Projekt „***Nowa oferta edukacyjna Uniwersytetu Wrocławskiego odpowiedzią na współczesne potrzeby rynku pracy i gospodarki opartej na wiedzy***”

Dane:

2000 największych spółek światowych z 2004 (*Forbes Magazine*)

```
data("Forbes2000",package="HSAUR2")
attach(Forbes2000)
```

```
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 Group"
"ExxonMobil" ...
 $ country    : Factor w/ 61 levels "Africa","Australia",..: 60 60 60 60 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 ...
```

```
summary(sales)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.010   2.018   4.365   9.697   9.547 256.300
```

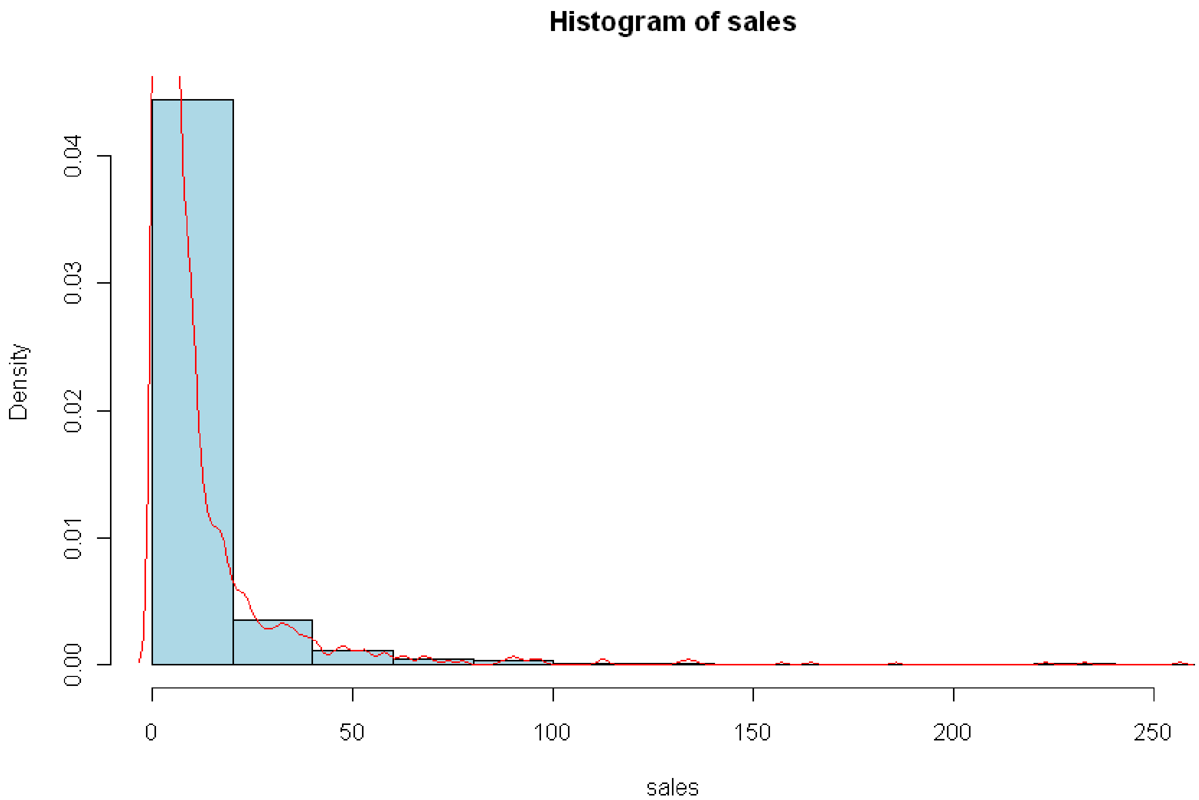
```
name[country=="Poland"]
```

```
[1] "PKN Orlen"
```

```
sales[country=="Poland"]
```

```
[1] 4.41
```

```
hist(sales,prob=T,col="lightblue")
lines(density(sales),col=2)
```



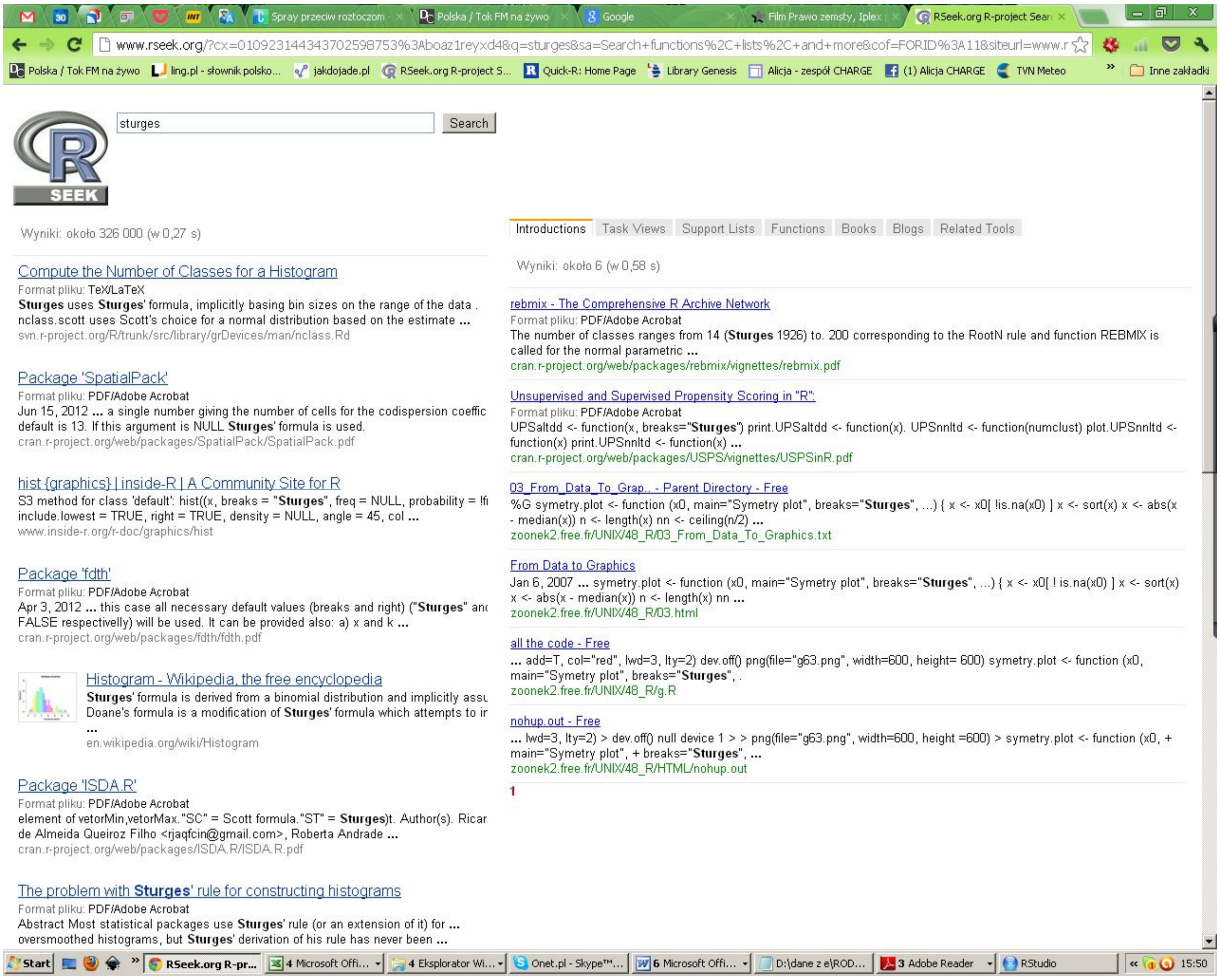
Dobór liczby klas:

Sturges: $\lceil \log_2 n + 1 \rceil$

Scott: $3.5 \frac{s}{\sqrt[3]{n}}$ (nie dla dużych n!)

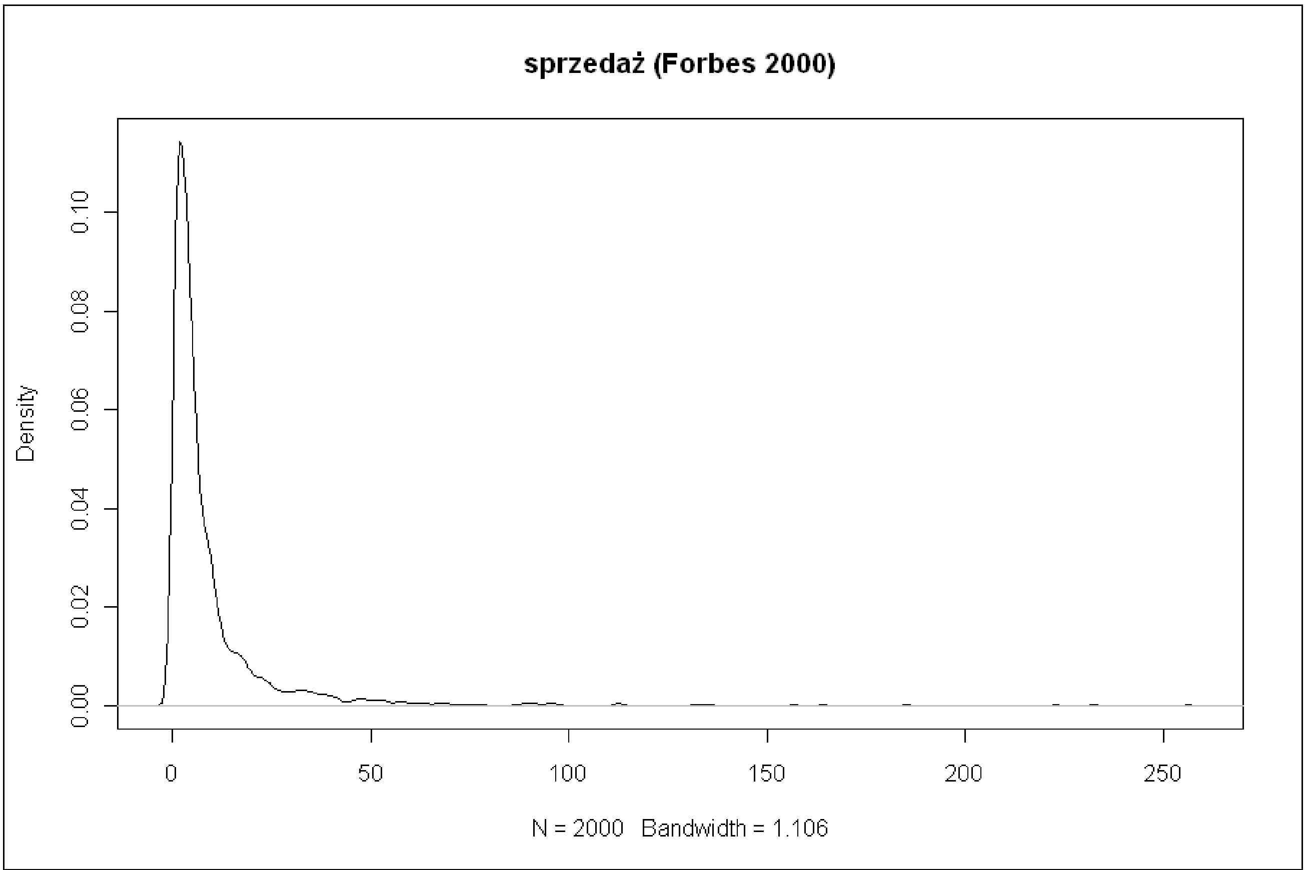
Freedman, Diaconis: $2 \frac{IQ}{\sqrt[3]{n}}$

http://www.rseek.org/



http://www.statmethods.net/index.html

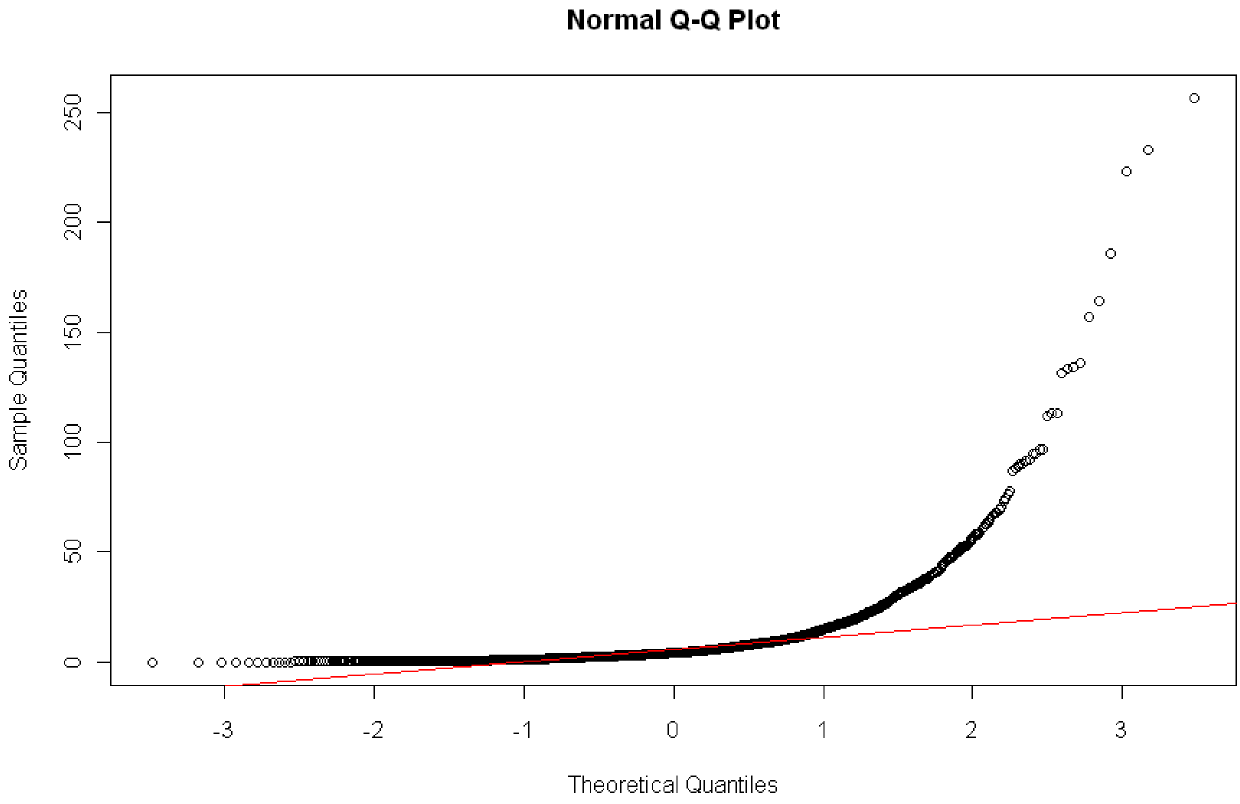
```
plot(density(sales), main="sprzedaż (Forbes 2000)")
```




```
cat(" średnia sprzedaż [mld$]=",mean(sales),"\n",
   "mediana sprzedaży [mld$]=",median(sales))
```

```
średnia sprzedaż [mld$]= 9.69701
mediana sprzedaży [mld$]= 4.365
```

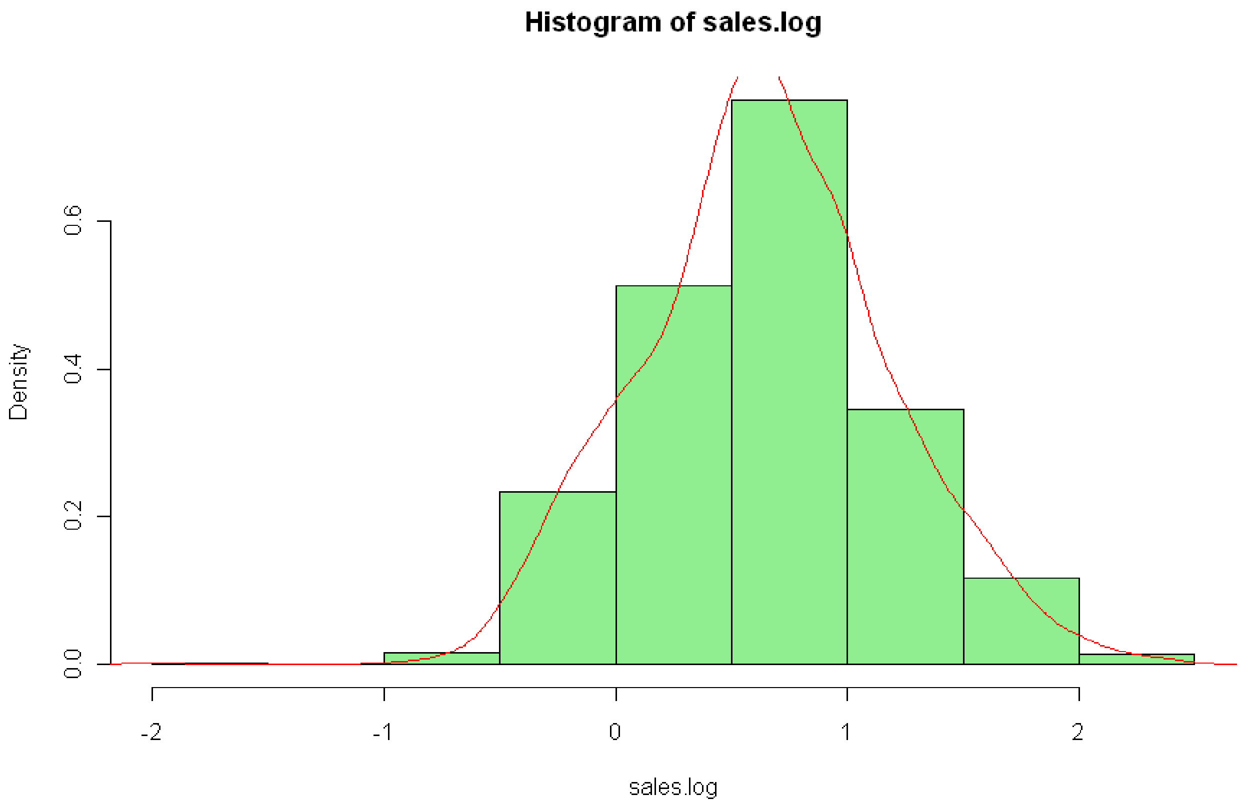
```
qqnorm(sales)
qqline(sales,col=2)
```



*

```
sales.log <- log10(sales)
```

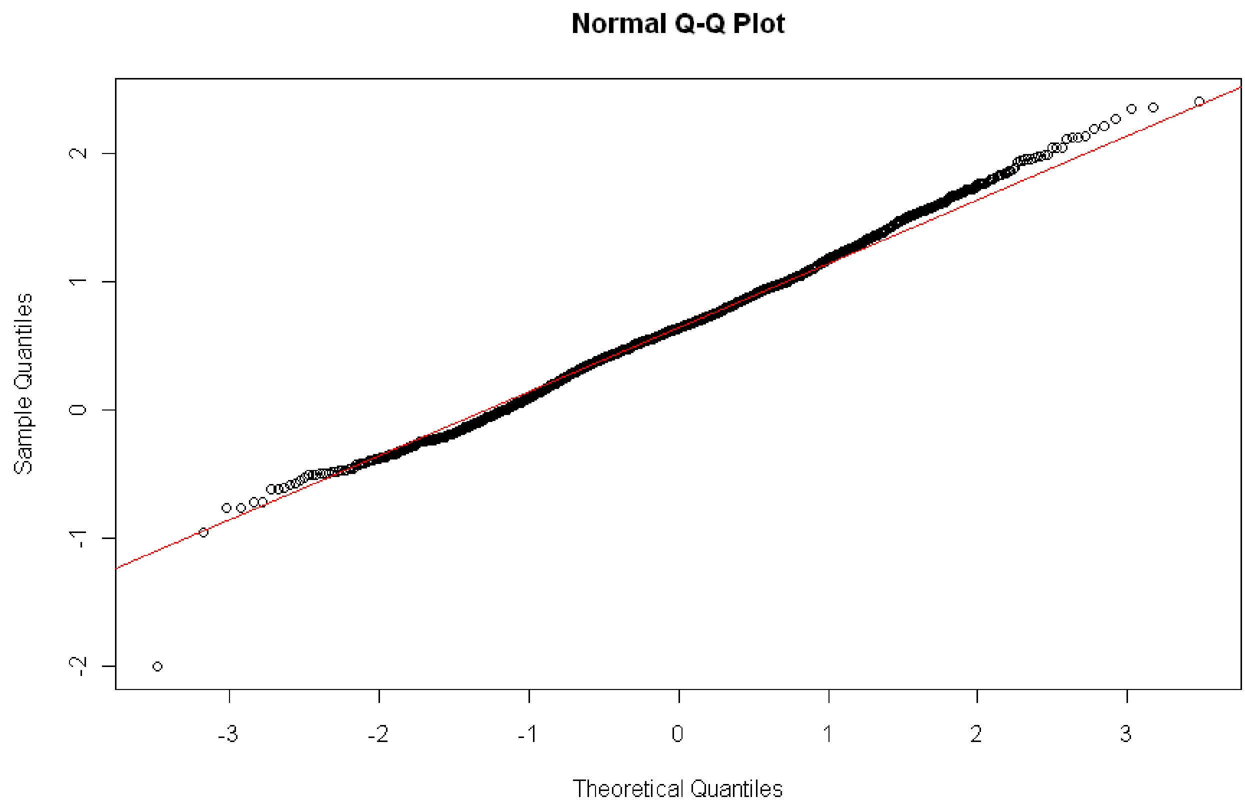
```
hist(sales.log,prob=T,col="lightgreen")
lines(density(sales.log),col=2)
```




```
cat("SKALA LOGARYTMICZNA","\n", "średnia sprzedaż [mld$]=",mean(sales.log),"\n",
   "mediana sprzedaży [mld$]=",median(sales.log))
```

```
SKALA LOGARYTMICZNA
 średnia sprzedaż [mld$]= 0.646992
 mediana sprzedaży [mld$]= 0.639984
```

```
sales.qq<-qqnorm(sales.log)
qqline(sales.log,col=2)
```




```
cat("MEDIANA W SKALI ORYGINALNEJ" ,"\n",median(sales),"\n",
   "POWRÓT ZE SKALI LOGARYTMICZNEJ","\n",10^median(sales.log))
```

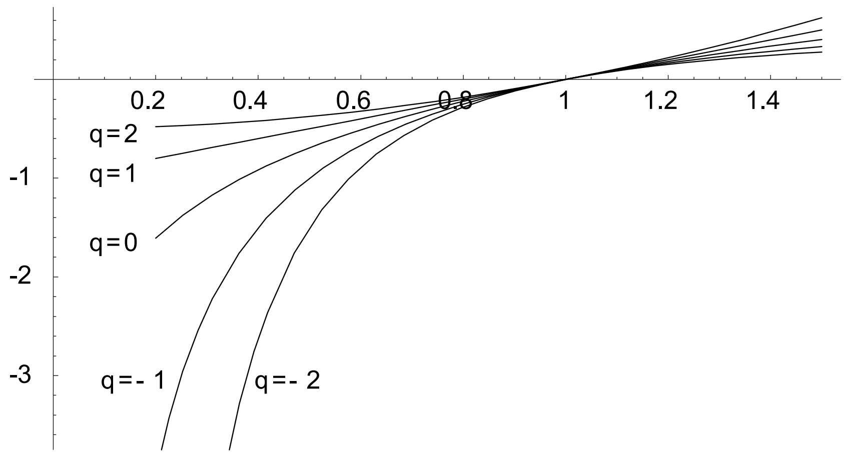
```
MEDIANA W SKALI ORYGINALNEJ
 4.365
 POWRÓT ZE SKALI LOGARYTMICZNEJ
 4.364997
```

```
summary(lm(sales.qq$y~sales.qq$x))
```

```
Call:
lm(formula = sales.qq$y ~ sales.qq$x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.79024 -0.02007 -0.00356  0.01685  0.19621 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.6469920  0.0007160   903.6   <2e-16 ***
sales.qq$x  0.5334340  0.0007163   744.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.03202 on 1998 degrees of freedom
Multiple R-squared: 0.9964,     Adjusted R-squared: 0.9964 
F-statistic: 5.547e+05 on 1 and 1998 DF,  p-value: < 2.2e-16
```

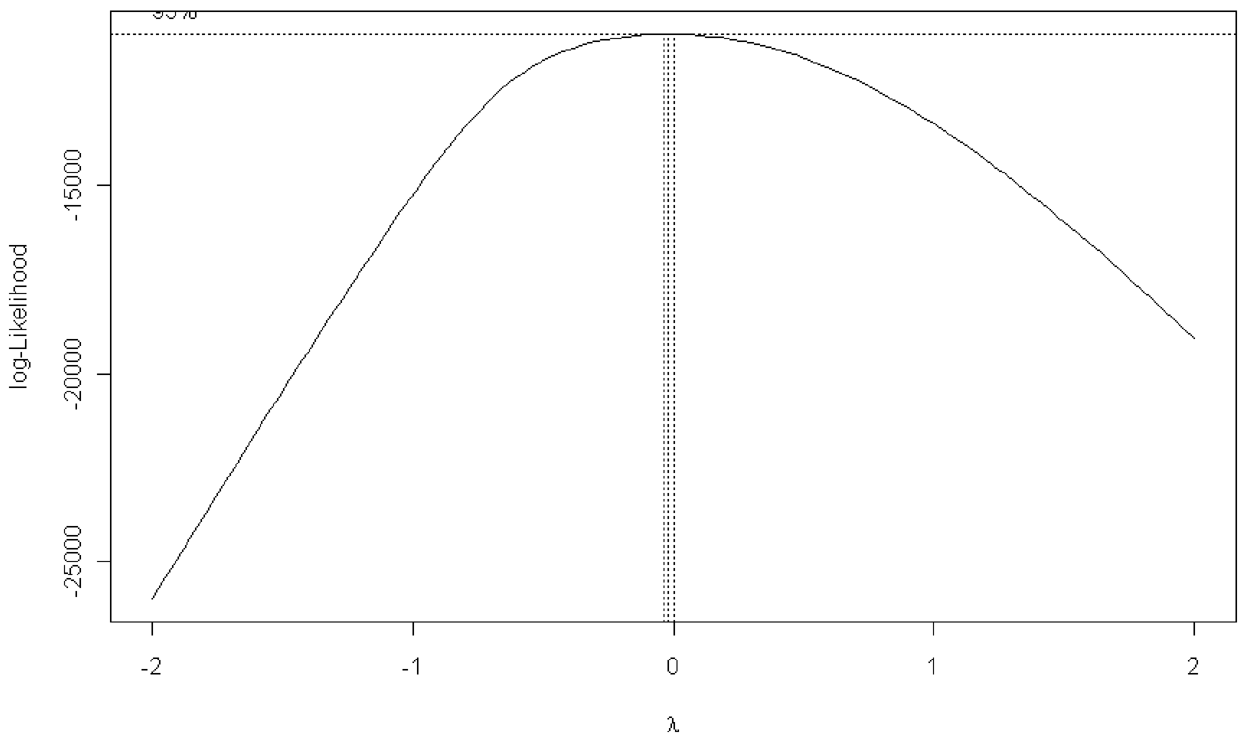
Rodzina przekształceń Boxa-Coxa

$$h_q(x) = \begin{cases} \dfrac{x^q - 1}{q} & q \neq 0 \\ \ln(x) & q = 0 \end{cases}$$

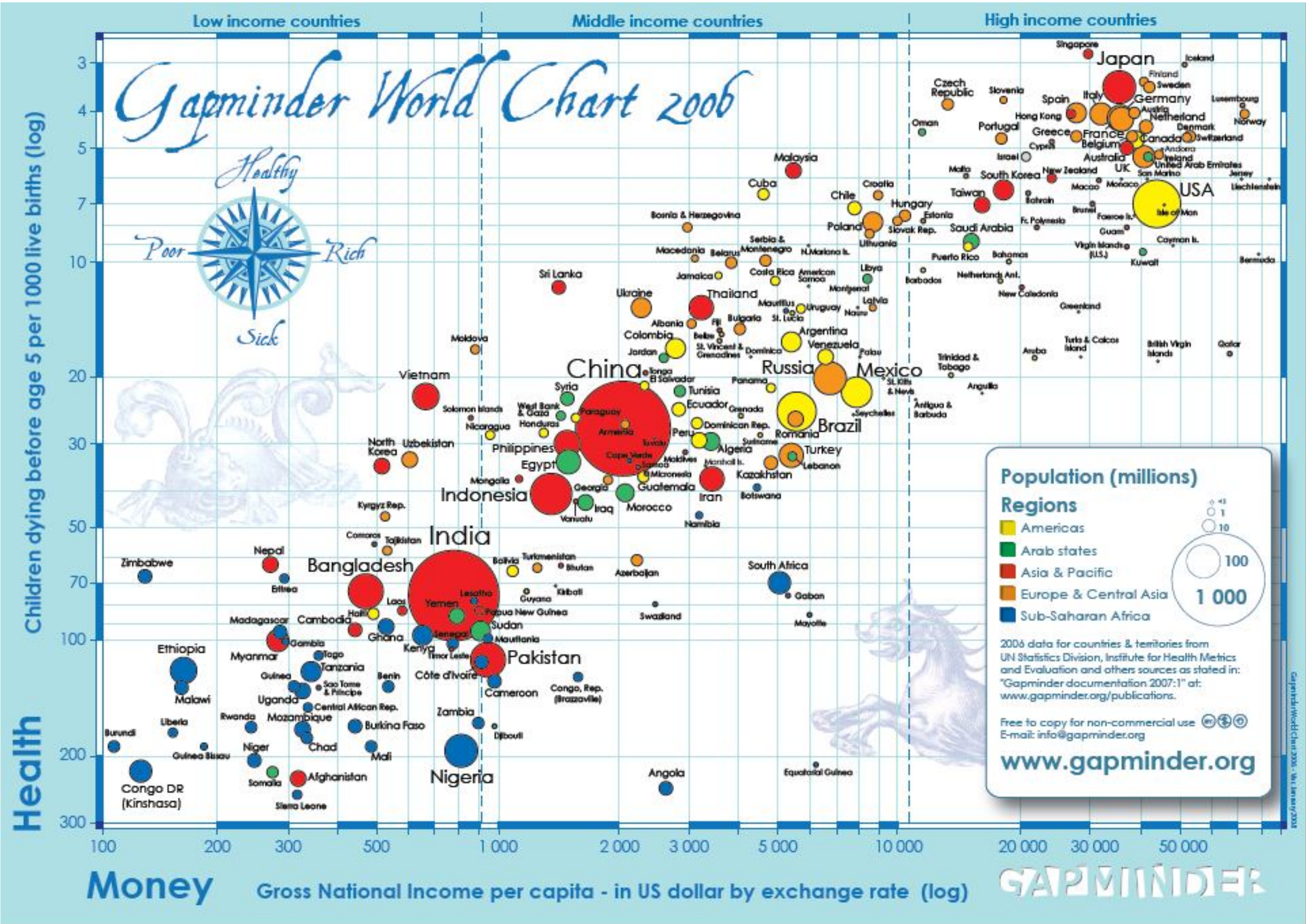


*

```
library(MASS)
sales.bc<-boxcox(sales~sales)
sales.bc$x[which.max(sales.bc$y)]
```




```
[1] -0.02020202
```


Gapminder World Chart 2006
Low income countries
Middle income countries
High income countries
Healthy
Poor
Rich
Sick
Children dying before age 5 per 1000 live births (log)
Health
Money
Gross National Income per capita - in US dollar by exchange rate (log)
Population (millions)
Regions
Americas
Arab states
Asia & Pacific
Europe & Central Asia
Sub-Saharan Africa
100
1 000
2006 data for countries & territories from UN Statistics Division, Institute for Health Metrics and Evaluation and others sources as stated in: "Gapminder documentation 2007:1" at: www.gapminder.org/publications.
Free to copy for non-commercial use
E-mail: info@gapminder.org
www.gapminder.org
GAPMINDER
Japan
USA
China
India
Russia
Mexico
Brazil
Indonesia
Bangladesh
Pakistan
Nigeria
Ethiopia
Congo DR (Kinshasa)
Vietnam
Thailand
Iran
Turkey
Egypt
Philippines
South Africa
Angola
Zimbabwe
Afghanistan
Germany
France
Italy
Spain
UK
Canada
Australia
South Korea
Malaysia
Sri Lanka
Ukraine
Argentina
Colombia
3
4
5
7
10
20
30
50
70
100
200
300
100
200
300
500
1 000
2 000
3 000
5 000
10 000
20 000
30 000
50 000

```
levels(category)
```

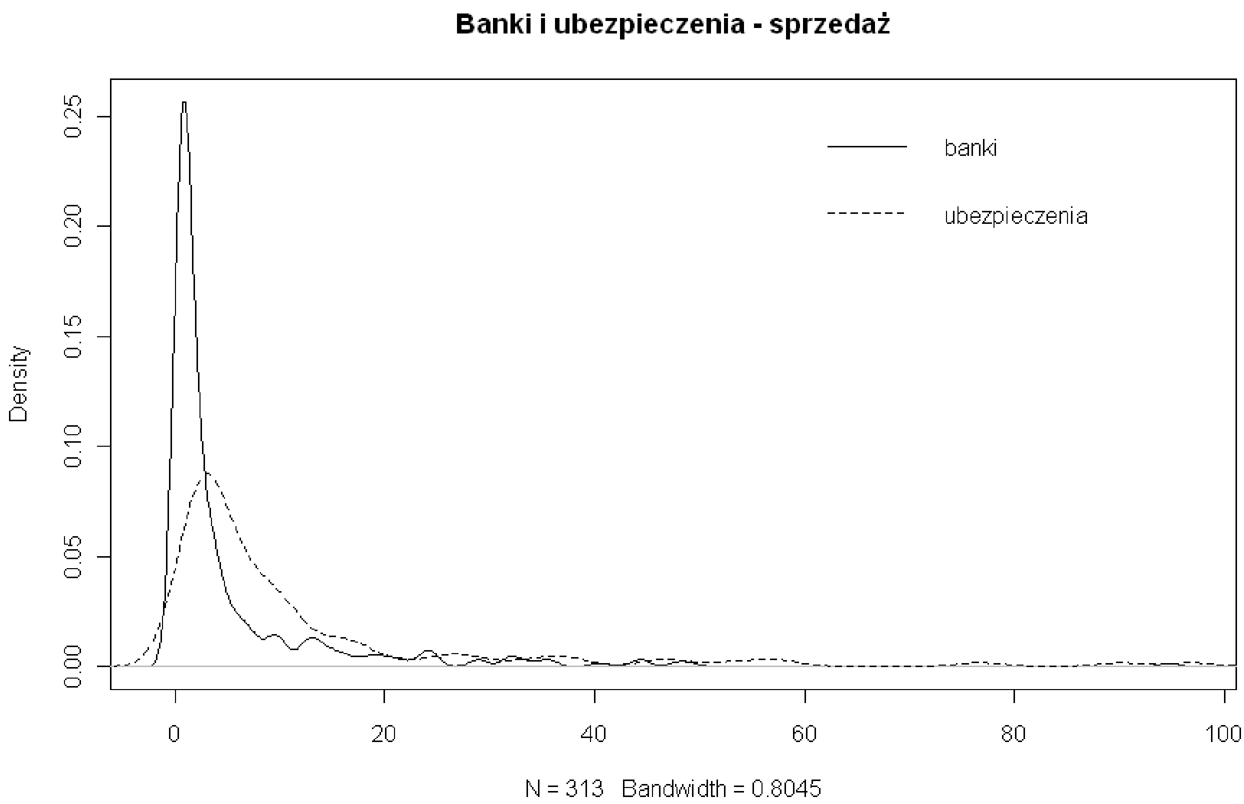
```
[1]  "Aerospace & defense"                "Banking"
 [3] "Business services & supplies"     "Capital goods"
 [5] "Chemicals"                        "Conglomerates"
 [7] "Construction"                     "Consumer durables"
 [9] "Diversified financials"           "Drugs & biotechnology"
[11] "Food 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"
```

```
sales.Banking <- sales[category= ="Banking"]
sales.Insurance <- sales[category= ="Insurance"]
```

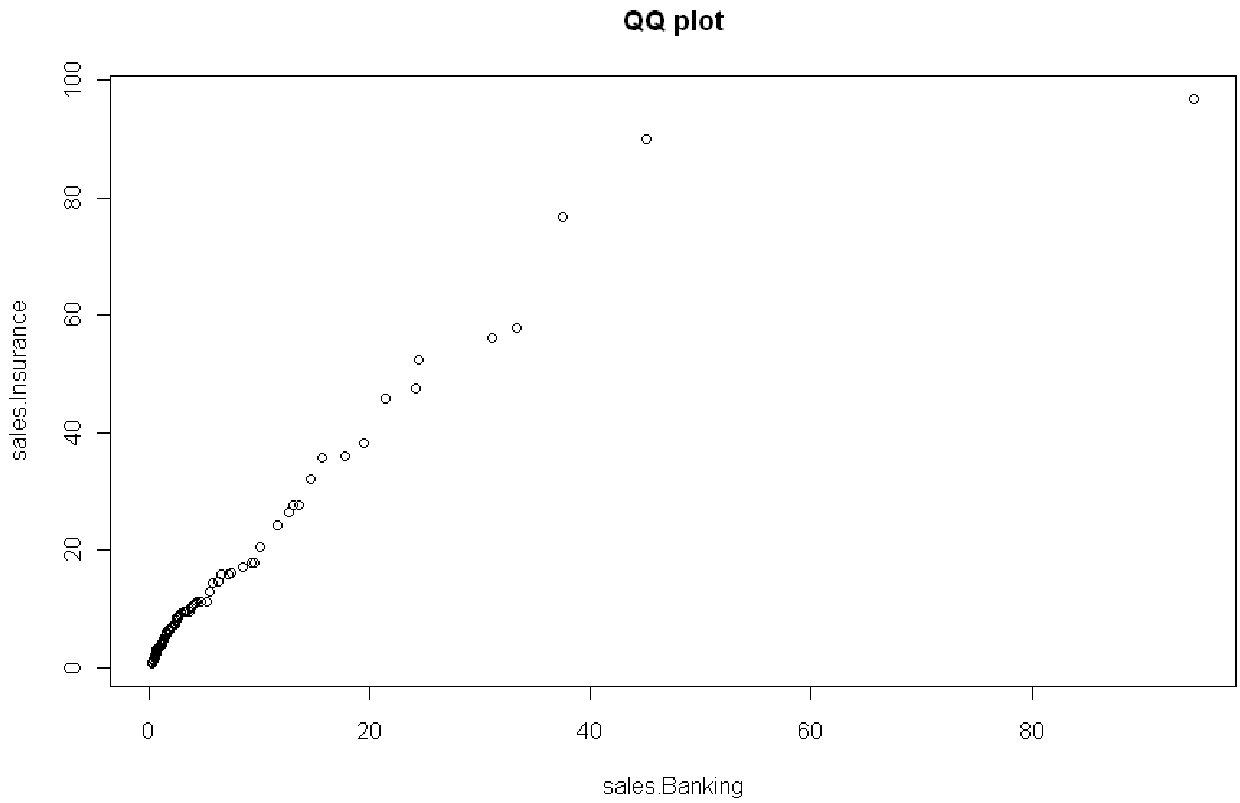
```
cat(" banki",length(sales.Banking),"\n","ubezpieczenia",length(sales.Insurance))
```

```
 banki 313
 ubezpieczenia 112
```

```
dBanking <- density(sales.Banking)
dInsurance <- density(sales.Insurance)
plot(dBanking,lty=1,main="Banki i ubezpieczenia - sprzedaż")
lines(dInsurance,lty=2)
legend("topright",lty=1:2,legend=c("banki","ubezpieczenia"),bty="n")
```



```
bi.qq <- qqplot(sales.Banking,sales.Insurance,main="QQ plot")
```



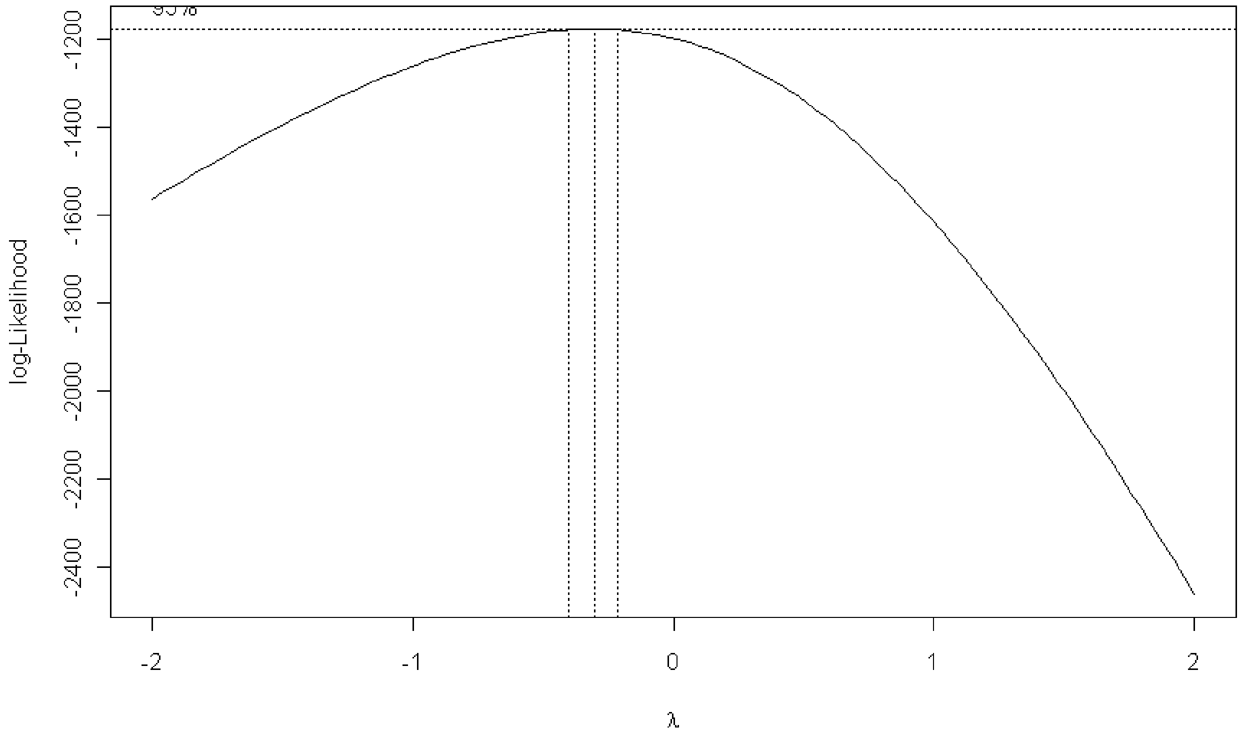

```
name[sales>80 & category= ="Banking"]
```

```
[1] "Citigroup"
```

```
summary(lm(bi.qq$y~bi.qq$x))
```

```
Call:
lm(formula = bi.qq$y ~ bi.qq$x)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.926  -2.130  -0.918   0.982  22.510 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.96172    0.61679   6.423 3.51e-09 ***
bi.qq$x      1.41321    0.04755  29.719  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 5.873 on 110 degrees of freedom
Multiple R-squared: 0.8893,	Adjusted R-squared: 0.8882 
F-statistic: 883.2 on 1 and 110 DF,  p-value: < 2.2e-16
```

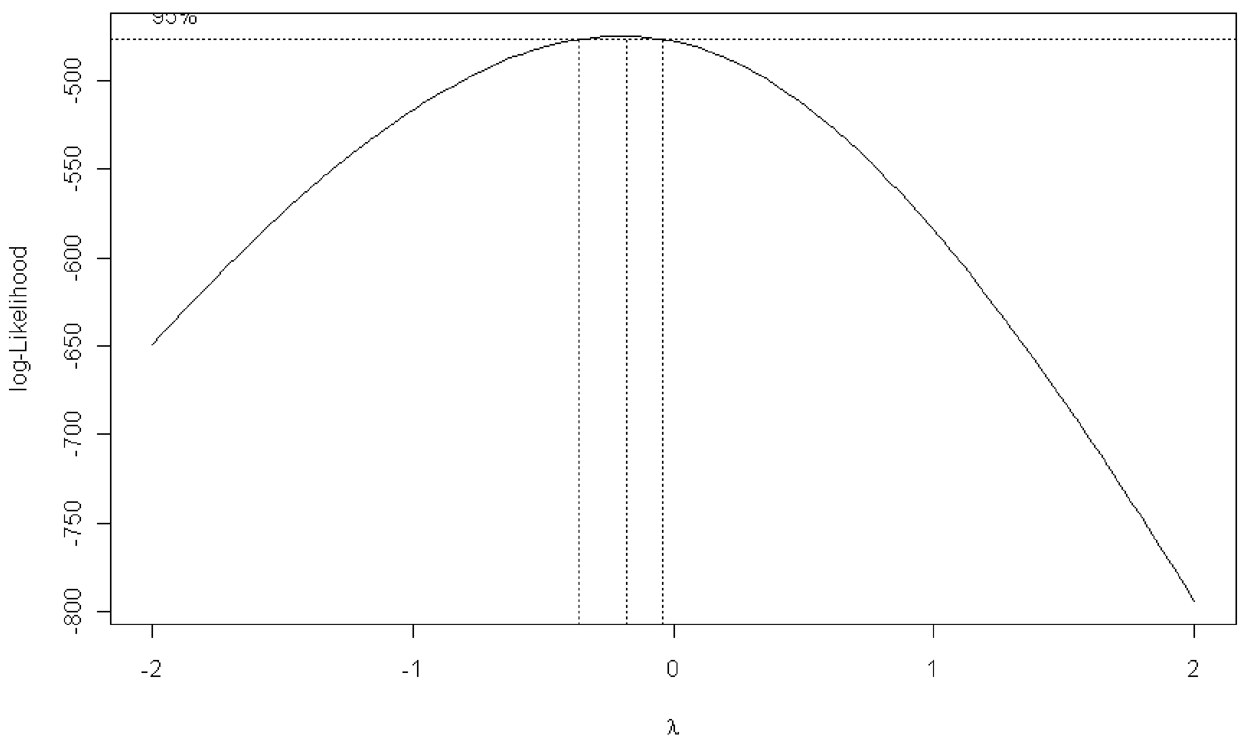
```
salesB.bc<-boxcox(sales.Banking~sales.Banking)
```




```
salesB.bc$x[which.max(salesB.bc$y)]
```

```
[1] -0.3030303
```

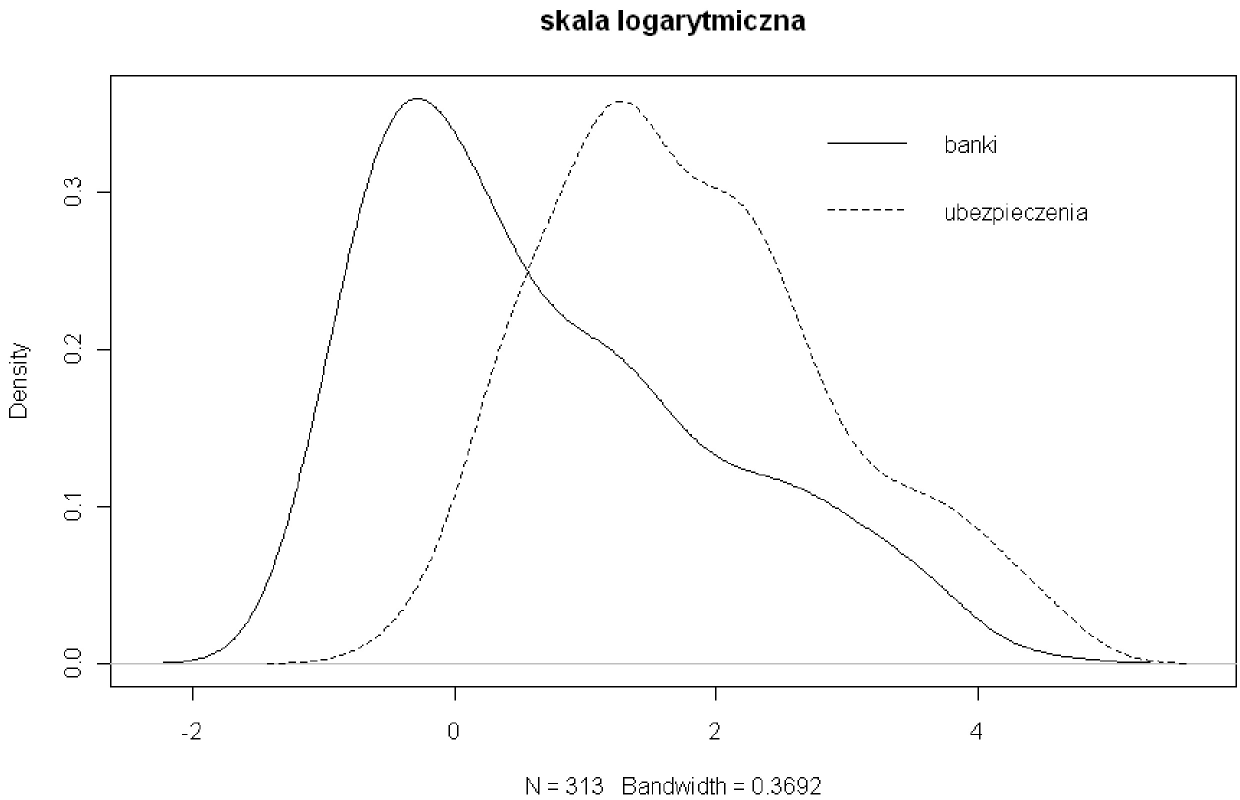
```
salesI.bc <- boxcox(sales.Insurance~sales.Insurance)
```



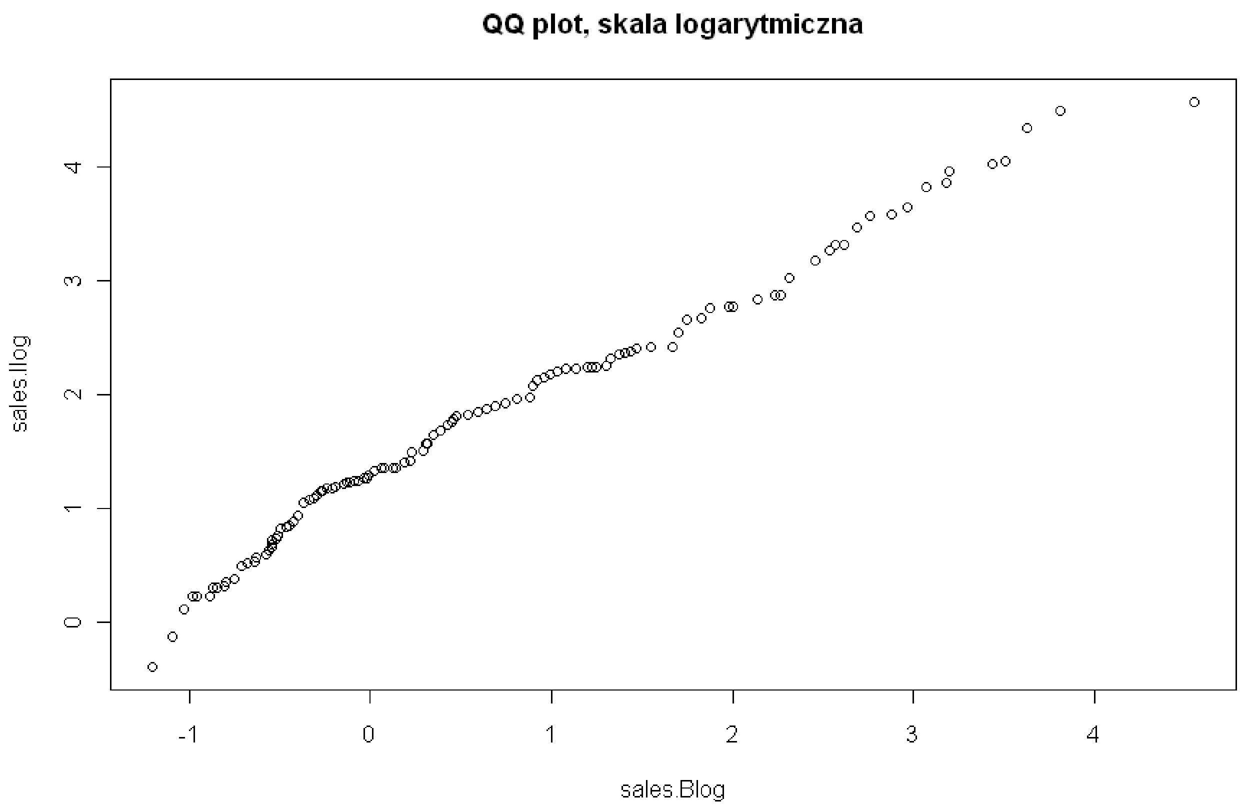

```
salesI.bc$x[which.max(salesI.bc$y)]
```

```
[1] -0.1818182
```

```
sales.Blog <- log(sales.Banking)
sales.Ilog <- log(sales.Insurance)
dBankingLog <- density(sales.Blog)
dInsuranceLog <- density(sales.Ilog)
plot(dBankingLog,lty=1,main="skala logarytmiczna")
lines(dInsuranceLog,lty=2)
legend("topright",lty=1:2,legend=c("banki","ubezpieczenia"),bty="n")
```



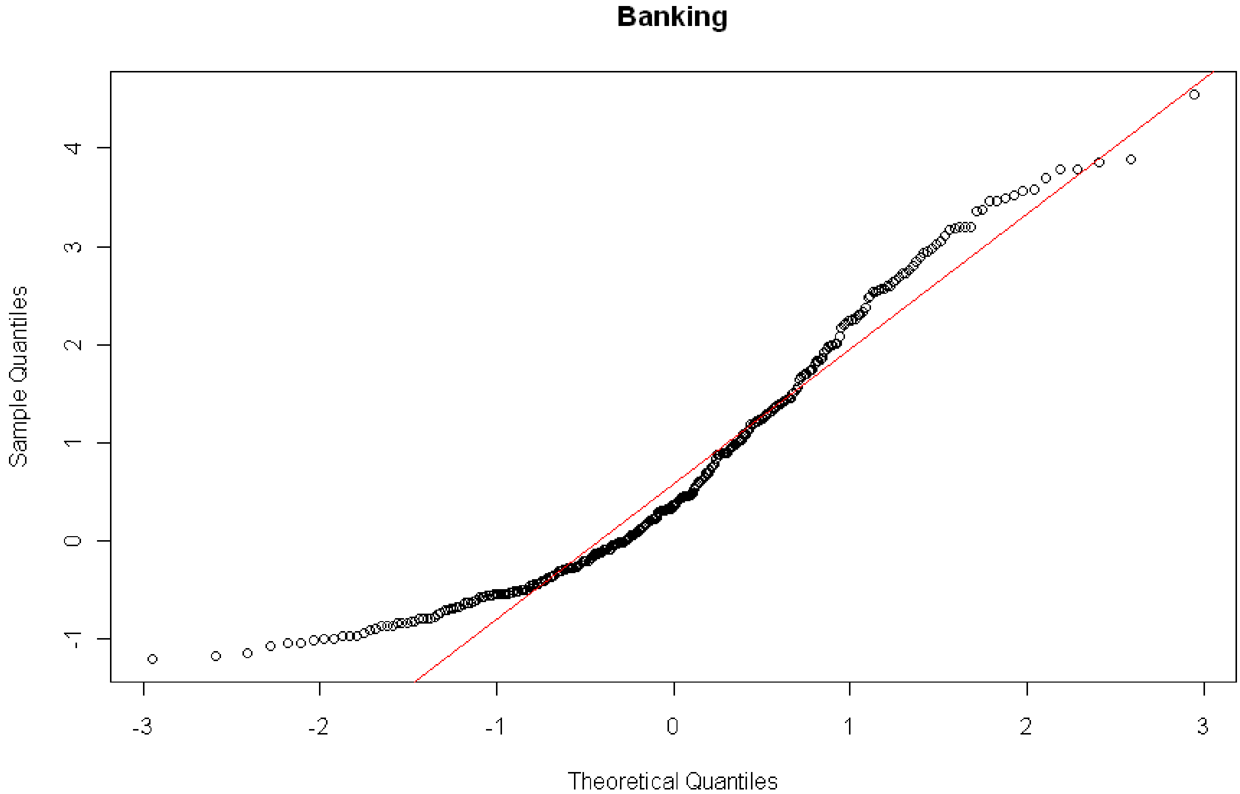
```
bilog.qq <- qqplot(sales.Blog,sales.Ilog,main="QQ plot, skala logarytmiczna")
```




```
summary(lm(bilog.qq$y~bilog.qq$x))
```

```
Call:
lm(formula = bilog.qq$y ~ bilog.qq$x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.59671 -0.08138  0.02436  0.11100  0.20260 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.21536    0.01526   79.67   <2e-16 ***
bilog.qq$x   0.83417    0.01022   81.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.1421 on 110 degrees of freedom
Multiple R-squared: 0.9837,	Adjusted R-squared: 0.9836 
F-statistic:  6656 on 1 and 110 DF,  p-value: < 2.2e-16
```

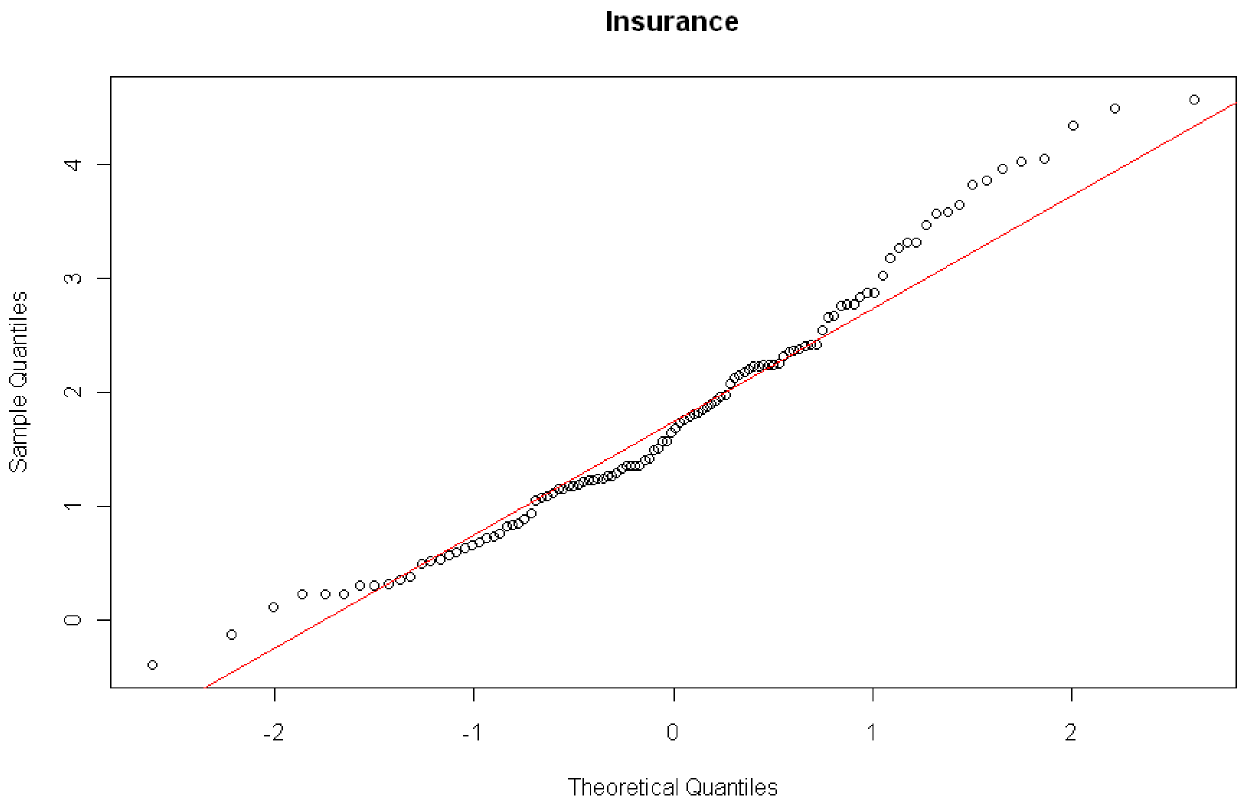
```
salesBlog.qq<-qqnorm(sales.Blog, main="Banking")
qqline(sales.Blog,col=2)
```




```
summary(lm(salesBlog.qq$y~salesBlog.qq$x))
```

```
Call:
lm(formula = salesBlog.qq$y ~ salesBlog.qq$x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.36679 -0.28317 -0.08461  0.21483  1.78173 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     0.70247    0.01900   36.97   <2e-16 ***
salesBlog.qq$x  1.25094    0.01904   65.70   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.3362 on 311 degrees of freedom
Multiple R-squared: 0.9328,    Adjusted R-squared: 0.9326 
F-statistic:  4316 on 1 and 311 DF,  p-value: < 2.2e-16
```

```
salesIlog.qq<-qqnorm(sales.Ilog, main="Insurance")
qqline(sales.Ilog,col=2)
```




```
summary(lm(salesIlog.qq$y~salesIlog.qq$x))
```

```
Call:
lm(formula = salesIlog.qq$y ~ salesIlog.qq$x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25854 -0.10931 -0.05309  0.03546  0.67075 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.80674    0.01749  103.33   <2e-16 ***
salesIlog.qq$x  1.09499    0.01759   62.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.185 on 110 degrees of freedom
Multiple R-squared: 0.9724,     Adjusted R-squared: 0.9722 
F-statistic:  3877 on 1 and 110 DF,  p-value: < 2.2e-16
```

```
shapiro.test(sales.Ilog)
```

```
        Shapiro-Wilk normality test

data:  sales.Ilog
W = 0.9701, p-value = 0.01303
```

PARADOKS SZKŁA POWIĘKSZAJĄCEGO

```
shapiro.test(sales.Blog)
```

```
        Shapiro-Wilk normality test

data:  sales.Blog
W = 0.9316, p-value = 8.237e-11
```

```
var.test(sales.Blog,sales.Ilog)
```

```
        F test to compare two variances

data:  sales.Blog and sales.Ilog
F = 1.3626, num df = 312, denom df = 111, p-value = 0.05714
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9905142 1.8323970
sample estimates:
ratio of variances
          1.362628
```

```
t.test(sales.Blog,sales.Ilog, var.equal=T, paired=F)
```

```
        Two Sample t-test

data:  sales.Blog and sales.Ilog
t = -8.0323, df = 423, p-value = 9.599e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.374499 -0.834045
sample estimates:
mean of x mean of y
0.7024724 1.8067445
```

```
c(exp(-1.374499),exp(-0.834045))
```

```
[1] 0.2529663 0.4342890
```

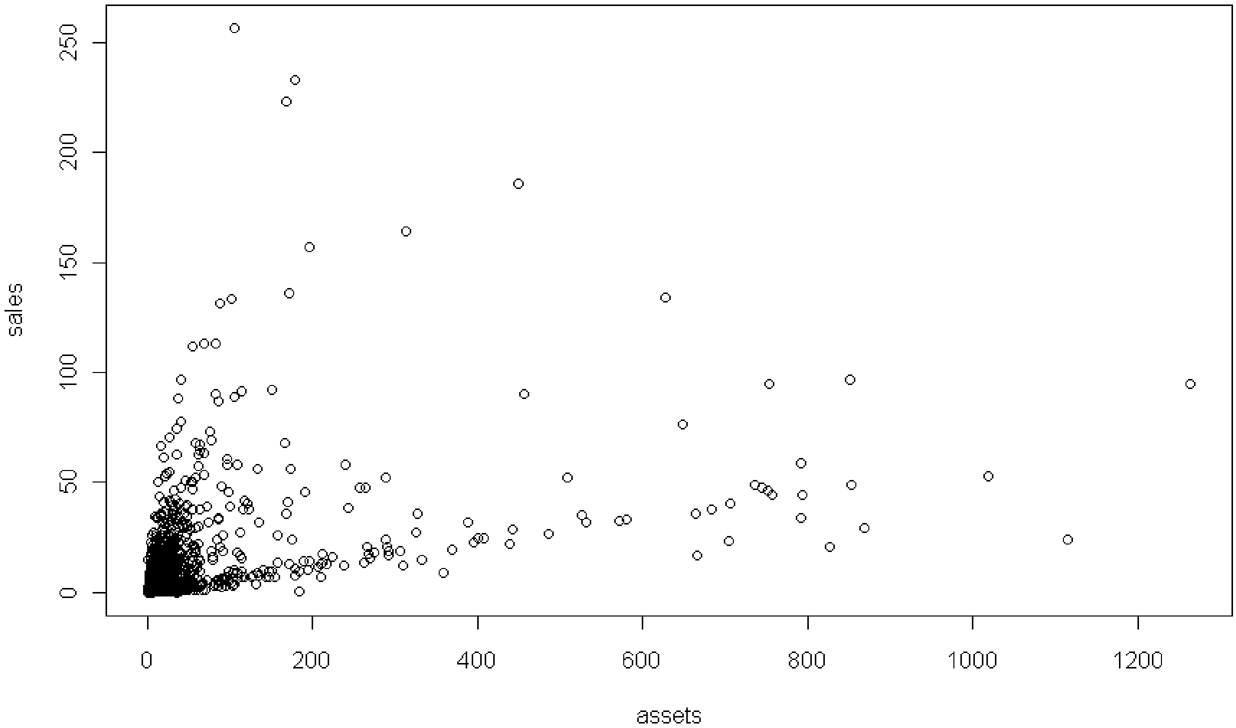
RELACJE MIĘDZY ZMIENNYMI

Majątek (assets) i wysokość sprzedaży (sales)

```
summary(assets)
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.270    4.025    9.345   34.040   22.790 1264.000
```

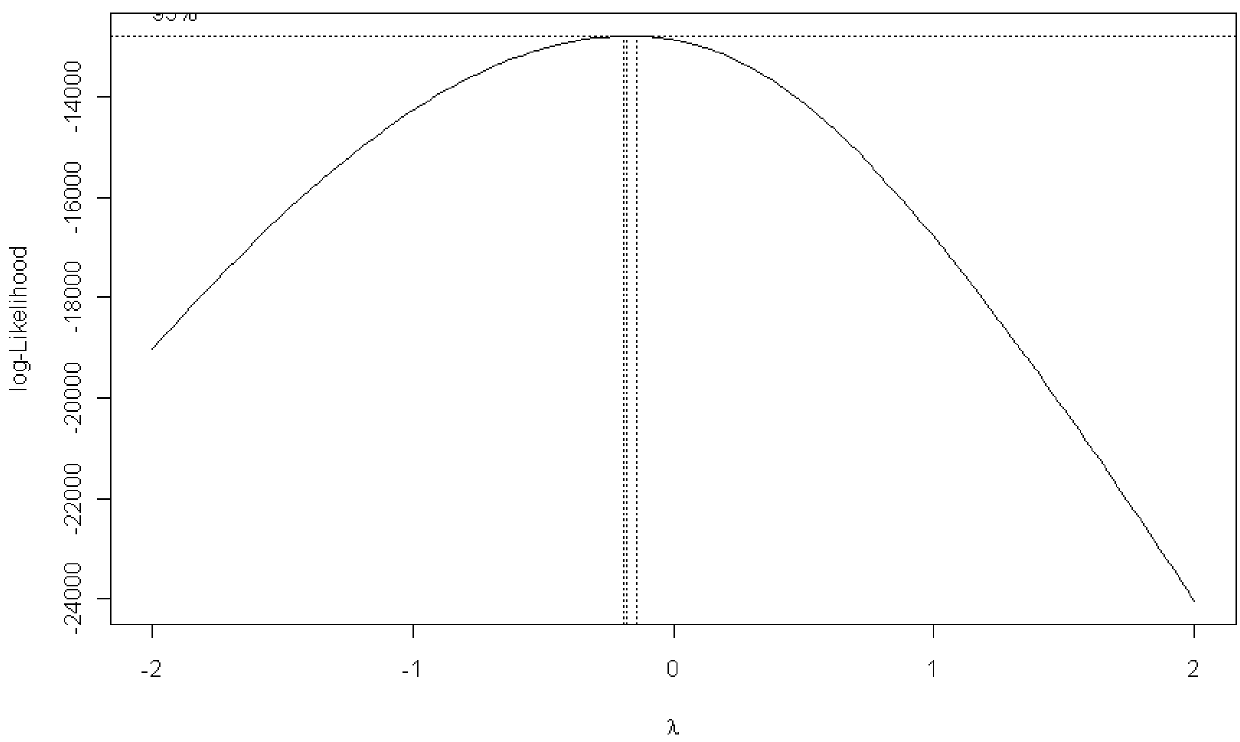
```
plot(assets,sales)
```




```
name[assets>1200]
```

```
[1] "Citigroup"
```

```
assets.bc <- boxcox(assets~assets)
```

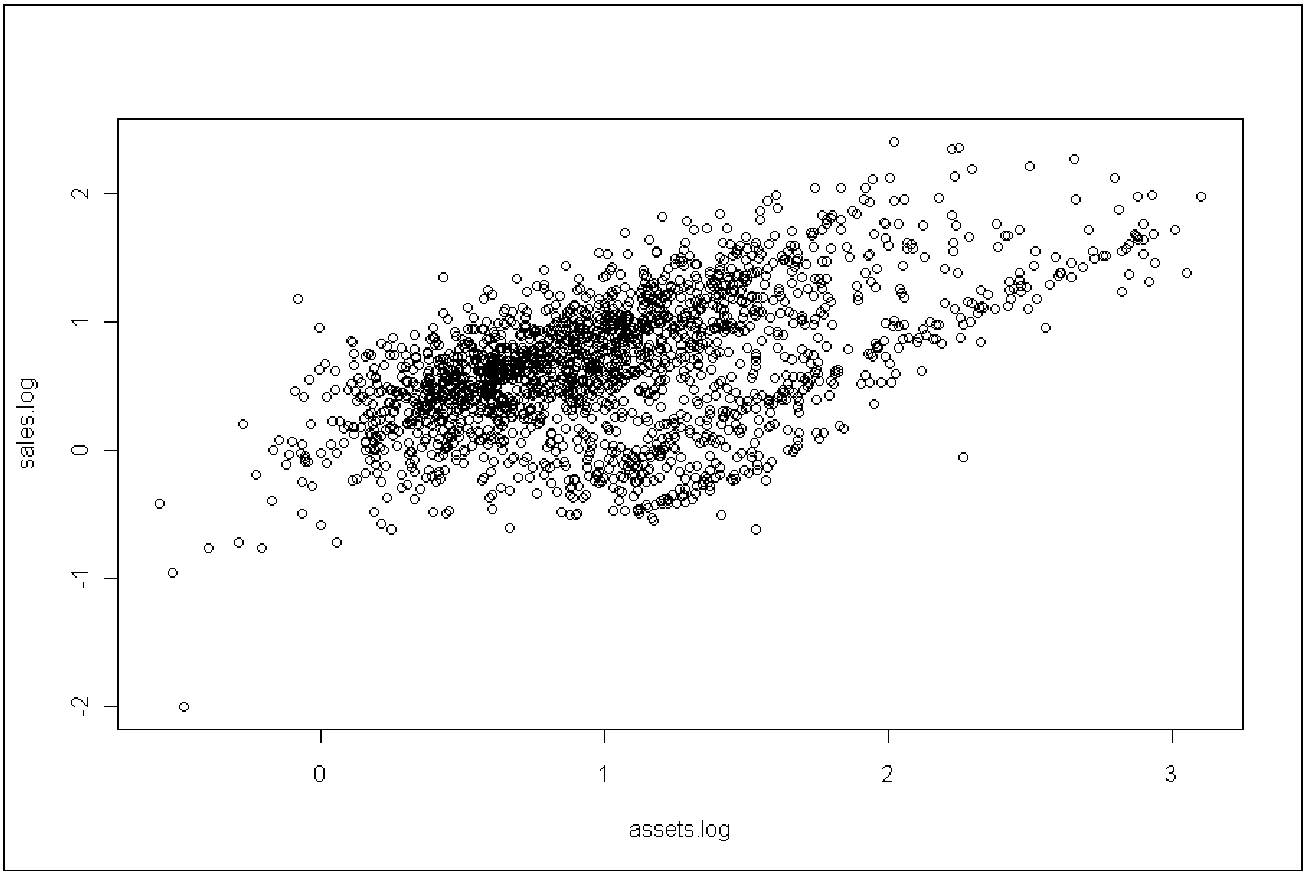



```
assets.bc$x[which.max(assets.bc$y)]
```

```
[1] -0.1818182
```

```
assets.log  <-  log10(assets)
```

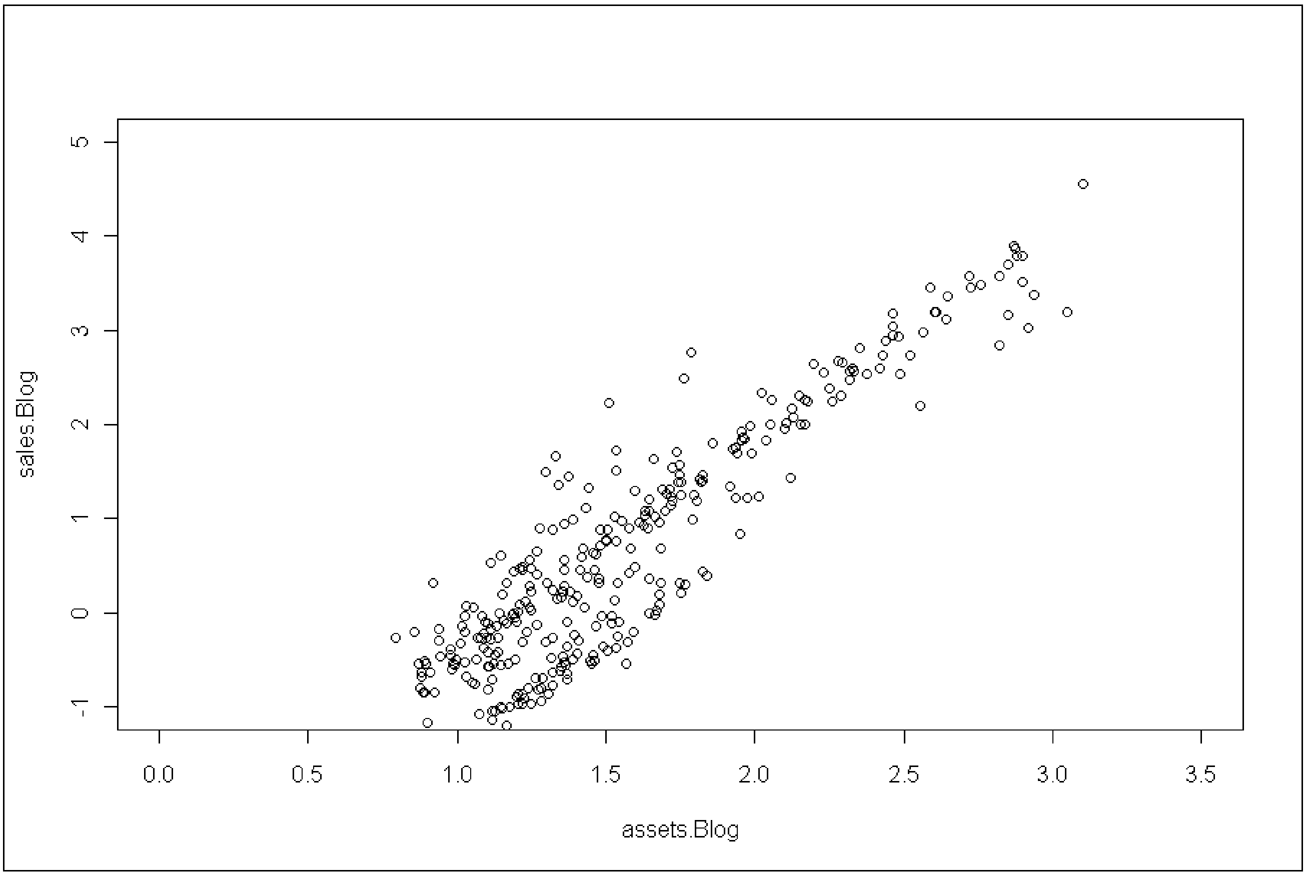
```
plot(assets.log,sales.log)
```



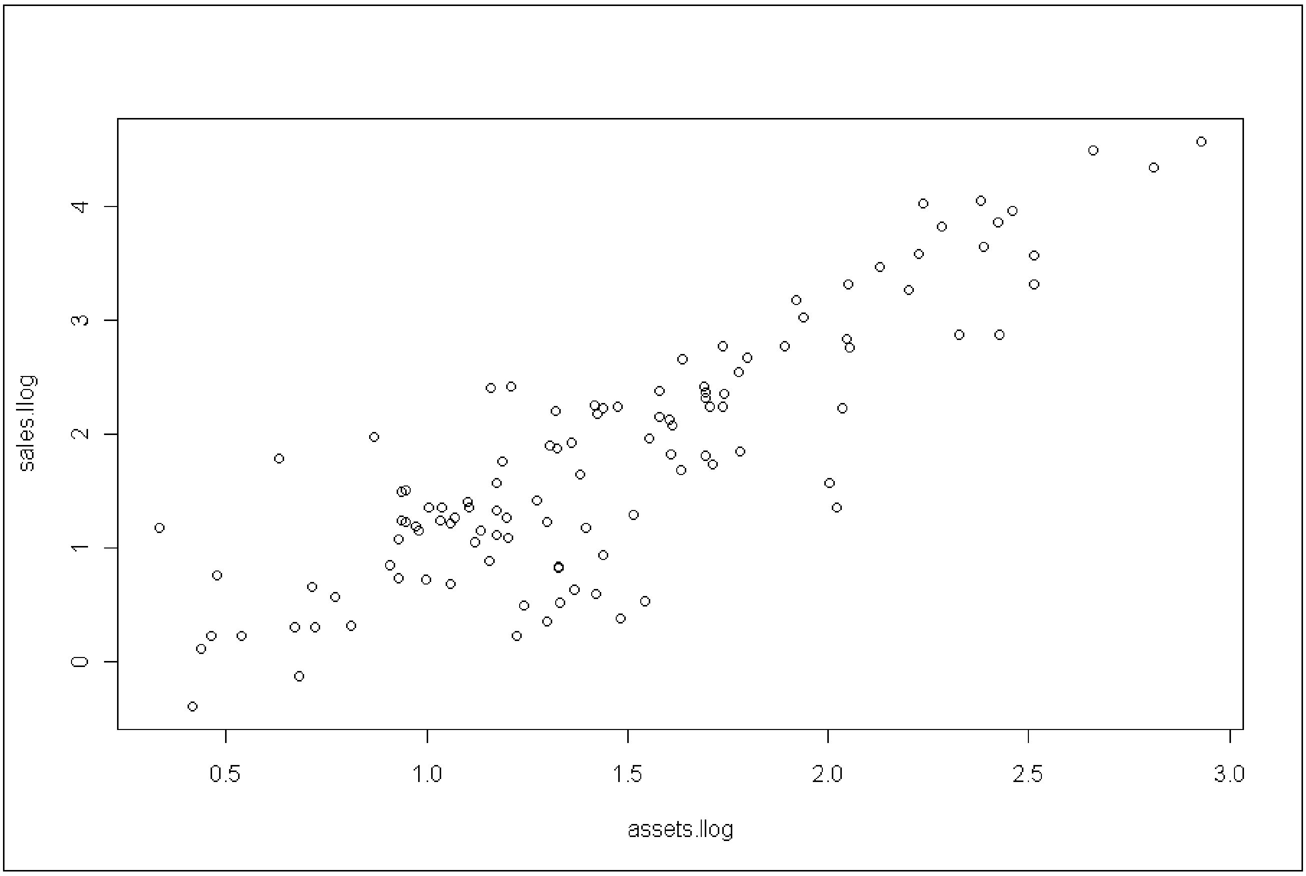

```
assets.Blog <- assets.log[category= ="Banking"]
```

```
assets.Ilog <- assets.log[category= ="Insurance"]
```

```
plot(assets.Blog,sales.Blog)
```



plot(assets.Ilog,sales.Ilog)



```
minmax <- function(x) c(min(x),max(x))
```

```
minmax(assets.Blog)
```

```
[1] 0.7916906 3.1017574
```

```
minmax(assets.Ilog)
```

```
[1] 0.3344538 2.9300520
```

```
minmax(sales.Blog)
```

```
[1] -1.203973  4.550820
```

```
minmax(sales.Ilog)
```

```
[1] -0.3856625  4.5734731
```

```
plot(assets.Blog, sales.Blog, xlim=c(0,3.5), ylim=c(-1.5,5), col=2,
    xlab="assets[log]", ylab="sales[log]")
```

```
points(assets.Ilog, sales.Ilog, col=3)
```

```
legend("bottomright", col=c(2,3), pch=c(1,1),
      legend=c("banki","ubezpieczenia"), bty="n")
```