Data Mining Practice Part.3

1. Package

```
# install.packages('rpart')

library(rpart)
```

2. Data Objects in R

```
# data import
data(iris)
data(kyphosis, package = "rpart")
data(Forbes2000, package = "HSAUR")
ls()
```

```
## [1] "Forbes2000" "iris"       "kyphosis"
```

```
str(Forbes2000)
```

```
## 'data.frame':    2000 obs. of  8 variables:
##  $ rank       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ name       : chr  "Citigroup" "General Electric" "American Intl Grou
p" "ExxonMobil" ...
##  $ country    : Factor w/ 61 levels "Africa","Australia",..: 60 60 60 6
0 56 60 56 28 60 60 ...
##  $ category   : Factor w/ 27 levels "Aerospace & defense",..: 2 6 16 19
 19 2 2 8 9 20 ...
##  $ sales      : num  94.7 134.2 76.7 222.9 232.6 ...
##  $ profits    : num  17.85 15.59 6.46 20.96 10.27 ...
##  $ assets     : num  1264 627 648 167 178 ...
##  $ marketvalue: num  255 329 195 277 174 ...
```

```
help(Forbes2000)
```

```
## No documentation for 'Forbes2000' in specified packages and libraries:
## you could try '??Forbes2000'
```

```
class(Forbes2000)
```

```
## [1] "data.frame"
```

```
dim(Forbes2000)
```

```
## [1] 2000    8
```

```
nrow(Forbes2000)
```

```
## [1] 2000
```

```
ncol(Forbes2000)
```

```
## [1] 8
names(Forbes2000)
## [1] "rank"        "name"        "country"     "category"    "sales"
## [6] "profits"     "assets"      "marketvalue"

class(Forbes2000$rank)
## [1] "integer"
length(Forbes2000$rank)
## [1] 2000

class(Forbes2000$name)
## [1] "character"

class(Forbes2000$category)
## [1] "factor"
nlevels(Forbes2000$category)
## [1] 27
levels(Forbes2000$category)
##  [1] "Aerospace & defense"              "Banking"
##  [3] "Business services & supplies"     "Capital goods"
##  [5] "Chemicals"                        "Conglomerates"
##  [7] "Construction"                     "Consumer durables"
##  [9] "Diversified financials"           "Drugs & biotechnology"
## [11] "Food drink & tobacco"             "Food markets"
## [13] "Health care equipment & services" "Hotels restaurants & leisure"
## [15] "Household & personal products"    "Insurance"
## [17] "Materials"                        "Media"
## [19] "Oil & gas operations"             "Retailing"
## [21] "Semiconductors"                   "Software & services"
```

```
## [23] "Technology hardware & equipment"  "Telecommunications services"

## [25] "Trading companies"                "Transportation"

## [27] "Utilities"
```

```
unique(Forbes2000$category)
```

```
##  [1] Banking                          Conglomerates
##  [3] Insurance                        Oil & gas operations
##  [5] Consumer durables                Diversified financials
##  [7] Retailing                        Technology hardware & equipment
##  [9] Food drink & tobacco             Drugs & biotechnology
## [11] Telecommunications services      Software & services
## [13] Media                            Household & personal products
## [15] Semiconductors                   Utilities
## [17] Transportation                   Food markets
## [19] Chemicals                        Aerospace & defense
## [21] Materials                        Capital goods
## [23] Business services & supplies     Hotels restaurants & leisure
## [25] Construction                     Health care equipment & services
## [27] Trading companies
## 27 Levels: Aerospace & defense Banking ... Utilities
```

```
table(Forbes2000$category)
```

```
##
##              Aerospace & defense                          Banking
##                               19                              313
##     Business services & supplies                  Capital goods
##                               70                               53
##                         Chemicals                  Conglomerates
##                               50                               31
##                      Construction              Consumer durables
##                               79                               74
##           Diversified financials           Drugs & biotechnology
##                              158                               45
##             Food drink & tobacco                    Food markets
##                               83                               33
## Health care equipment & services     Hotels restaurants & leisure
##                               65                               37
##    Household & personal products                        Insurance
##                               44                              112
##                        Materials                            Media
##                               97                               61
##             Oil & gas operations                        Retailing
##                               90                               88
##                   Semiconductors              Software & services
##                               26                               31
##  Technology hardware & equipment      Telecommunications services
##                               59                               67
```

```
##                Trading companies                     Transportation
##                               25                                 80
##                        Utilities
##                              110
```

```
class(Forbes2000$sales)
```

```
## [1] "numeric"
```

```
median(Forbes2000$sales)
```

```
## [1] 4.365
```

```
mean(Forbes2000$sales)
```

```
## [1] 9.697
```

```
range(Forbes2000$sales)
```

```
## [1]   0.01 256.33
```

```
summary(Forbes2000$sales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.01    2.02    4.36    9.70    9.55  256.00
```

3. Basic Data Manipulation

```
Forbes2000[1:3, c("name", "sales", "profits", "assets")]
```

```
##                  name  sales profits assets
## 1           Citigroup  94.71   17.85 1264.0
## 2    General Electric 134.19   15.59  626.9
## 3 American Intl Group  76.66    6.46  647.7
```

```
order.sales <- order(Forbes2000$sales)
Forbes2000[order.sales[c(2000, 1999, 1998)], c("name", "sales", "profits",
 "assets")]
```

```
##               name sales profits assets
## 10 Wal-Mart Stores 256.3    9.05  104.9
## 5               BP 232.6   10.27  177.6
## 4       ExxonMobil 222.9   20.96  167.0
```

```
Forbes2000[Forbes2000$assets > 1000, c("name", "sales", "profits", "assets
")]
```

```
##                 name sales profits assets
## 1          Citigroup 94.71   17.85   1264
## 9         Fannie Mae 53.13    6.48   1019
## 403 Mizuho Financial 24.40  -20.11   1116
```

```
table(Forbes2000$assets > 1000)
```

```
##
## FALSE  TRUE
##  1997     3
```

```
na.profits <- is.na(Forbes2000$profits)
table(na.profits)
```

```
## na.profits
## FALSE  TRUE
##  1995     5
```

```
Forbes2000[na.profits, c("name", "sales", "profits", "assets")]
```

```
##                        name sales profits assets
## 772                     AMP  5.40      NA  42.94
## 1085                    HHG  5.68      NA  51.65
## 1091                    NTL  3.50      NA  10.59
## 1425      US Airways Group  5.50      NA   8.58
## 1909 Laidlaw International  4.48      NA   3.98
```

```
table(complete.cases(Forbes2000))
```

```
##
## FALSE  TRUE
##     5  1995
```

```
uk <- subset(Forbes2000, country == "United Kingdom")
dim(uk)
```

```
## [1] 137   8
```

4. Simple Summary Statistics

```
summary(Forbes2000)
```

```
##       rank          name                     country
##  Min.   :   1   Length:2000        United States :751
##  1st Qu.: 501   Class :character   Japan         :316
##  Median :1000   Mode  :character   United Kingdom:137
##  Mean   :1000                      Germany       : 65
##  3rd Qu.:1500                      France        : 63
##  Max.   :2000                      Canada        : 56
##                                    (Other)       :612
##                    category        sales            profits
##  Banking                : 313   Min.   :  0.01   Min.   :-25.830
##  Diversified financials: 158   1st Qu.:  2.02   1st Qu.:  0.080
##  Insurance              : 112   Median :  4.37   Median :  0.200
##  Utilities              : 110   Mean   :  9.70   Mean   :  0.381
##  Materials              :  97   3rd Qu.:  9.55   3rd Qu.:  0.440
##  Oil & gas operations   :  90   Max.   :256.33   Max.   : 20.960
##  (Other)                :1120                    NA's   :5
##      assets        marketvalue
##  Min.   :   0.3   Min.   :  0.0
```

```
##  1st Qu.:   4.0   1st Qu.:  2.7
##  Median :   9.3   Median :  5.2
##  Mean   :  34.0   Mean   : 11.9
##  3rd Qu.:  22.8   3rd Qu.: 10.6
##  Max.   :1264.0   Max.   :328.5
##
```

```
lapply(Forbes2000, summary)
```

```
## $rank
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1     501    1000    1000    1500    2000
##
## $name
##    Length     Class      Mode
##      2000 character character
##
## $country
##                        Africa                     Australia
##                             2                            37
##    Australia/ United Kingdom                       Austria
##                             2                             8
##                       Bahamas                       Belgium
##                             1                             9
##                       Bermuda                        Brazil
##                            20                            15
##                        Canada                Cayman Islands
##                            56                             5
##                         Chile                         China
##                             4                            25
##               Czech Republic                       Denmark
##                             2                            10
##                       Finland                        France
##                            11                            63
##       France/ United Kingdom                       Germany
##                             1                            65
##                        Greece                Hong Kong/China
##                            12                            20
##                       Hungary                         India
##                             2                            27
##                     Indonesia                       Ireland
##                             7                             8
##                       Islands                        Israel
##                             1                             8
##                         Italy                         Japan
##                            41                           316
##                        Jordan                     Kong/China
##                             1                             4
##                         Korea                       Liberia
##                             4                             1
##                    Luxembourg                      Malaysia
##                             2                            16
##                        Mexico                   Netherlands
```

```
##                            17                            28
##  Netherlands/ United Kingdom                   New Zealand
##                             2                             1
##                        Norway                      Pakistan
##                             8                             1
##       Panama/ United Kingdom                          Peru
##                             1                             1
##                   Philippines                        Poland
##                             2                             1
##                      Portugal                        Russia
##                             7                            12
##                     Singapore                  South Africa
##                            16                            15
##                   South Korea                         Spain
##                            45                            29
##                        Sweden                   Switzerland
##                            26                            34
##                        Taiwan                      Thailand
##                            35                             9
##                        Turkey                United Kingdom
##                            12                           137
##    United Kingdom/ Australia  United Kingdom/ Netherlands
##                             1                             1
## United Kingdom/ South Africa                United States
##                             1                           751
##                     Venezuela
##                             1
##
## $category
##               Aerospace & defense                       Banking
##                                19                           313
##     Business services & supplies                Capital goods
##                                70                            53
##                          Chemicals                Conglomerates
##                                50                            31
##                      Construction            Consumer durables
##                                79                            74
##           Diversified financials          Drugs & biotechnology
##                               158                            45
##             Food drink & tobacco                  Food markets
##                                83                            33
## Health care equipment & services    Hotels restaurants & leisure
##                                65                            37
##    Household & personal products                       Insurance
##                                44                           112
##                         Materials                         Media
##                                97                            61
##             Oil & gas operations                     Retailing
##                                90                            88
##                   Semiconductors           Software & services
##                                26                            31
##  Technology hardware & equipment     Telecommunications services
##                                59                            67
```

```
##                Trading companies                    Transportation
##                               25                                80
##                        Utilities
##                              110
##
## $sales
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.01    2.02    4.36    9.70    9.55  256.00
##
## $profits
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -25.800   0.080   0.200   0.381   0.440  21.000       5
##
## $assets
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.3     4.0     9.3    34.0    22.8  1260.0
##
## $marketvalue
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     2.7     5.2    11.9    10.6   329.0
```

```
mprofits <- tapply(Forbes2000$profits, Forbes2000$category, median, na.rm
= TRUE)
```

```
median(Forbes2000$profits)
```

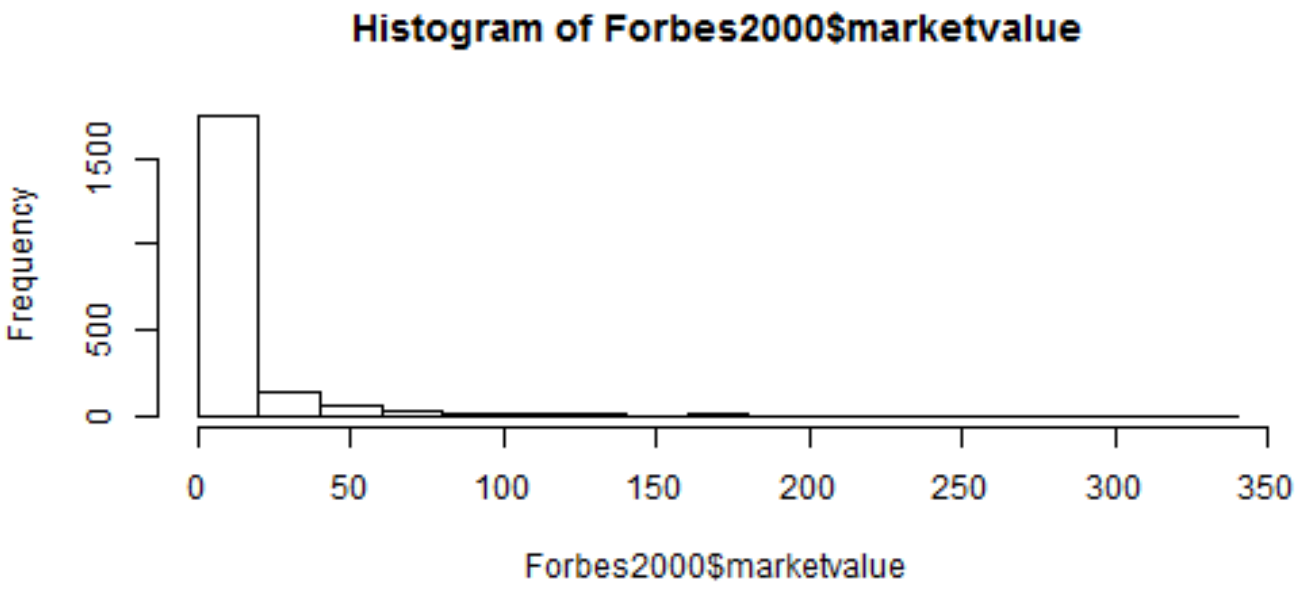
```
## [1] NA
```

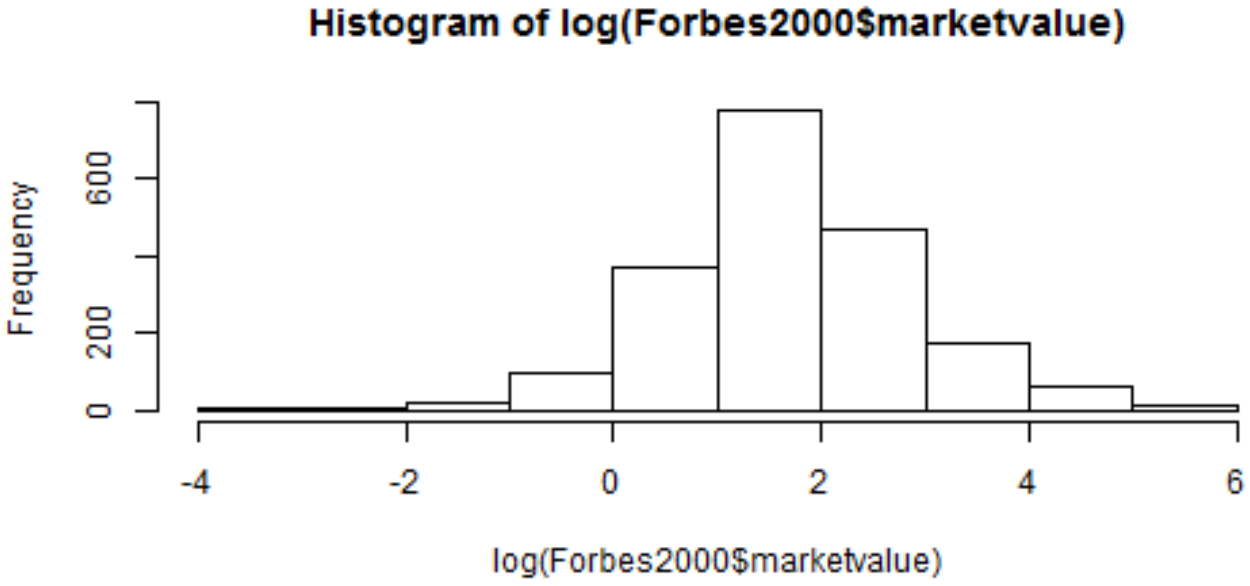
4.1 Simple Graphics

```
fm <- marketvalue ~ sales
class(fm)
```

```
## [1] "formula"
```

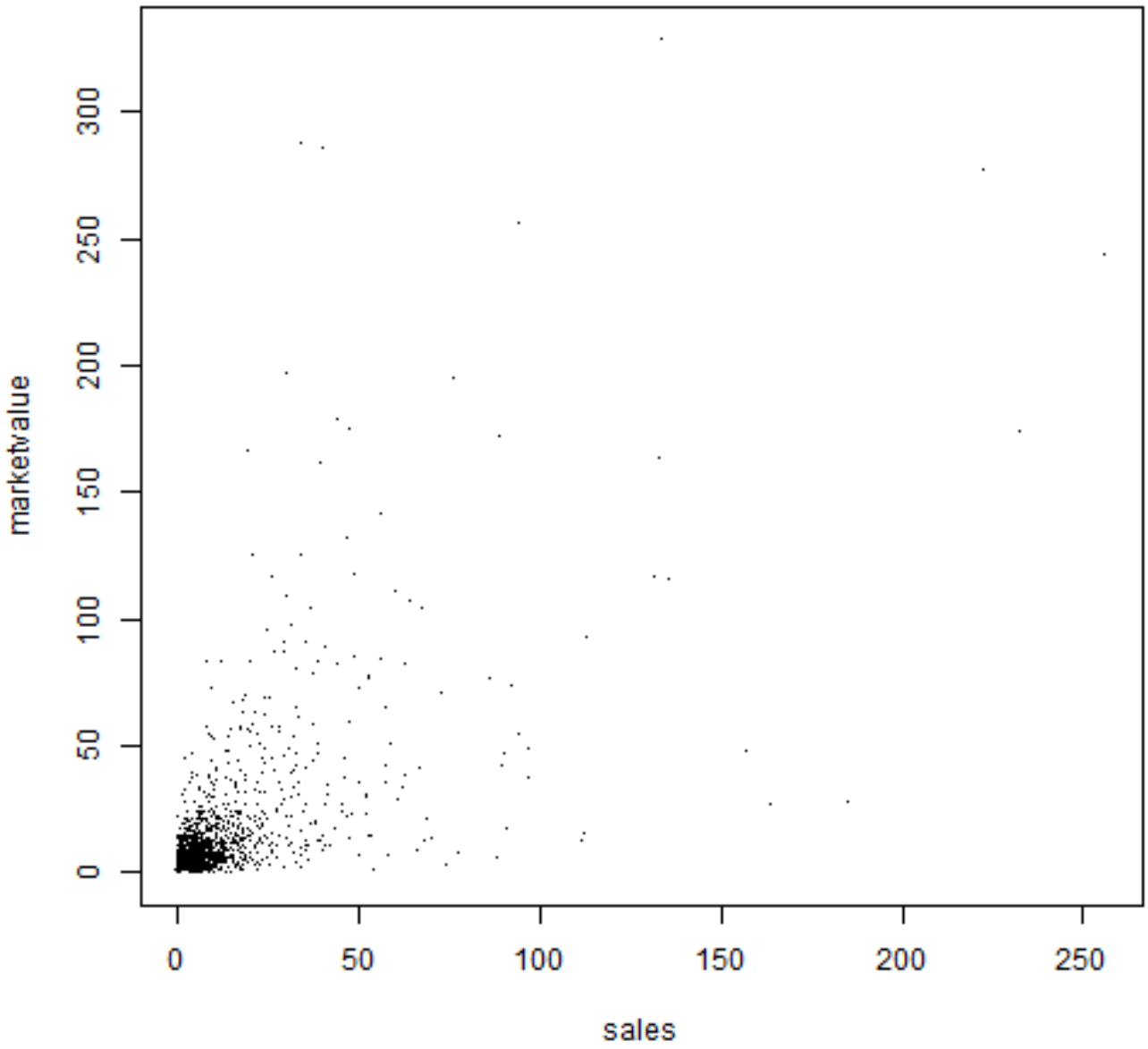
```
par(mfrow = c(2, 1))
hist(Forbes2000$marketvalue)
hist(log(Forbes2000$marketvalue))
```





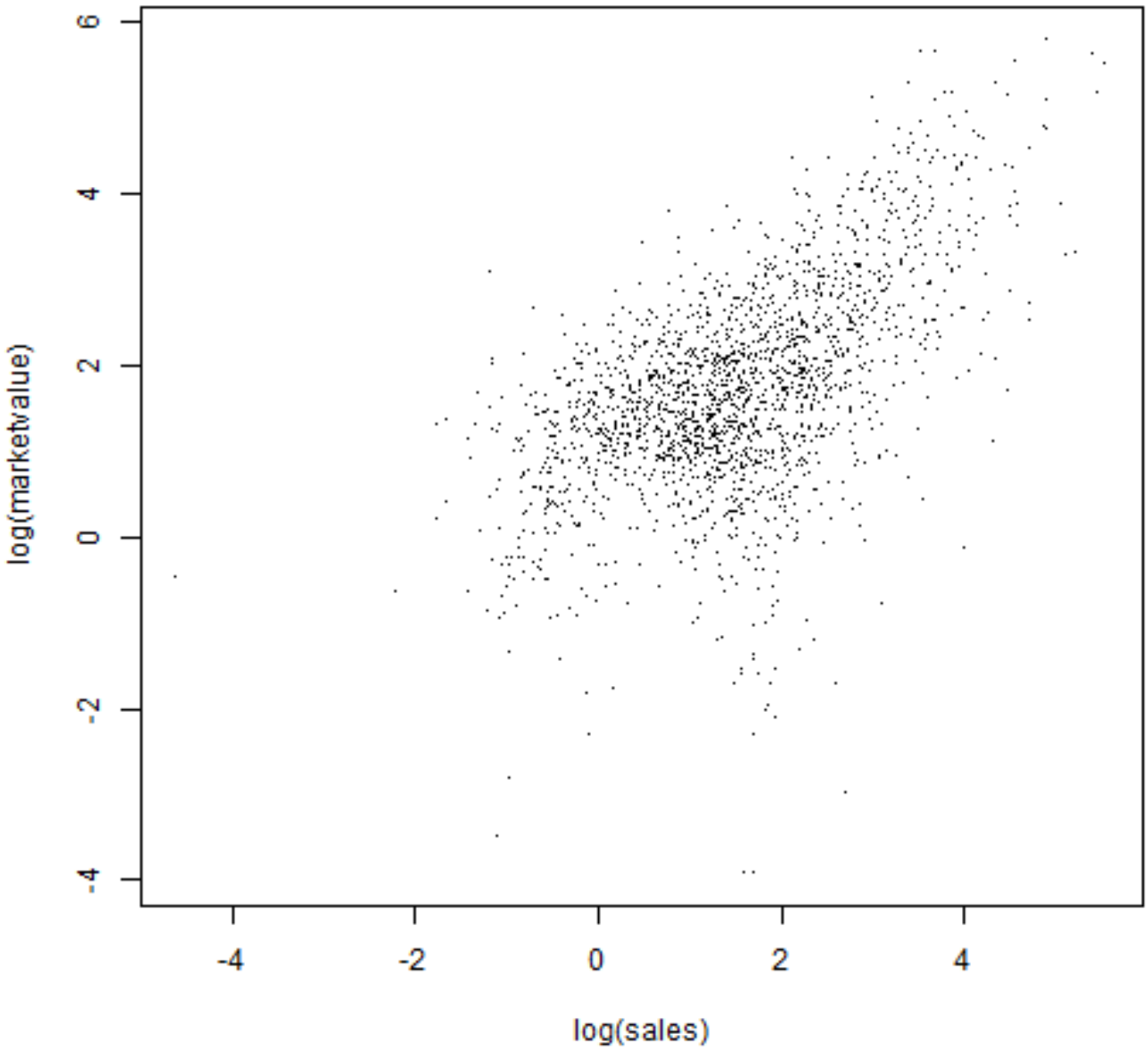
plot of chunk unnamed-chunk-5

```
par(mfrow = c(1, 1))
plot(fm, data = Forbes2000, pch = ".")
```



plot of chunk unnamed-chunk-5

```
plot(log(marketvalue) ~ log(sales), data = Forbes2000, pch = ".")
```



plot of chunk unnamed-chunk-5

5. Markdown document

6. Exercise

Ex 1. Calculate the median profit for the companies in the United States and the median profit for the companies in the UK, France and Germany.

Ex 2. Find all German companies with negative profits.

Ex 3. Find the average value of sales for the companies in each country in the Forbes data set, and find the number of companies in each country with profits above 5 billion US dollars.