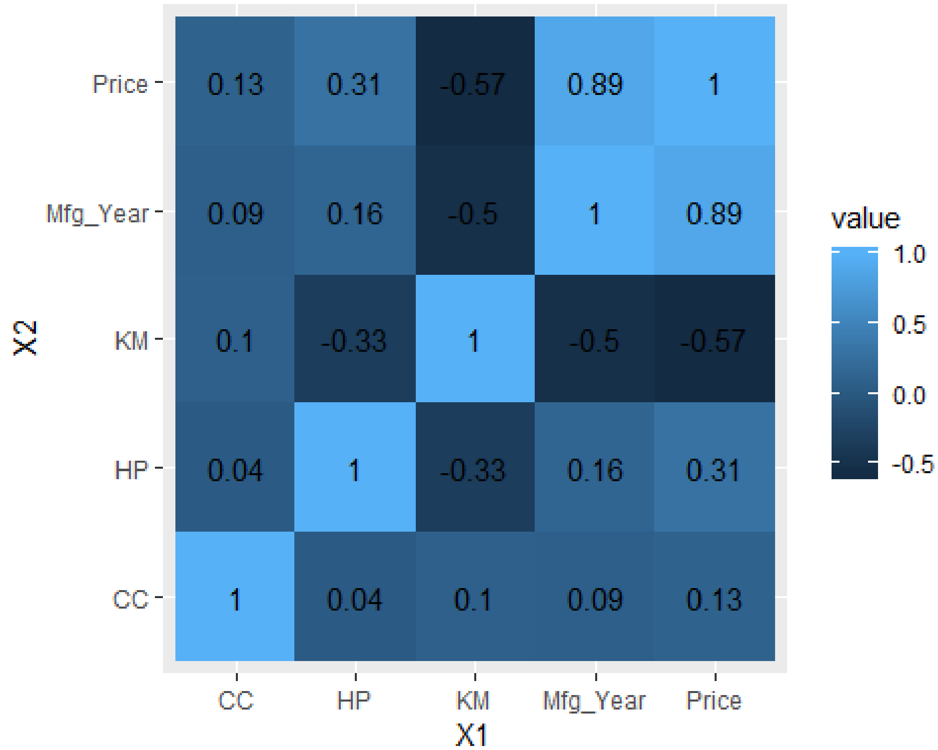
Homework1.R

cmrump

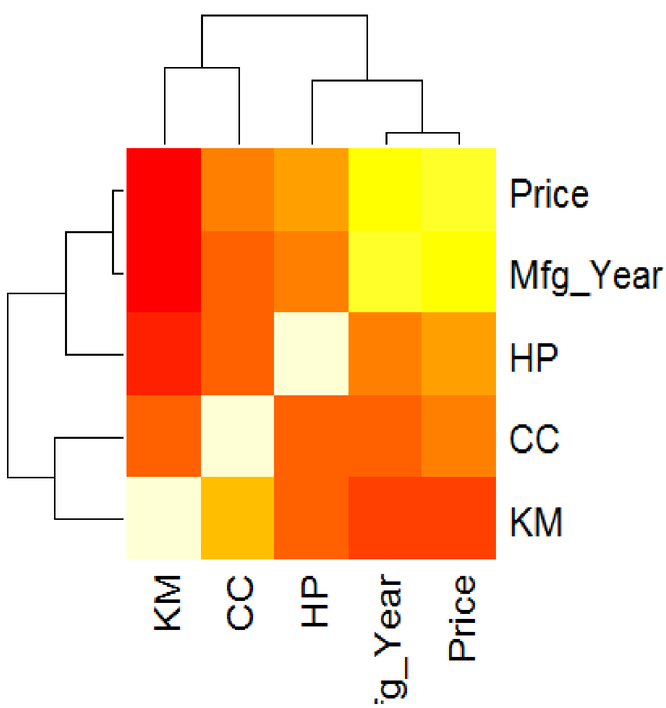
Wed Feb 13 17:55:12 2019

```
setwd("~/My Courses/Data Mining/Datasets/DMBA-3eR-datasets")

toyota.df <- read.csv("ToyotaCorolla.csv")
unique(toyota.df$Mfg_Year)

## [1] 2002 2003 2004 2001 2000 1999 1998

unique(toyota.df$CC) #Notice the outlier

##  [1]  2000  1800  1900  1600  1400  1598 16000  1995  1398  1300  1587
## [12]  1975  1332

select.var <- c(3, 6, 7, 8, 9, 13)
head(toyota.df[,select.var],10)

##    Price Mfg_Year    KM Fuel_Type  HP   CC
## 1  13500     2002 46986    Diesel  90 2000
## 2  13750     2002 72937    Diesel  90 2000
## 3  13950     2002 41711    Diesel  90 2000
## 4  14950     2002 48000    Diesel  90 2000
## 5  13750     2002 38500    Diesel  90 2000
## 6  12950     2002 61000    Diesel  90 2000
## 7  16900     2002 94612    Diesel  90 2000
## 8  18600     2002 75889    Diesel  90 2000
## 9  21500     2002 19700    Petrol 192 1800
## 10 12950     2002 71138    Diesel  69 1900

toyota.cor <- round(cor(na.omit(toyota.df[,c(3, 6, 7, 9, 13)])),2) #correlati
on submatrix
toyota.cor

##          Price Mfg_Year    KM    HP   CC
## Price     1.00     0.89 -0.57  0.31 0.13
## Mfg_Year  0.89     1.00 -0.50  0.16 0.09
## KM       -0.57    -0.50  1.00 -0.33 0.10
## HP        0.31     0.16 -0.33  1.00 0.04
## CC        0.13     0.09  0.10  0.04 1.00

# alternative heatmap with ggplot
library(ggplot2)
library(reshape) # to generate input for the plot
melted.cor.mat <- melt(toyota.cor)
ggplot(melted.cor.mat, aes(x = X1, y = X2, fill = value)) +
```

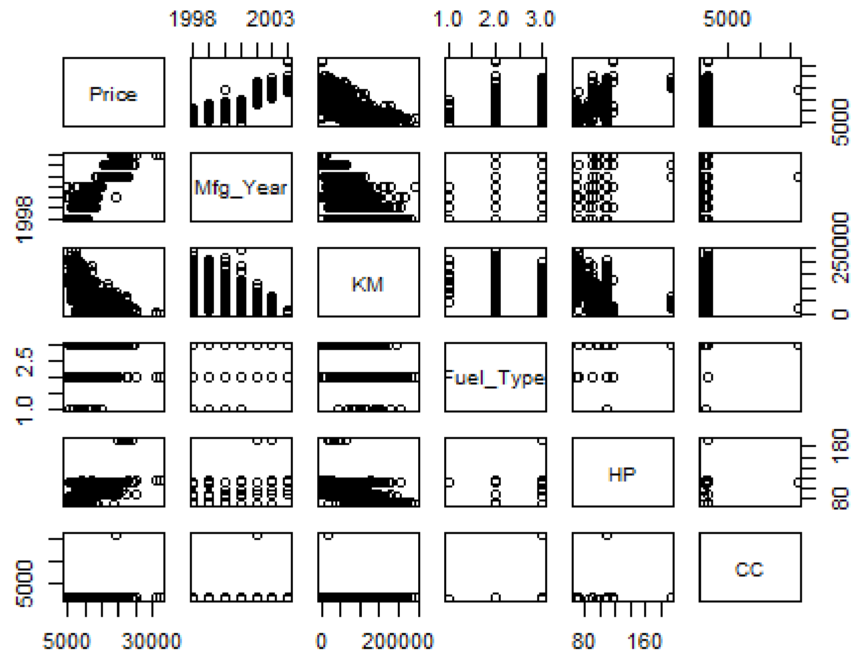
```r
  geom_tile() +
  geom_text(aes(x = X1, y = X2, label = value))
```



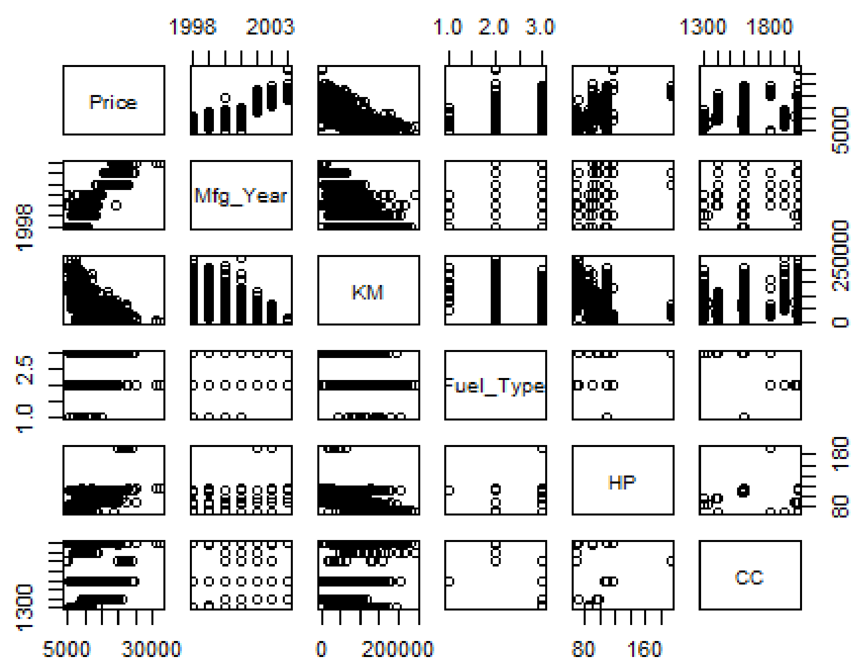

```r
#standard heatmap with dendrogram clustering
heatmap(toyota.cor)
```



```
plot(na.omit(toyota.df[,select.var])) #matrix plot
```




```
plot(na.omit(toyota.df[-c(81),select.var])) #matrix plot without outlier
```



```
#install.packages(dummies)
library(dummies)

## dummies-1.5.6 provided by Decision Patterns

toyota.dum.df <- dummy.data.frame(toyota.df[,-2], sep = ".")
names(toyota.dum.df)

##  [1] "Id"               "Price"            "Age_08_04"
##  [4] "Mfg_Month"        "Mfg_Year"         "KM"
##  [7] "Fuel_Type.CNG"    "Fuel_Type.Diesel" "Fuel_Type.Petrol"
## [10] "HP"               "Met_Color"        "Color.Beige"
## [13] "Color.Black"      "Color.Blue"       "Color.Green"
## [16] "Color.Grey"       "Color.Red"        "Color.Silver"
## [19] "Color.Violet"     "Color.White"      "Color.Yellow"
## [22] "Automatic"        "CC"               "Doors"
## [25] "Cylinders"        "Gears"            "Quarterly_Tax"
## [28] "Weight"           "Mfr_Guarantee"    "BOVAG_Guarantee"
## [31] "Guarantee_Period" "ABS"              "Airbag_1"
## [34] "Airbag_2"         "Airco"            "Automatic_airco"
## [37] "Boardcomputer"    "CD_Player"        "Central_Lock"
## [40] "Powered_Windows"  "Power_Steering"   "Radio"
## [43] "Mistlamps"        "Sport_Model"      "Backseat_Divider"
## [46] "Metallic_Rim"     "Radio_cassette"   "Parking_Assistant"
## [49] "Tow_Bar"

head(toyota.dum.df[,c(1:12)],10)

##    Id Price Age_08_04 Mfg_Month Mfg_Year    KM Fuel_Type.CNG
## 1   1 13500        23        10     2002 46986             0
## 2   2 13750        23        10     2002 72937             0
## 3   3 13950        24         9     2002 41711             0
## 4   4 14950        26         7     2002 48000             0
## 5   5 13750        30         3     2002 38500             0
## 6   6 12950        32         1     2002 61000             0
## 7   7 16900        27         6     2002 94612             0
## 8   8 18600        30         3     2002 75889             0
## 9   9 21500        27         6     2002 19700             0
## 10 10 12950        23        10     2002 71138             0
##    Fuel_Type.Diesel Fuel_Type.Petrol  HP Met_Color Color.Beige
## 1                 1                0  90         1           0
## 2                 1                0  90         1           0
## 3                 1                0  90         1           0
## 4                 1                0  90         0           0
## 5                 1                0  90         0           0
## 6                 1                0  90         0           0
## 7                 1                0  90         1           0
## 8                 1                0  90         1           0
## 9                 0                1 192         0           0
## 10                1                0  69         0           0
```

```
# use set.seed() to get the same partitions when re-running the R code.
set.seed(1)

## partitioning into training (50%), validation (30%), test (20%)
# randomly sample 50% of the row IDs for training
train.rows <- sample(rownames(toyota.dum.df), dim(toyota.dum.df)[1]*0.5)

# sample 30% of the row IDs into the validation set, drawing only from record
s
# not already in the training set via setdiff()
valid.rows <- sample(setdiff(rownames(toyota.dum.df), train.rows), dim(toyota
.dum.df)[1]*0.3)

# assign the remaining 20% row IDs serve as test
test.rows <- setdiff(rownames(toyota.dum.df), union(train.rows, valid.rows))

# create the 3 data frames by collecting all columns from the appropriate row
s
train.data <- toyota.dum.df[train.rows, ]
valid.data <- toyota.dum.df[valid.rows, ]
test.data <- toyota.dum.df[test.rows, ]

head(train.data[,c(1:12)],10)
```

```
##        Id Price Age_08_04 Mfg_Month Mfg_Year     KM Fuel_Type.CNG
## 382   384  7750        54         3     2000 174139             0
## 534   536 11895        52         5     2000  47689             0
## 822   825  8450        64         5     1999  70116             0
## 1302 1308  6900        80         1     1998  70939             0
## 289   290 11895        44         1     2001  44218             0
## 1286 1292  7950        77         4     1998  72703             0
## 1351 1357  7750        76         5     1998  60833             0
## 945   948 10250        57        12     1999  54000             0
## 899   902  8950        65         4     1999  60000             0
## 89     89 15950        19         2     2003  51884             0
##      Fuel_Type.Diesel Fuel_Type.Petrol  HP Met_Color Color.Beige
## 382                 1                0  72         1           0
## 534                 0                1 110         0           0
## 822                 0                1 110         1           0
## 1302                0                1 110         1           0
## 289                 0                1  97         1           0
## 1286                0                1 110         1           0
## 1351                0                1 110         1           0
## 945                 0                1 110         1           0
## 899                 0                1  86         1           0
## 89                  0                1  97         1           0
```

```
head(valid.data[,c(1:12)],10)
```

```
##         Id Price Age_08_04 Mfg_Month Mfg_Year     KM Fuel_Type.CNG
## 604   607  6950        58        11     1999 205000             0
## 1094 1099  5250        72         9     1998 126478             0
## 5       5 13750        30         3     2002  38500             0
## 849   852  9950        65         4     1999  65513             0
## 1283 1289  7500        80         1     1998  73200             0
## 995   999  7750        64         5     1999  43000             0
## 343   345 14950        42         3     2001  29640             0
## 1219 1225  9450        70        11     1998  85470             0
## 463   465 10750        46        11     2000  69574             0
## 813   816  8950        65         4     1999  71317             0
##       Fuel_Type.Diesel Fuel_Type.Petrol  HP Met_Color Color.Beige
## 604                  1                0  72         1           0
## 1094                 0                1 110         1           0
## 5                    1                0  90         0           0
## 849                  0                1 110         1           0
## 1283                 0                1 110         1           0
## 995                  0                1  86         0           0
## 343                  0                1 110         0           0
## 1219                 0                1 107         0           0
## 463                  0                1  97         0           0
## 813                  0                1 110         0           0
```

```
head(test.data[,c(1:12)],10)
```

```
##    Id Price Age_08_04 Mfg_Month Mfg_Year     KM Fuel_Type.CNG
## 3   3 13950        24         9     2002  41711             0
## 14 14 21500        31         2     2002  23000             0
## 16 16 22000        28         5     2002  18739             0
## 23 23 15950        28         5     2002  56349             0
## 24 24 16950        28         5     2002  32220             0
## 26 26 15950        25         8     2002  28450             0
## 38 38 14950        23        10     2002  10000             0
## 40 40 14750        27         6     2002  27500             0
## 43 43 13950        22        11     2002  46961             0
## 45 45 16950        22        11     2002 100250             0
##    Fuel_Type.Diesel Fuel_Type.Petrol  HP Met_Color Color.Beige
## 3                 1                0  90         1           0
## 14                0                1 192         1           0
## 16                0                1 192         0           0
## 23                0                1 110         1           0
## 24                0                1 110         1           0
## 26                0                1 110         1           0
## 38                0                1  97         1           0
## 40                0                1  97         0           0
## 43                0                1  97         0           0
## 45                1                0  90         0           0
```

Problem 2.11

a. Notice in the matrix plot (atop p. 3 of knitted R code) the outlier CC of 16000. Assuming that this was really 1600, the plot was redone at bottom of p.3 to show more meaningful relationships to CC variable.

As for correlation patterns, the plots in row 1, columns 2 & 3 show prices increase for newer cars (Mfg_Year) and decrease for cars with higher mileage (KM), as expected. Also expected, mileage decreases for newer cars as shown in plot of row 3, column 2.

b. Dummy Variables (see p. 4 of knitted R code)

i. The categorical fuel_type variable has three categories: petrol, diesel and compressed natural gas (CNG), i.e. methane. To convert these variables into dummy variables, we use only need keep two variables. The binary variable Petrol gets the value 1 if Fuel Type=Petrol and otherwise it gets the value 0. The binary variable Diesel gets the value 1 if Fuel Type=Diesel and otherwise 0. Deleting CNG would designate this fuel as the "reference category." If Fuel type is CNG, both of the other binary variables take the value 0.

ii. Partitioning (see pp. 5-6 of knitted R code)

Training dataset

The training dataset is used to train or build models. For example, in a linear regression, the training dataset is used to fit the linear regression model, i.e. to compute the regression coefficients. This is usually the largest partition.

Validation dataset

Once a model is built on training data, we assess the accuracy of the model on unseen data. For this, the model should be used on a dataset that was not used in the training process. In the validation data we know the actual value of the response variable, and can therefore examine the difference between the actual value and the predicted value to determine the error in prediction. Based on this performance, sometimes the validation dataset is used to tweak the model, or to choose between multiple fitted models.

Test dataset

The validation dataset is often used to select a model with minimum error. Testing that model on completely unseen data gives a realistic estimate of the performance of the model. When a model is finally chosen, its accuracy with the validation dataset is still an optimistic estimate of how it would perform with unseen data. This is because (1) the final model has come out as the winner among the competing models based on the fact that its accuracy with the validation dataset is highest, and/or (2) the validation set was used to help build one or more models. Thus, you need to set aside yet another portion of data, which is used neither in training nor in validation, which is called the test dataset. The accuracy of the model on the test data gives a realistic estimate of the performance of the model on completely unseen data.

Problem 4.3

a. As discussed in Problem 2.11 b.i., fuel type and color are the categorical variables.

b. See Problem 2.11 b.i.

c. Only need keep *N*-1 variables; for fuel type, 3-1 = 2 variables needed.

d. See p.4 of knitted R code.

e. As shown on pp. 1-2 of knitted R code, Price is highly positively correlated ($r = 0.89$) with year of manufacture (Mfg_Year) and negatively correlated ($r = -0.57$) with mileage (KM). As expected mileage is also negatively correlated ($r = -0.5$) with year of manufacture as newer cars have been driven less over their shorter lifetime. Increasing horsepower (HP) also increases price ($r = 0.31$).