CKME136_Capstone_Projects

Ebunoluwa Odeniyi

Load required packages

```
require(plyr)
```

```
## Loading required package: plyr
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

Load data

```
accs <- read.csv("C:/Users/YENN/Desktop/UST/FARS2016N/accident2016.csv", header = T, s
tringsAsFactors = F)
```

summay

```
summary(accs)
```

```
##      STATE          ST_CASE          VE_TOTAL         VE_FORMS
##  Min.   : 1.00   Min.   : 10001   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:12.00   1st Qu.:122032   1st Qu.: 1.000   1st Qu.: 1.000
##  Median :26.00   Median :260878   Median : 1.000   Median : 1.000
##  Mean   :27.14   Mean   :272117   Mean   : 1.556   Mean   : 1.517
##  3rd Qu.:42.00   3rd Qu.:420377   3rd Qu.: 2.000   3rd Qu.: 2.000
##  Max.   :56.00   Max.   :560100   Max.   :64.000   Max.   :64.000
##     PVH_INVL             PEDS           PERNOTMVIT         PERMVIT
##  Min.   : 0.00000   Min.   : 0.0000   Min.   : 0.0000   Min.   :  0.000
##  1st Qu.: 0.00000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:  1.000
##  Median : 0.00000   Median : 0.0000   Median : 0.0000   Median :  2.000
##  Mean   : 0.03969   Mean   : 0.2186   Mean   : 0.2289   Mean   :  2.254
##  3rd Qu.: 0.00000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:  3.000
##  Max.   :11.00000   Max.   :11.0000   Max.   :11.0000   Max.   :120.000
##     PERSONS            COUNTY           CITY           DAY
##  Min.   :  0.000   Min.   :  0.0   Min.   :   0   Min.   : 1.00
##  1st Qu.:  1.000   1st Qu.: 31.0   1st Qu.:   0   1st Qu.: 8.00
##  Median :  2.000   Median : 71.0   Median :   0   Median :16.00
##  Mean   :  2.264   Mean   : 91.4   Mean   :1227   Mean   :15.75
##  3rd Qu.:  3.000   3rd Qu.:115.0   3rd Qu.:1980   3rd Qu.:23.00
##  Max.   :120.000   Max.   :999.0   Max.   :9999   Max.   :31.00
##      MONTH             YEAR         DAY_WEEK          HOUR
##  Min.   : 1.000   Min.   :2016   Min.   :1.000   Min.   : 0.00
##  1st Qu.: 4.000   1st Qu.:2016   1st Qu.:2.000   1st Qu.: 7.00
##  Median : 7.000   Median :2016   Median :4.000   Median :14.00
##  Mean   : 6.745   Mean   :2016   Mean   :4.135   Mean   :13.41
##  3rd Qu.:10.000   3rd Qu.:2016   3rd Qu.:6.000   3rd Qu.:19.00
##  Max.   :12.000   Max.   :2016   Max.   :7.000   Max.   :99.00
##      MINUTE           NHS            RUR_URB         FUNC_SYS
##  Min.   : 0.00   Min.   :0.0000   Min.   :1.000   Min.   : 1.000
##  1st Qu.:14.00   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 3.000
##  Median :30.00   Median :0.0000   Median :2.000   Median : 4.000
##  Mean   :29.09   Mean   :0.4151   Mean   :1.708   Mean   : 7.107
##  3rd Qu.:45.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.: 5.000
##  Max.   :99.00   Max.   :9.0000   Max.   :9.000   Max.   :99.000
##     RD_OWNER         ROUTE         TWAY_ID            TWAY_ID2
##  Min.   : 1.00   Min.   :1.000   Length:34439       Length:34439
##  1st Qu.: 1.00   1st Qu.:2.000   Class :character   Class :character
##  Median : 1.00   Median :3.000   Mode  :character   Mode  :character
##  Mean   :17.63   Mean   :3.593
##  3rd Qu.: 4.00   3rd Qu.:5.000
##  Max.   :99.00   Max.   :9.000
##      MILEPT         LATITUDE         LONGITUD           SP_JUR
##  Min.   :    0   Min.   : 19.10   Min.   :-174.20   Min.   :0.00000
##  1st Qu.:    0   1st Qu.: 33.02   1st Qu.: -97.95   1st Qu.:0.00000
##  Median :   58   Median : 36.25   Median : -87.78   Median :0.00000
##  Mean   :15195   Mean   : 36.91   Mean   : -85.37   Mean   :0.04504
##  3rd Qu.:  417   3rd Qu.: 40.55   3rd Qu.: -81.48   3rd Qu.:0.00000
```

```
##  Max.   :99999   Max.   :100.00   Max.   :1000.00   Max.   :9.00000
##     HARM_EV         MAN_COLL         RELJCT1           RELJCT2
##  Min.   : 1.00   Min.   : 0.000   Min.   :0.00000   Min.   : 1.000
##  1st Qu.: 8.00   1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.: 1.000
##  Median :12.00   Median : 0.000   Median :0.00000   Median : 1.000
##  Mean   :17.97   Mean   : 1.865   Mean   :0.04663   Mean   : 2.217
##  3rd Qu.:30.00   3rd Qu.: 2.000   3rd Qu.:0.00000   3rd Qu.: 2.000
##  Max.   :99.00   Max.   :99.000   Max.   :9.00000   Max.   :99.000
##     TYP_INT         WRK_ZONE         REL_ROAD         LGT_COND
##  Min.   : 1.00   Min.   :0.0000   Min.   : 1.000   Min.   :1.000
##  1st Qu.: 1.00   1st Qu.:0.0000   1st Qu.: 1.000   1st Qu.:1.000
##  Median : 1.00   Median :0.0000   Median : 1.000   Median :2.000
##  Mean   : 1.61   Mean   :0.0367   Mean   : 2.401   Mean   :1.899
##  3rd Qu.: 1.00   3rd Qu.:0.0000   3rd Qu.: 4.000   3rd Qu.:2.000
##  Max.   :99.00   Max.   :4.0000   Max.   :99.000   Max.   :9.000
##     WEATHER1         WEATHER2           WEATHER          SCH_BUS
##  Min.   : 1.000   Min.   : 0.00000   Min.   : 1.000   Min.   :0.000000
##  1st Qu.: 1.000   1st Qu.: 0.00000   1st Qu.: 1.000   1st Qu.:0.000000
##  Median : 1.000   Median : 0.00000   Median : 1.000   Median :0.000000
##  Mean   : 7.604   Mean   : 0.08246   Mean   : 7.584   Mean   :0.003049
##  3rd Qu.: 2.000   3rd Qu.: 0.00000   3rd Qu.: 2.000   3rd Qu.:0.000000
##  Max.   :99.000   Max.   :99.00000   Max.   :99.000   Max.   :1.000000
##      RAIL              NOT_HOUR        NOT_MIN         ARR_HOUR
##  Length:34439       Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
##  Class :character   1st Qu.:14.00   1st Qu.:30.00   1st Qu.:15.00
##  Mode  :character   Median :88.00   Median :88.00   Median :99.00
##                     Mean   :56.63   Mean   :64.71   Mean   :58.85
##                     3rd Qu.:99.00   3rd Qu.:99.00   3rd Qu.:99.00
##                     Max.   :99.00   Max.   :99.00   Max.   :99.00
##     ARR_MIN         HOSP_HR         HOSP_MN           CF1
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:32.00   1st Qu.:22.00   1st Qu.:55.00   1st Qu.: 0.000
##  Median :98.00   Median :88.00   Median :88.00   Median : 0.000
##  Mean   :66.52   Mean   :72.08   Mean   :76.37   Mean   : 1.574
##  3rd Qu.:99.00   3rd Qu.:99.00   3rd Qu.:99.00   3rd Qu.: 0.000
##  Max.   :99.00   Max.   :99.00   Max.   :99.00   Max.   :99.000
##       CF2               CF3              FATALS         DRUNK_DR
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:1.000   1st Qu.:0.0000
##  Median : 0.0000   Median : 0.0000   Median :1.000   Median :0.0000
##  Mean   : 0.5913   Mean   : 0.4926   Mean   :1.088   Mean   :0.2604
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :99.0000   Max.   :99.0000   Max.   :9.000   Max.   :3.0000
```

Merge YEAR, MONTH, DAY, HOUR, MINUTE into Timestamps

```
accs$TIMESTAMP <- with(accs, ISOdatetime(YEAR, MONTH, DAY, HOUR, MINUTE, sec = 0, tz
="""))
```

Remove TWAY_ID2 attribute, the only variable with missing values: <sum(is.na (accs$TWAY_ID2))> and TWAY_ID, not appropiate for the research project

Remove YEAR, MONTH, DAY, HOUR, MINUTE attributes - it's been merged into Timestamps 12:14

Remove WEATHER1, WEATHER2 attributes, are duplicate of the original WEATHER

Remove RAIL attribute, no relevant to the research

accs2016 <- accs[,-c (1:2,10:11,12:14,16:17,23:24,37:38,41)]

```
accs2016 <- accs[,-c(1:2,10:14,16:17,23:24,37:38,41,53)]
```

Values of the FATALS attributes

```
table(accs2016$FATALS)
```

```
##
##     1     2     3     4     5     6     9
## 31984  2033   315    80    19     7     1
```

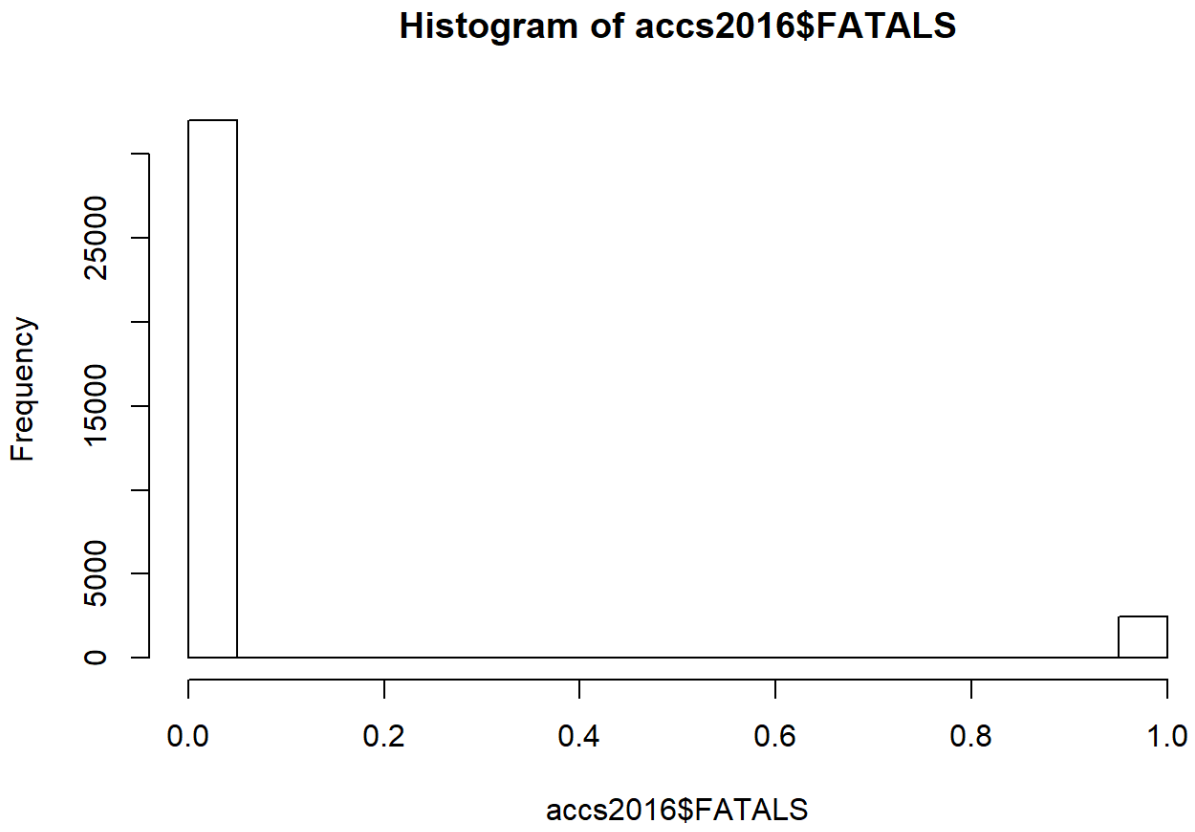
Reduce the levels of values for FATALS to two binary levels:: single death (1) = 0, and multiple deaths (2-9) = 1.

```
accs2016$FATALS <- mapvalues(accs2016$FATALS, from = c("1", "2", "3", "4", "5", "6",
"9"), to = c(0,1,1,1,1,1,1))

table(accs2016$FATALS)
```

```
##
##     0     1
## 31984  2455
```

Graph the frequency distribution of FATALS variable

```
hist(accs2016$FATALS, freq = T)
```




```
accs_LM <- lm(formula = FATALS ~ ., data = accs2016)

summary(accs_LM)
```

```
##
## Call:
## lm(formula = FATALS ~ ., data = accs2016)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.33485 -0.08415 -0.04826 -0.02151  1.03579
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.518e-02  1.532e-02   2.296 0.021670 *
## VE_TOTAL     5.260e-03  4.749e-03   1.108 0.267978
## VE_FORMS    -2.987e-02  5.283e-03  -5.653 1.59e-08 ***
## PVH_INVL            NA         NA      NA       NA
## PEDS        -3.453e-02  1.003e-02  -3.443 0.000576 ***
## PERNOTMVIT   3.496e-02  9.447e-03   3.701 0.000215 ***
## PERMVIT      4.120e-02  9.845e-04  41.850  < 2e-16 ***
## PERSONS             NA         NA      NA       NA
## DAY_WEEK     9.611e-04  6.391e-04   1.504 0.132642
## NHS          5.044e-03  2.185e-03   2.309 0.020947 *
## RUR_URB     -1.753e-02  2.376e-03  -7.378 1.64e-13 ***
## FUNC_SYS     7.811e-04  1.682e-04   4.643 3.45e-06 ***
## RD_OWNER     1.498e-04  4.421e-05   3.388 0.000705 ***
## ROUTE       -4.455e-03  8.554e-04  -5.208 1.92e-07 ***
## MILEPT      -6.028e-08  4.174e-08  -1.444 0.148709
## LATITUDE    -1.821e-04  2.924e-04  -0.623 0.533371
## LONGITUD    -7.848e-06  2.333e-05  -0.336 0.736542
## SP_JUR       8.039e-03  2.934e-03   2.740 0.006146 **
## HARM_EV      1.502e-04  1.077e-04   1.395 0.162885
## MAN_COLL     9.436e-04  2.470e-04   3.820 0.000134 ***
## RELJCT1      2.528e-03  5.797e-03   0.436 0.662803
## RELJCT2     -6.497e-04  3.983e-04  -1.631 0.102871
## TYP_INT      4.961e-04  3.867e-04   1.283 0.199509
## WRK_ZONE     6.063e-04  4.256e-03   0.142 0.886715
## REL_ROAD    -2.265e-04  3.545e-04  -0.639 0.522964
## LGT_COND     1.559e-03  1.227e-03   1.270 0.203967
## WEATHER     -1.561e-04  6.372e-05  -2.450 0.014289 *
## SCH_BUS     -4.925e-02  2.426e-02  -2.030 0.042354 *
## NOT_HOUR     1.154e-04  1.132e-04   1.020 0.307910
## NOT_MIN     -1.382e-04  1.112e-04  -1.243 0.213777
## ARR_HOUR    -1.461e-04  1.167e-04  -1.252 0.210563
## ARR_MIN      1.189e-04  1.144e-04   1.039 0.298975
## HOSP_HR      1.075e-04  1.222e-04   0.879 0.379346
## HOSP_MN     -1.012e-05  1.401e-04  -0.072 0.942421
## CF1          1.802e-04  3.184e-04   0.566 0.571429
## CF2         -3.246e-03  9.389e-04  -3.457 0.000546 ***
## CF3          2.947e-03  9.449e-04   3.119 0.001816 **
## DRUNK_DR     3.641e-02  3.059e-03  11.902  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2475 on 34403 degrees of freedom
## Multiple R-squared:  0.07563,    Adjusted R-squared:  0.07469
## F-statistic: 80.42 on 35 and 34403 DF,  p-value: < 2.2e-16
```

The selected variables are:

*** VE_FORMS, PEDS, PERNOTMVIT, PERMVIT, RUR_URB, FUNC_SYS, RD_OWNER, ROUTE, MAN_COLL, CF2, DRUNK_DR,

** SP_JUR, CF3,

* NHS, WEATHER, SCH_BUS

USing the selected features

```
accs_features <- select(accs2016, VE_FORMS, PEDS, PERNOTMVIT, PERMVIT, RUR_URB, FUNC_S
YS, RD_OWNER, ROUTE, MAN_COLL, CF2, DRUNK_DR,  SP_JUR, CF3, WEATHER, FATALS)
```

```
cor(accs_features)
```

```
##                VE_FORMS         PEDS   PERNOTMVIT      PERMVIT
## VE_FORMS    1.000000000 -0.222932739 -0.217231350  0.685345598
## PEDS       -0.222932739  1.000000000  0.954730435 -0.193597709
## PERNOTMVIT -0.217231350  0.954730435  1.000000000 -0.187598332
## PERMVIT     0.685345598 -0.193597709 -0.187598332  1.000000000
## RUR_URB    -0.020325056  0.128033453  0.123267835 -0.025777296
## FUNC_SYS   -0.057862331  0.040194646  0.035854294 -0.035928826
## RD_OWNER   -0.007164953  0.011035772  0.008747208 -0.002472209
## ROUTE      -0.120867886  0.121236490  0.102551827 -0.104213436
## MAN_COLL    0.269518577 -0.129497184 -0.126834814  0.178105772
## CF2         0.013595577  0.017238806  0.022172992  0.007391173
## DRUNK_DR   -0.043790023 -0.176780245 -0.163780617 -0.026692382
## SP_JUR     -0.008622483  0.005808173  0.003661580  0.009350063
## CF3        -0.009399729  0.006574471  0.009521669 -0.006238776
## WEATHER    -0.020521497 -0.004251603 -0.008065947 -0.024258561
## FATALS      0.129410281 -0.067001868 -0.057494445  0.249936364
##                   RUR_URB     FUNC_SYS     RD_OWNER       ROUTE
## VE_FORMS   -0.0203250562 -0.057862331 -0.007164953 -0.12086789
## PEDS        0.1280334529  0.040194646  0.011035772  0.12123649
## PERNOTMVIT  0.1232678350  0.035854294  0.008747208  0.10255183
## PERMVIT    -0.0257772958 -0.035928826 -0.002472209 -0.10421344
## RUR_URB     1.0000000000  0.872986880  0.354289194  0.26603388
## FUNC_SYS    0.8729868800  1.000000000  0.398863273  0.26589673
## RD_OWNER    0.3542891942  0.398863273  1.000000000  0.28823733
## ROUTE       0.2660338769  0.265896731  0.288237331  1.00000000
## MAN_COLL    0.0035062321 -0.008949828  0.015904144  0.02545610
## CF2        -0.0002109704 -0.012470834  0.105921565  0.01184705
## DRUNK_DR    0.0021416577  0.017150637  0.006205862  0.03009841
## SP_JUR      0.1652026689  0.191347407  0.090372675  0.10359414
## CF3        -0.0005897542 -0.009921756  0.110938097  0.02345277
## WEATHER    -0.0311707891 -0.005821064 -0.042347989  0.01198376
## FATALS     -0.0403655458 -0.022551512  0.004527961 -0.05611442
##                  MAN_COLL           CF2      DRUNK_DR       SP_JUR
## VE_FORMS    0.2695185775  0.0135955773 -0.0437900232 -0.008622483
## PEDS       -0.1294971837  0.0172388058 -0.1767802448  0.005808173
## PERNOTMVIT -0.1268348139  0.0221729917 -0.1637806175  0.003661580
## PERMVIT     0.1781057720  0.0073911728 -0.0266923821  0.009350063
## RUR_URB     0.0035062321 -0.0002109704  0.0021416577  0.165202669
## FUNC_SYS   -0.0089498278 -0.0124708343  0.0171506370  0.191347407
## RD_OWNER    0.0159041437  0.1059215648  0.0062058622  0.090372675
## ROUTE       0.0254561002  0.0118470454  0.0300984053  0.103594142
## MAN_COLL    1.0000000000 -0.0004204331 -0.0291157001  0.046382793
## CF2        -0.0004204331  1.0000000000 -0.0014379643  0.002784372
## DRUNK_DR   -0.0291157001 -0.0014379643  1.0000000000  0.014701189
## SP_JUR      0.0463827928  0.0027843722  0.0147011891  1.000000000
## CF3        -0.0022922252  0.9778521808 -0.0007606222  0.002694350
## WEATHER     0.0602953242 -0.0056301285  0.0184236791  0.027981894
## FATALS      0.0477157142 -0.0031700924  0.0613891359  0.017871154
```

```
##                      CF3      WEATHER       FATALS
## VE_FORMS   -0.0093997289 -0.020521497  0.129410281
## PEDS        0.0065744707 -0.004251603 -0.067001868
## PERNOTMVIT  0.0095216693 -0.008065947 -0.057494445
## PERMVIT    -0.0062387760 -0.024258561  0.249936364
## RUR_URB    -0.0005897542 -0.031170789 -0.040365546
## FUNC_SYS   -0.0099217557 -0.005821064 -0.022551512
## RD_OWNER    0.1109380972 -0.042347989  0.004527961
## ROUTE       0.0234527652  0.011983759 -0.056114425
## MAN_COLL   -0.0022922252  0.060295324  0.047715714
## CF2         0.9778521808 -0.005630128 -0.003170092
## DRUNK_DR   -0.0007606222  0.018423679  0.061389136
## SP_JUR      0.0026943500  0.027981894  0.017871154
## CF3         1.0000000000 -0.003598776 -0.002297333
## WEATHER    -0.0035987764  1.000000000 -0.015402083
## FATALS     -0.0022973334 -0.015402083  1.000000000
```

What is the correlation between the attributes other than FATALS variable?

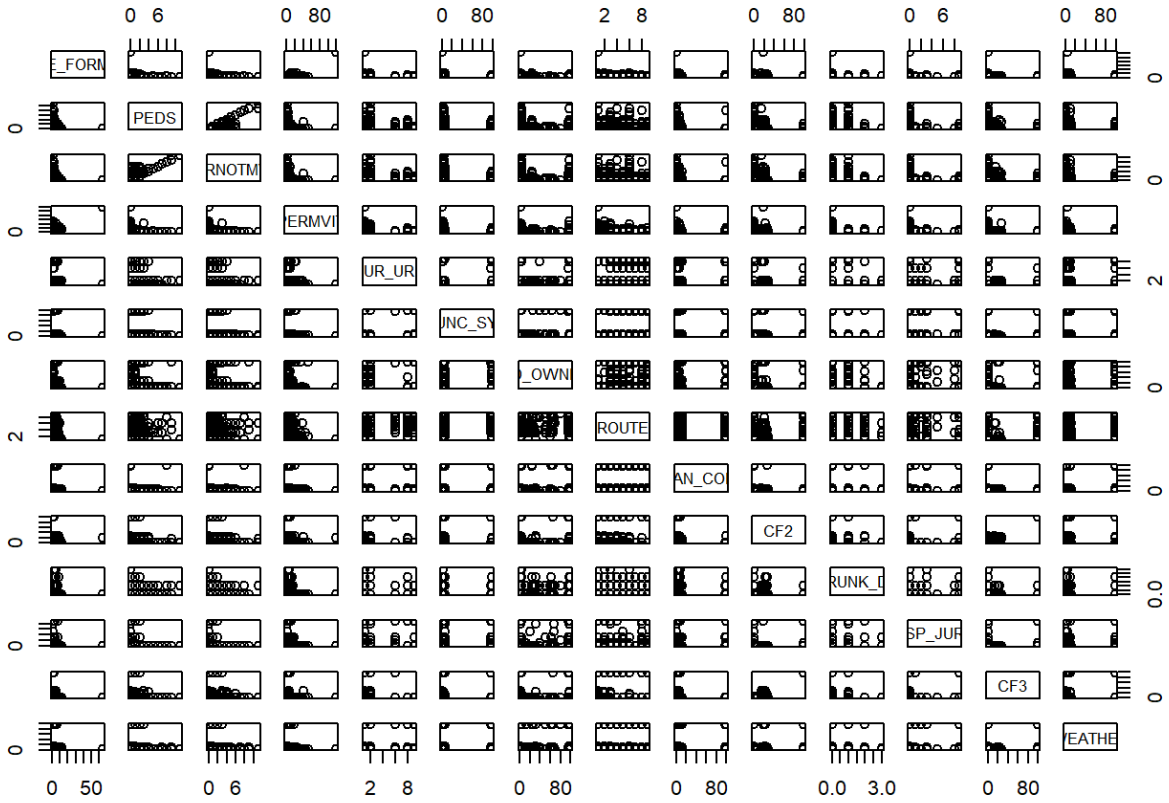
Remove FATALS

```
library(corrplot)

accs2016_f <- accs_features[,-c(15)]

cor(accs2016_f)
```

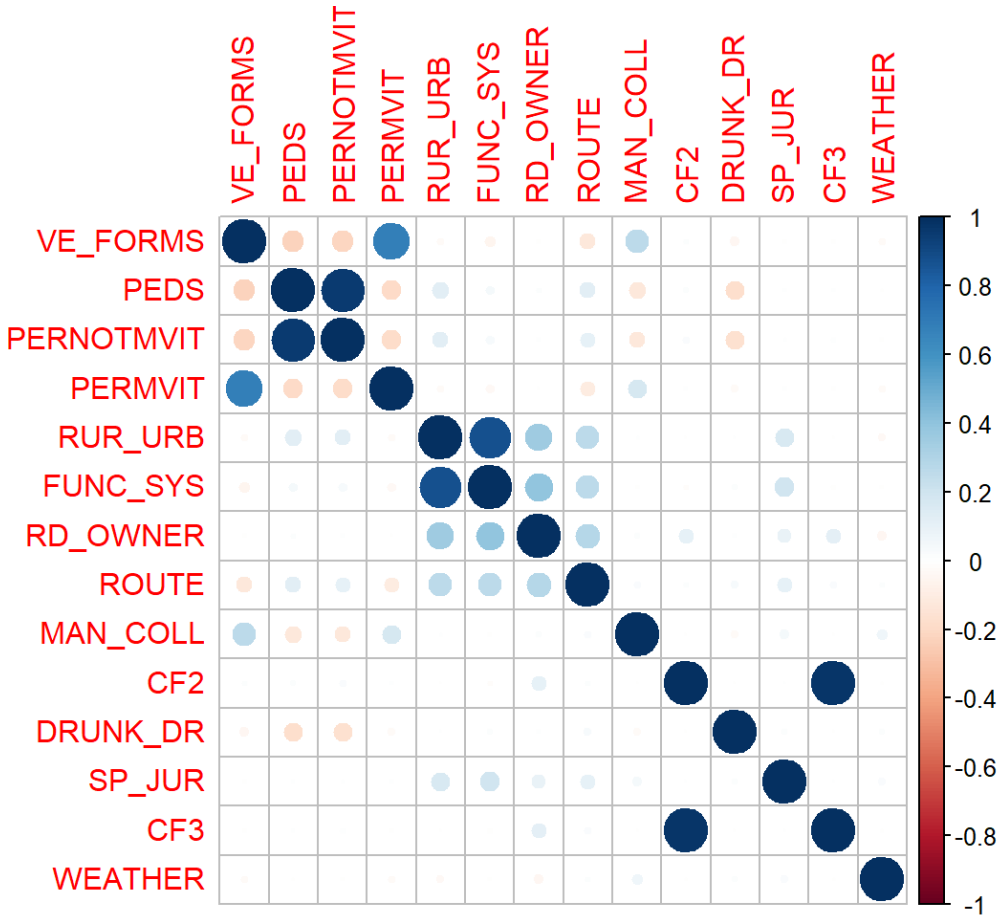
```
##                VE_FORMS         PEDS   PERNOTMVIT      PERMVIT
## VE_FORMS    1.000000000 -0.222932739 -0.217231350  0.685345598
## PEDS       -0.222932739  1.000000000  0.954730435 -0.193597709
## PERNOTMVIT -0.217231350  0.954730435  1.000000000 -0.187598332
## PERMVIT     0.685345598 -0.193597709 -0.187598332  1.000000000
## RUR_URB    -0.020325056  0.128033453  0.123267835 -0.025777296
## FUNC_SYS   -0.057862331  0.040194646  0.035854294 -0.035928826
## RD_OWNER   -0.007164953  0.011035772  0.008747208 -0.002472209
## ROUTE      -0.120867886  0.121236490  0.102551827 -0.104213436
## MAN_COLL    0.269518577 -0.129497184 -0.126834814  0.178105772
## CF2         0.013595577  0.017238806  0.022172992  0.007391173
## DRUNK_DR   -0.043790023 -0.176780245 -0.163780617 -0.026692382
## SP_JUR     -0.008622483  0.005808173  0.003661580  0.009350063
## CF3        -0.009399729  0.006574471  0.009521669 -0.006238776
## WEATHER    -0.020521497 -0.004251603 -0.008065947 -0.024258561
##                   RUR_URB     FUNC_SYS     RD_OWNER       ROUTE
## VE_FORMS   -0.0203250562 -0.057862331 -0.007164953 -0.12086789
## PEDS        0.1280334529  0.040194646  0.011035772  0.12123649
## PERNOTMVIT  0.1232678350  0.035854294  0.008747208  0.10255183
## PERMVIT    -0.0257772958 -0.035928826 -0.002472209 -0.10421344
## RUR_URB     1.0000000000  0.872986880  0.354289194  0.26603388
## FUNC_SYS    0.8729868800  1.000000000  0.398863273  0.26589673
## RD_OWNER    0.3542891942  0.398863273  1.000000000  0.28823733
## ROUTE       0.2660338769  0.265896731  0.288237331  1.00000000
## MAN_COLL    0.0035062321 -0.008949828  0.015904144  0.02545610
## CF2        -0.0002109704 -0.012470834  0.105921565  0.01184705
## DRUNK_DR    0.0021416577  0.017150637  0.006205862  0.03009841
## SP_JUR      0.1652026689  0.191347407  0.090372675  0.10359414
## CF3        -0.0005897542 -0.009921756  0.110938097  0.02345277
## WEATHER    -0.0311707891 -0.005821064 -0.042347989  0.01198376
##                 MAN_COLL           CF2      DRUNK_DR       SP_JUR
## VE_FORMS    0.2695185775  0.0135955773 -0.0437900232 -0.008622483
## PEDS       -0.1294971837  0.0172388058 -0.1767802448  0.005808173
## PERNOTMVIT -0.1268348139  0.0221729917 -0.1637806175  0.003661580
## PERMVIT     0.1781057720  0.0073911728 -0.0266923821  0.009350063
## RUR_URB     0.0035062321 -0.0002109704  0.0021416577  0.165202669
## FUNC_SYS   -0.0089498278 -0.0124708343  0.0171506370  0.191347407
## RD_OWNER    0.0159041437  0.1059215648  0.0062058622  0.090372675
## ROUTE       0.0254561002  0.0118470454  0.0300984053  0.103594142
## MAN_COLL    1.0000000000 -0.0004204331 -0.0291157001  0.046382793
## CF2        -0.0004204331  1.0000000000 -0.0014379643  0.002784372
## DRUNK_DR   -0.0291157001 -0.0014379643  1.0000000000  0.014701189
## SP_JUR      0.0463827928  0.0027843722  0.0147011891  1.000000000
## CF3        -0.0022922252  0.9778521808 -0.0007606222  0.002694350
## WEATHER     0.0602953242 -0.0056301285  0.0184236791  0.027981894
##                     CF3      WEATHER
## VE_FORMS   -0.0093997289 -0.020521497
## PEDS        0.0065744707 -0.004251603
```

```
## PERNOTMVIT  0.0095216693 -0.008065947
## PERMVIT    -0.0062387760 -0.024258561
## RUR_URB    -0.0005897542 -0.031170789
## FUNC_SYS   -0.0099217557 -0.005821064
## RD_OWNER    0.1109380972 -0.042347989
## ROUTE       0.0234527652  0.011983759
## MAN_COLL   -0.0022922252  0.060295324
## CF2         0.9778521808 -0.005630128
## DRUNK_DR   -0.0007606222  0.018423679
## SP_JUR      0.0026943500  0.027981894
## CF3         1.0000000000 -0.003598776
## WEATHER    -0.0035987764  1.000000000
```

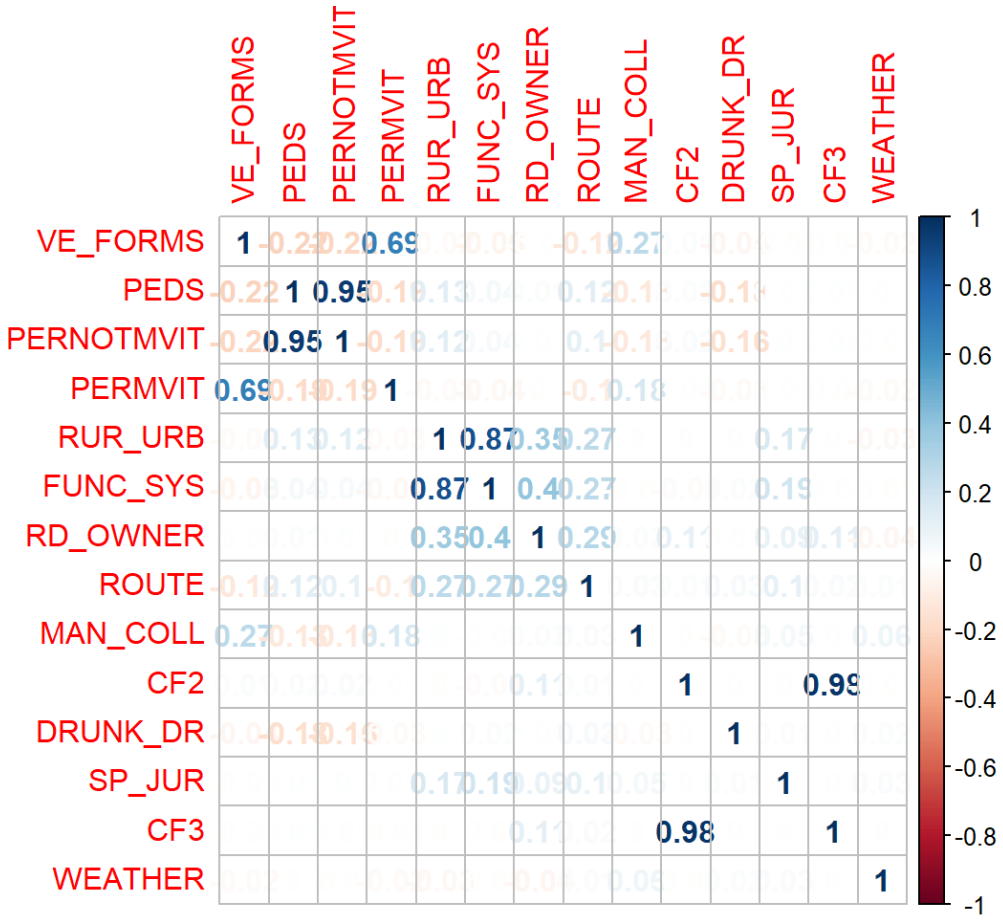
```
plot(accs2016_f)
```



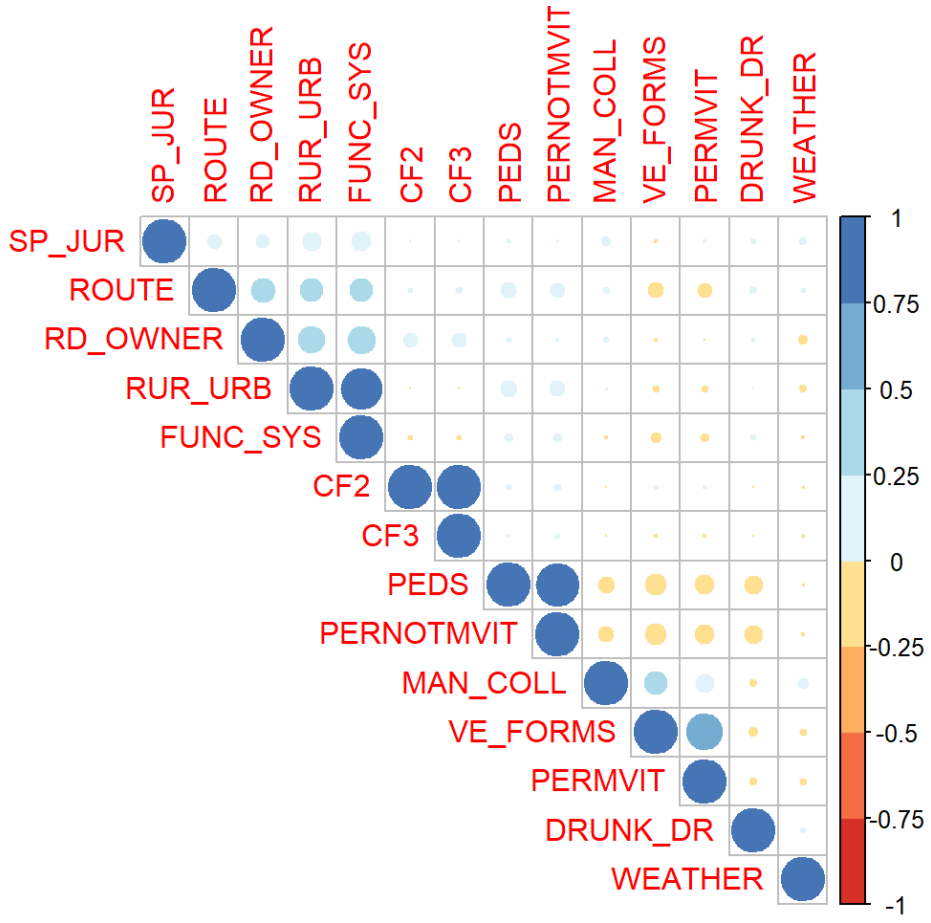

```
corrplot(cor(accs2016_f))
```




```
corrplot(cor(accs2016_f), method = c("number"))
```




```
library(RColorBrewer)

M <-cor(accs2016_f)
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))
```




```
mcor<-round(cor(accs2016_f),2)

mcor
```

```
##            VE_FORMS  PEDS PERNOTMVIT PERMVIT RUR_URB FUNC_SYS RD_OWNER
## VE_FORMS       1.00 -0.22      -0.22    0.69   -0.02    -0.06    -0.01
## PEDS          -0.22  1.00       0.95   -0.19    0.13     0.04     0.01
## PERNOTMVIT    -0.22  0.95       1.00   -0.19    0.12     0.04     0.01
## PERMVIT        0.69 -0.19      -0.19    1.00   -0.03    -0.04     0.00
## RUR_URB       -0.02  0.13       0.12   -0.03    1.00     0.87     0.35
## FUNC_SYS      -0.06  0.04       0.04   -0.04    0.87     1.00     0.40
## RD_OWNER      -0.01  0.01       0.01    0.00    0.35     0.40     1.00
## ROUTE         -0.12  0.12       0.10   -0.10    0.27     0.27     0.29
## MAN_COLL       0.27 -0.13      -0.13    0.18    0.00    -0.01     0.02
## CF2            0.01  0.02       0.02    0.01    0.00    -0.01     0.11
## DRUNK_DR      -0.04 -0.18      -0.16   -0.03    0.00     0.02     0.01
## SP_JUR        -0.01  0.01       0.00    0.01    0.17     0.19     0.09
## CF3           -0.01  0.01       0.01   -0.01    0.00    -0.01     0.11
## WEATHER       -0.02  0.00      -0.01   -0.02   -0.03    -0.01    -0.04
##             ROUTE MAN_COLL   CF2 DRUNK_DR SP_JUR   CF3 WEATHER
## VE_FORMS    -0.12     0.27  0.01    -0.04  -0.01 -0.01   -0.02
## PEDS         0.12    -0.13  0.02    -0.18   0.01  0.01    0.00
## PERNOTMVIT   0.10    -0.13  0.02    -0.16   0.00  0.01   -0.01
## PERMVIT     -0.10     0.18  0.01    -0.03   0.01 -0.01   -0.02
## RUR_URB      0.27     0.00  0.00     0.00   0.17  0.00   -0.03
## FUNC_SYS     0.27    -0.01 -0.01     0.02   0.19 -0.01   -0.01
## RD_OWNER     0.29     0.02  0.11     0.01   0.09  0.11   -0.04
## ROUTE        1.00     0.03  0.01     0.03   0.10  0.02    0.01
## MAN_COLL     0.03     1.00  0.00    -0.03   0.05  0.00    0.06
## CF2          0.01     0.00  1.00     0.00   0.00  0.98   -0.01
## DRUNK_DR     0.03    -0.03  0.00     1.00   0.01  0.00    0.02
## SP_JUR       0.10     0.05  0.00     0.01   1.00  0.00    0.03
## CF3          0.02     0.00  0.98     0.00   0.00  1.00    0.00
## WEATHER      0.01     0.06 -0.01     0.02   0.03  0.00    1.00
```

```
# cor(accs2016_f, method="pearson")
```

Normalize the data set.

I use the code below for normalization

```
normalize <- function(x) {
               return ((x - min(x)) / (max(x) - min(x))) }
```

Once we run this code, we are required to normalize the numeric features in the data set. Instead of normalizing each of the ? individual variables we use:

```
accident_n <- as.data.frame(lapply(accs_features[1:14], normalize))
accident_n$FATALS <- accs_features$FATALS

summary(accident_n)
```

```
##     VE_FORMS            PEDS           PERNOTMVIT         PERMVIT
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.008333
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.016667
##  Mean   :0.00820   Mean   :0.01988   Mean   :0.02081   Mean   :0.018780
##  3rd Qu.:0.01587   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.025000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000000
##     RUR_URB           FUNC_SYS          RD_OWNER           ROUTE
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.02041   1st Qu.:0.00000   1st Qu.:0.1250
##  Median :0.12500   Median :0.03061   Median :0.00000   Median :0.2500
##  Mean   :0.08854   Mean   :0.06231   Mean   :0.16970   Mean   :0.3242
##  3rd Qu.:0.12500   3rd Qu.:0.04082   3rd Qu.:0.03061   3rd Qu.:0.5000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##     MAN_COLL            CF2              DRUNK_DR           SP_JUR
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.00000   Median :0.000000   Median :0.00000   Median :0.000000
##  Mean   :0.01884   Mean   :0.005973   Mean   :0.08681   Mean   :0.005004
##  3rd Qu.:0.02020   3rd Qu.:0.000000   3rd Qu.:0.33333   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.00000   Max.   :1.000000
##       CF3              WEATHER            FATALS
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.000000   Median :0.00000   Median :0.00000
##  Mean   :0.004976   Mean   :0.06719   Mean   :0.07129
##  3rd Qu.:0.000000   3rd Qu.:0.01020   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.00000   Max.   :1.00000
```

Divide the data to training and testing groups.

```
acc_trainindex <- sample(1:nrow(accident_n), 0.7 * nrow(accident_n))

acc_trainset <- accident_n[acc_trainindex,]
acc_testset  <- accident_n[-acc_trainindex,]

str(acc_trainset)
```

```
## 'data.frame':    24107 obs. of  15 variables:
##  $ VE_FORMS  : num  0.0317 0.0159 0 0 0 ...
##  $ PEDS      : num  0 0 0 0 0.0909 ...
##  $ PERNOTMVIT: num  0 0 0 0 0.0909 ...
##  $ PERMVIT   : num  0.025 0.025 0.01667 0.00833 0.00833 ...
##  $ RUR_URB   : num  0.125 0 0 0.125 0 0 0.125 0.125 0.125 0.125 ...
##  $ FUNC_SYS  : num  0 0.0408 0.0408 0.0306 0.0612 ...
##  $ RD_OWNER  : num  0.3061 0 0.9898 0.9898 0.0204 ...
##  $ ROUTE     : num  0 0.125 0.875 0.625 0.875 0.25 0 0.25 0 0.125 ...
##  $ MAN_COLL  : num  0 0.0202 0 0 0 ...
##  $ CF2       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DRUNK_DR  : num  0 0 0.333 0.333 0 ...
##  $ SP_JUR    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CF3       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ WEATHER   : num  0 0.0102 0.0918 0.0918 0.0102 ...
##  $ FATALS    : num  0 0 1 0 0 0 0 0 0 0 ...
```

```
str(acc_testset)
```

```
## 'data.frame':    10332 obs. of  15 variables:
##  $ VE_FORMS  : num  0 0 0 0.0159 0 ...
##  $ PEDS      : num  0 0 0 0 0 ...
##  $ PERNOTMVIT: num  0 0 0 0 0 ...
##  $ PERMVIT   : num  0.00833 0.00833 0.00833 0.03333 0.01667 ...
##  $ RUR_URB   : num  0.125 0 0 0 0 0.125 0 0 0 0 ...
##  $ FUNC_SYS  : num  0 0 0.0306 0.0204 0.0408 ...
##  $ RD_OWNER  : num  0 0 0 0 0.0102 ...
##  $ ROUTE     : num  0 0 0.25 0.25 0.375 0.625 0.375 0 0.375 0.125 ...
##  $ MAN_COLL  : num  0 0 0 0.0202 0 ...
##  $ CF2       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DRUNK_DR  : num  0.333 0.333 0 0 0.333 ...
##  $ SP_JUR    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CF3       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ WEATHER   : num  0 0 0.0918 0 0 ...
##  $ FATALS    : num  0 0 0 1 1 0 0 0 0 0 ...
```

Using the KNN algorithm to predict FATALITIES using its attributes.

```
library(class)
library(gmodels)

##Let's remove the response variables

acc_trainset_new <- acc_trainset[-15]
acc_testset_new <- acc_testset[-15]


#Let's store labels from train and test datasets

acc_trainlabels <- acc_trainset$FATALS
acc_testlabels <- acc_testset$FATALS


#For k=3, let's make our prediction on the test set.
acc_pred <- knn(train = acc_trainset_new, test = acc_testset_new, cl = acc_trainlabel
s, k=3)
```

Evaluate the model performance.

```
CrossTable(x=acc_testlabels, y=acc_pred, prop.chisq = F)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  10332
##
##
##                | acc_pred
## acc_testlabels |         0 |         1 | Row Total |
## ---------------|-----------|-----------|-----------|
##              0 |      9442 |       158 |      9600 |
##                |     0.984 |     0.016 |     0.929 |
##                |     0.932 |     0.790 |           |
##                |     0.914 |     0.015 |           |
## ---------------|-----------|-----------|-----------|
##              1 |       690 |        42 |       732 |
##                |     0.943 |     0.057 |     0.071 |
##                |     0.068 |     0.210 |           |
##                |     0.067 |     0.004 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |     10132 |       200 |     10332 |
##                |     0.981 |     0.019 |           |
## ---------------|-----------|-----------|-----------|
##
##
```