exam-R-kalman.R

mkalm

Tue Feb 02 20:52:55 2016

```r
## Name: Miklos Kalman
## KEMBA 2017-II
library(data.table)

## Transform the mtcars dataset to data.table and store as a new object
dt <- data.table(mtcars, keep.rownames = TRUE)

## Count the number of cars with less than 4 gears
dt[gear<4,.N]
```

```
## [1] 15
```

```r
## Count the number of cars with more than 4 gears and less than 100 horsepower
dt[gear>4 & hp<100,.N]
```

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

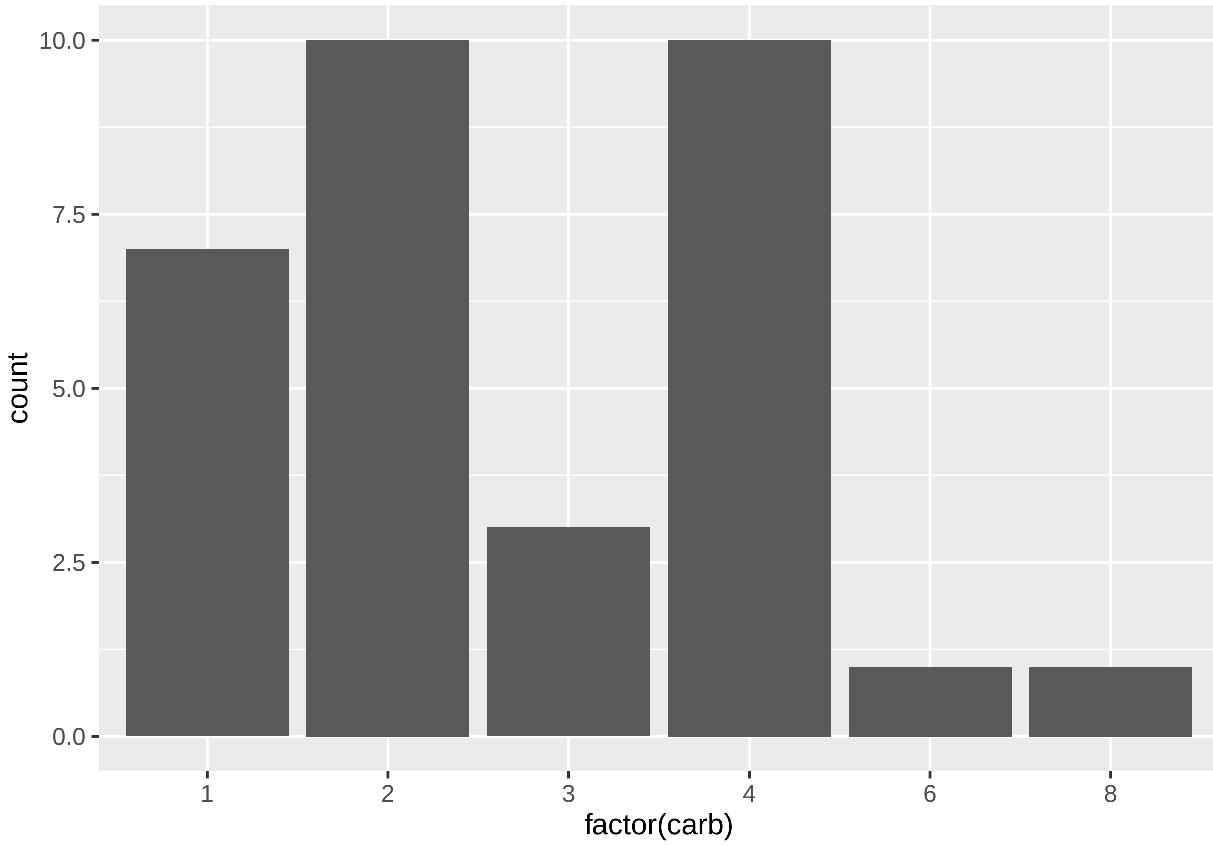
```r
## What's the average weight of cars with 4 cylinders?
dt[cyl==4,mean(wt)]
```

```
## [1] 2.285727
```

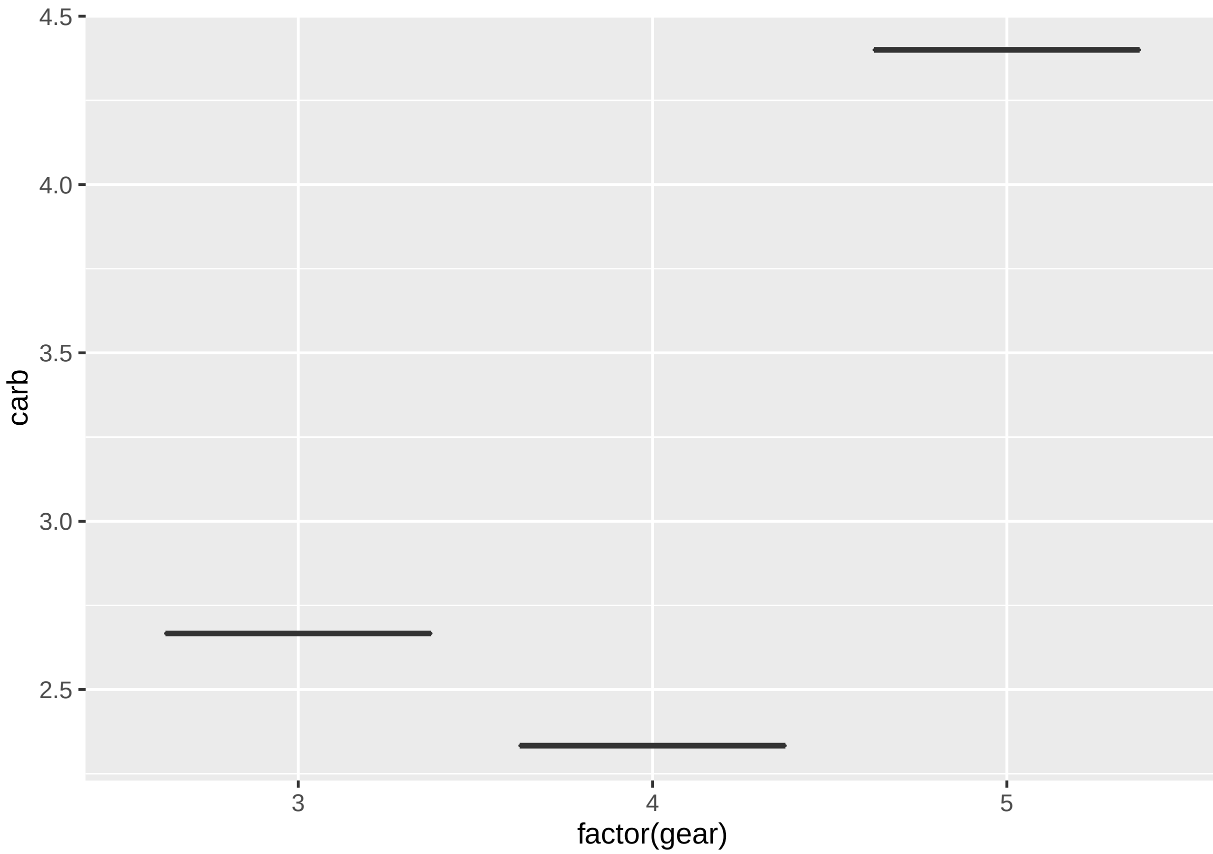
```r
## Which car has the best fuel consumption?
dt[order(mpg, decreasing = TRUE)][1]
```

```
##                rn  mpg cyl disp hp drat    wt qsec vs am gear carb
## 1: Toyota Corolla 33.9   4 71.1 65 4.22 1.835 19.9  1  1    4    1
```

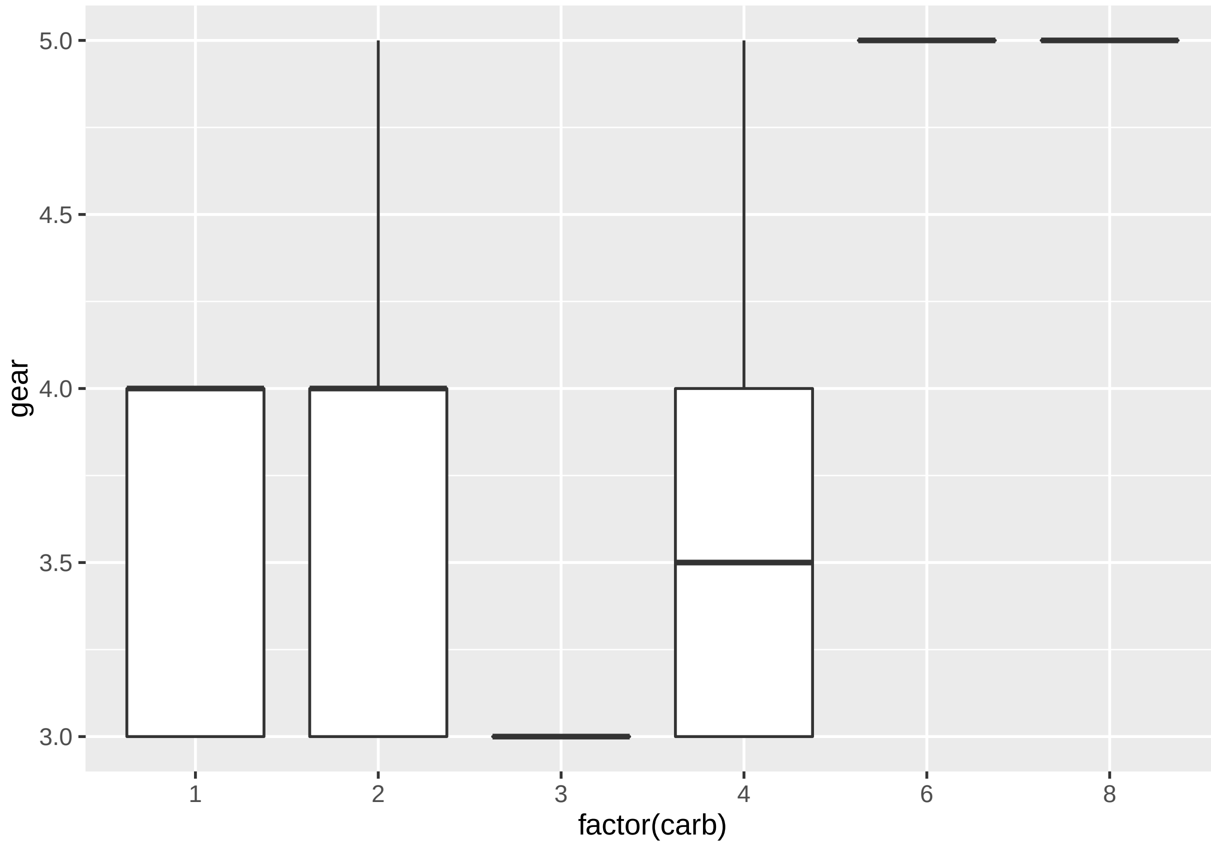
```r
## Plot the distribution of the number of carburetors
library(ggplot2)
ggplot(dt, aes(x = factor(carb) )) + geom_bar()
```



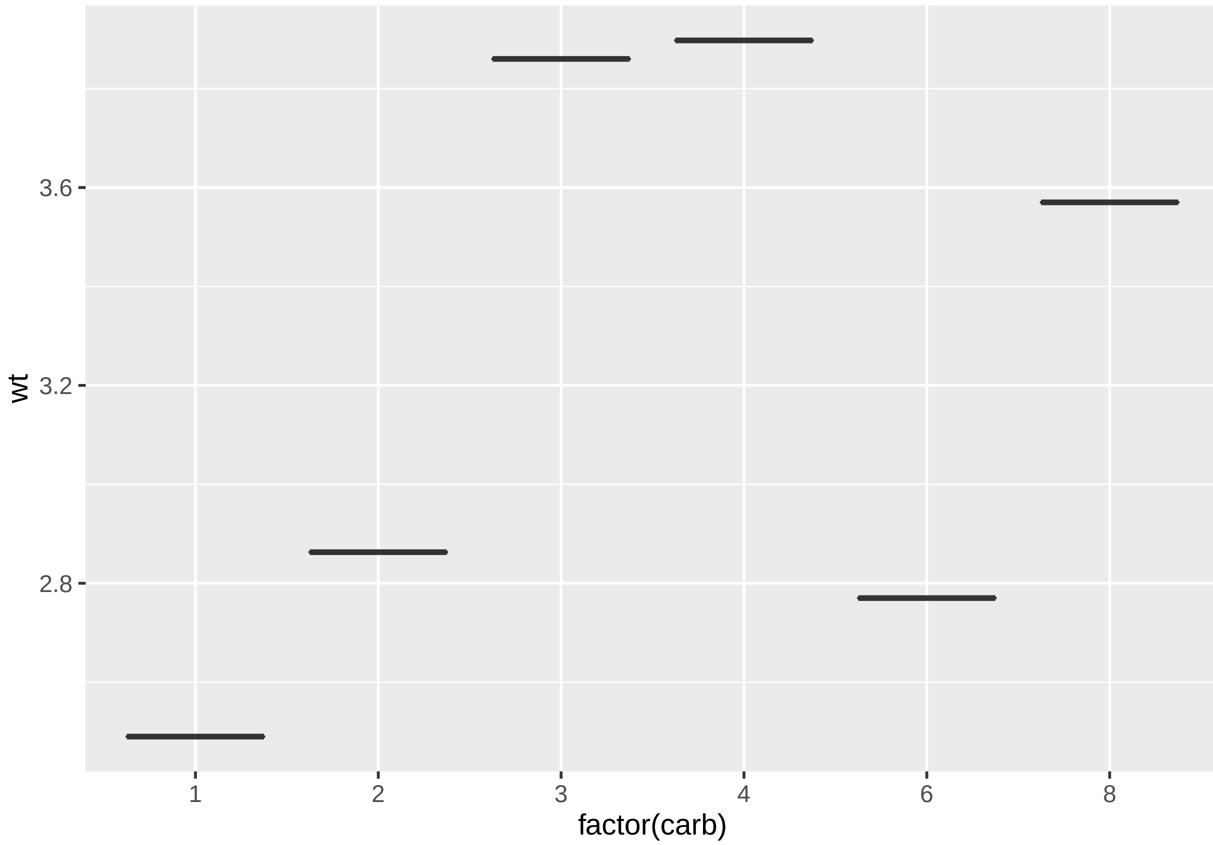

```
## Plot the distribution of the number of carburetors grouped by gears
agg = aggregate(carb ~ gear, FUN = mean, data = dt)
ggplot(agg, aes(x=factor(gear), y=carb))+geom_boxplot()
```



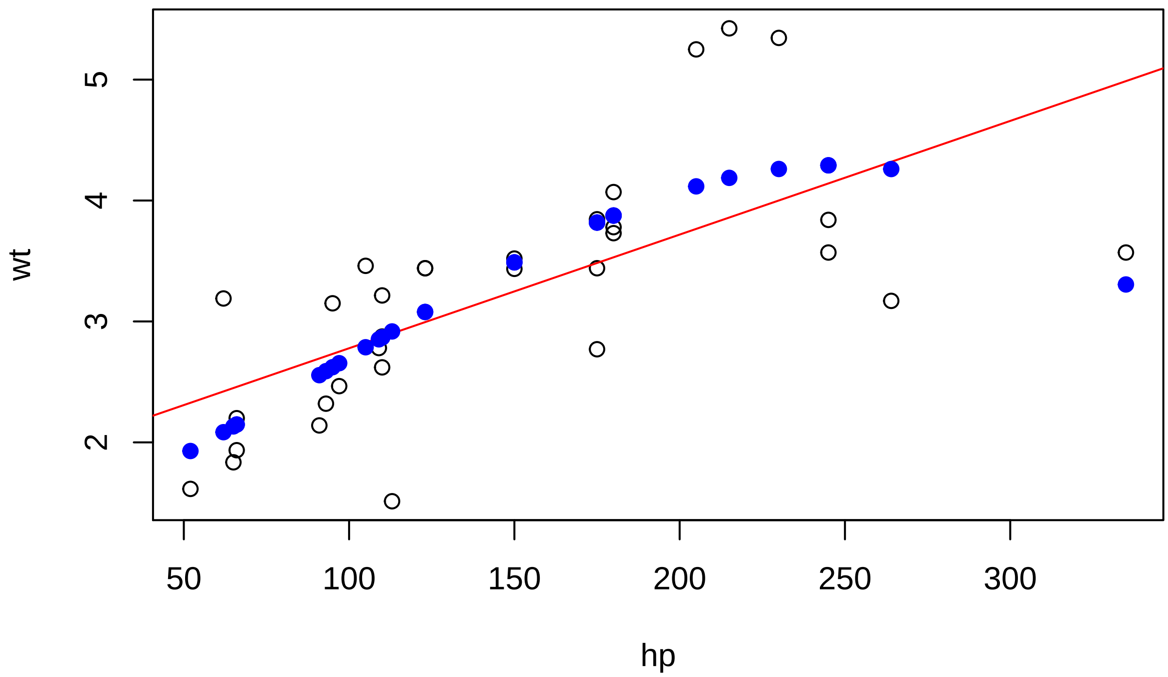

```
##alternate with boxplot
ggplot(dt, aes(x = factor(carb), y = gear)) + geom_boxplot()
```




```
## Plot the average weight grouped by the number of carburetors
agg = aggregate(wt ~ carb, FUN = mean, data=dt)
ggplot(agg, aes(x=factor(carb), y=wt))+geom_boxplot()
```




```
## Plot the weight and horsepower of cars
plot(wt~hp,dt)


## Add a linear trend line to the above plot
fit <- lm(wt ~ hp, data=dt)
abline(fit, col = 'red')


## Add a 3rd degree polynomial model to the above plot
fit3 <- lm (wt  ~ poly(hp, 3, raw=TRUE)  , data=dt)

predfit3 <- predict(fit3)
points(dt$hp, predfit3, col='blue',pch=19)
```




```
## Fit a linear model on the weight of cars to predict fuel consumption
fitfuel <- lm(mpg~wt, data=dt)

## What's the estimated fuel consumption of a car with wt = 5?
predict(fitfuel, newdata = data.frame(wt=5))
```

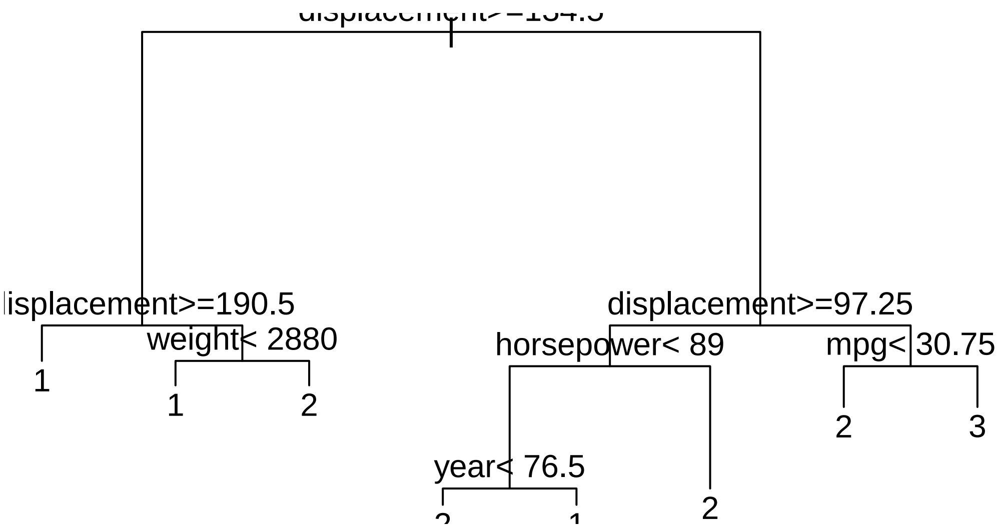
```
##        1
## 10.56277
```

```
## Install the ISLR package and use its Auto for the below exercises
#install.packages("ISLR")
library(ISLR)
## Build and visualize a decision tree to tell if a car was made in America, Europe or Japan


dta <- data.table(Auto, keep.rownames = TRUE)
library(rpart)
ct<- rpart(factor(origin) ~ mpg+cylinders+displacement+horsepower+weight+acceleration+year,data =dta,mi

str(dta)
```

```
## Classes 'data.table' and 'data.frame':   392 obs. of  10 variables:
##  $ rn          : chr  "1" "2" "3" "4" ...
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
```

```
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 2
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
#  Visualize the above decision tree
plot(ct);text(ct)
```




```
## Apply k-means or hierarchical clustering on the dataset to split the observations into 3 groups
kc <- kmeans(dta[,1:7], 3)
str(kc$cluster)
```

```
##  int [1:7] 2 2 1 1 1 3 3
```

```
## Bonus exercise: train a reasonable k-NN or other ML model classifying cars as American VS other orig
#aut<-data.table(Auto)
#aut[origin=3]
```