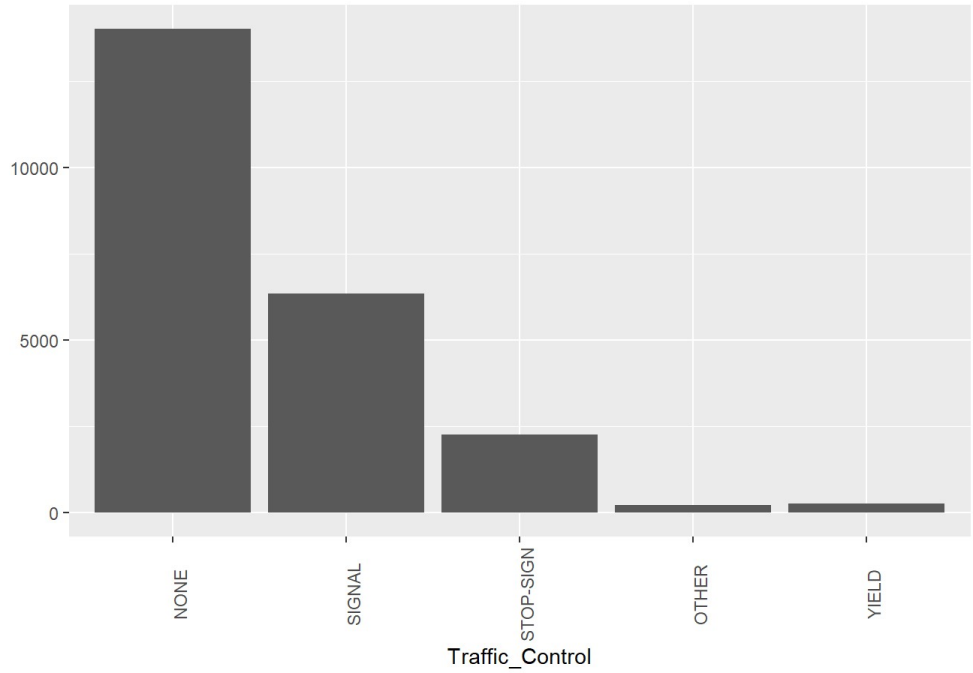
Light Barplot

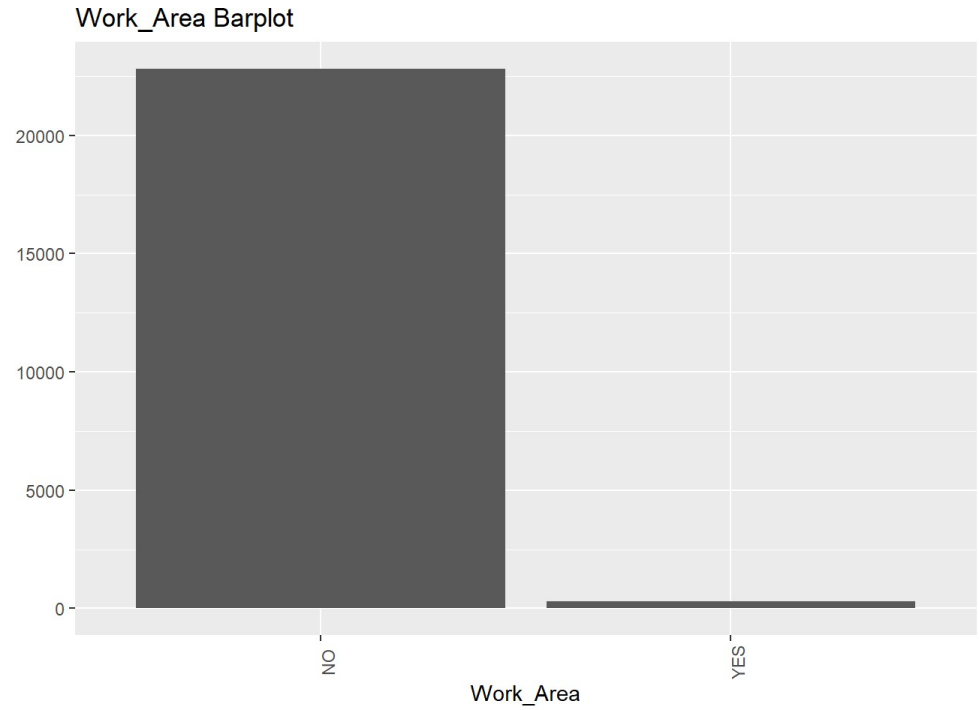


Weather Barplot



Traffic_Control Barplot



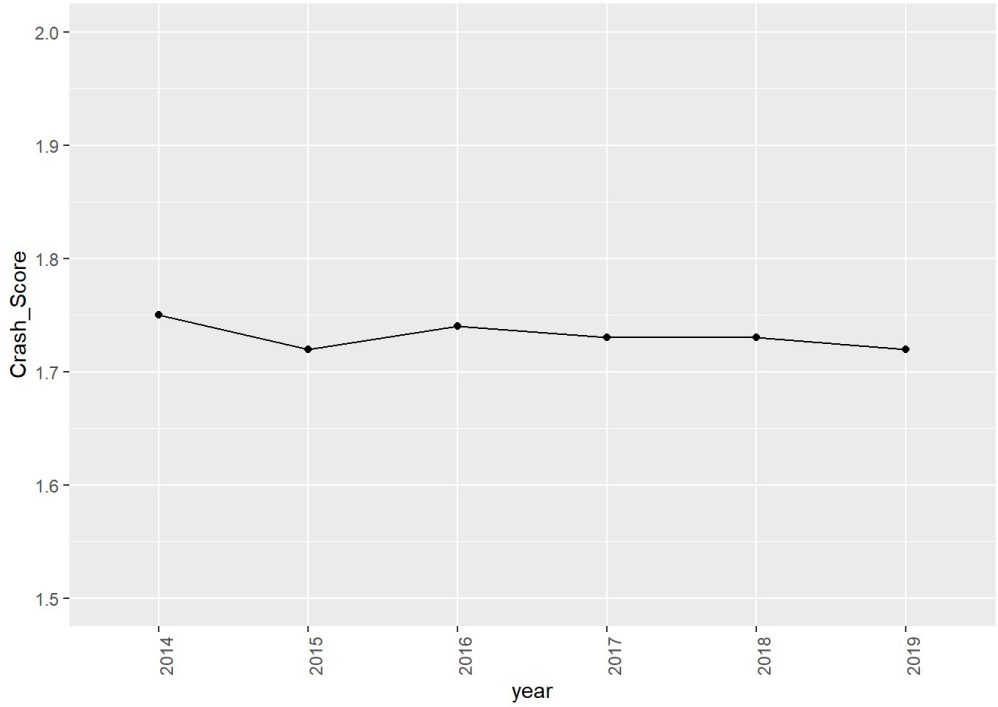



```
# as we can observe,
# there are more car crashes during time 2(4am to 8am), 3(8am to 12pm), 5(4pm to 8pm), which is during rush hour
# there are more car crashes with no traffic control
```

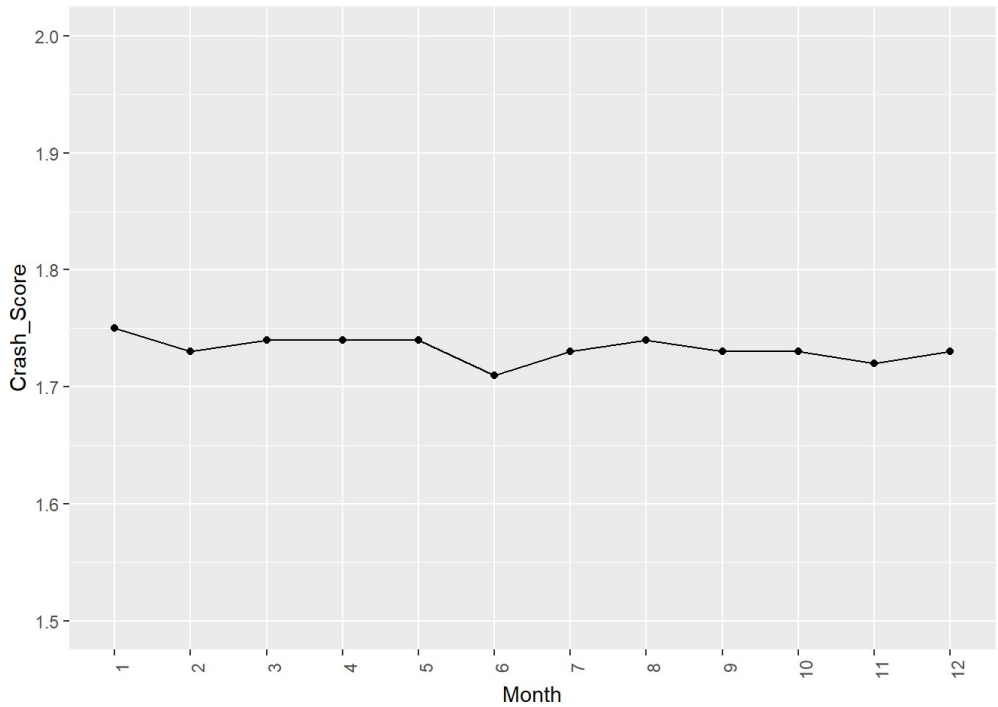
```
# line plot for mean and median
vars <- colnames(data)[-1]
for (i in vars) {
  x <- data %>% group_by_(i)%>%
    summarise(
    mean=round(mean(log(Crash_Score)),2),
    median=round(median(log(Crash_Score)),2)
    )
  print(x)
  x <- as.data.frame(x)
  print(
    qplot(x[, 1], x[, 3], data=x, geom=c("point", "line")) +
    theme(axis.text.x=element_text(angle = 90)) +
    scale_x_discrete(name=i, limits=x[, 1]) +
    scale_y_continuous(name="Crash_Score", limits=c(1.5, 2))
    )
}
```

```
## Warning: group_by_() is deprecated.
## Please use group_by() instead
##
## The 'programming' vignette or the tidyeval book can help you
## to program with group_by() : https://tidyeval.tidyverse.org
## This warning is displayed once per session.
```

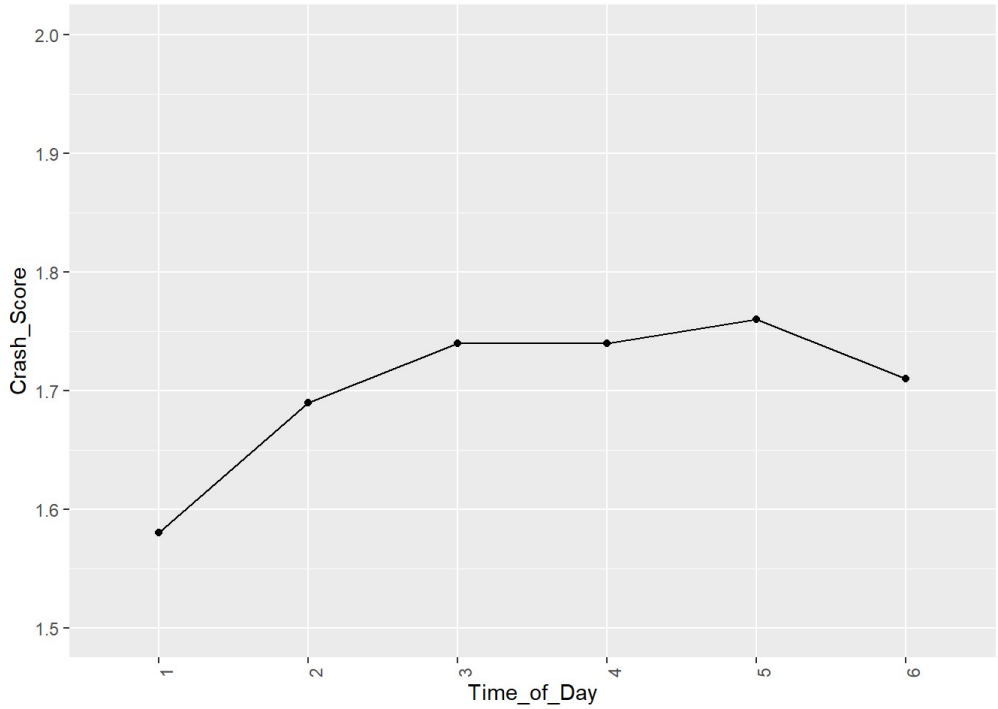
```
## # A tibble: 6 x 3
##    year  mean median
##   <int> <dbl>  <dbl>
## 1  2014  1.68   1.75
## 2  2015  1.67   1.72
## 3  2016  1.67   1.74
## 4  2017  1.67   1.73
## 5  2018  1.66   1.73
## 6  2019  1.67   1.72
```




```
## # A tibble: 12 x 3
##    Month  mean median
##    <int> <dbl>  <dbl>
##  1     1  1.68   1.75
##  2     2  1.68   1.73
##  3     3  1.68   1.74
##  4     4  1.67   1.74
##  5     5  1.66   1.74
##  6     6  1.67   1.71
##  7     7  1.66   1.73
##  8     8  1.67   1.74
##  9     9  1.67   1.73
## 10    10  1.69   1.73
## 11    11  1.66   1.72
## 12    12  1.68   1.73
```

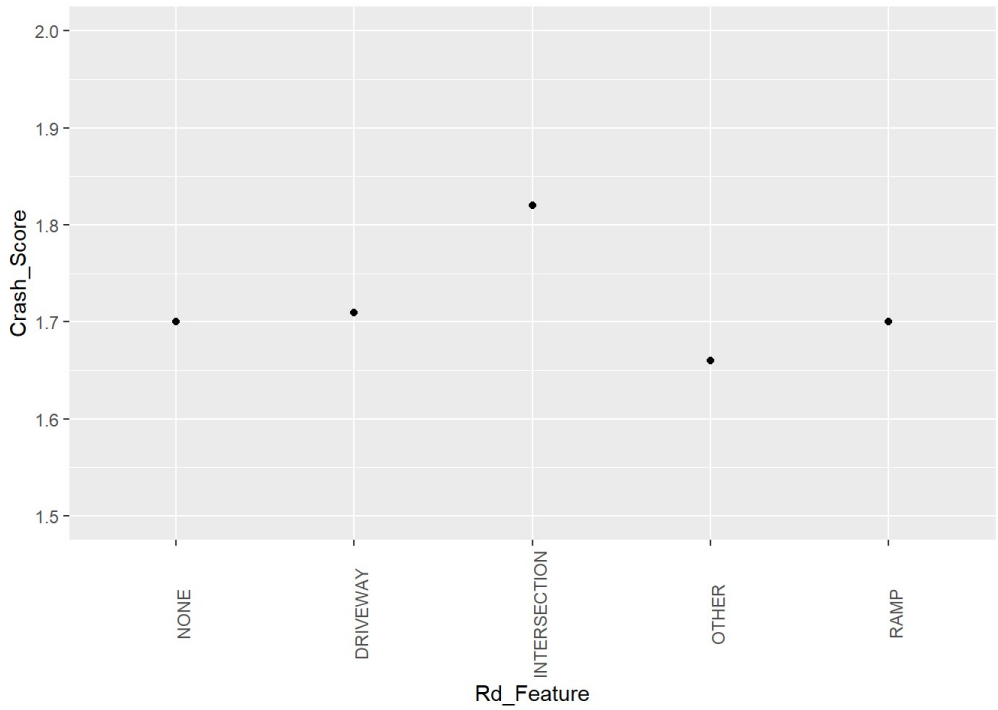


```
## # A tibble: 6 x 3
##   Time_of_Day  mean median
##         <int> <dbl>  <dbl>
## 1           1  1.51   1.58
## 2           2  1.63   1.69
## 3           3  1.68   1.74
## 4           4  1.68   1.74
## 5           5  1.7    1.76
## 6           6  1.64   1.71
```




```
## # A tibble: 5 x 3
##   Rd_Feature    mean median
##   <fct>        <dbl>  <dbl>
## 1 NONE          1.64   1.7
## 2 DRIVEWAY      1.63   1.71
## 3 INTERSECTION  1.76   1.82
## 4 OTHER         1.56   1.66
## 5 RAMP          1.64   1.7
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```

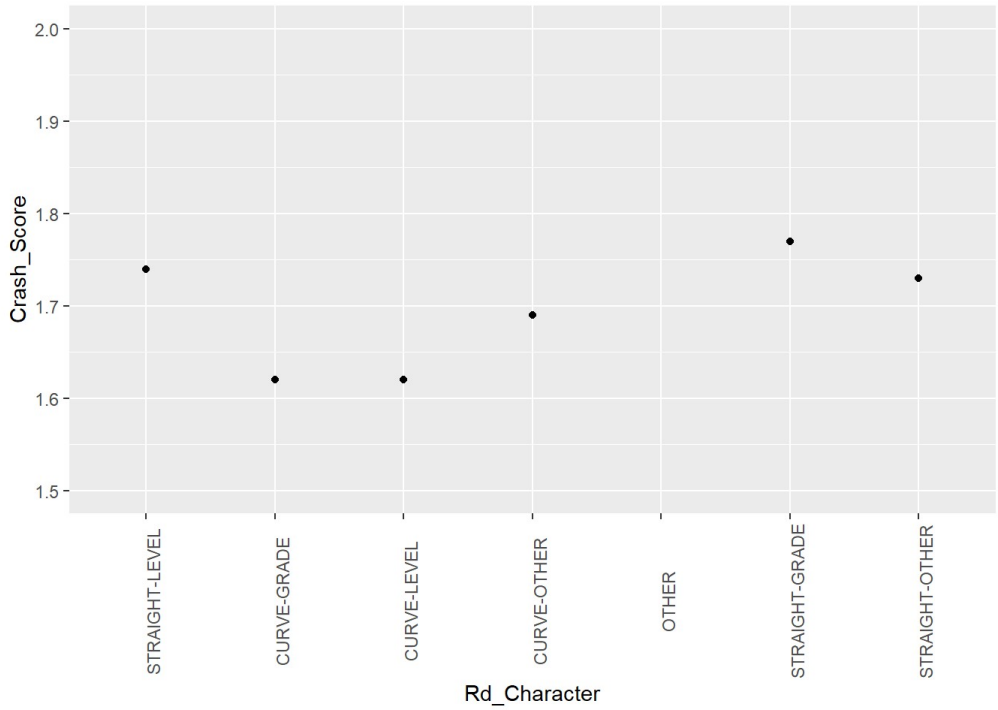



```
## # A tibble: 7 x 3
##   Rd_Character    mean median
##   <fct>          <dbl>  <dbl>
## 1 STRAIGHT-LEVEL  1.68   1.74
## 2 CURVE-GRADE     1.57   1.62
## 3 CURVE-LEVEL     1.57   1.62
## 4 CURVE-OTHER     1.69   1.69
## 5 OTHER           1.29   1.43
## 6 STRAIGHT-GRADE  1.68   1.77
## 7 STRAIGHT-OTHER  1.66   1.73
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

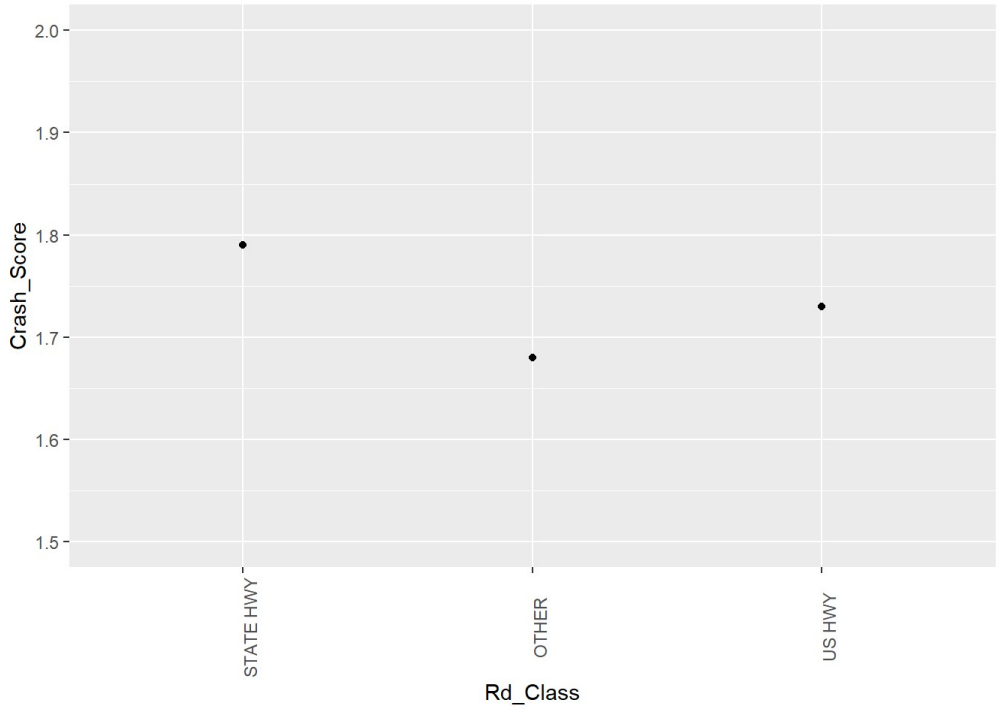
```
## Warning: Removed 1 rows containing missing values (geom_path).
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



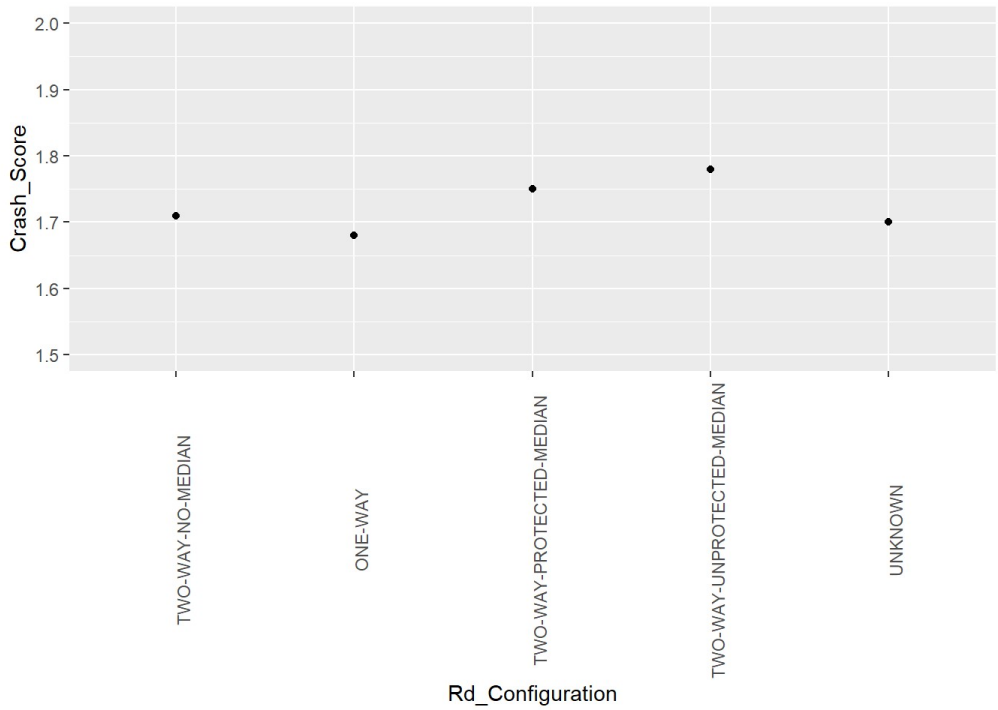
```
## # A tibble: 3 x 3
##   Rd_Class   mean median
##   <fct>     <dbl>  <dbl>
## 1 STATE HWY  1.72   1.79
## 2 OTHER      1.61   1.68
## 3 US HWY     1.69   1.73
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



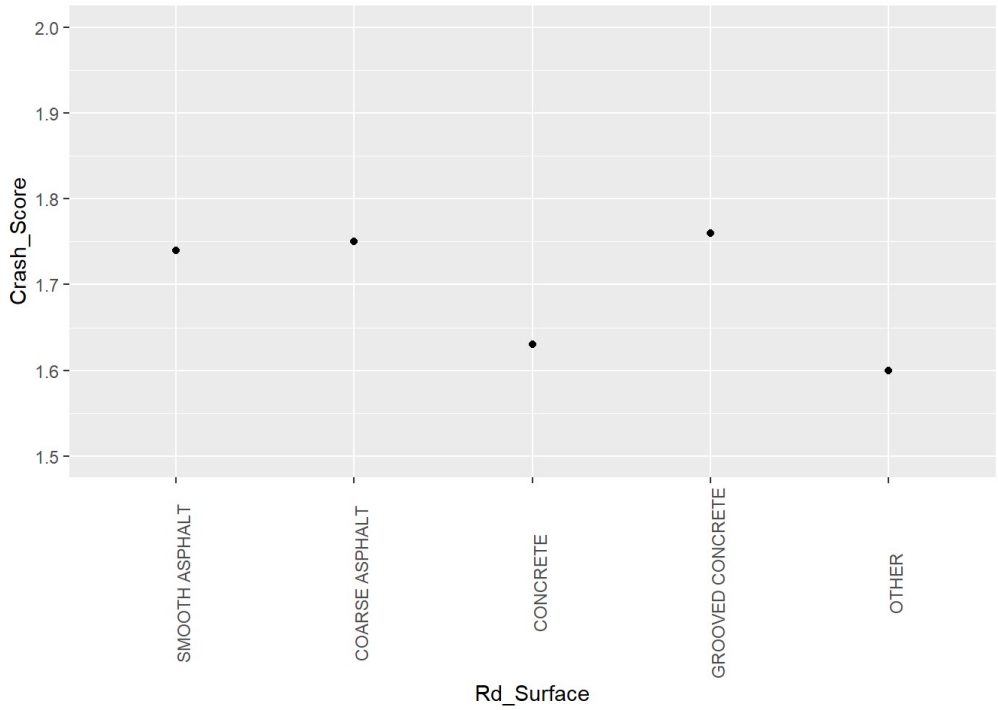

```
## # A tibble: 5 x 3
##   Rd_Configuration            mean median
##   <fct>                      <dbl>  <dbl>
## 1 TWO-WAY-NO-MEDIAN           1.65   1.71
## 2 ONE-WAY                     1.64   1.68
## 3 TWO-WAY-PROTECTED-MEDIAN    1.7    1.75
## 4 TWO-WAY-UNPROTECTED-MEDIAN  1.71   1.78
## 5 UNKNOWN                     1.65   1.7
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



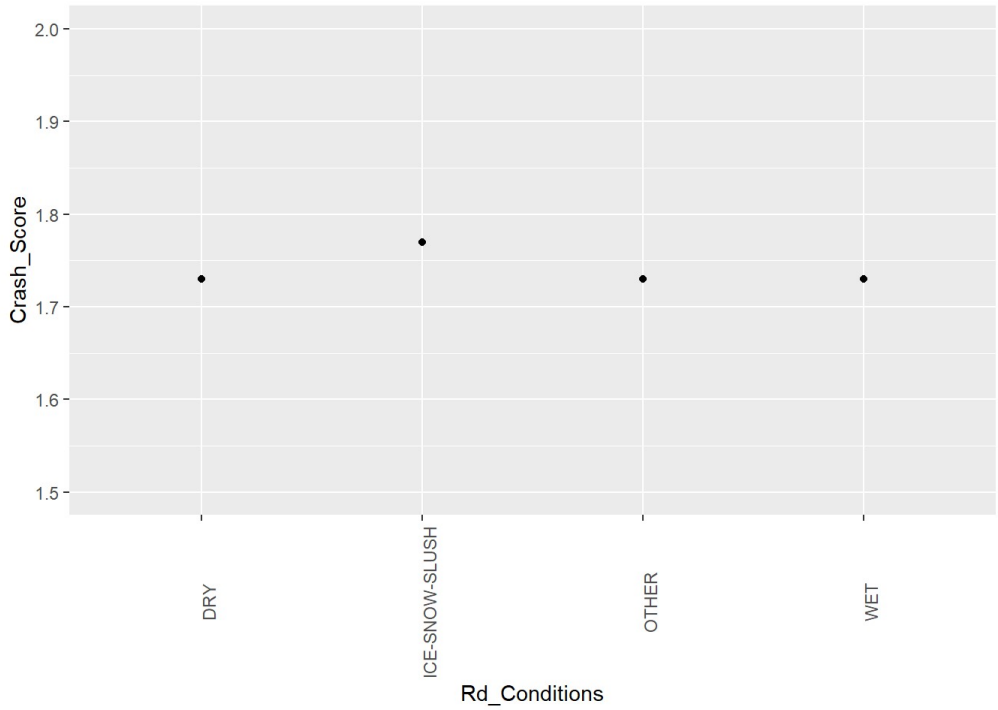

```
## # A tibble: 5 x 3
##   Rd_Surface        mean median
##   <fct>            <dbl>  <dbl>
## 1 SMOOTH ASPHALT    1.67   1.74
## 2 COARSE ASPHALT    1.69   1.75
## 3 CONCRETE          1.59   1.63
## 4 GROOVED CONCRETE  1.66   1.76
## 5 OTHER             1.38   1.6
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



```
## # A tibble: 4 x 3
##   Rd_Conditions   mean median
##   <fct>          <dbl>  <dbl>
## 1 DRY             1.67   1.73
## 2 ICE-SNOW-SLUSH  1.68   1.77
## 3 OTHER           1.57   1.73
## 4 WET             1.67   1.73
```

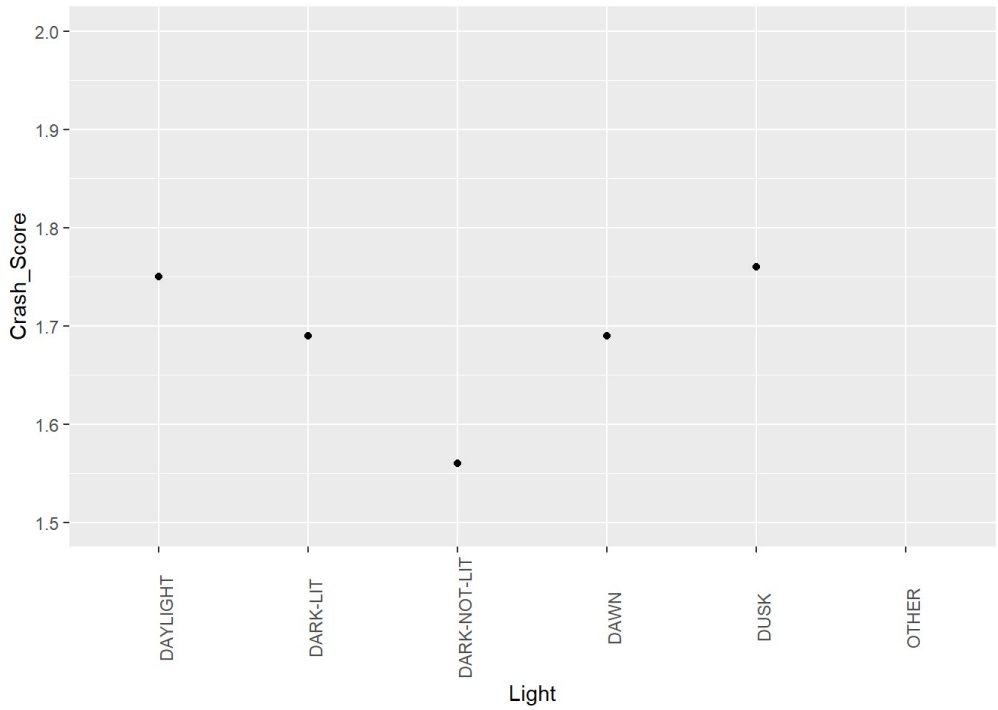
```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```




```
## # A tibble: 6 x 3
##   Light         mean median
##   <fct>        <dbl>  <dbl>
## 1 DAYLIGHT      1.69   1.75
## 2 DARK-LIT      1.61   1.69
## 3 DARK-NOT-LIT  1.55   1.56
## 4 DAWN          1.65   1.69
## 5 DUSK          1.69   1.76
## 6 OTHER         1.44   1.47
```

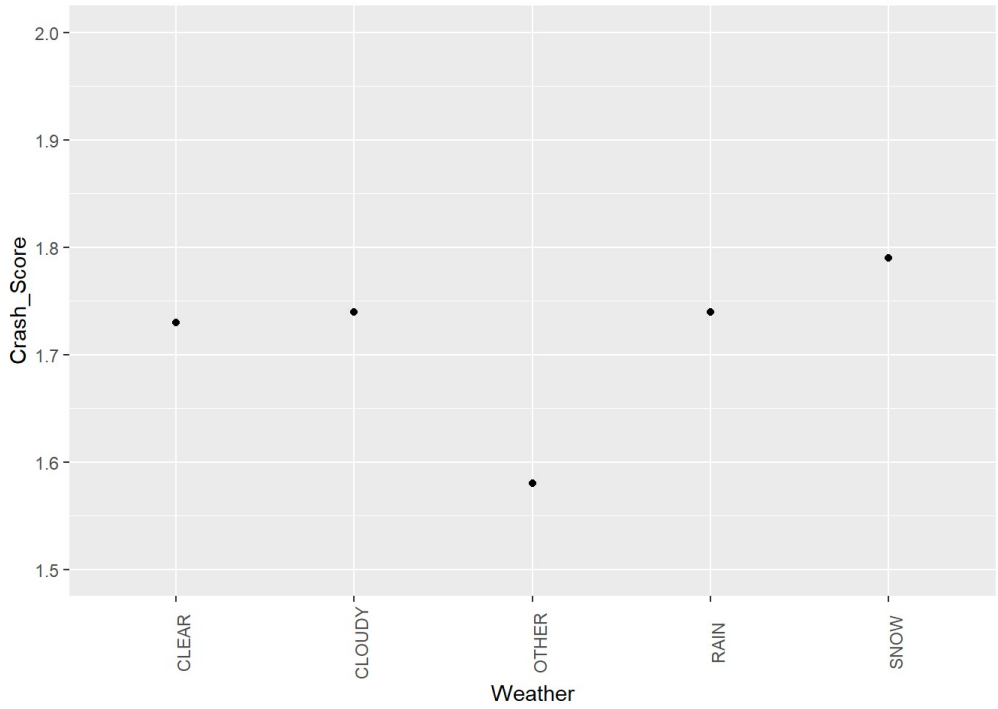
```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_path).
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



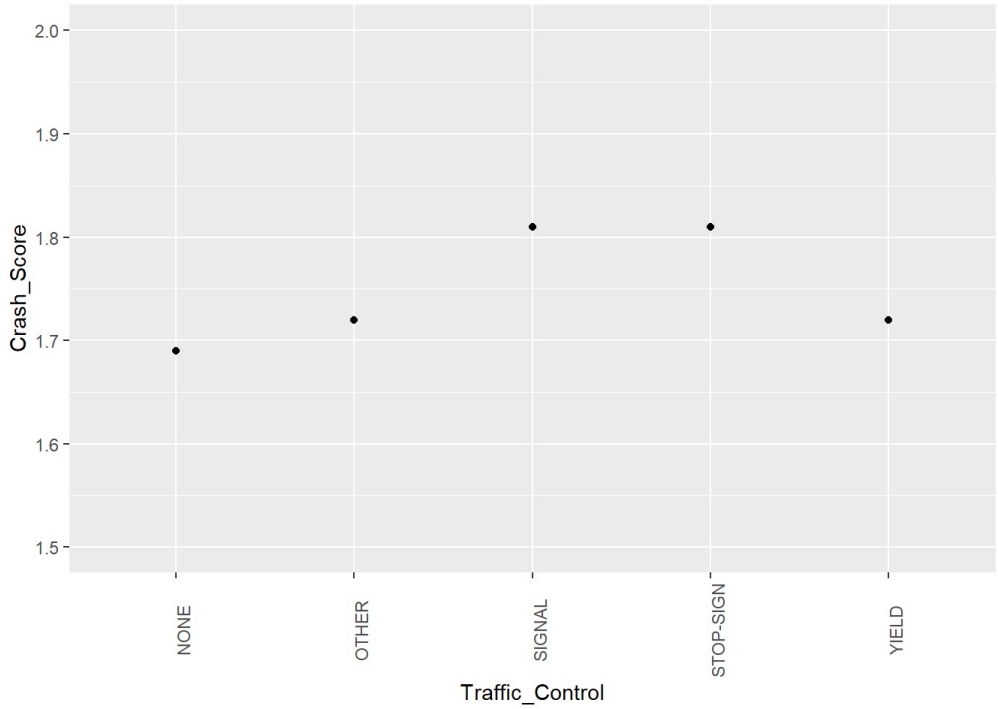

```
## # A tibble: 5 x 3
##   Weather  mean median
##   <fct>   <dbl>  <dbl>
## 1 CLEAR    1.67   1.73
## 2 CLOUDY   1.67   1.74
## 3 OTHER    1.51   1.58
## 4 RAIN     1.69   1.74
## 5 SNOW     1.68   1.79
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



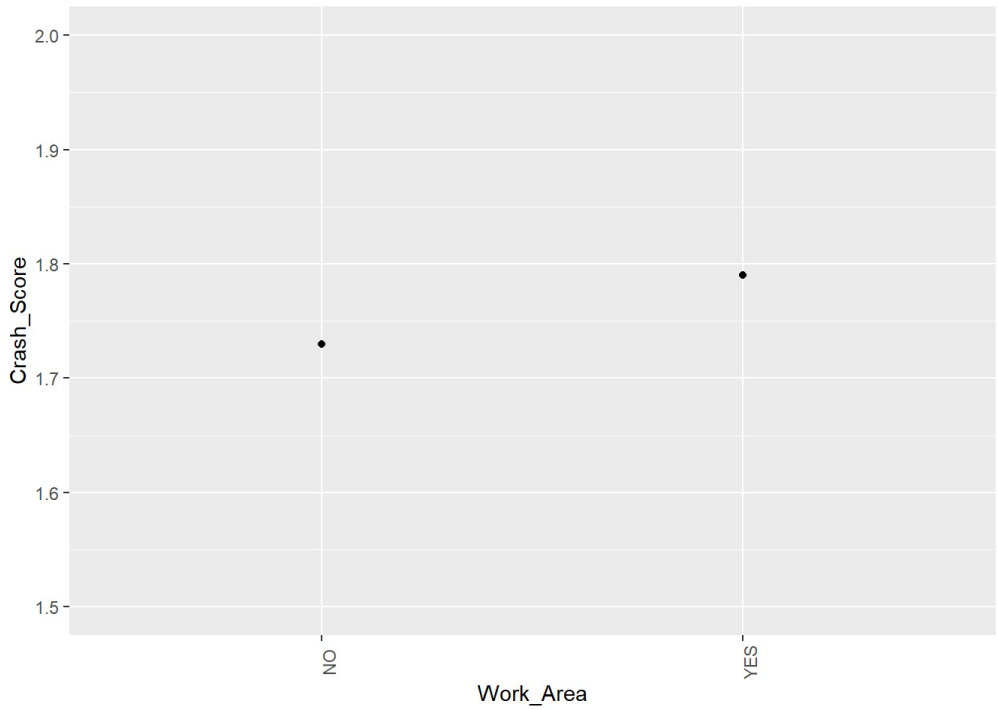
```
## # A tibble: 5 x 3
##   Traffic_Control  mean median
##   <fct>           <dbl>  <dbl>
## 1 NONE             1.62   1.69
## 2 OTHER            1.65   1.72
## 3 SIGNAL           1.76   1.81
## 4 STOP-SIGN        1.74   1.81
## 5 YIELD            1.68   1.72
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



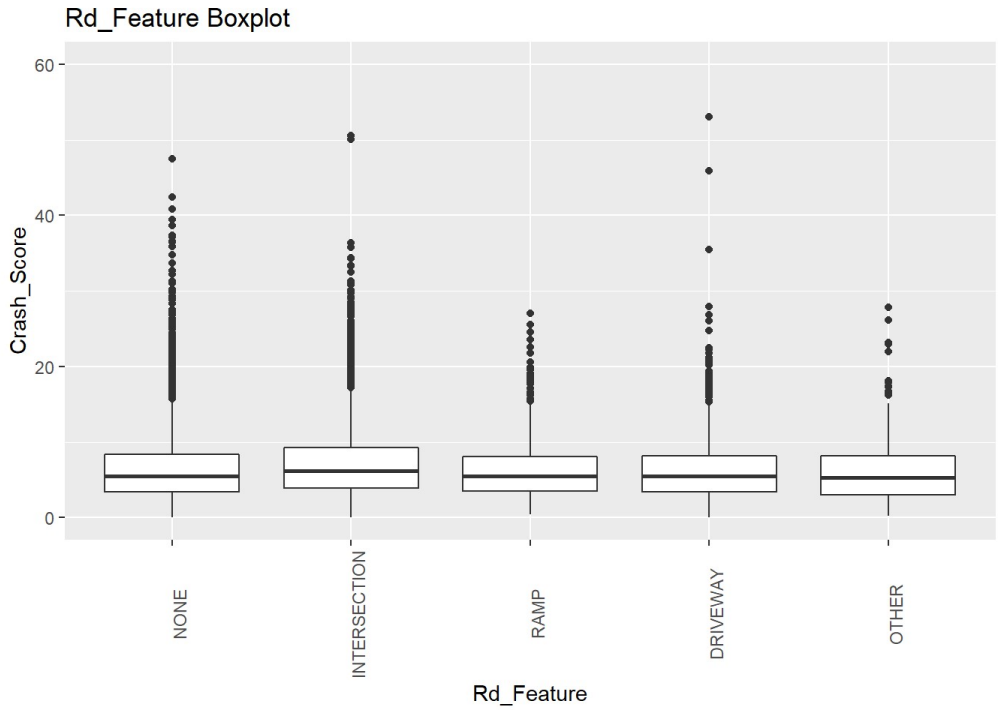

```
## # A tibble: 2 x 3
##   Work_Area  mean median
##   <fct>     <dbl>  <dbl>
## 1 NO         1.67   1.73
## 2 YES        1.7    1.79
```

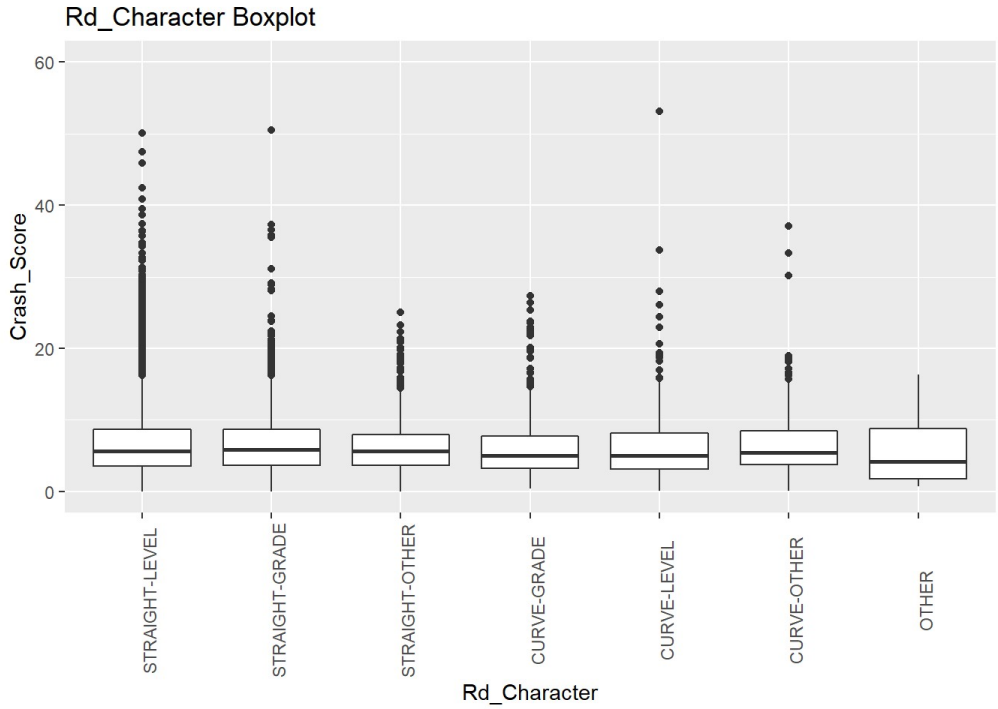
```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



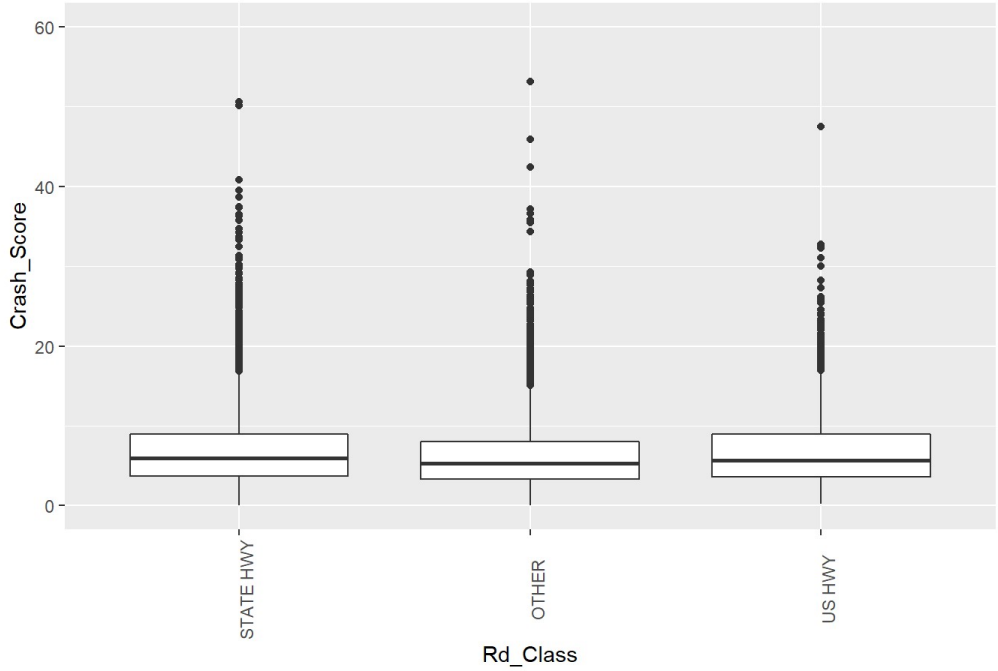

```
# as we can observe,
# there is no significant change over years (from 2014-2019), nor significant change over the year (from Jan. to Dec.), in other words, no seasonality for car crashes
# the crash score is affected by the time of the day, the median score is higher durig the daytime (8am to 8pm), the median score is significant lower for time 1 (midnight to 4am), and relatively lower for time 2 (4am to 8am) and time 6 (8pm to midnight)
# for the raod feature, intersection has significantly higher median score
# for the road character, straight level has significantly higher median score
# for the road class, state hwy has the highest score, followed by us hwy and other hwy
# surprisingly, for the road configuration, two way protected median, two way unprotected median has higher median score than two way no median
# for the road surface, concrete has significantly lower median score
# for the road condition, ice snow slush has a relatively higher median score, dry and wet have similar median score, and other has the lowest median score
# surprisingly, for light, daylight has the highest median score, followed by dusk and down, and then dark light and dark not light
# for the weather, rain and snow has a relatively higher median score, other has the lowest median score
# surprisingly, for the road configuration, signal and stop sign have the highest median socre none has the lowest median score
# work area has a higher median score than non work area
# the boxplot and violin plot are not as intuitive as the line plot for the median
```

```
# boxplots for categorical variables
vars <- colnames(data)[-(1:4)]
for (i in vars) {
  print(
    qplot(as.factor(data[,i]), data$Crash_Score, geom="boxplot", main = paste(i, "Boxplot")) +
    theme(axis.text.x = element_text(angle = 90)) +
    scale_x_discrete(name = i, limits = unique(data[, i])) +
    scale_y_continuous(name="Crash_Score", limits=c(0, 60))
    )
}
```

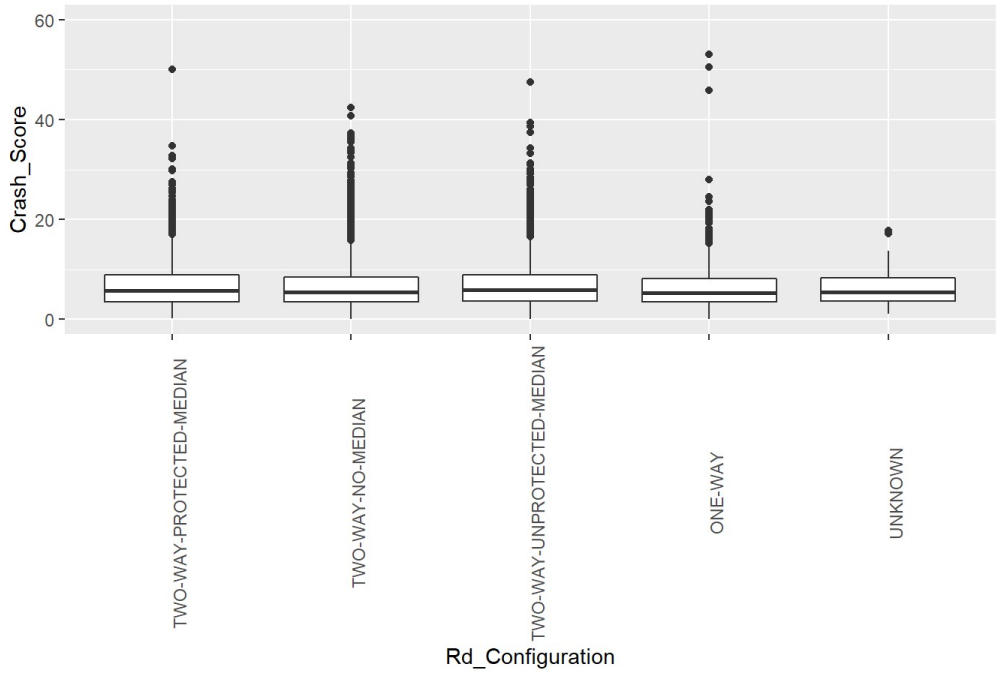

Rd_Feature Boxplot
Crash_Score
60
40
20
0
NONE
INTERSECTION
RAMP
DRIVEWAY
OTHER
Rd_Feature

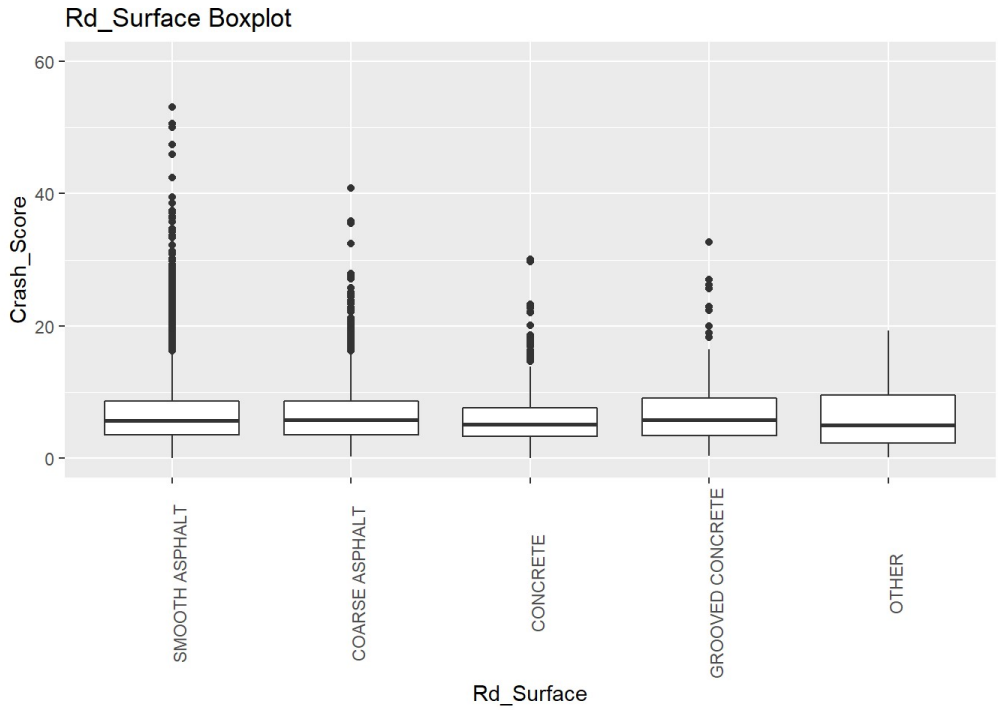

Rd_Character Boxplot
Crash_Score
60
40
20
0
STRAIGHT-LEVEL
STRAIGHT-GRADE
STRAIGHT-OTHER
CURVE-GRADE
CURVE-LEVEL
CURVE-OTHER
OTHER
Rd_Character

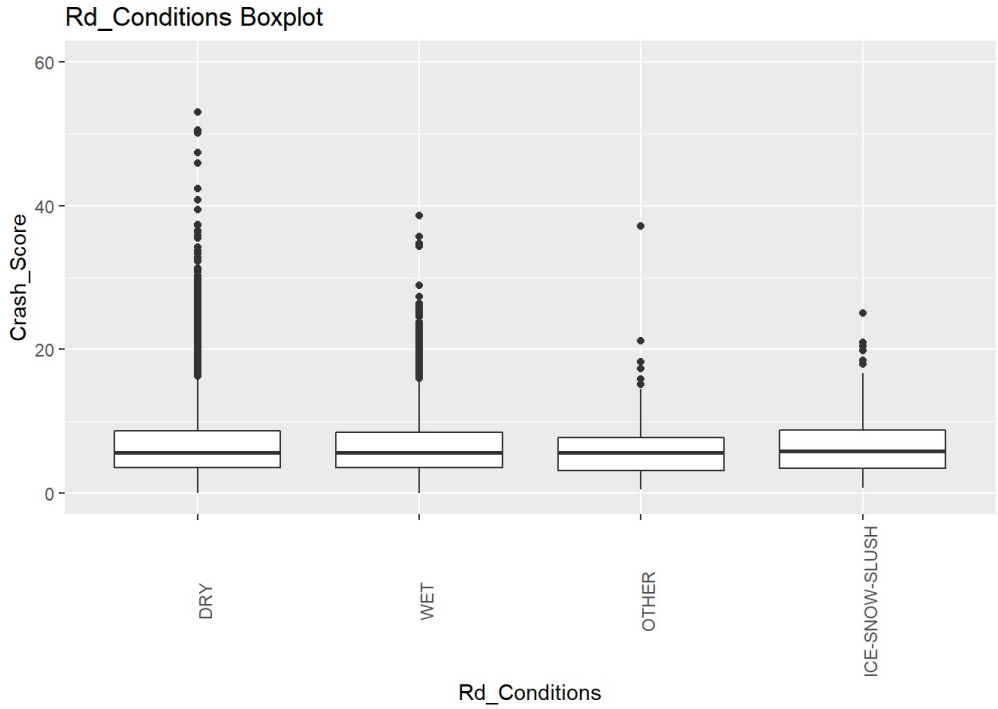
Rd_Class Boxplot



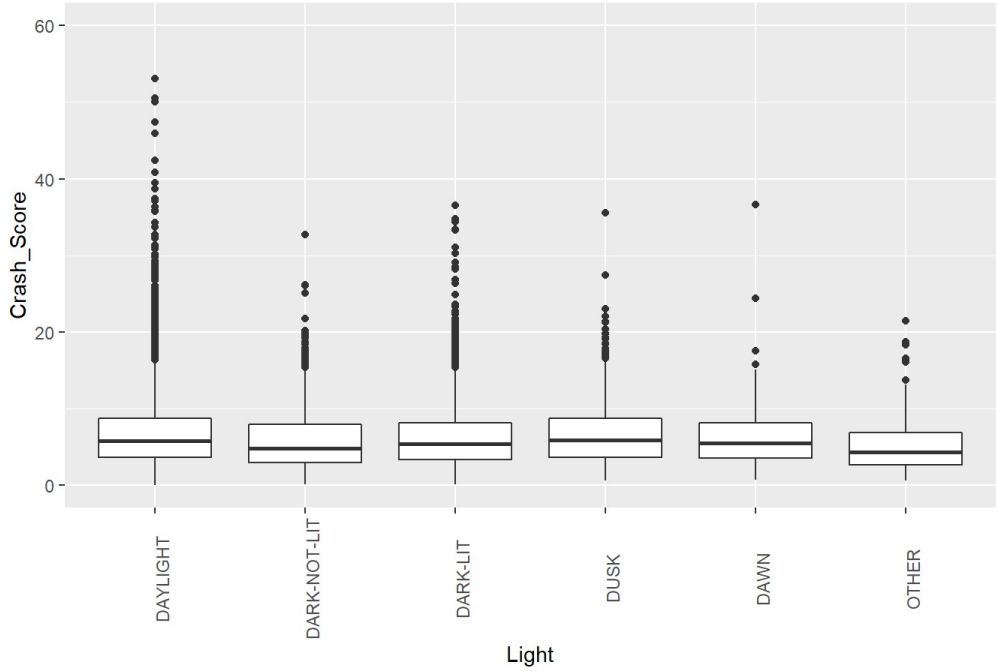
Rd_Configuration Boxplot



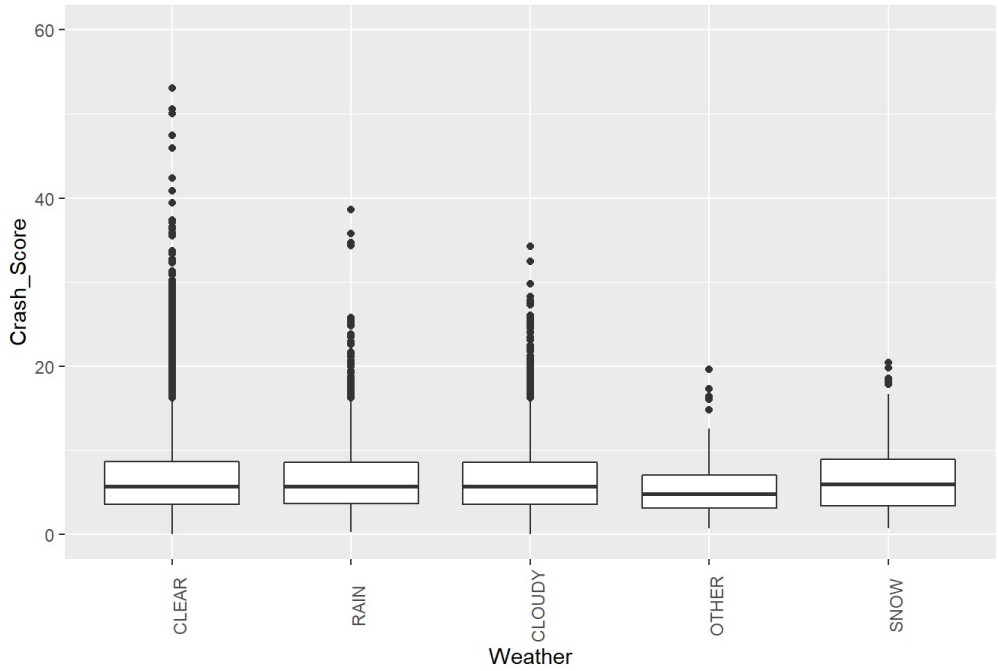

Rd_Surface Boxplot
Crash_Score
60
40
20
0
SMOOTH ASPHALT
COARSE ASPHALT
CONCRETE
GROOVED CONCRETE
OTHER
Rd_Surface

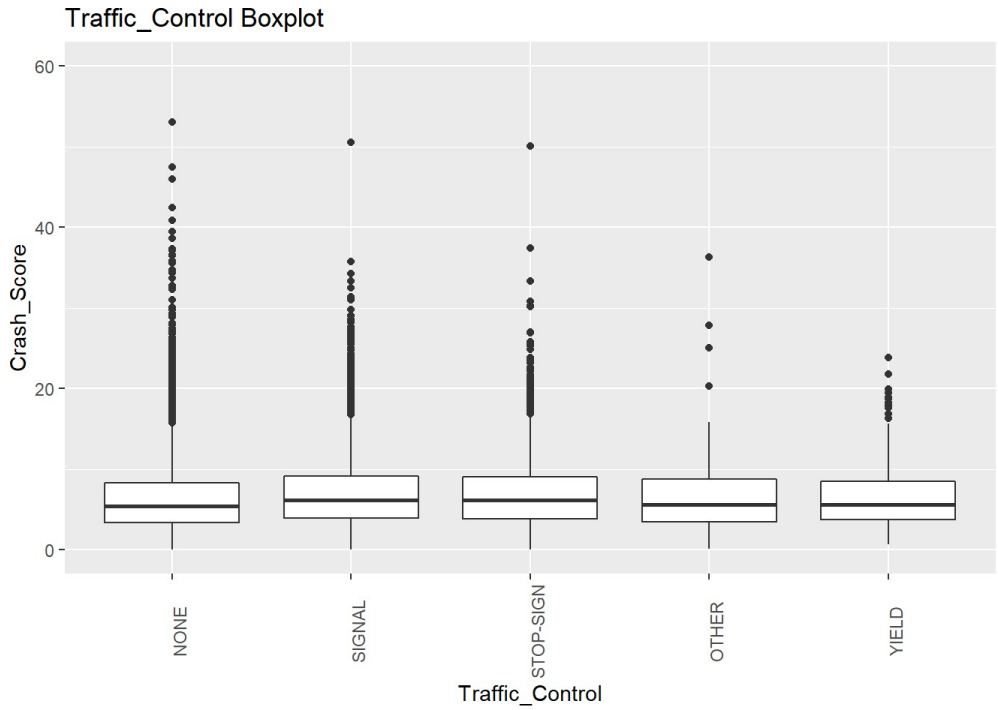

Rd_Conditions Boxplot
Crash_Score
60
40
20
0
DRY
WET
OTHER
ICE-SNOW-SLUSH
Rd_Conditions

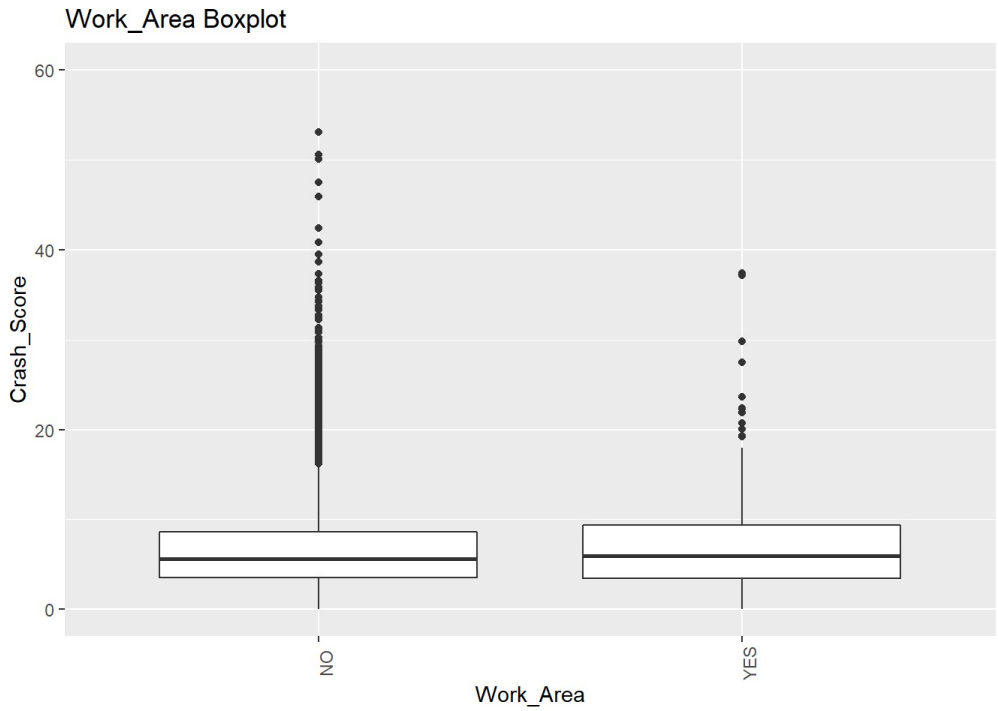
Light Boxplot



Weather Boxplot

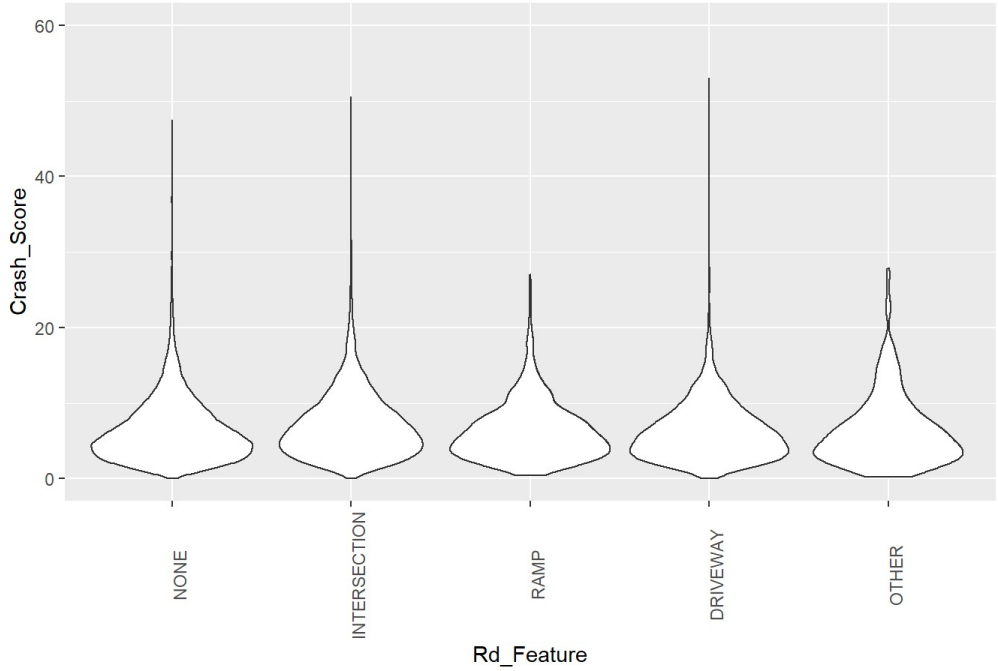




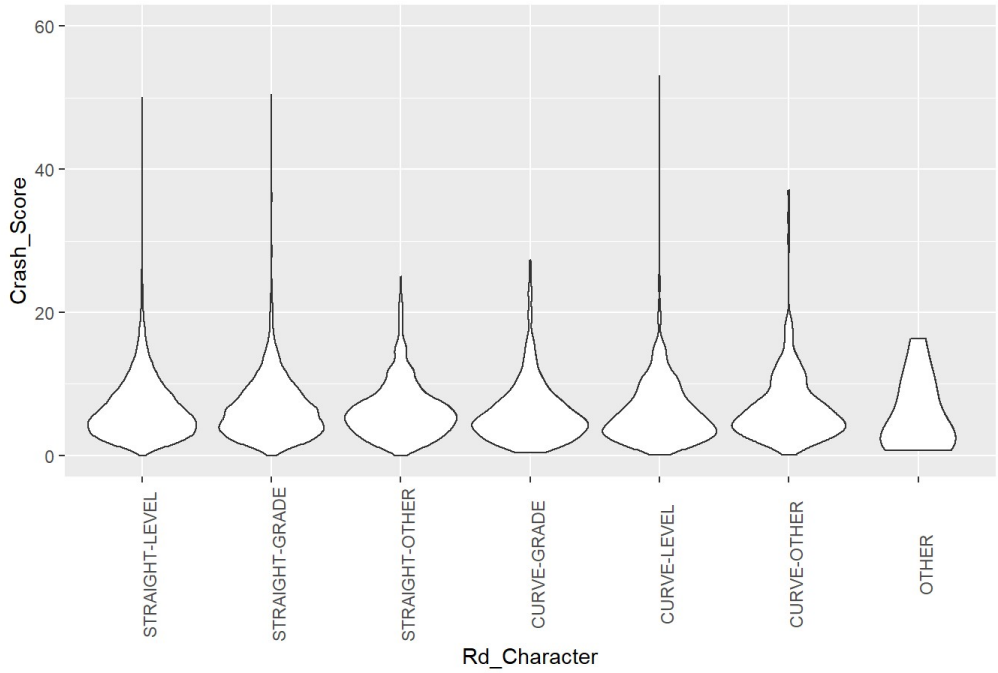



```
# the violin plot for different variables
vars <- colnames(data)[-(1:4)]
for (i in vars) {
  print(
    qplot(as.factor(data[,i]), data$Crash_Score, geom="violin", main = paste(i, "Violin plot")) +
    theme(axis.text.x = element_text(angle = 90)) +
    scale_x_discrete(name = i, limits = unique(data[, i])) +
    scale_y_continuous(name="Crash_Score", limits=c(0, 60))
    )
}
```

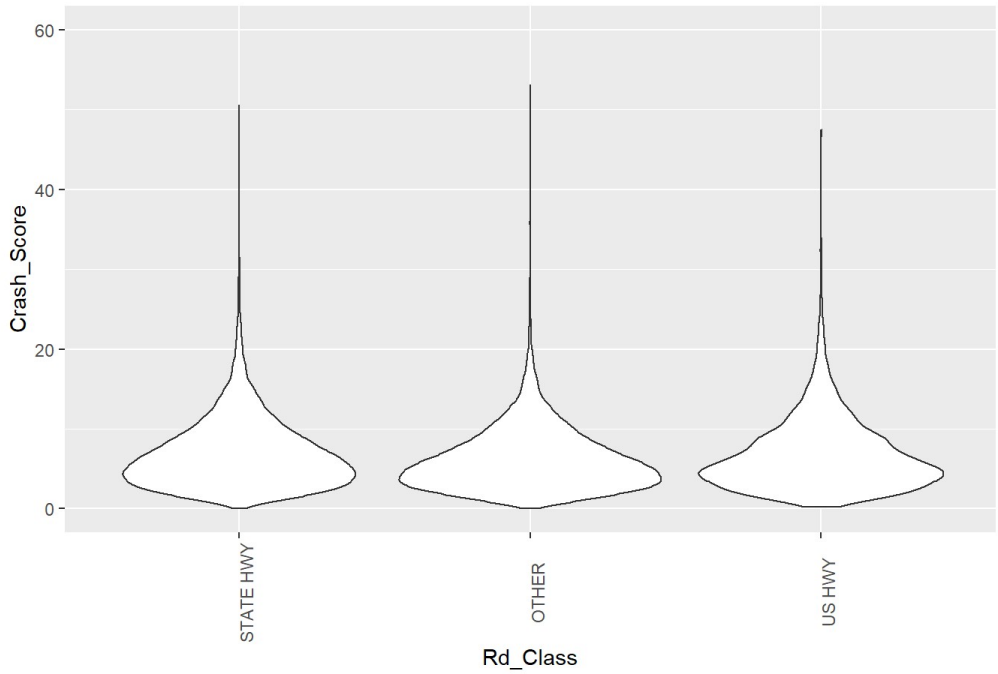
Rd_Feature Violin plot



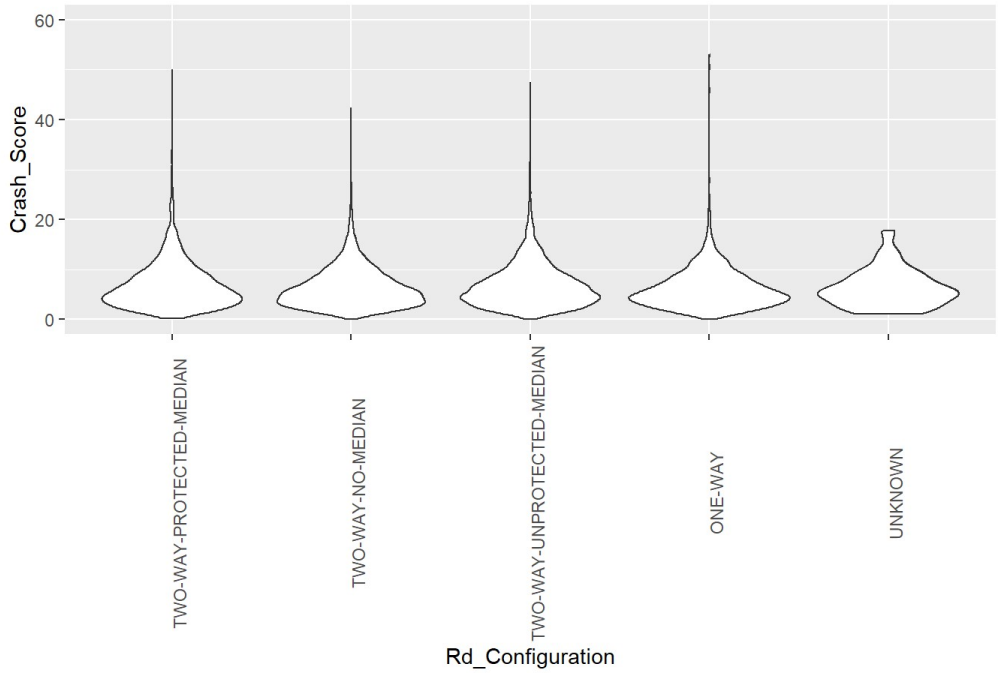
Rd_Character Violin plot



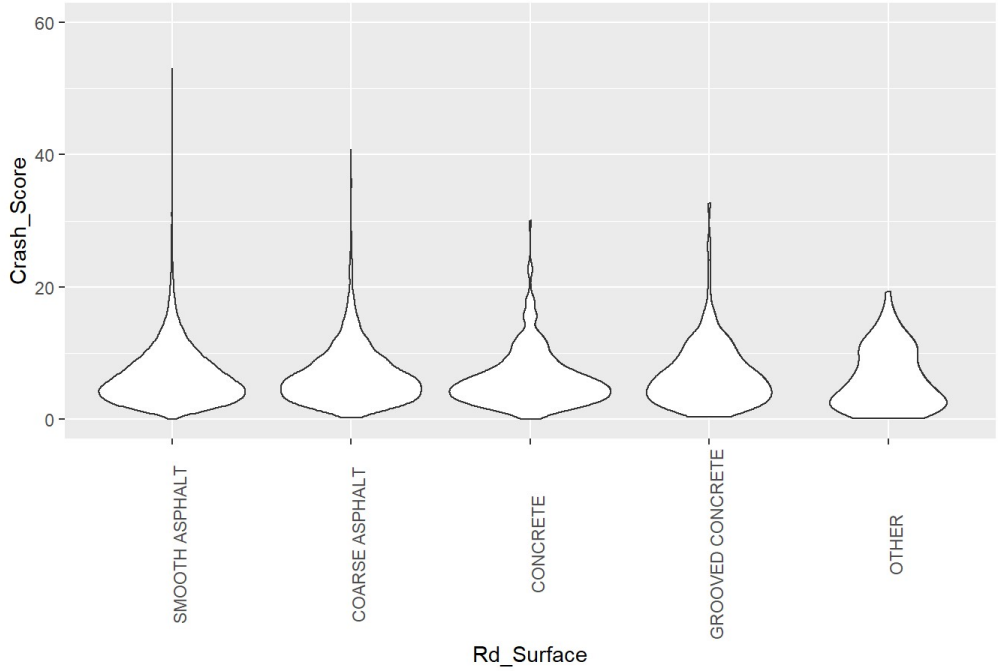
Rd_Class Violin plot



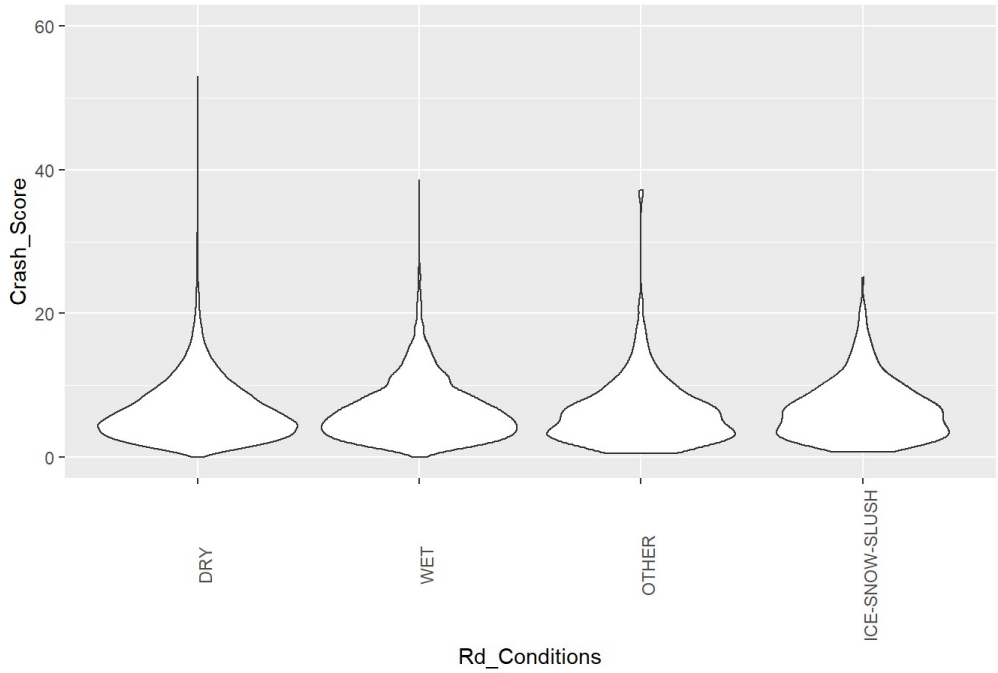
Rd_Configuration Violin plot



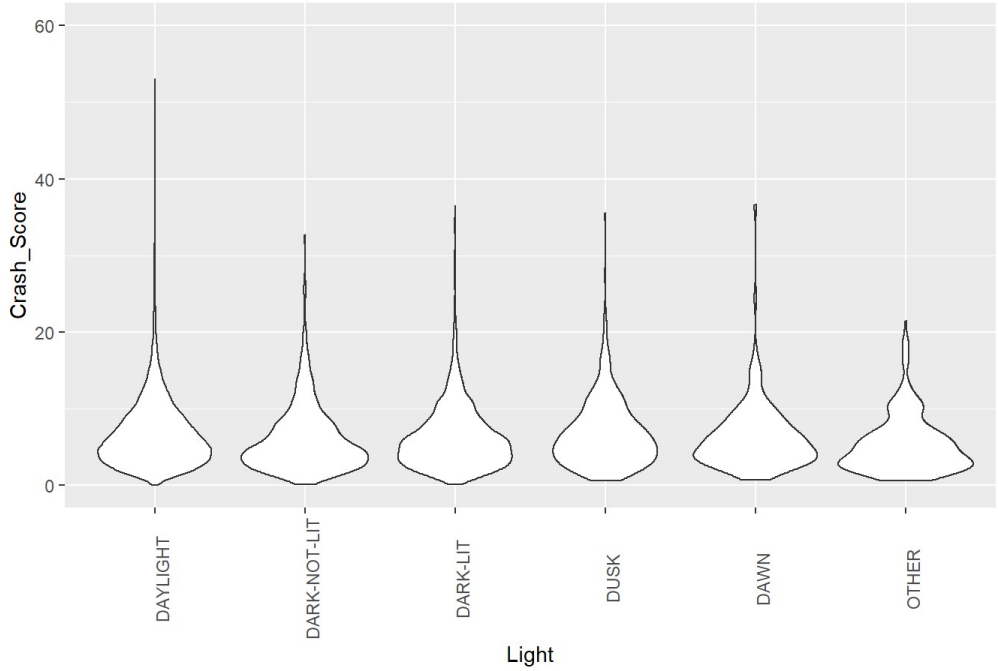
Rd_Surface Violin plot



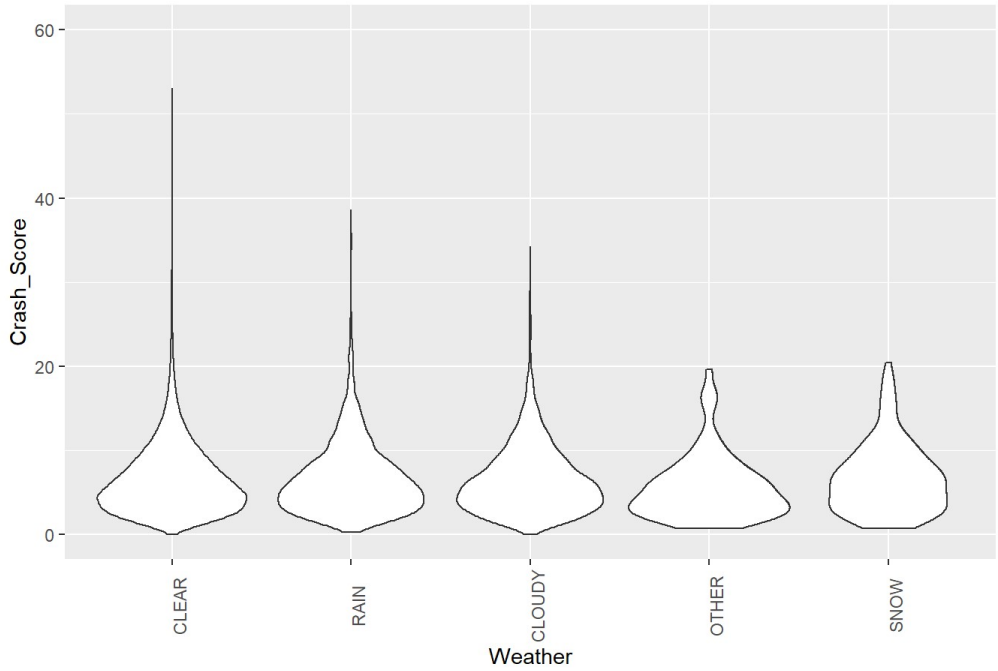
Rd_Conditions Violin plot



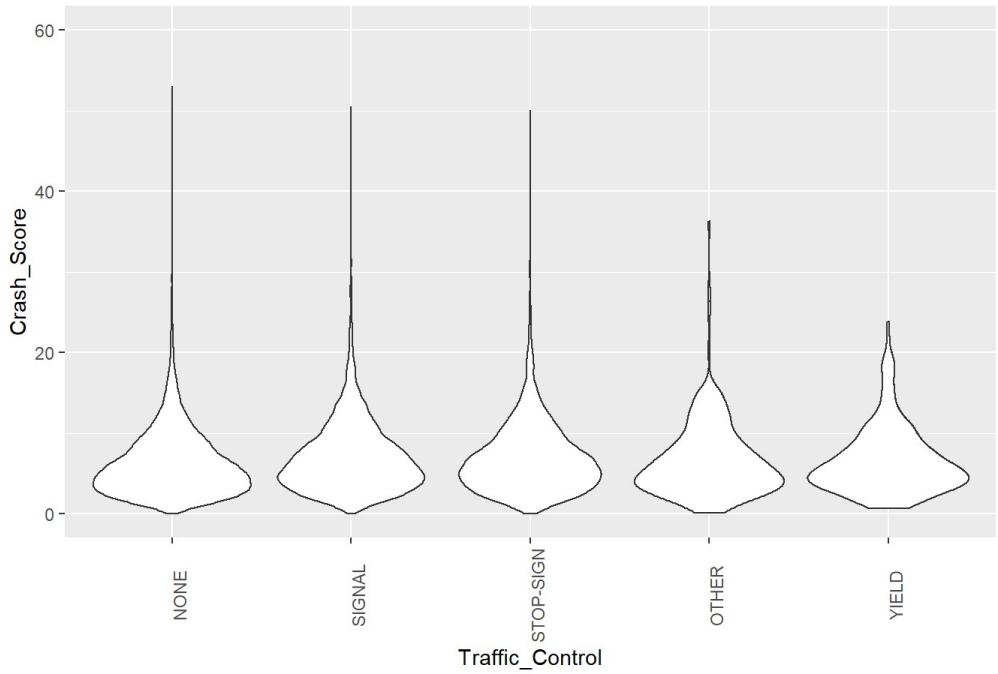
Light Violin plot



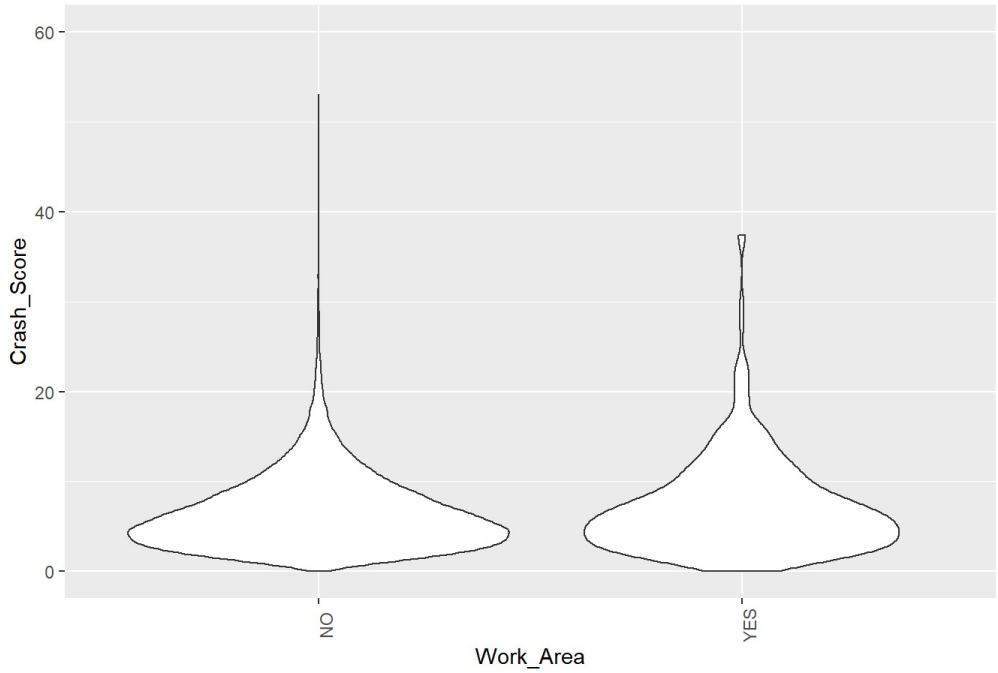
Weather Violin plot



Traffic_Control Violin plot



Work_Area Violin plot



model development

```
benchmarking <- function(formula, family, data){
  # 10 fold cross validation
  set.seed(123)
  data <- data[sample(nrow(data)),]
  folds <- cut(seq(1,nrow(data)),breaks=10,labels=FALSE)
  # create AIC, RMSE, R2, Adj_R2 list
  AIC_list <- c()
  RMSE_list <- c()
  R2_list <- c()
  OOS_R2_list <- c()
  for(i in 1:10){
    # train test split
    testIndexes <- which(folds==i,arr.ind=TRUE)
    testData <- data[testIndexes, ]
    trainData <- data[-testIndexes, ]
    # model
    glm <- glm(formula, family, data=trainData)
    # model summary
    summary <- summary(glm)
    # model prediction
    predict <- predict(glm,newdata=testData,type="response")
    # AIC
    AIC <- summary$aic
    # RMSE
    RMSE <- sqrt(sum((testData$Crash_Score-predict)^2)/nrow(testData))
    # R2
    R2 <- 1 - (summary$deviance/summary$null.deviance)
    # OOS R2
    OOS_R2 <- 1 - sum((testData$Crash_Score-predict)^2) /sum((testData$Crash_Score-mean(testData$Crash_Score))^2)
    # append to RMSE list
    AIC_list <- append(AIC_list, AIC)
    # append to RMSE list
    RMSE_list <- append(RMSE_list, RMSE)
    # append to R2 list
    R2_list <- append(R2_list, R2)
    # append to OOS R2 list
    OOS_R2_list <- append(OOS_R2_list, OOS_R2)
  }
  # mean AIC
  print("AIC")
  print(round(mean(AIC_list), 5))
  # mean RMSE
  print("RMSE")
  print(round(mean(RMSE_list), 5))
  # mean R2
  print("R2")
  print(round(mean(R2_list), 5))
  # mean Adj R2
  print("OOS_R2")
  print(round(mean(OOS_R2_list), 5))
}
```

```
benchmarking(Crash_Score ~ ., gaussian(), data)
```

```
## [1] "AIC"
## [1] 119370.4
## [1] "RMSE"
## [1] 4.25029
## [1] "R2"
## [1] 0.01658
## [1] "OOS_R2"
## [1] 0.01245
```

```
benchmarking(Crash_Score ~ ., Gamma(), data)
```

```
## [1] "AIC"
## [1] 111851.2
## [1] "RMSE"
## [1] 4.251
## [1] "R2"
## [1] 0.01687
## [1] "OOS_R2"
## [1] 0.01212
```

```
benchmarking(Crash_Score ~ ., gaussian(link="log"), data)
```

```
## [1] "AIC"
## [1] 119372.3
## [1] "RMSE"
## [1] 4.25057
## [1] "R2"
## [1] 0.0165
## [1] "OOS_R2"
## [1] 0.01231
```

```
benchmarking(Crash_Score ~ ., Gamma(link="log"), data)
```

```
## [1] "AIC"
## [1] 111847.2
## [1] "RMSE"
## [1] 4.25062
## [1] "R2"
## [1] 0.01705
## [1] "OOS_R2"
## [1] 0.0123
```

```
# conclusion 1
# the gaussian distribution has the highest OOS R2 and the lowest RMSE
# the gamma distribution has the highest R2 and the lowest AIC
# conclusion 2
# there are too many variables and too many varaible levels
# as a result, we will do a feature transformation
```

feature transfromation

```
data2 <- data
```

```
# feature transfromation
re_level <- function(var, re_levels){
  # level
  data2[,var] <- as.factor(data2[,var])
  var.levels <- levels(data2[,var])
  data2[,var] <-  mapvalues(data2[,var],var.levels, re_levels)
  # relevel
  table <- as.data.frame(table(data2[,var]))
  max <- which.max(table[,2])
  level.name <- as.character(table[max,1])
  data2[,var] <- relevel(data2[,var], ref=level.name)
  return(data2)
}
```

```
# time of day
data2 <- re_level("Time_of_Day", c("OVERNIGHT","LATE-EARLY","DAYTIME","DAYTIME","DAYTIME", "LATE-EARLY"))
# we relevel the time of day into day time (8am - 8pm), overnight (midnight to 4am), and late early (4am to 8am, 8pm to mi
dnight)
table(data2[, "Time_of_Day"])
```

```
##
##    DAYTIME  OVERNIGHT LATE-EARLY
##      18345        808       3984
```

```
# road feature
# level
data2 <- re_level("Rd_Feature" , c("OTHER","OTHER","INTERSECTION","OTHER","OTHER"))
# we relevel the road feature into interaction and other
table(data2[, "Rd_Feature"])
```

```
##
##        OTHER INTERSECTION
##        16435         6702
```

```
# road character
# level
data2 <- re_level("Rd_Character", c("STRAIGHT","CURVE","CURVE","CURVE","CURVE","STRAIGHT","STRAIGHT"))
# we relevel the road character into straight and curve, note that other is classfied as curve
table(data2[, "Rd_Character"])
```

```
##
## STRAIGHT    CURVE
##    21517     1620
```

```
# road surface
# level
data2 <- re_level("Rd_Surface", c("ASPHALT","ASPHALT","OTHER","OTHER","OTHER"))
# we relevel the road surface into asohalt and other
table(data2[, "Rd_Surface"])
```

```
##
## ASPHALT   OTHER
##   22004    1133
```

```
# weather
# level
data2 <- re_level("Weather", c("CLEAR-CLOUDY","CLEAR-CLOUDY","OTHER","RAIN-SNOW ","RAIN-SNOW "))
# we relevel the weather into clear couldy, rain snow, and other
table(data2[, "Weather"])
```

```
##
## CLEAR-CLOUDY        OTHER   RAIN-SNOW
##        20627           85         2425
```

```
# traffic control
# level
data2 <- re_level("Traffic_Control", c("OTHER","OTHER","SIGNAL-STOP","SIGNAL-STOP","OTHER"))
# we relevel the traffic control into signal-stop and other
table(data2[, "Traffic_Control"])
```

```
##
##       OTHER SIGNAL-STOP
##       14516        8621
```

```
summary(data2)
```

```
##   Crash_Score          year          Month           Time_of_Day
##  Min.   : 0.010   Min.   :2014   Min.   : 1.00   DAYTIME   :18345
##  1st Qu.: 3.540   1st Qu.:2015   1st Qu.: 3.00   OVERNIGHT :  808
##  Median : 5.660   Median :2016   Median : 7.00   LATE-EARLY: 3984
##  Mean   : 6.567   Mean   :2016   Mean   : 6.56
##  3rd Qu.: 8.600   3rd Qu.:2017   3rd Qu.:10.00
##  Max.   :53.070   Max.   :2019   Max.   :12.00
##         Rd_Feature      Rd_Character        Rd_Class
##  OTHER       :16435   STRAIGHT:21517   STATE HWY:10603
##  INTERSECTION: 6702   CURVE   : 1620   OTHER    : 9960
##                                        US HWY   : 2574
##
##
##
##                    Rd_Configuration   Rd_Surface           Rd_Conditions
##  TWO-WAY-NO-MEDIAN         :12076   ASPHALT:22004   DRY           :19262
##  ONE-WAY                   : 1496   OTHER  : 1133   ICE-SNOW-SLUSH:  322
##  TWO-WAY-PROTECTED-MEDIAN  : 2627                   OTHER         :  134
##  TWO-WAY-UNPROTECTED-MEDIAN: 6882                   WET           : 3419
##  UNKNOWN                   :   56
##
##           Light               Weather         Traffic_Control  Work_Area
##  DAYLIGHT    :18262   CLEAR-CLOUDY:20627   OTHER      :14516   NO :22823
##  DARK-LIT    : 3219   OTHER       :   85   SIGNAL-STOP: 8621   YES:  314
##  DARK-NOT-LIT:  708   RAIN-SNOW   : 2425
##  DAWN        :  140
##  DUSK        :  602
##  OTHER       :  206
```

```
write.csv(data2, "dataset2.csv")
```

```
benchmarking(Crash_Score ~ ., gaussian(), data2)
```

```
## [1] "AIC"
## [1] 119369.6
## [1] "RMSE"
## [1] 4.25016
## [1] "R2"
## [1] 0.01521
## [1] "OOS_R2"
## [1] 0.01251
```

```
benchmarking(Crash_Score ~ ., Gamma(), data2)
```

```
## [1] "AIC"
## [1] 111852.5
## [1] "RMSE"
## [1] 4.25075
## [1] "R2"
## [1] 0.01549
## [1] "OOS_R2"
## [1] 0.01223
```

```
benchmarking(Crash_Score ~ ., gaussian(link="log"), data2)
```

```
## [1] "AIC"
## [1] 119371.6
## [1] "RMSE"
## [1] 4.25036
## [1] "R2"
## [1] 0.01511
## [1] "OOS_R2"
## [1] 0.01241
```

```
benchmarking(Crash_Score ~ ., Gamma(link="log"), data2)
```

```
## [1] "AIC"
## [1] 111848.9
## [1] "RMSE"
## [1] 4.25046
## [1] "R2"
## [1] 0.01565
## [1] "OOS_R2"
## [1] 0.01237
```

```
# conclusion 1
# the gaussian distribution has the highest OOS R2 and the lowest RMSE
# the gamma distribution has the highest R2 and the lowest AIC
# conclusion 2
# feature transformation slightly improves the AIC, RMSE, and OOS R2
# conclusion 3
# there are still too many variables and too many varaible levels
# as a result, we will do a feature selection
```

feature selection

```
set.seed(123)
data2 <- data2[sample(nrow(data2)),]
folds <- cut(seq(1,nrow(data2)),breaks=10,labels=FALSE)
testIndexes <- which(folds==1,arr.ind=TRUE)
trainIndexes <- which(folds!=1,arr.ind=TRUE)
testData2 <- data2[testIndexes, ]
trainData2 <- data2[trainIndexes, ]
```

forward BIC

```
glm <- glm(Crash_Score ~ ., gaussian(), data = trainData2)
glm_1 <- glm(Crash_Score ~ 1, gaussian(), data = trainData2)
stepAIC(glm_1, direction="forward", k=log(nrow(trainData2)), scope=list(upper = glm, lower = glm_1))
```

```
## Start:  AIC=119540.7
## Crash_Score ~ 1
##
##                    Df Deviance    AIC
## + Rd_Class          2   376443 119401
## + Traffic_Control   1   376932 119419
## + Rd_Feature        1   377306 119439
## + Rd_Configuration  4   378245 119521
## + Time_of_Day       2   378874 119536
## + Rd_Character      1   379100 119538
## <none>                  379331 119541
## + Light             5   378480 119544
## + Work_Area         1   379246 119546
## + Rd_Surface        1   379283 119548
## + year              1   379312 119550
## + Month             1   379331 119551
## + Weather           2   379202 119554
## + Rd_Conditions     3   379316 119570
##
## Step:  AIC=119401.5
## Crash_Score ~ Rd_Class
##
##                    Df Deviance    AIC
## + Traffic_Control   1   375340 119350
## + Rd_Feature        1   375406 119354
## + Time_of_Day       2   375896 119391
## + Rd_Character      1   376176 119397
## <none>                  376443 119401
## + Light             5   375565 119403
## + Rd_Surface        1   376341 119406
## + Work_Area         1   376387 119408
## + year              1   376412 119410
## + Month             1   376442 119411
## + Weather           2   376349 119416
## + Rd_Conditions     3   376405 119429
## + Rd_Configuration  4   376348 119436
##
## Step:  AIC=119350.3
## Crash_Score ~ Rd_Class + Traffic_Control
##
##                    Df Deviance    AIC
## + Time_of_Day       2   374861 119344
## + Rd_Feature        1   375081 119346
## + Rd_Character      1   375111 119348
## <none>                  375340 119350
## + Work_Area         1   375267 119356
## + Light             5   374561 119357
## + Rd_Surface        1   375296 119358
## + year              1   375310 119359
## + Month             1   375338 119360
## + Weather           2   375247 119365
## + Rd_Conditions     3   375304 119378
## + Rd_Configuration  4   375261 119386
##
## Step:  AIC=119343.6
## Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day
##
##                    Df Deviance    AIC
## + Rd_Feature        1   374597 119339
## + Rd_Character      1   374654 119342
## <none>                  374861 119344
## + Work_Area         1   374791 119350
## + Rd_Surface        1   374819 119351
## + year              1   374826 119352
## + Month             1   374859 119353
## + Weather           2   374773 119359
## + Light             5   374452 119371
## + Rd_Conditions     3   374837 119372
## + Rd_Configuration  4   374773 119379
##
## Step:  AIC=119338.9
## Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day + Rd_Feature
##
```

```
##                     Df Deviance    AIC
## + Rd_Character      1   374398 119338
## <none>                  374597 119339
## + Work_Area         1   374527 119345
## + Rd_Surface        1   374558 119347
## + year              1   374566 119347
## + Month             1   374595 119349
## + Weather           2   374508 119354
## + Light             5   374186 119366
## + Rd_Conditions     3   374573 119367
## + Rd_Configuration  4   374520 119374
##
## Step:  AIC=119337.8
## Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day + Rd_Feature +
##     Rd_Character
##
##                     Df Deviance    AIC
## <none>                  374398 119338
## + Work_Area         1   374327 119344
## + Rd_Surface        1   374354 119345
## + year              1   374364 119346
## + Month             1   374396 119348
## + Weather           2   374313 119353
## + Light             5   373995 119365
## + Rd_Conditions     3   374372 119366
## + Rd_Configuration  4   374352 119375
```

```
##
## Call:  glm(formula = Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day +
##     Rd_Feature + Rd_Character, family = gaussian(), data = trainData2)
##
## Coefficients:
##                (Intercept)               Rd_ClassOTHER
##                     6.6557                     -0.5540
##             Rd_ClassUS HWY  Traffic_ControlSIGNAL-STOP
##                     0.1751                      0.3169
##       Time_of_DayOVERNIGHT       Time_of_DayLATE-EARLY
##                    -0.6200                     -0.2855
##     Rd_FeatureINTERSECTION           Rd_CharacterCURVE
##                     0.3189                     -0.3840
##
## Degrees of Freedom: 20822 Total (i.e. Null);  20815 Residual
## Null Deviance:       379300
## Residual Deviance: 374400    AIC: 119300
```

```
# reasoning
# we use forward BIC, as BIC is a more conservative approach compared to AIC, and there is a greater penalty for each para
meter added
```

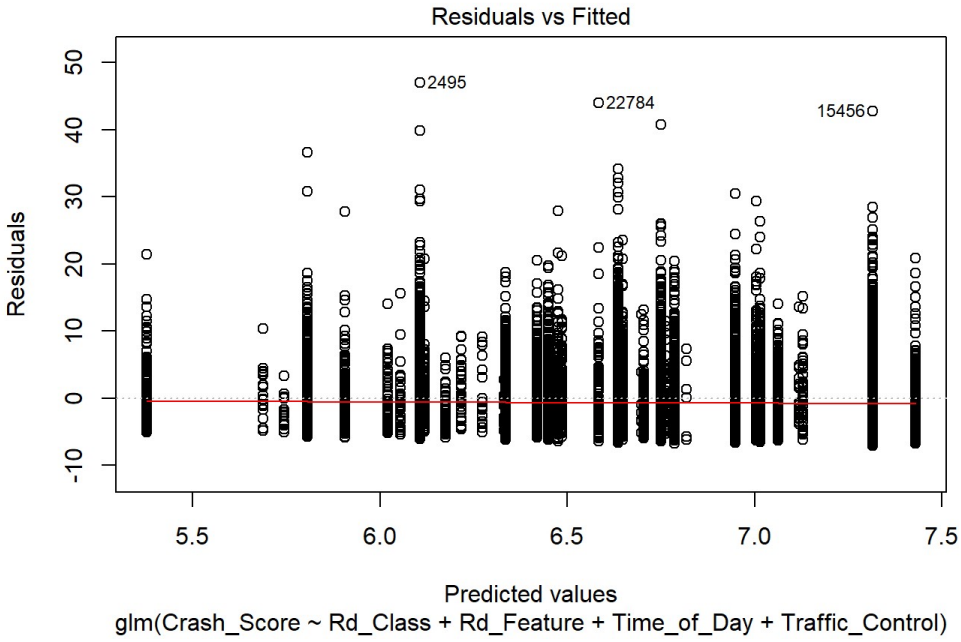
```
benchmarking(Crash_Score ~ Rd_Class + Rd_Feature + Time_of_Day + Traffic_Control, gaussian(), data2)
```

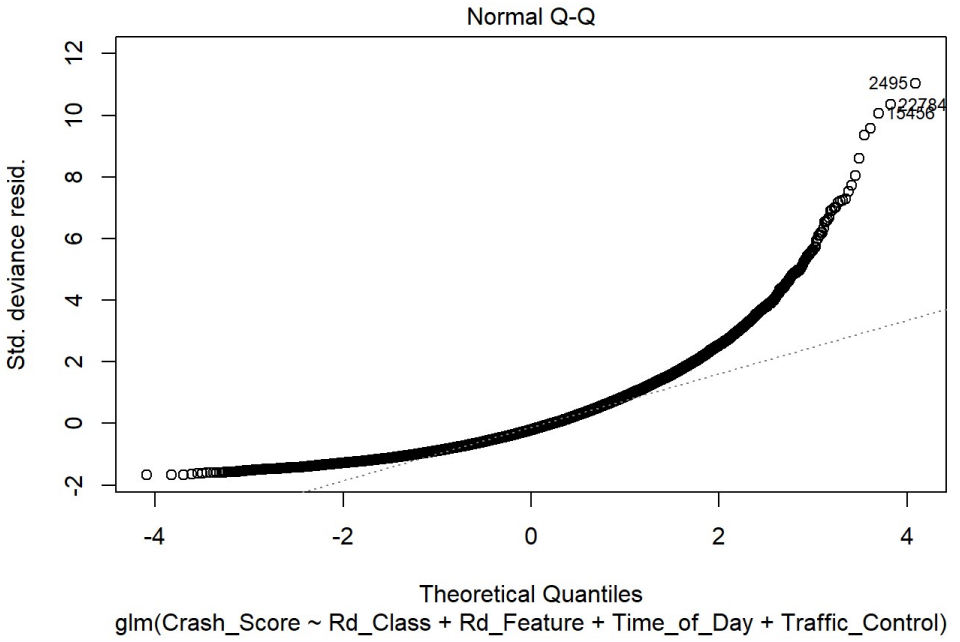
```
## [1] "AIC"
## [1] 119385.3
## [1] "RMSE"
## [1] 4.25146
## [1] "R2"
## [1] 0.01266
## [1] "OOS_R2"
## [1] 0.01172
```

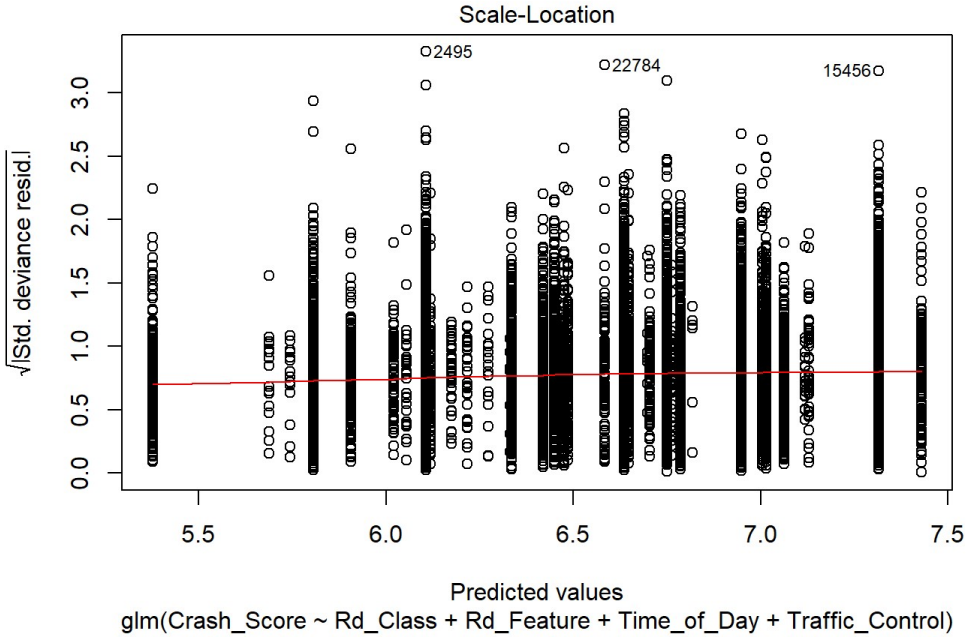
```
# visualization
glm_selected <- glm(Crash_Score ~ Rd_Class + Rd_Feature + Time_of_Day + Traffic_Control, gaussian(), data = data2)
summary(glm_selected)
```

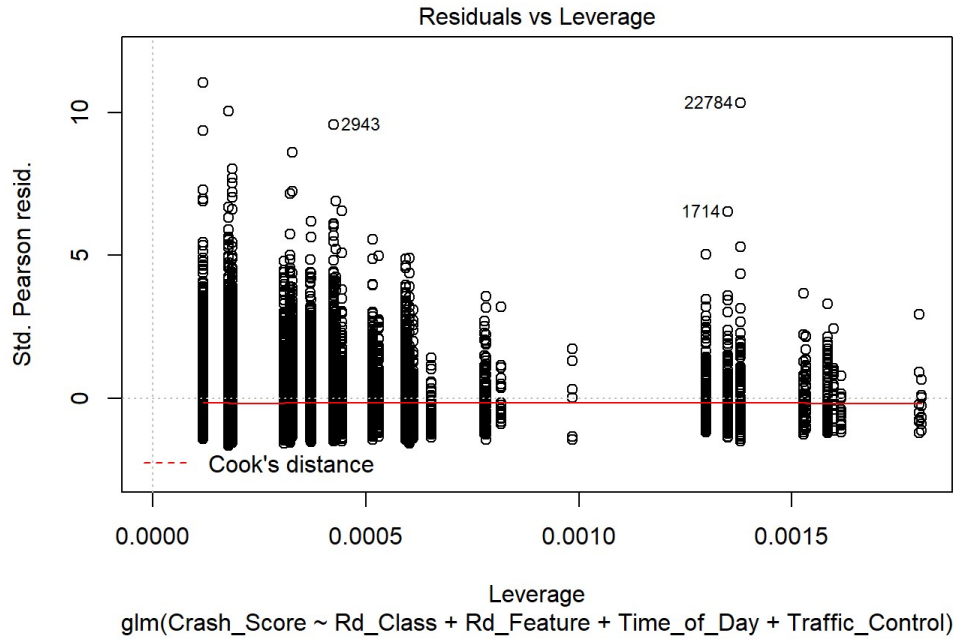
```
##
## Call:
## glm(formula = Crash_Score ~ Rd_Class + Rd_Feature + Time_of_Day +
##     Traffic_Control, family = gaussian(), data = data2)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -7.116  -2.990  -0.888   1.993  46.962
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.63688    0.05809 114.247  < 2e-16 ***
## Rd_ClassOTHER              -0.52929    0.06449  -8.207 2.38e-16 ***
## Rd_ClassUS HWY              0.11414    0.09647   1.183    0.237
## Rd_FeatureINTERSECTION      0.36731    0.08050   4.563 5.07e-06 ***
## Time_of_DayOVERNIGHT       -0.73084    0.15299  -4.777 1.79e-06 ***
## Time_of_DayLATE-EARLY      -0.30067    0.07449  -4.036 5.45e-05 ***
## Traffic_ControlSIGNAL-STOP  0.31143    0.07598   4.099 4.17e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 18.08283)
##
##     Null deviance: 423606  on 23136  degrees of freedom
## Residual deviance: 418256  on 23130  degrees of freedom
## AIC: 132650
##
## Number of Fisher Scoring iterations: 2
```

```
plot(glm_selected)
```


Residuals vs Fitted
Residuals
Predicted values
glm(Crash_Score ~ Rd_Class + Rd_Feature + Time_of_Day + Traffic_Control)
2495
22784
15456


Normal Q-Q
Std. deviance resid.
Theoretical Quantiles
glm(Crash_Score ~ Rd_Class + Rd_Feature + Time_of_Day + Traffic_Control)
2495
22784
15456






```
glm <- glm(Crash_Score ~ ., Gamma(link="log"), data = trainData2)
glm_1 <- glm(Crash_Score ~ 1, Gamma(link="log"), data = trainData2)
stepAIC(glm_1, direction="forward", k=log(nrow(trainData2)), scope=list(upper = glm, lower = glm_1))
```

```
## Start:  AIC=112078.1
## Crash_Score ~ 1
##
##                     Df Deviance    AIC
## + Rd_Class           2   8710.5 111937
## + Traffic_Control    1   8723.6 111959
## + Rd_Feature         1   8732.7 111980
## + Rd_Configuration   4   8753.4 112059
## + Time_of_Day        2   8767.5 112072
## + Rd_Character       1   8773.0 112075
## <none>                   8778.6 112078
## + Light              5   8757.5 112078
## + Work_Area          1   8776.7 112084
## + Rd_Surface         1   8777.4 112085
## + year               1   8778.1 112087
## + Month              1   8778.6 112088
## + Weather            2   8775.2 112090
## + Rd_Conditions      3   8778.2 112107
##
## Step:  AIC=111924.6
## Crash_Score ~ Rd_Class
##
##                     Df Deviance    AIC
## + Traffic_Control    1   8684.9 111874
## + Rd_Feature         1   8687.1 111879
## + Time_of_Day        2   8697.5 111914
## + Rd_Character       1   8704.3 111920
## <none>                   8710.5 111925
## + Light              5   8689.7 111925
## + Rd_Surface         1   8707.8 111928
## + Work_Area          1   8709.3 111932
## + year               1   8709.6 111933
## + Month              1   8710.4 111935
## + Weather            2   8707.8 111938
## + Rd_Conditions      3   8709.7 111953
## + Rd_Configuration   4   8708.1 111959
##
## Step:  AIC=111869.2
## Crash_Score ~ Rd_Class + Traffic_Control
##
##                     Df Deviance    AIC
## + Time_of_Day        2   8672.9 111860
## + Rd_Feature         1   8679.3 111866
## + Rd_Character       1   8679.8 111867
## <none>                   8684.9 111869
## + Light              5   8665.6 111873
## + Work_Area          1   8683.3 111875
## + Rd_Surface         1   8683.7 111876
## + year               1   8684.1 111877
## + Month              1   8684.9 111879
## + Weather            2   8682.4 111883
## + Rd_Conditions      3   8684.1 111897
## + Rd_Configuration   4   8683.3 111905
##
## Step:  AIC=111858.2
## Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day
##
##                     Df Deviance    AIC
## + Rd_Feature         1   8667.3 111855
## + Rd_Character       1   8668.2 111857
## <none>                   8672.9 111858
## + Work_Area          1   8671.2 111864
## + Rd_Surface         1   8671.7 111865
## + year               1   8671.9 111866
## + Month              1   8672.8 111868
## + Weather            2   8670.4 111872
## + Light              5   8662.4 111883
## + Rd_Conditions      3   8672.3 111887
## + Rd_Configuration   4   8671.1 111894
##
## Step:  AIC=111853.8
## Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day + Rd_Feature
##
```

```
##                    Df Deviance    AIC
## + Rd_Character      1   8662.7 111853
## <none>                 8667.3 111854
## + Work_Area         1   8665.6 111860
## + Rd_Surface        1   8666.2 111861
## + year              1   8666.4 111862
## + Month             1   8667.2 111864
## + Weather           2   8664.8 111868
## + Light             5   8656.7 111878
## + Rd_Conditions     3   8666.6 111882
## + Rd_Configuration  4   8665.6 111890
##
## Step:  AIC=111852.1
## Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day + Rd_Feature +
##     Rd_Character
##
##                    Df Deviance    AIC
## <none>                 8662.7 111852
## + Work_Area         1   8660.9 111858
## + Rd_Surface        1   8661.6 111859
## + year              1   8661.8 111860
## + Month             1   8662.7 111862
## + Weather           2   8660.4 111866
## + Light             5   8652.4 111877
## + Rd_Conditions     3   8662.0 111880
## + Rd_Configuration  4   8661.7 111889
```

```
##
## Call:  glm(formula = Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day +
##     Rd_Feature + Rd_Character, family = Gamma(link = "log"),
##     data = trainData2)
##
## Coefficients:
##                (Intercept)               Rd_ClassOTHER
##                    1.89363                    -0.08543
##             Rd_ClassUS HWY  Traffic_ControlSIGNAL-STOP
##                    0.02677                     0.04996
##       Time_of_DayOVERNIGHT       Time_of_DayLATE-EARLY
##                   -0.10061                    -0.04462
##     Rd_FeatureINTERSECTION           Rd_CharacterCURVE
##                    0.04639                    -0.05843
##
## Degrees of Freedom: 20822 Total (i.e. Null);  20815 Residual
## Null Deviance:       8779
## Residual Deviance: 8663  AIC: 111800
```

```
# reasoning
# we use forward BIC, as BIC is a more conservative approach compared to AIC, and there is a greater penalty for each para
meter added
```

```
benchmarking(Crash_Score ~ Rd_Class + Rd_Feature + Time_of_Day + Traffic_Control, Gamma(link="log"), data2)
```

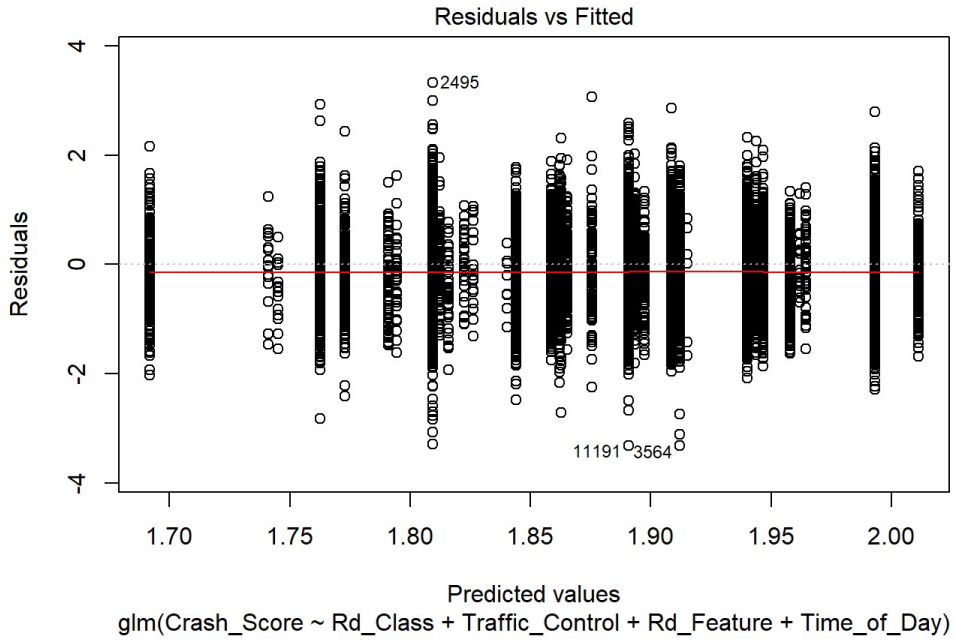
```
## [1] "AIC"
## [1] 111873.9
## [1] "RMSE"
## [1] 4.25163
## [1] "R2"
## [1] 0.01285
## [1] "OOS_R2"
## [1] 0.01163
```

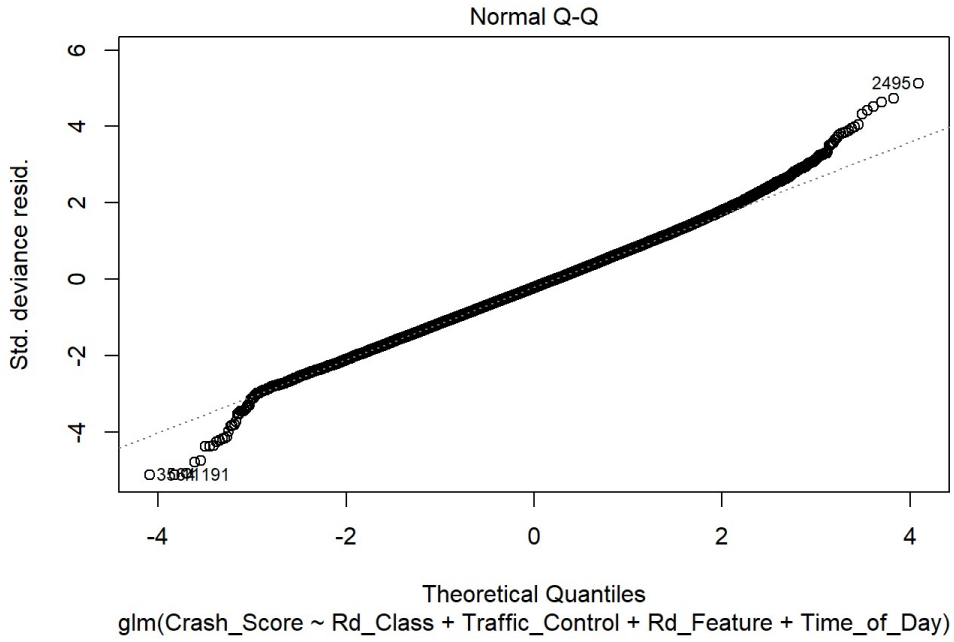
visualization

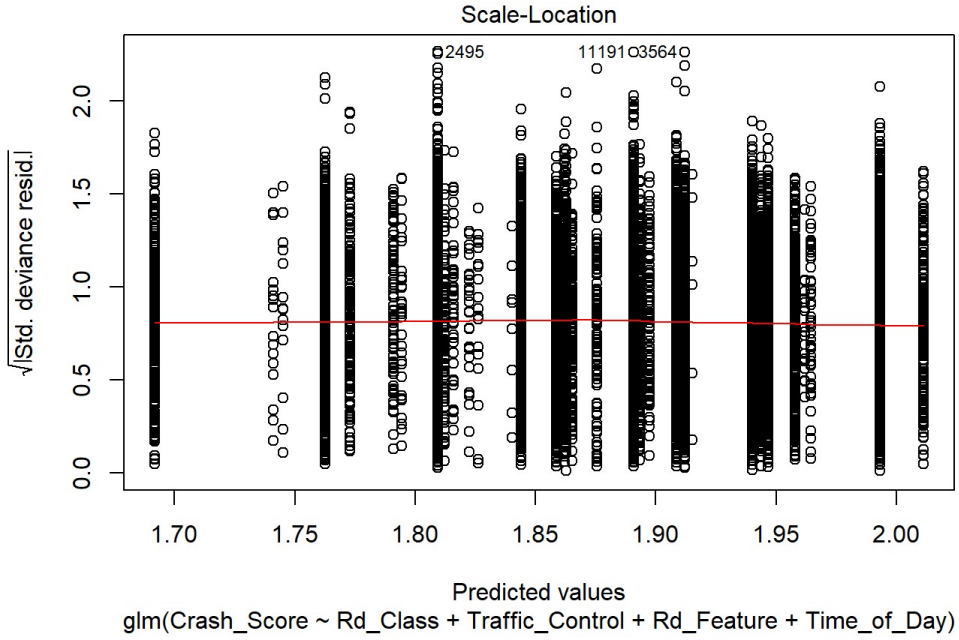
```
# visualization
glm_selected <- glm(Crash_Score ~ Rd_Class + Traffic_Control + Rd_Feature +
    Time_of_Day, Gamma(link="log"), data = data2)
summary(glm_selected)
```

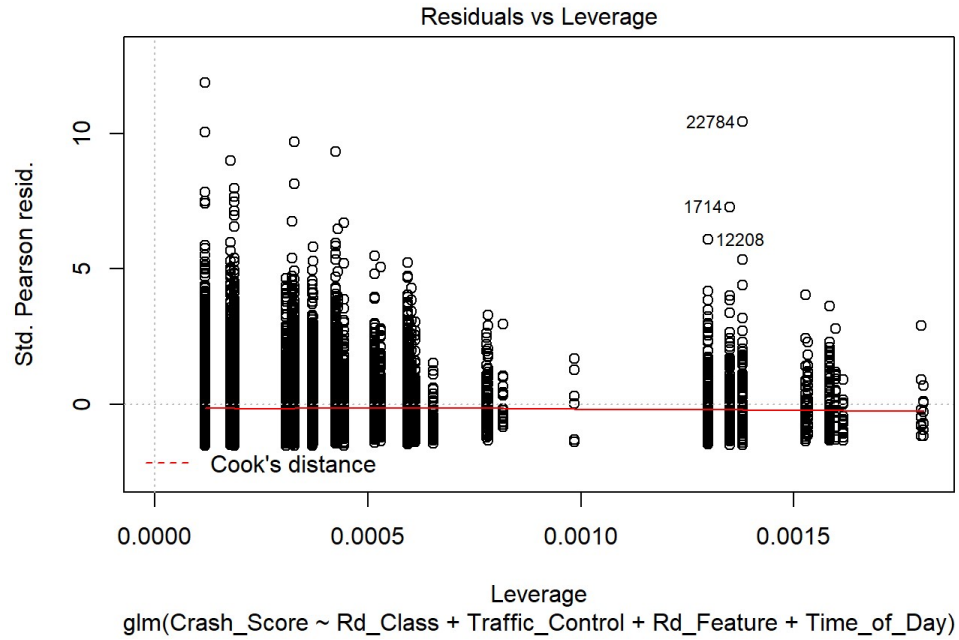
```
##
## Call:
## glm(formula = Crash_Score ~ Rd_Class + Traffic_Control + Rd_Feature +
##     Time_of_Day, family = Gamma(link = "log"), data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3222  -0.5540  -0.1431   0.2772   3.3251
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.890659   0.008845 213.758  < 2e-16 ***
## Rd_ClassOTHER              -0.081263   0.009819  -8.276  < 2e-16 ***
## Rd_ClassUS HWY              0.017964   0.014688   1.223    0.221
## Traffic_ControlSIGNAL-STOP  0.049331   0.011569   4.264 2.02e-05 ***
## Rd_FeatureINTERSECTION      0.053241   0.012257   4.344 1.41e-05 ***
## Time_of_DayOVERNIGHT       -0.117649   0.023294  -5.051 4.44e-07 ***
## Time_of_DayLATE-EARLY      -0.046650   0.011342  -4.113 3.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.4191914)
##
##     Null deviance: 9740.7  on 23136  degrees of freedom
## Residual deviance: 9615.7  on 23130  degrees of freedom
## AIC: 124303
##
## Number of Fisher Scoring iterations: 5
```

```
plot(glm_selected)
```


Residuals vs Fitted
Residuals
Predicted values
glm(Crash_Score ~ Rd_Class + Traffic_Control + Rd_Feature + Time_of_Day)
2495
11191
3564
1.70
1.75
1.80
1.85
1.90
1.95
2.00
4
2
0
-2
-4


Normal Q-Q
Std. deviance resid.
Theoretical Quantiles
glm(Crash_Score ~ Rd_Class + Traffic_Control + Rd_Feature + Time_of_Day)
2495
191
-4
-2
0
2
4
6





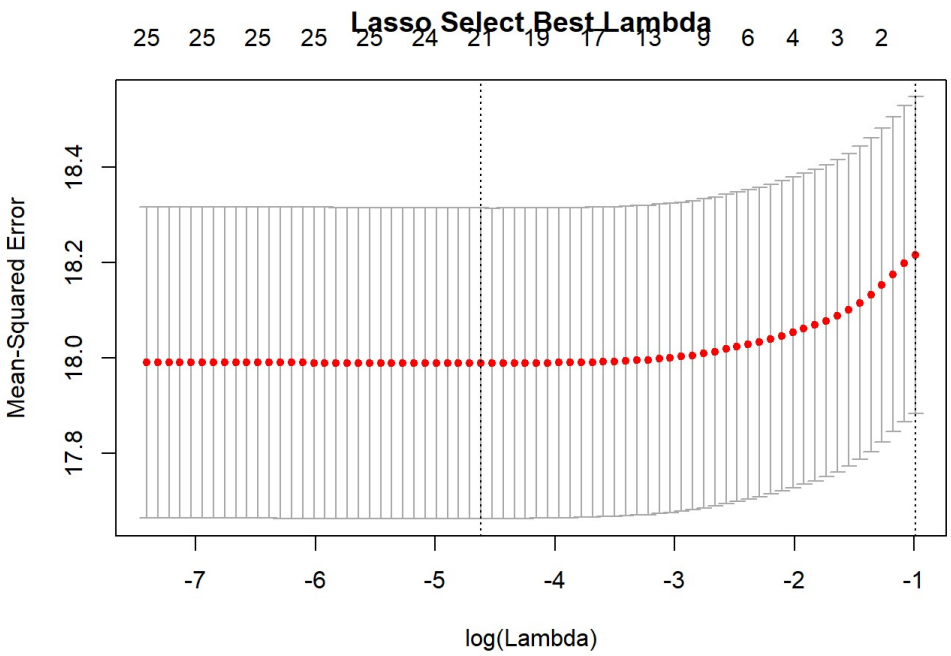

```
# the residuals vs fitted model indicates that the residuals has zero means and constant variables
# the q-q plot indicates that there is normal distribution for most residuals, but not for extremes
# it could fit better for a heavy tailed model
```

```
# conclusion 1
# feature selection with BIC forward does not improve the model performance
# however, feature selection identifies 4 key features, Rd_Class, Traffic_Control, Rd_Feature, and Time_of_Day
# conclusion 2
# gaussian distribution and gamma distribution with log link have similar model performance
# according to viusalization, gamma distribution with log link is a better choice
```

regularization

```
lasso <- function(x_train, x_test, y_train, y_test){
  # lasso regression
  lasso = glmnet(x_train, y_train, alpha = 1)
  # select best lambda
  cvlasso = cv.glmnet(x_train, y_train, type.measure="mse", nfolds = 10, alpha = 1)
  plot(cvlasso, main = "Lasso Select Best Lambda")
  lasso.lam.min = cvlasso$lambda.min
  # lasso regression with best lambda
  lasso_best <- glmnet(x_train, y_train, "gaussian", lambda = lasso.lam.min, alpha = 1)
  predict <- predict(lasso_best, x_test)

  # RMSE
  print("RMSE")
  RMSE <- sqrt(sum((y_test-predict)^2)/nrow(x_test))
  print(RMSE)
  # R2
  print("R2")
  R2 <- lasso_best$dev.ratio
  print(R2)
  # OOS R2
  print("OOS_R2")
  OOS_R2 <- 1 - sum((y_test-predict)^2) /sum((y_test-mean(y_test))^2)
  print(OOS_R2)
  # coefficient
  print("Coefficient")
  lasso.coef = coef(lasso, s = lasso.lam.min)
  print(lasso.coef)
}
# reasoning
# we use lasso regression, as there are useless variables
```

```
x_train <- model.matrix(Crash_Score ~ ., trainData2)
x_test <- model.matrix(Crash_Score ~ ., testData2)
y_train <- trainData2$Crash_Score
y_test <- testData2$Crash_Score

lasso(x_train, x_test, y_train, y_test)
```



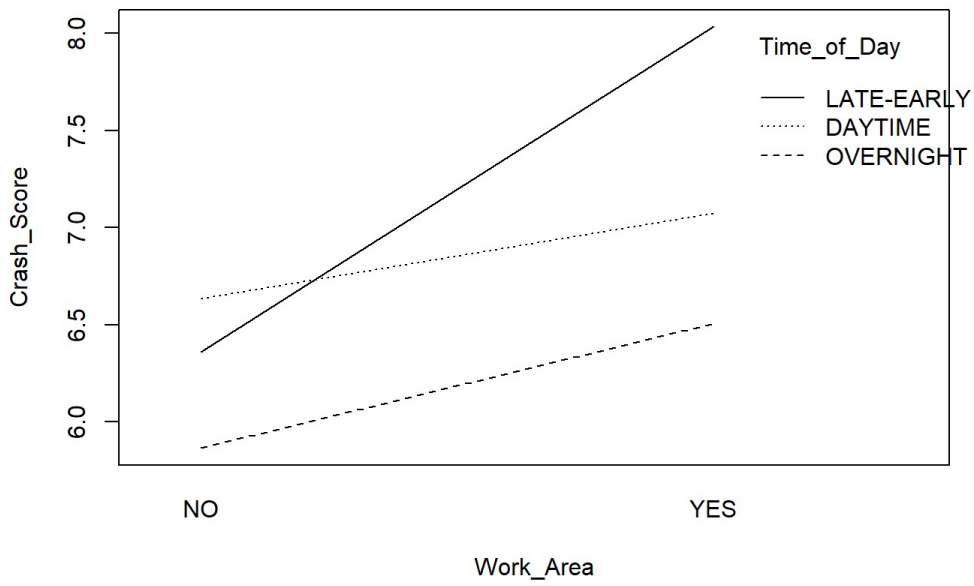
```
## [1] "RMSE"
## [1] 4.3442
## [1] "R2"
## [1] 0.0149278
## [1] "OOS_R2"
## [1] 0.0125859
## [1] "Coefficient"
## 27 x 1 sparse Matrix of class "dgCMatrix"
##                                                      1
## (Intercept)                                 46.74682364
## (Intercept)                                  .
## year                                        -0.01987011
## Month                                        .
## Time_of_DayOVERNIGHT                        -0.32920560
## Time_of_DayLATE-EARLY                       -0.12105863
## Rd_FeatureINTERSECTION                       0.30130566
## Rd_CharacterCURVE                           -0.32116857
## Rd_ClassOTHER                               -0.53137105
## Rd_ClassUS HWY                               0.15798595
## Rd_ConfigurationONE-WAY                     -0.10699815
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN     0.09526021
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN   .
## Rd_ConfigurationUNKNOWN                      0.07843682
## Rd_SurfaceOTHER                             -0.20296062
## Rd_ConditionsICE-SNOW-SLUSH                  .
## Rd_ConditionsOTHER                           0.51209521
## Rd_ConditionsWET                            -0.11392122
## LightDARK-LIT                               -0.32803771
## LightDARK-NOT-LIT                           -0.39816990
## LightDAWN                                   -0.06072986
## LightDUSK                                    .
## LightOTHER                                  -0.73964621
## WeatherOTHER                                -0.85475554
## WeatherRAIN-SNOW                             0.13708819
## Traffic_ControlSIGNAL-STOP                   0.29912913
## Work_AreaYES                                 0.39325957
```

```
# conclusion 1
# feature selection with lasso regression does not improve the model performance
```

interaction

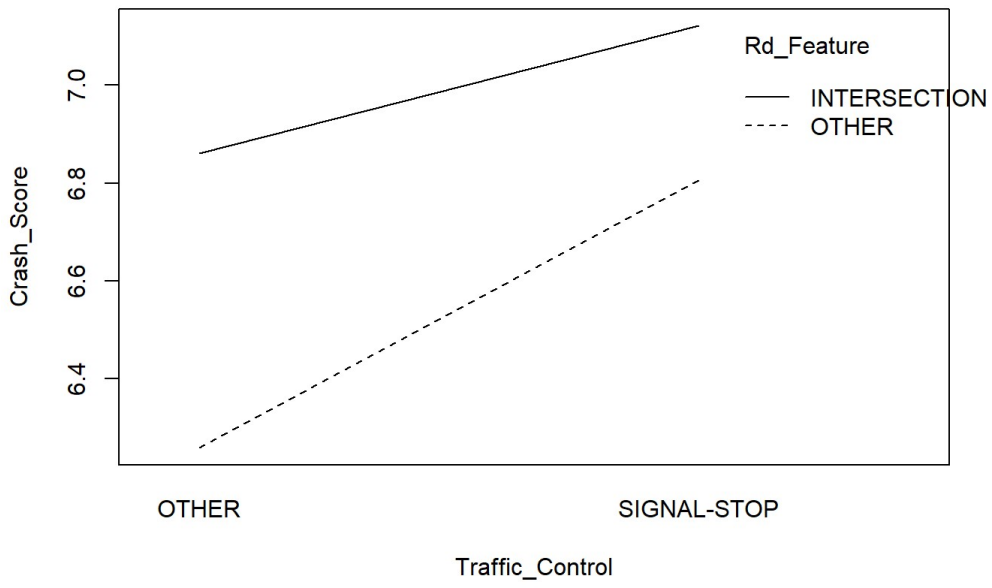
```
# an interaction is  when changing the level of one variable changes how levels of the other variables affect the dependen
t variable
```

```
interaction.plot(data2$Work_Area, data2$Time_of_Day, data2$Crash_Score, ylab="Crash_Score", xlab="Work_Area", trace.label
="Time_of_Day")
```




```
# as we can observe, the relative impact for work area is higher for late early, and lower for daytime and overnight
# as during the rush hour, there are more traffic in the work area
```

```
interaction.plot(data2$Traffic_Control, data2$Rd_Feature, data2$Crash_Score, ylab="Crash_Score", xlab="Traffic_Control", t
race.label="Rd_Feature")
```




```
# as we can observe, the relative impact for interaction is lower for signal stop
# this indicate that signal and stop could be a good traffic control practice for intersection
```

```
# conclusion 1
# as we can observe, there are lots of interaction between variables
# so we will taking interaction into consideration
```

```
# test
benchmarking(Crash_Score ~ . + (.)^2, Gamma(link="log"), data2)
```

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

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

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

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

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

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

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

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

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

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

```
## [1] "AIC"
## [1] 111936.3
## [1] "RMSE"
## [1] 4.31503
## [1] "R2"
## [1] 0.03526
## [1] "OOS_R2"
## [1] -0.01931
```

```
# conclusion 1
# as we can observe, taking interaction into consideration significantly improves R2 but not OOS R2
# as a result, we will do a feature selection
```

feature selection

forward BIC

```
glm <- glm(Crash_Score ~ . + (.)^2, Gamma(link="log"), data = trainData2)
glm_1 <- glm(Crash_Score ~ 1, Gamma(link="log"), data = trainData2)
stepAIC(glm_1, direction="forward", k=log(nrow(trainData2)), scope=list(upper = glm, lower = glm_1))
```

```
## Start:  AIC=112078.1
## Crash_Score ~ 1
##
##                     Df Deviance    AIC
## + Rd_Class           2   8710.5 111937
## + Traffic_Control    1   8723.6 111959
## + Rd_Feature         1   8732.7 111980
## + Rd_Configuration   4   8753.4 112059
## + Time_of_Day        2   8767.5 112072
## + Rd_Character       1   8773.0 112075
## <none>                   8778.6 112078
## + Light              5   8757.5 112078
## + Work_Area          1   8776.7 112084
## + Rd_Surface         1   8777.4 112085
## + year               1   8778.1 112087
## + Month              1   8778.6 112088
## + Weather            2   8775.2 112090
## + Rd_Conditions      3   8778.2 112107
##
## Step:  AIC=111924.6
## Crash_Score ~ Rd_Class
##
##                     Df Deviance    AIC
## + Traffic_Control    1   8684.9 111874
## + Rd_Feature         1   8687.1 111879
## + Time_of_Day        2   8697.5 111914
## + Rd_Character       1   8704.3 111920
## <none>                   8710.5 111925
## + Light              5   8689.7 111925
## + Rd_Surface         1   8707.8 111928
## + Work_Area          1   8709.3 111932
## + year               1   8709.6 111933
## + Month              1   8710.4 111935
## + Weather            2   8707.8 111938
## + Rd_Conditions      3   8709.7 111953
## + Rd_Configuration   4   8708.1 111959
##
## Step:  AIC=111869.2
## Crash_Score ~ Rd_Class + Traffic_Control
##
##                            Df Deviance    AIC
## + Time_of_Day               2   8672.9 111860
## + Rd_Feature                1   8679.3 111866
## + Rd_Character              1   8679.8 111867
## <none>                          8684.9 111869
## + Light                     5   8665.6 111873
## + Work_Area                 1   8683.3 111875
## + Rd_Surface                1   8683.7 111876
## + year                      1   8684.1 111877
## + Month                     1   8684.9 111879
## + Rd_Class:Traffic_Control  2   8680.8 111879
## + Weather                   2   8682.4 111883
## + Rd_Conditions             3   8684.1 111897
## + Rd_Configuration          4   8683.3 111905
##
## Step:  AIC=111858.2
## Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day
##
##                               Df Deviance    AIC
## + Rd_Feature                   1   8667.3 111855
## + Rd_Character                 1   8668.2 111857
## <none>                             8672.9 111858
## + Time_of_Day:Traffic_Control  2   8665.2 111860
## + Work_Area                    1   8671.2 111864
## + Rd_Surface                   1   8671.7 111865
## + year                         1   8671.9 111866
## + Rd_Class:Traffic_Control     2   8668.5 111868
## + Month                        1   8672.8 111868
## + Weather                      2   8670.4 111872
## + Light                        5   8662.4 111883
## + Time_of_Day:Rd_Class         4   8667.9 111886
## + Rd_Conditions                3   8672.3 111887
## + Rd_Configuration             4   8671.1 111894
```

```
##
## Step:  AIC=111853.8
## Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day + Rd_Feature
##
##                              Df Deviance    AIC
## + Rd_Character                 1   8662.7 111853
## <none>                            8667.3 111854
## + Time_of_Day:Traffic_Control  2   8660.0 111856
## + Work_Area                    1   8665.6 111860
## + Rd_Surface                   1   8666.2 111861
## + year                         1   8666.4 111862
## + Rd_Feature:Traffic_Control   1   8667.2 111863
## + Month                        1   8667.2 111864
## + Rd_Class:Traffic_Control     2   8663.6 111865
## + Weather                      2   8664.8 111868
## + Time_of_Day:Rd_Feature       2   8666.2 111871
## + Rd_Feature:Rd_Class          2   8666.7 111872
## + Light                        5   8656.7 111878
## + Time_of_Day:Rd_Class         4   8662.3 111882
## + Rd_Conditions                3   8666.6 111882
## + Rd_Configuration             4   8665.6 111890
##
## Step:  AIC=111852.1
## Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day + Rd_Feature +
##     Rd_Character
##
##                                Df Deviance    AIC
## <none>                              8662.7 111852
## + Time_of_Day:Traffic_Control   2   8655.9 111856
## + Work_Area                     1   8660.9 111858
## + Rd_Surface                    1   8661.6 111859
## + Rd_Character:Traffic_Control  1   8661.7 111860
## + year                          1   8661.8 111860
## + Rd_Feature:Traffic_Control    1   8662.5 111861
## + Month                         1   8662.7 111862
## + Rd_Feature:Rd_Character       1   8662.7 111862
## + Rd_Class:Traffic_Control      2   8659.7 111865
## + Rd_Character:Rd_Class         2   8660.3 111866
## + Weather                       2   8660.4 111866
## + Time_of_Day:Rd_Feature        2   8661.8 111870
## + Time_of_Day:Rd_Character      2   8662.0 111870
## + Rd_Feature:Rd_Class           2   8662.1 111871
## + Light                         5   8652.4 111877
## + Time_of_Day:Rd_Class          4   8657.7 111880
## + Rd_Conditions                 3   8662.0 111880
## + Rd_Configuration              4   8661.7 111889
```

```
##
## Call:  glm(formula = Crash_Score ~ Rd_Class + Traffic_Control + Time_of_Day +
##     Rd_Feature + Rd_Character, family = Gamma(link = "log"),
##     data = trainData2)
##
## Coefficients:
##                (Intercept)               Rd_ClassOTHER
##                    1.89363                    -0.08543
##             Rd_ClassUS HWY  Traffic_ControlSIGNAL-STOP
##                    0.02677                     0.04996
##       Time_of_DayOVERNIGHT       Time_of_DayLATE-EARLY
##                   -0.10061                    -0.04462
##     Rd_FeatureINTERSECTION           Rd_CharacterCURVE
##                    0.04639                    -0.05843
##
## Degrees of Freedom: 20822 Total (i.e. Null);  20815 Residual
## Null Deviance:       8779
## Residual Deviance: 8663  AIC: 111800
```

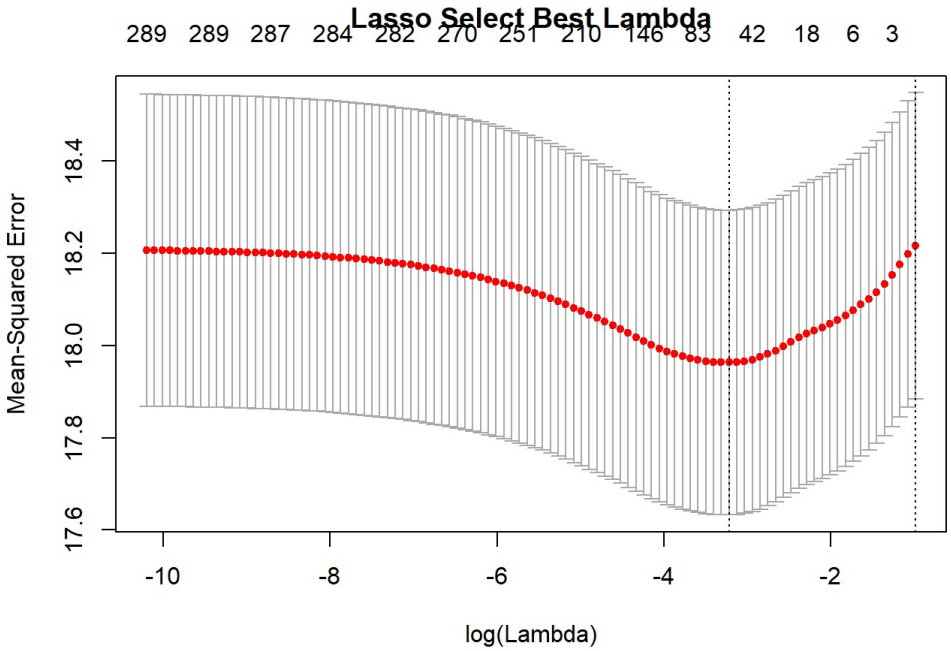
```
benchmarking(Crash_Score ~ Rd_Class + Rd_Feature + Time_of_Day + Traffic_Control, Gamma(link="log"), data2)
```

```
## [1] "AIC"
## [1] 111873.9
## [1] "RMSE"
## [1] 4.25163
## [1] "R2"
## [1] 0.01285
## [1] "OOS_R2"
## [1] 0.01163
```

```
# conclusion 1
# the interaction is not reflected in the BIC forward selection
```

regularization

```
x_train <- model.matrix(Crash_Score ~ . + (.)^2, trainData2)
x_test <- model.matrix(Crash_Score ~ . + (.)^2, testData2)
y_train <- trainData2$Crash_Score
y_test <- testData2$Crash_Score
lasso(x_train, x_test, y_train, y_test)
```



```
## [1] "RMSE"
## [1] 4.358811
## [1] "R2"
## [1] 0.02076226
## [1] "OOS_R2"
## [1] 0.005932683
## [1] "Coefficient"
## 305 x 1 sparse Matrix of class "dgCMatrix"
##                                                                            1
## (Intercept)                                                     6.669354e+00
## (Intercept)                                                     .
## year                                                            .
## Month                                                           .
## Time_of_DayOVERNIGHT                                            .
## Time_of_DayLATE-EARLY                                           .
## Rd_FeatureINTERSECTION                                          1.470619e-01
## Rd_CharacterCURVE                                              -7.640142e-02
## Rd_ClassOTHER                                                  -5.042379e-01
## Rd_ClassUS HWY                                                  .
## Rd_ConfigurationONE-WAY                                         .
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN                        .
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN                      .
## Rd_ConfigurationUNKNOWN                                         .
## Rd_SurfaceOTHER                                                 .
## Rd_ConditionsICE-SNOW-SLUSH                                     .
## Rd_ConditionsOTHER                                              .
## Rd_ConditionsWET                                                .
## LightDARK-LIT                                                   .
## LightDARK-NOT-LIT                                              -2.814100e-01
## LightDAWN                                                       .
## LightDUSK                                                       .
## LightOTHER                                                     -3.415025e-01
## WeatherOTHER                                                   -1.801047e-01
## WeatherRAIN-SNOW                                                .
## Traffic_ControlSIGNAL-STOP                                      2.755850e-01
## Work_AreaYES                                                    .
## year:Month                                                      .
## year:Time_of_DayOVERNIGHT                                       .
## year:Time_of_DayLATE-EARLY                                      .
## year:Rd_FeatureINTERSECTION                                     .
## year:Rd_CharacterCURVE                                         -6.350639e-06
## year:Rd_ClassOTHER                                             -7.110371e-06
## year:Rd_ClassUS HWY                                             .
## year:Rd_ConfigurationONE-WAY                                    .
## year:Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN                   .
## year:Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN                 .
## year:Rd_ConfigurationUNKNOWN                                    .
## year:Rd_SurfaceOTHER                                            .
## year:Rd_ConditionsICE-SNOW-SLUSH                                .
## year:Rd_ConditionsOTHER                                         .
## year:Rd_ConditionsWET                                           .
## year:LightDARK-LIT                                              .
## year:LightDARK-NOT-LIT                                         -2.549550e-07
## year:LightDAWN                                                  .
## year:LightDUSK                                                  .
## year:LightOTHER                                                -2.016596e-06
## year:WeatherOTHER                                              -5.315008e-06
## year:WeatherRAIN-SNOW                                           .
## year:Traffic_ControlSIGNAL-STOP                                 .
## year:Work_AreaYES                                               .
## Month:Time_of_DayOVERNIGHT                                      .
## Month:Time_of_DayLATE-EARLY                                     .
## Month:Rd_FeatureINTERSECTION                                    .
## Month:Rd_CharacterCURVE                                         .
## Month:Rd_ClassOTHER                                             .
## Month:Rd_ClassUS HWY                                            .
## Month:Rd_ConfigurationONE-WAY                                   .
## Month:Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN                  1.333912e-02
## Month:Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN                .
## Month:Rd_ConfigurationUNKNOWN                                   .
## Month:Rd_SurfaceOTHER                                           .
## Month:Rd_ConditionsICE-SNOW-SLUSH                               .
## Month:Rd_ConditionsOTHER                                        .
```

```
## Month:Rd_ConditionsWET                                                .
## Month:LightDARK-LIT                                                  -8.729910e-03
## Month:LightDARK-NOT-LIT                                               .
## Month:LightDAWN                                                       .
## Month:LightDUSK                                                       .
## Month:LightOTHER                                                      .
## Month:WeatherOTHER                                                    .
## Month:WeatherRAIN-SNOW                                                .
## Month:Traffic_ControlSIGNAL-STOP                                      .
## Month:Work_AreaYES                                                    .
## Time_of_DayOVERNIGHT:Rd_FeatureINTERSECTION                           .
## Time_of_DayLATE-EARLY:Rd_FeatureINTERSECTION                          .
## Time_of_DayOVERNIGHT:Rd_CharacterCURVE                                .
## Time_of_DayLATE-EARLY:Rd_CharacterCURVE                              -1.641321e-01
## Time_of_DayOVERNIGHT:Rd_ClassOTHER                                    .
## Time_of_DayLATE-EARLY:Rd_ClassOTHER                                   .
## Time_of_DayOVERNIGHT:Rd_ClassUS HWY                                   .
## Time_of_DayLATE-EARLY:Rd_ClassUS HWY                                  .
## Time_of_DayOVERNIGHT:Rd_ConfigurationONE-WAY                          .
## Time_of_DayLATE-EARLY:Rd_ConfigurationONE-WAY                         .
## Time_of_DayOVERNIGHT:Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN         -1.226415e+00
## Time_of_DayLATE-EARLY:Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN        .
## Time_of_DayOVERNIGHT:Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN       .
## Time_of_DayLATE-EARLY:Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN      -3.887772e-02
## Time_of_DayOVERNIGHT:Rd_ConfigurationUNKNOWN                          .
## Time_of_DayLATE-EARLY:Rd_ConfigurationUNKNOWN                         .
## Time_of_DayOVERNIGHT:Rd_SurfaceOTHER                                  .
## Time_of_DayLATE-EARLY:Rd_SurfaceOTHER                                 .
## Time_of_DayOVERNIGHT:Rd_ConditionsICE-SNOW-SLUSH                      .
## Time_of_DayLATE-EARLY:Rd_ConditionsICE-SNOW-SLUSH                    -8.857857e-03
## Time_of_DayOVERNIGHT:Rd_ConditionsOTHER                               .
## Time_of_DayLATE-EARLY:Rd_ConditionsOTHER                              .
## Time_of_DayOVERNIGHT:Rd_ConditionsWET                                 .
## Time_of_DayLATE-EARLY:Rd_ConditionsWET                               -2.487585e-01
## Time_of_DayOVERNIGHT:LightDARK-LIT                                   -6.660436e-01
## Time_of_DayLATE-EARLY:LightDARK-LIT                                  -2.568340e-01
## Time_of_DayOVERNIGHT:LightDARK-NOT-LIT                               -6.302672e-03
## Time_of_DayLATE-EARLY:LightDARK-NOT-LIT                               .
## Time_of_DayOVERNIGHT:LightDAWN                                        .
## Time_of_DayLATE-EARLY:LightDAWN                                       .
## Time_of_DayOVERNIGHT:LightDUSK                                        .
## Time_of_DayLATE-EARLY:LightDUSK                                       .
## Time_of_DayOVERNIGHT:LightOTHER                                       .
## Time_of_DayLATE-EARLY:LightOTHER                                      .
## Time_of_DayOVERNIGHT:WeatherOTHER                                    -1.383097e+00
## Time_of_DayLATE-EARLY:WeatherOTHER                                    .
## Time_of_DayOVERNIGHT:WeatherRAIN-SNOW                                 .
## Time_of_DayLATE-EARLY:WeatherRAIN-SNOW                               -9.565979e-02
## Time_of_DayOVERNIGHT:Traffic_ControlSIGNAL-STOP                       2.839178e-01
## Time_of_DayLATE-EARLY:Traffic_ControlSIGNAL-STOP                      1.233154e-01
## Time_of_DayOVERNIGHT:Work_AreaYES                                     .
## Time_of_DayLATE-EARLY:Work_AreaYES                                    3.600130e-01
## Rd_FeatureINTERSECTION:Rd_CharacterCURVE                              .
## Rd_FeatureINTERSECTION:Rd_ClassOTHER                                  .
## Rd_FeatureINTERSECTION:Rd_ClassUS HWY                                 .
## Rd_FeatureINTERSECTION:Rd_ConfigurationONE-WAY                        .
## Rd_FeatureINTERSECTION:Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN       .
## Rd_FeatureINTERSECTION:Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN     1.387518e-01
## Rd_FeatureINTERSECTION:Rd_ConfigurationUNKNOWN                        .
## Rd_FeatureINTERSECTION:Rd_SurfaceOTHER                                .
## Rd_FeatureINTERSECTION:Rd_ConditionsICE-SNOW-SLUSH                    .
## Rd_FeatureINTERSECTION:Rd_ConditionsOTHER                             .
## Rd_FeatureINTERSECTION:Rd_ConditionsWET                               .
## Rd_FeatureINTERSECTION:LightDARK-LIT                                  .
## Rd_FeatureINTERSECTION:LightDARK-NOT-LIT                              .
## Rd_FeatureINTERSECTION:LightDAWN                                      .
## Rd_FeatureINTERSECTION:LightDUSK                                      .
## Rd_FeatureINTERSECTION:LightOTHER                                     .
## Rd_FeatureINTERSECTION:WeatherOTHER                                   .
## Rd_FeatureINTERSECTION:WeatherRAIN-SNOW                               1.386209e-01
## Rd_FeatureINTERSECTION:Traffic_ControlSIGNAL-STOP                     .
## Rd_FeatureINTERSECTION:Work_AreaYES                                   .
## Rd_CharacterCURVE:Rd_ClassOTHER                                       .
```

```
## Rd_CharacterCURVE:Rd_ClassUS HWY                                          -9.061925e-02
## Rd_CharacterCURVE:Rd_ConfigurationONE-WAY                                   .
## Rd_CharacterCURVE:Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN                  .
## Rd_CharacterCURVE:Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN               -1.089713e-02
## Rd_CharacterCURVE:Rd_ConfigurationUNKNOWN                                   .
## Rd_CharacterCURVE:Rd_SurfaceOTHER                                          .
## Rd_CharacterCURVE:Rd_ConditionsICE-SNOW-SLUSH                              .
## Rd_CharacterCURVE:Rd_ConditionsOTHER                                       .
## Rd_CharacterCURVE:Rd_ConditionsWET                                         .
## Rd_CharacterCURVE:LightDARK-LIT                                           -5.037818e-02
## Rd_CharacterCURVE:LightDARK-NOT-LIT                                        .
## Rd_CharacterCURVE:LightDAWN                                                .
## Rd_CharacterCURVE:LightDUSK                                                .
## Rd_CharacterCURVE:LightOTHER                                               .
## Rd_CharacterCURVE:WeatherOTHER                                             .
## Rd_CharacterCURVE:WeatherRAIN-SNOW                                        -3.048711e-01
## Rd_CharacterCURVE:Traffic_ControlSIGNAL-STOP                               .
## Rd_CharacterCURVE:Work_AreaYES                                             2.433929e-01
## Rd_ClassOTHER:Rd_ConfigurationONE-WAY                                      .
## Rd_ClassUS HWY:Rd_ConfigurationONE-WAY                                     .
## Rd_ClassOTHER:Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN                      8.562199e-02
## Rd_ClassUS HWY:Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN                    .
## Rd_ClassOTHER:Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN                   .
## Rd_ClassUS HWY:Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN                   3.525367e-01
## Rd_ClassOTHER:Rd_ConfigurationUNKNOWN                                      .
## Rd_ClassUS HWY:Rd_ConfigurationUNKNOWN                                     .
## Rd_ClassOTHER:Rd_SurfaceOTHER                                             -4.245063e-01
## Rd_ClassUS HWY:Rd_SurfaceOTHER                                             .
## Rd_ClassOTHER:Rd_ConditionsICE-SNOW-SLUSH                                  2.167308e-01
## Rd_ClassUS HWY:Rd_ConditionsICE-SNOW-SLUSH                                 .
## Rd_ClassOTHER:Rd_ConditionsOTHER                                           .
## Rd_ClassUS HWY:Rd_ConditionsOTHER                                          .
## Rd_ClassOTHER:Rd_ConditionsWET                                             .
## Rd_ClassUS HWY:Rd_ConditionsWET                                            .
## Rd_ClassOTHER:LightDARK-LIT                                                .
## Rd_ClassUS HWY:LightDARK-LIT                                               .
## Rd_ClassOTHER:LightDARK-NOT-LIT                                            .
## Rd_ClassUS HWY:LightDARK-NOT-LIT                                           .
## Rd_ClassOTHER:LightDAWN                                                    .
## Rd_ClassUS HWY:LightDAWN                                                  -3.504681e-01
## Rd_ClassOTHER:LightDUSK                                                    .
## Rd_ClassUS HWY:LightDUSK                                                   .
## Rd_ClassOTHER:LightOTHER                                                   .
## Rd_ClassUS HWY:LightOTHER                                                 -2.527679e-01
## Rd_ClassOTHER:WeatherOTHER                                                 .
## Rd_ClassUS HWY:WeatherOTHER                                                .
## Rd_ClassOTHER:WeatherRAIN-SNOW                                             .
## Rd_ClassUS HWY:WeatherRAIN-SNOW                                            .
## Rd_ClassOTHER:Traffic_ControlSIGNAL-STOP                                   2.024663e-03
## Rd_ClassUS HWY:Traffic_ControlSIGNAL-STOP                                  .
## Rd_ClassOTHER:Work_AreaYES                                                 .
## Rd_ClassUS HWY:Work_AreaYES                                                6.153946e-01
## Rd_ConfigurationONE-WAY:Rd_SurfaceOTHER                                    .
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:Rd_SurfaceOTHER                   .
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:Rd_SurfaceOTHER                 .
## Rd_ConfigurationUNKNOWN:Rd_SurfaceOTHER                                    .
## Rd_ConfigurationONE-WAY:Rd_ConditionsICE-SNOW-SLUSH                        .
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:Rd_ConditionsICE-SNOW-SLUSH       -7.839814e-03
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:Rd_ConditionsICE-SNOW-SLUSH     .
## Rd_ConfigurationUNKNOWN:Rd_ConditionsICE-SNOW-SLUSH                        .
## Rd_ConfigurationONE-WAY:Rd_ConditionsOTHER                                 .
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:Rd_ConditionsOTHER                -4.699878e-01
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:Rd_ConditionsOTHER              .
## Rd_ConfigurationUNKNOWN:Rd_ConditionsOTHER                                 .
## Rd_ConfigurationONE-WAY:Rd_ConditionsWET                                   .
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:Rd_ConditionsWET                  .
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:Rd_ConditionsWET                .
## Rd_ConfigurationUNKNOWN:Rd_ConditionsWET                                   .
## Rd_ConfigurationONE-WAY:LightDARK-LIT                                      .
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:LightDARK-LIT                     .
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:LightDARK-LIT                   -3.186050e-01
## Rd_ConfigurationUNKNOWN:LightDARK-LIT                                      .
## Rd_ConfigurationONE-WAY:LightDARK-NOT-LIT                                  .
```

```
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:LightDARK-NOT-LIT             .
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:LightDARK-NOT-LIT           .
## Rd_ConfigurationUNKNOWN:LightDARK-NOT-LIT                              .
## Rd_ConfigurationONE-WAY:LightDAWN                                      .
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:LightDAWN                     .
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:LightDAWN                   .
## Rd_ConfigurationUNKNOWN:LightDAWN                                      .
## Rd_ConfigurationONE-WAY:LightDUSK                                      3.715807e-01
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:LightDUSK                     .
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:LightDUSK                   .
## Rd_ConfigurationUNKNOWN:LightDUSK                                      .
## Rd_ConfigurationONE-WAY:LightOTHER                                     .
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:LightOTHER                    -7.459324e-01
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:LightOTHER                  .
## Rd_ConfigurationUNKNOWN:LightOTHER                                     .
## Rd_ConfigurationONE-WAY:WeatherOTHER                                   -2.872357e-01
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:WeatherOTHER                  -1.226187e+00
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:WeatherOTHER                .
## Rd_ConfigurationUNKNOWN:WeatherOTHER                                   .
## Rd_ConfigurationONE-WAY:WeatherRAIN-SNOW                               .
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:WeatherRAIN-SNOW              .
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:WeatherRAIN-SNOW            .
## Rd_ConfigurationUNKNOWN:WeatherRAIN-SNOW                               .
## Rd_ConfigurationONE-WAY:Traffic_ControlSIGNAL-STOP                     -4.879065e-01
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:Traffic_ControlSIGNAL-STOP    .
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:Traffic_ControlSIGNAL-STOP  .
## Rd_ConfigurationUNKNOWN:Traffic_ControlSIGNAL-STOP                     .
## Rd_ConfigurationONE-WAY:Work_AreaYES                                   .
## Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN:Work_AreaYES                  2.848331e-01
## Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN:Work_AreaYES                .
## Rd_ConfigurationUNKNOWN:Work_AreaYES                                   .
## Rd_SurfaceOTHER:Rd_ConditionsICE-SNOW-SLUSH                            .
## Rd_SurfaceOTHER:Rd_ConditionsOTHER                                     .
## Rd_SurfaceOTHER:Rd_ConditionsWET                                       -1.201060e-01
## Rd_SurfaceOTHER:LightDARK-LIT                                          .
## Rd_SurfaceOTHER:LightDARK-NOT-LIT                                      .
## Rd_SurfaceOTHER:LightDAWN                                              .
## Rd_SurfaceOTHER:LightDUSK                                              .
## Rd_SurfaceOTHER:LightOTHER                                             .
## Rd_SurfaceOTHER:WeatherOTHER                                           .
## Rd_SurfaceOTHER:WeatherRAIN-SNOW                                       .
## Rd_SurfaceOTHER:Traffic_ControlSIGNAL-STOP                             .
## Rd_SurfaceOTHER:Work_AreaYES                                           .
## Rd_ConditionsICE-SNOW-SLUSH:LightDARK-LIT                              .
## Rd_ConditionsOTHER:LightDARK-LIT                                       .
## Rd_ConditionsWET:LightDARK-LIT                                         .
## Rd_ConditionsICE-SNOW-SLUSH:LightDARK-NOT-LIT                          1.326312e+00
## Rd_ConditionsOTHER:LightDARK-NOT-LIT                                   .
## Rd_ConditionsWET:LightDARK-NOT-LIT                                     .
## Rd_ConditionsICE-SNOW-SLUSH:LightDAWN                                  .
## Rd_ConditionsOTHER:LightDAWN                                           .
## Rd_ConditionsWET:LightDAWN                                             .
## Rd_ConditionsICE-SNOW-SLUSH:LightDUSK                                  .
## Rd_ConditionsOTHER:LightDUSK                                           .
## Rd_ConditionsWET:LightDUSK                                             .
## Rd_ConditionsICE-SNOW-SLUSH:LightOTHER                                 .
## Rd_ConditionsOTHER:LightOTHER                                          .
## Rd_ConditionsWET:LightOTHER                                            .
## Rd_ConditionsICE-SNOW-SLUSH:WeatherOTHER                               .
## Rd_ConditionsOTHER:WeatherOTHER                                        .
## Rd_ConditionsWET:WeatherOTHER                                          .
## Rd_ConditionsICE-SNOW-SLUSH:WeatherRAIN-SNOW                           .
## Rd_ConditionsOTHER:WeatherRAIN-SNOW                                    .
## Rd_ConditionsWET:WeatherRAIN-SNOW                                      .
## Rd_ConditionsICE-SNOW-SLUSH:Traffic_ControlSIGNAL-STOP                 .
## Rd_ConditionsOTHER:Traffic_ControlSIGNAL-STOP                          .
## Rd_ConditionsWET:Traffic_ControlSIGNAL-STOP                            .
## Rd_ConditionsICE-SNOW-SLUSH:Work_AreaYES                               .
## Rd_ConditionsOTHER:Work_AreaYES                                        1.028946e+01
## Rd_ConditionsWET:Work_AreaYES                                          .
## LightDARK-LIT:WeatherOTHER                                             .
## LightDARK-NOT-LIT:WeatherOTHER                                         .
## LightDAWN:WeatherOTHER                                                 .
```

```
## LightDUSK:WeatherOTHER                                    .
## LightOTHER:WeatherOTHER                                   .
## LightDARK-LIT:WeatherRAIN-SNOW                            5.700702e-02
## LightDARK-NOT-LIT:WeatherRAIN-SNOW                        .
## LightDAWN:WeatherRAIN-SNOW                                .
## LightDUSK:WeatherRAIN-SNOW                                .
## LightOTHER:WeatherRAIN-SNOW                               .
## LightDARK-LIT:Traffic_ControlSIGNAL-STOP                  1.289174e-01
## LightDARK-NOT-LIT:Traffic_ControlSIGNAL-STOP              .
## LightDAWN:Traffic_ControlSIGNAL-STOP                      .
## LightDUSK:Traffic_ControlSIGNAL-STOP                      .
## LightOTHER:Traffic_ControlSIGNAL-STOP                     .
## LightDARK-LIT:Work_AreaYES                                .
## LightDARK-NOT-LIT:Work_AreaYES                            .
## LightDAWN:Work_AreaYES                                    .
## LightDUSK:Work_AreaYES                                    .
## LightOTHER:Work_AreaYES                                   .
## WeatherOTHER:Traffic_ControlSIGNAL-STOP                   .
## WeatherRAIN-SNOW :Traffic_ControlSIGNAL-STOP              .
## WeatherOTHER:Work_AreaYES                                 .
## WeatherRAIN-SNOW :Work_AreaYES                            .
## Traffic_ControlSIGNAL-STOP:Work_AreaYES                   .
```

```
# conclusion 1
# feature selection with lasso regression significantly improves the R2 and slightly improves OOS R2
```

outliers

```
data3 <- filter(data2, Crash_Score <= 10)
```

```
benchmarking(Crash_Score ~ ., gaussian(), data3)
```

```
## [1] "AIC"
## [1] 78417.06
## [1] "RMSE"
## [1] 2.3559
## [1] "R2"
## [1] 0.01402
## [1] "OOS_R2"
## [1] 0.01051
```