Data cleaning

```
#install.packages("mice") mice - Multivariate Imputation by chained Equation
#install.packages("VIM")
library(mice)

## Warning: package 'mice' was built under R version 3.6.1

## Loading required package: lattice

##
## Attaching package: 'mice'

## The following objects are masked from 'package:base':
##
##     cbind, rbind

library(VIM)

## Warning: package 'VIM' was built under R version 3.6.1

## Loading required package: colorspace

## Warning: package 'colorspace' was built under R version 3.6.1

## Loading required package: grid

## Loading required package: data.table

## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4

## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at:
https://github.com/alexkowa/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

```
data <- read.csv("file:///C:/Users/badal/Desktop/datset_/vehicleMiss.csv" ,
header = T)
head(data)
```

```
##   vehicle fm Mileage  lh     lc     mc State
## 1       1  0     863 1.1  66.30 697.23    MS
## 2       2 10    4644 2.4 233.03 119.66    CA
## 3       3 15   16330 4.2 325.08 175.46    WI
## 4       4  0      13 1.0  66.64   0.00    OR
## 5       5 13   22537 4.5 328.66 175.46    AZ
## 6       6 21   40931 3.1 205.28 175.46    FL
```

```
any(is.na(data))
```

```
## [1] TRUE
```

```
summary(data)
```

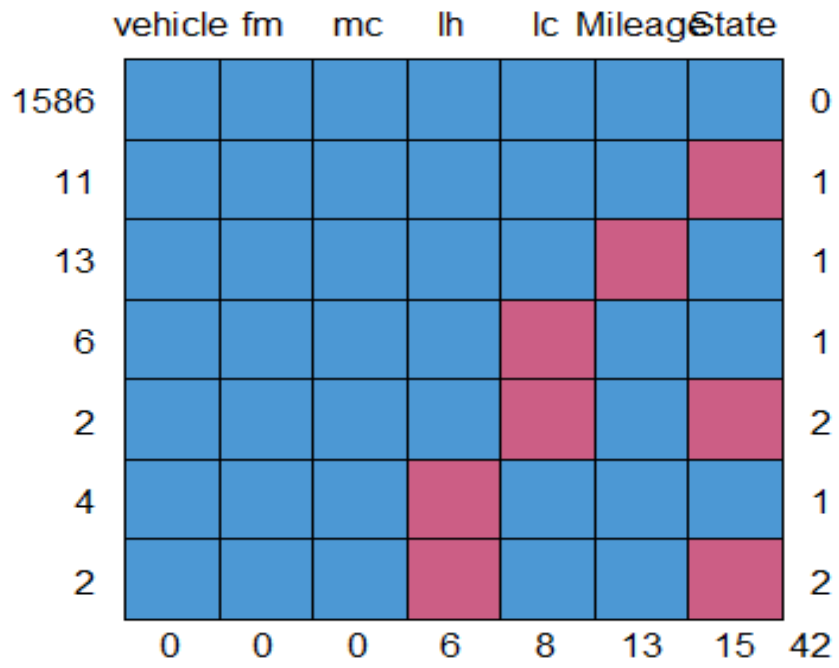
```
##     vehicle             fm            Mileage            lh
##  Min.   :   1.0   Min.   :-1.000   Min.   :    1   Min.   : 0.000
##  1st Qu.: 406.8   1st Qu.: 4.000   1st Qu.: 5778   1st Qu.: 1.500
##  Median : 812.5   Median :10.000   Median :17000   Median : 2.600
##  Mean   : 812.5   Mean   : 9.414   Mean   :20559   Mean   : 3.294
##  3rd Qu.:1218.2   3rd Qu.:14.000   3rd Qu.:30061   3rd Qu.: 4.300
##  Max.   :1624.0   Max.   :23.000   Max.   :99983   Max.   :35.200
##                                    NA's   :13      NA's   :6
##        lc               mc            State
##  Min.   :   0.0   Min.   :   0.0   TX     :290
##  1st Qu.: 106.5   1st Qu.: 119.7   CA     :199
##  Median : 195.4   Median : 119.7   FL     :167
##  Mean   : 242.8   Mean   : 179.4   GA     : 75
##  3rd Qu.: 317.8   3rd Qu.: 175.5   AZ     : 61
##  Max.   :3234.4   Max.   :3891.1   (Other):817
##  NA's   :8                         NA's   : 15
```

missing data

```
#percentage_missing_data
p <- function(x) {sum(is.na(x))/length(x)*100}
apply(data, 2,p)
```

```
##   vehicle        fm   Mileage        lh        lc        mc     State
## 0.0000000 0.0000000 0.8004926 0.3694581 0.4926108 0.0000000 0.9236453
```

```
md.pattern(data)
```



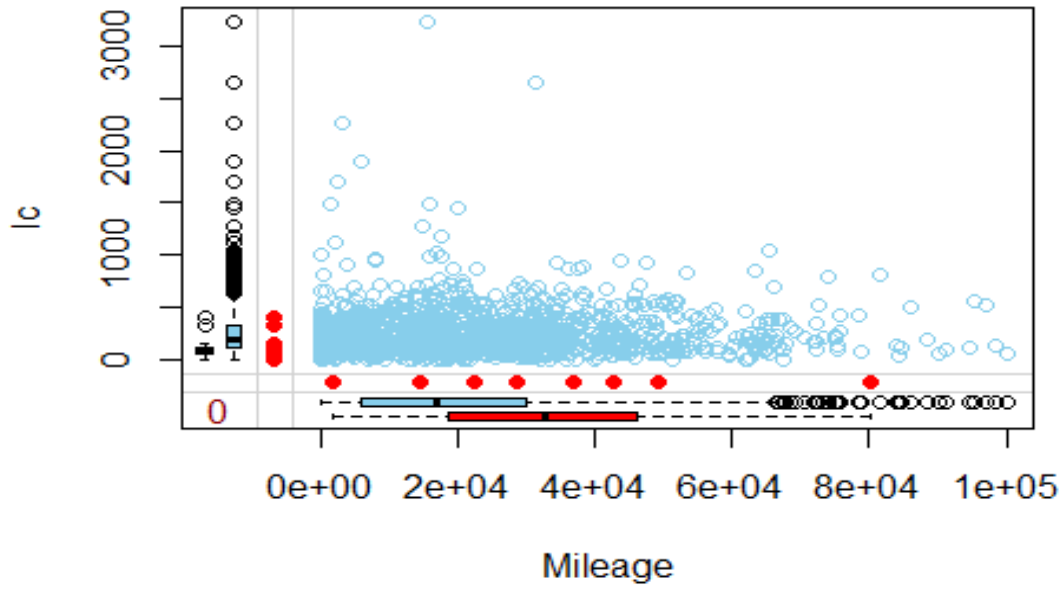

```
##      vehicle fm mc lh lc Mileage State
## 1586       1  1  1  1  1       1     1  0
## 11         1  1  1  1  1       1     0  1
## 13         1  1  1  1  1       0     1  1
## 6          1  1  1  1  0       1     1  1
## 2          1  1  1  1  0       1     0  2
## 4          1  1  1  0  1       1     1  1
## 2          1  1  1  0  1       1     0  2
##            0  0  0  6  8      13    15 42
```

```
md.pairs(data)
```

```
## $rr
##         vehicle   fm Mileage   lh   lc   mc State
## vehicle    1624 1624    1611 1618 1616 1624  1609
## fm         1624 1624    1611 1618 1616 1624  1609
## Mileage    1611 1611    1611 1605 1603 1611  1596
## lh         1618 1618    1605 1618 1610 1618  1605
## lc         1616 1616    1603 1610 1616 1616  1603
## mc         1624 1624    1611 1618 1616 1624  1609
## State      1609 1609    1596 1605 1603 1609  1609
##
## $rm
##         vehicle fm Mileage lh lc mc State
## vehicle       0  0      13  6  8  0    15
## fm            0  0      13  6  8  0    15
## Mileage       0  0       0  6  8  0    15
```

```
## lh             0  0      13  0  8  0    13
## lc             0  0      13  6  0  0    13
## mc             0  0      13  6  8  0    15
## State          0  0      13  4  6  0     0
##
## $mr
##         vehicle fm Mileage lh lc mc State
## vehicle       0  0       0  0  0  0     0
## fm            0  0       0  0  0  0     0
## Mileage      13 13       0 13 13 13    13
## lh            6  6       6  0  6  6     4
## lc            8  8       8  8  0  8     6
## mc            0  0       0  0  0  0     0
## State        15 15      15 13 13 15     0
##
## $mm
##         vehicle fm Mileage lh lc mc State
## vehicle       0  0       0  0  0  0     0
## fm            0  0       0  0  0  0     0
## Mileage       0  0      13  0  0  0     0
## lh            0  0       0  6  0  0     2
## lc            0  0       0  0  8  0     2
## mc            0  0       0  0  0  0     0
## State         0  0       0  2  2  0    15
```

```
marginplot(data[,c("Mileage", "lc")])
```



impute……..polyreg: multinominal logstic regression

```r
impute <- mice(data[,-1], m=4, seed = 123)
```

```
##
##  iter imp variable
##   1   1  Mileage  lh  lc  State
##   1   2  Mileage  lh  lc  State
##   1   3  Mileage  lh  lc  State
##   1   4  Mileage  lh  lc  State
##   2   1  Mileage  lh  lc  State
##   2   2  Mileage  lh  lc  State
##   2   3  Mileage  lh  lc  State
##   2   4  Mileage  lh  lc  State
##   3   1  Mileage  lh  lc  State
##   3   2  Mileage  lh  lc  State
##   3   3  Mileage  lh  lc  State
##   3   4  Mileage  lh  lc  State
##   4   1  Mileage  lh  lc  State
##   4   2  Mileage  lh  lc  State
##   4   3  Mileage  lh  lc  State
##   4   4  Mileage  lh  lc  State
##   5   1  Mileage  lh  lc  State
##   5   2  Mileage  lh  lc  State
##   5   3  Mileage  lh  lc  State
##   5   4  Mileage  lh  lc  State
```

```r
print(impute)
```

```
## Class: mids
## Number of multiple imputations:  4
## Imputation methods:
##        fm   Mileage        lh        lc        mc     State
##        ""     "pmm"     "pmm"     "pmm"        "" "polyreg"
## PredictorMatrix:
##         fm Mileage lh lc mc State
## fm       0       1  1  1  1     1
## Mileage  1       0  1  1  1     1
## lh       1       1  0  1  1     1
## lc       1       1  1  0  1     1
## mc       1       1  1  1  0     1
## State    1       1  1  1  1     0
```

```r
impute$imp$Mileage
```

```
##          1     2     3     4
## 19   34372 24008 10870 28782
## 20   12366   814 13800 24130
## 253   6060   221   622   713
## 254   2076 11051  3596 16068
## 255  28542 60202 88350 46519
## 256  22492 17204 35962 35963
## 861  14746  8541  9208 11932
```

```
## 862  20309   587  2863 10898
## 863  13785 51948 17201 20632
## 1568 32021 31716 22502 30104
## 1569 55137 10500  5065   487
## 1570   800   119   191     4
## 1571 11912  5082  1776 11565
```

```
data[253,]
```

```
##     vehicle fm Mileage  lh    lc     mc State
## 253     253  1      NA 1.4 89.89 119.66    FL
```

```
summary(data$Mileage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       1    5778   17000   20559   30061   99983      13
```

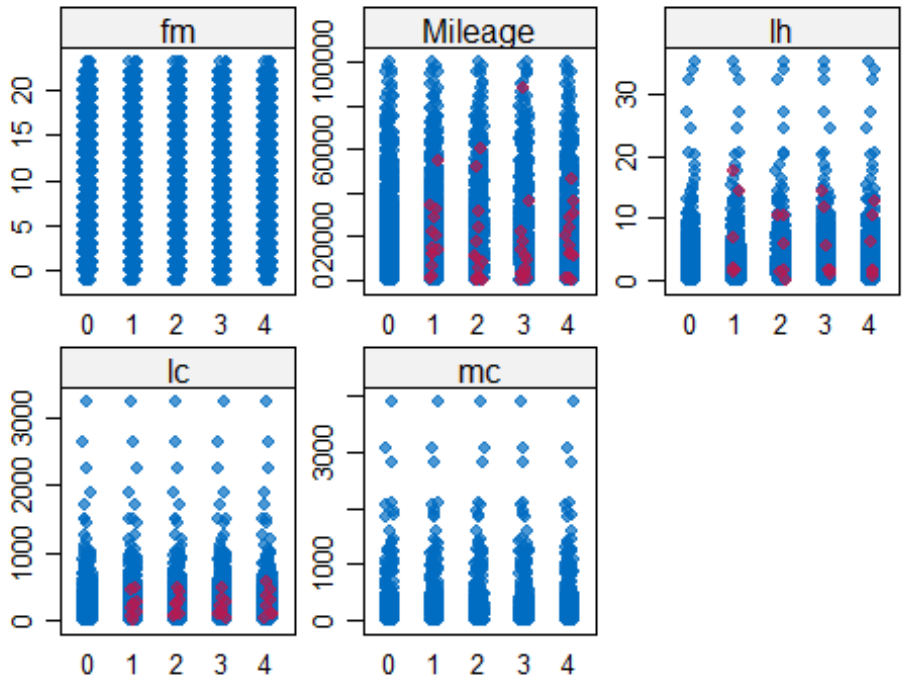
complete data set

```
data<- complete(impute,2)
```

```
summary(data)
```

```
##        fm              Mileage            lh               lc        
##  Min.   :-1.000   Min.   :    1   Min.   : 0.000   Min.   :   0.0  
##  1st Qu.: 4.000   1st Qu.: 5691   1st Qu.: 1.500   1st Qu.: 106.4  
##  Median :10.000   Median :16994   Median : 2.600   Median : 195.6  
##  Mean   : 9.414   Mean   :20531   Mean   : 3.301   Mean   : 242.8  
##  3rd Qu.:14.000   3rd Qu.:30057   3rd Qu.: 4.300   3rd Qu.: 317.8  
##  Max.   :23.000   Max.   :99983   Max.   :35.200   Max.   :3234.4  
##                                                                    
##        mc              State      
##  Min.   :   0.0   TX     :292  
##  1st Qu.: 119.7   CA     :200  
##  Median : 119.7   FL     :167  
##  Mean   : 179.4   GA     : 75  
##  3rd Qu.: 175.5   AZ     : 61  
##  Max.   :3891.1   LA     : 48  
##                   (Other):781
```

```
stripplot(impute,pch = 20 ,cex =1.2)
```




```
xyplot(impute, lc ~ lh | .imp, pch =20, cex =1.2)
```