Χρυσόστομος Πανής **ΑΕΜ: 3219**

Τα δεδομένα αποτελούνται από τις παρακάτω μεταβλητές **age, Sex, year_emp, income, debt_income, cred_debt, other_debt, default, level, granted, Scoring**.
Σε πρώτη φάση ανάλυσης θεωρούμαι τη Scoring την εξαρτημένη μεταβλητή και όλες τις υπόλοιπες ανεξάρτητες. Οι κατηγορικές μεταβλητές είναι οι **age, default, level, granted,** και ποσοτικές οι υπόλοιπες. Η σχέση της granted με τη Scoring είναι ξεκάθαρη, όταν η Scoring είναι μεγαλύτερη ή ίση του 65 τότε το granted είναι 1, διαφορετικά είναι 0.

Αρχικά κάνουμε ένα summary των δεδομένων

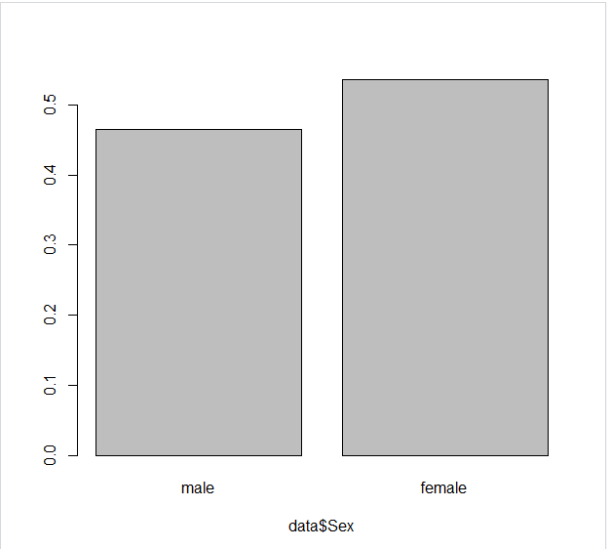
```
> summary(data)
      age             Sex           year_emp         income        debt_income
 Min.   :20.00   Min.   :1.000   Min.   : 0.000   Min.   : 14.00   Min.   : 0.400
 1st Qu.:29.00   1st Qu.:1.000   1st Qu.: 4.000   1st Qu.: 23.75   1st Qu.: 6.175
 Median :35.00   Median :2.000   Median : 7.000   Median : 36.00   Median :10.900
 Mean   :35.26   Mean   :1.536   Mean   : 8.664   Mean   : 43.81   Mean   :12.201
 3rd Qu.:41.00   3rd Qu.:2.000   3rd Qu.:14.000   3rd Qu.: 55.25   3rd Qu.:17.150
 Max.   :53.00   Max.   :2.000   Max.   :26.000   Max.   :176.00   Max.   :41.300
   cred_debt         other_debt         default       level          granted
 Min.   : 0.0000   Min.   : 0.0000   Min.   :0.0   Min.   :1.000   Min.   :0.0000
 1st Qu.: 0.3877   1st Qu.: 0.8067   1st Qu.:0.0   1st Qu.:1.000   1st Qu.:0.0000
 Median : 1.1099   Median : 1.8981   Median :0.0   Median :2.000   Median :1.0000
 Mean   : 2.6961   Mean   : 2.9138   Mean   :0.4   Mean   :2.086   Mean   :0.5286
 3rd Qu.: 2.8178   3rd Qu.: 4.0772   3rd Qu.:1.0   3rd Qu.:3.000   3rd Qu.:1.0000
 Max.   :15.0167   Max.   :14.7193   Max.   :1.0   Max.   :3.000   Max.   :1.0000
    Scoring
 Min.   : 11.00
 1st Qu.: 51.00
 Median : 65.00
 Mean   : 61.99
 3rd Qu.: 80.00
 Max.   :100.00
> |
```

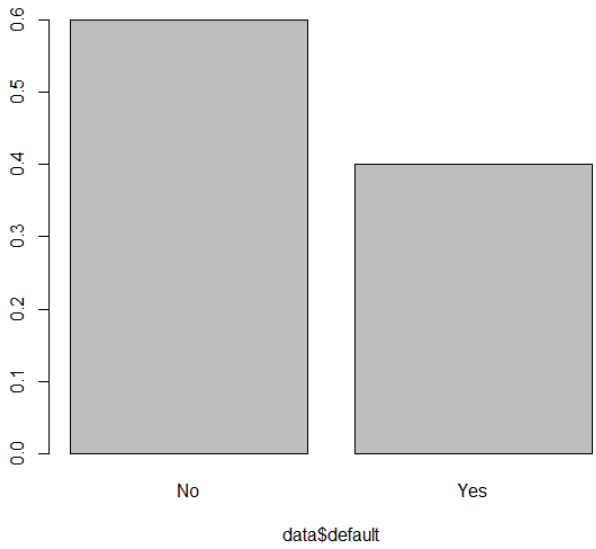
Στη συνέχεια δημιουργούμε πίνακες συχνοτήτων για τις κατηγορικές μεταβλητές.
Επιστρέφονται συχνότητες και ποσοστιαίες συχνότητες εμφάνισης μαζί με πληροφορίες σχετικά με τις ελλείπουσες παρατηρήσεις σε μορφή πίνακα

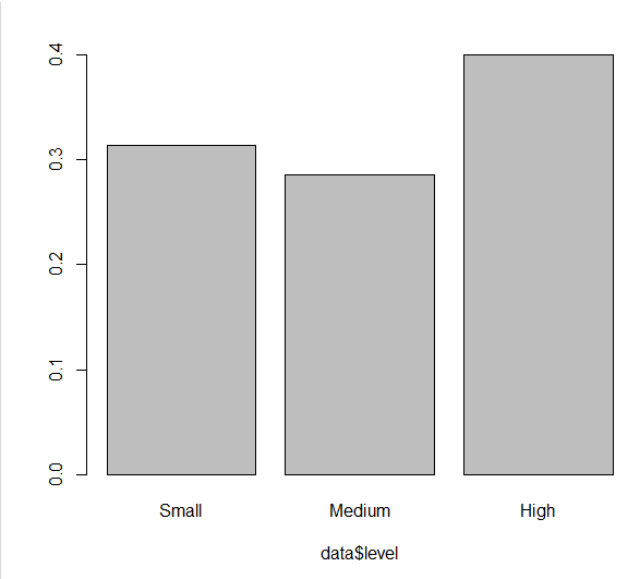
```
> freq(data$Sex)
Frequencies
data$Sex
Label: Sex
Type: Numeric

              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------- ------ --------- -------------- --------- --------------
          1     65     46.43          46.43     46.43          46.43
          2     75     53.57         100.00     53.57         100.00
      <NA>      0                               0.00         100.00
     Total    140    100.00         100.00    100.00         100.00
> freq(data$default)
Frequencies
data$default
Label: Previously defaulted
Type: Numeric

              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------- ------ --------- -------------- --------- --------------
          0     84     60.00          60.00     60.00          60.00
          1     56     40.00         100.00     40.00         100.00
      <NA>      0                               0.00         100.00
     Total    140    100.00         100.00    100.00         100.00
```

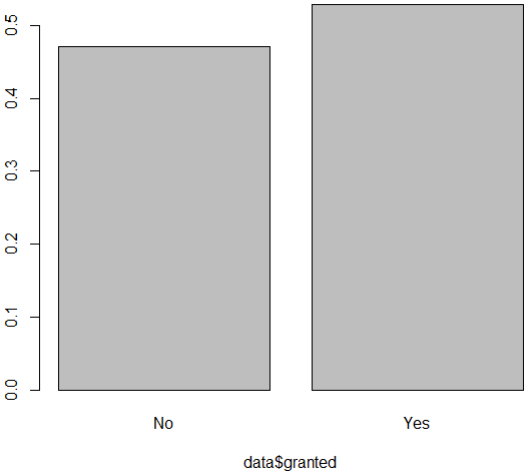
```
> freq(data$level)
Frequencies
data$level
Label: Loan amount
Type: Numeric

              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------- ------ --------- -------------- --------- --------------
          1     44     31.43          31.43     31.43          31.43
          2     40     28.57          60.00     28.57          60.00
          3     56     40.00         100.00     40.00         100.00
      <NA>      0                               0.00         100.00
     Total    140    100.00         100.00    100.00         100.00
> freq(data$granted)
Frequencies
data$granted
Label: Loan granted
Type: Numeric

              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------- ------ --------- -------------- --------- --------------
          0     66     47.14          47.14     47.14          47.14
          1     74     52.86         100.00     52.86         100.00
      <NA>      0                               0.00         100.00
     Total    140    100.00         100.00    100.00         100.00
> |
```

Ραβδογράμματα

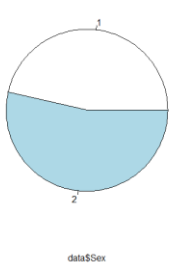


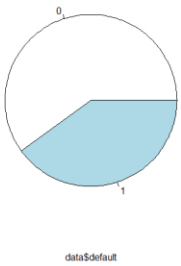


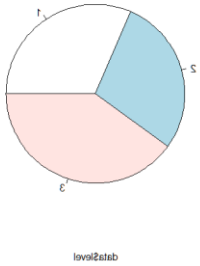


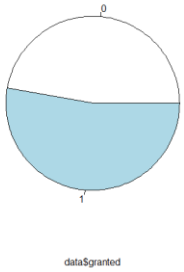


Κυκλικά διαγράμματα

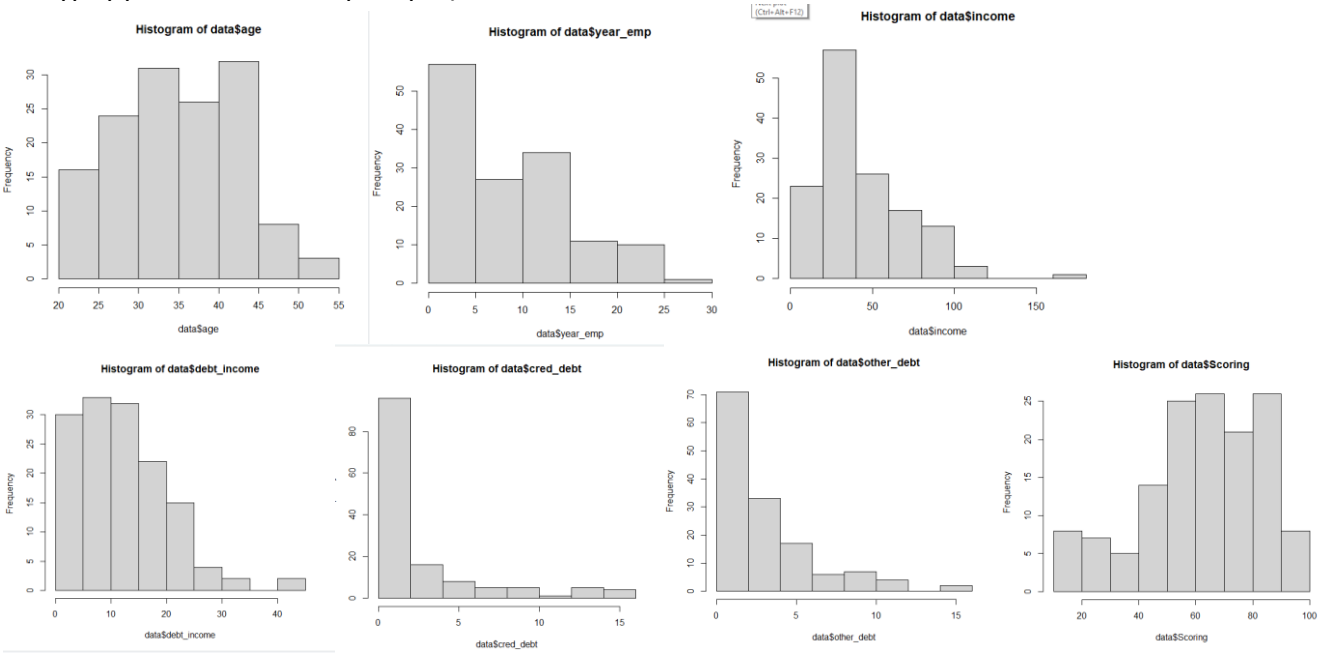






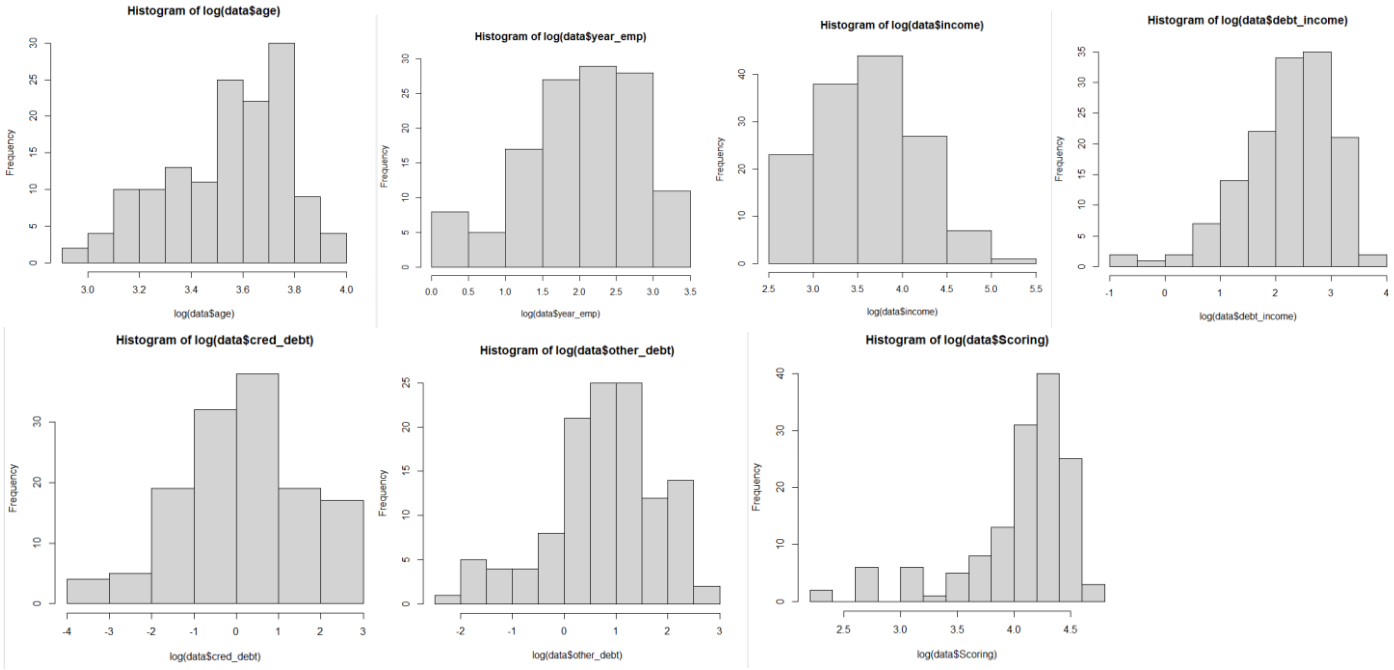


Ιστογράμματα ποσοτικών μεταβλητών



Παρατηρούμε μεγάλη θετική ασυμμετρία στο income, other_income, cred_income, debt_income, year_emp. Αρνητική ασυμμετρία στο Scoring.

Ιστογράμματα ποσοτικών μεταβλητών (λογαριθμικός μετασχηματισμός)
Ο λογαριθμικός μετασχηματισμός «διορθώνει» τη θετική ασυμμετρία.



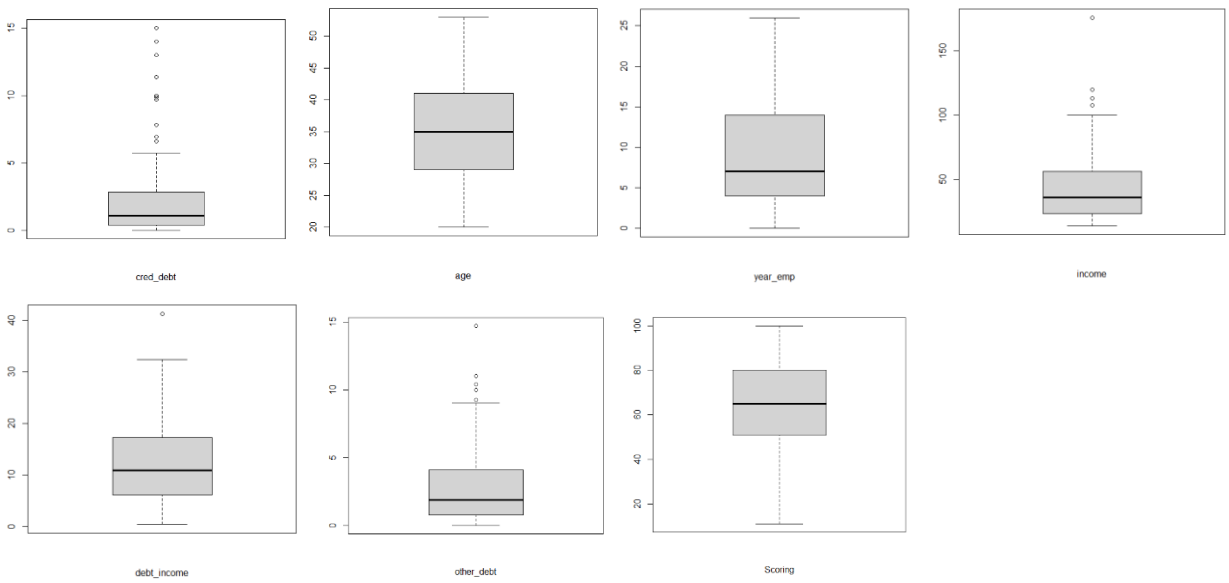
Χρησιμοποιώντας το πακέτο περιγραφικής στατιστικής psych:

```
> psych::describe(data[,c(1,3,4,5,6,7,11)])
            vars   n  mean    sd median trimmed   mad  min    max  range  skew kurtosis   se
age            1 140 35.26  7.52  35.00   35.26  8.90 20.0  53.00  33.00 -0.03    -0.57 0.64
year_emp       2 140  8.66  6.60   7.00    8.13  7.41  0.0  26.00  26.00  0.55    -0.57 0.56
income         3 140 43.81 26.77  36.00   39.99 22.24 14.0 176.00 162.00  1.55     3.44 2.26
debt_income    4 140 12.20  8.14  10.90   11.44  7.56  0.4  41.30  40.90  0.97     1.05 0.69
cred_debt      5 140  2.70  3.88   1.11    1.77  1.23  0.0  15.02  15.02  1.97     2.87 0.33
other_debt     6 140  2.91  3.11   1.90    2.35  2.36  0.0  14.72  14.72  1.59     2.37 0.26
Scoring        7 140 61.99 21.72  65.00   63.71 21.50 11.0 100.00  89.00 -0.58    -0.34 1.84
> summary(data[,c(1,3,4,5,6,7,11)])
      age           year_emp          income        debt_income      cred_debt         other_debt        Scoring
 Min.   :20.00   Min.   : 0.000   Min.   : 14.00   Min.   : 0.400   Min.   : 0.0000   Min.   : 0.0000   Min.   : 11.00
 1st Qu.:29.00   1st Qu.: 4.000   1st Qu.: 23.75   1st Qu.: 6.175   1st Qu.: 0.3877   1st Qu.: 0.8067   1st Qu.: 51.00
 Median :35.00   Median : 7.000   Median : 36.00   Median :10.900   Median : 1.1099   Median : 1.8981   Median : 65.00
 Mean   :35.26   Mean   : 8.664   Mean   : 43.81   Mean   :12.201   Mean   : 2.6961   Mean   : 2.9138   Mean   : 61.99
 3rd Qu.:41.00   3rd Qu.:14.000   3rd Qu.: 55.25   3rd Qu.:17.150   3rd Qu.: 2.8178   3rd Qu.: 4.0772   3rd Qu.: 80.00
 Max.   :53.00   Max.   :26.000   Max.   :176.00   Max.   :41.300   Max.   :15.0167   Max.   :14.7193   Max.   :100.00
>
```

Παρατηρούμε μεγάλες διαφορές ανάμεσα στη μέση τιμή και τη διάμεσο της Scoring και των υπολοίπων μεταβλητών. Τις τυπικές αποκλίσεις και τα εύρη τιμών τους(μεγάλο εύρος και τυπική απόκλιση οι income και scoring) και βγάζουμε συμπεράσματα για τις ασυμμετρίες (θετική ασυμμετρία: year_emp, income, debt_income, cred_income, other debt | αρνητική ασυμμετρία: age, Scoring) όπως επίσης και για τις κυρτώσεις (λεπτόκυρτη: income | πλατύκυρτη: όλες οι άλλες).

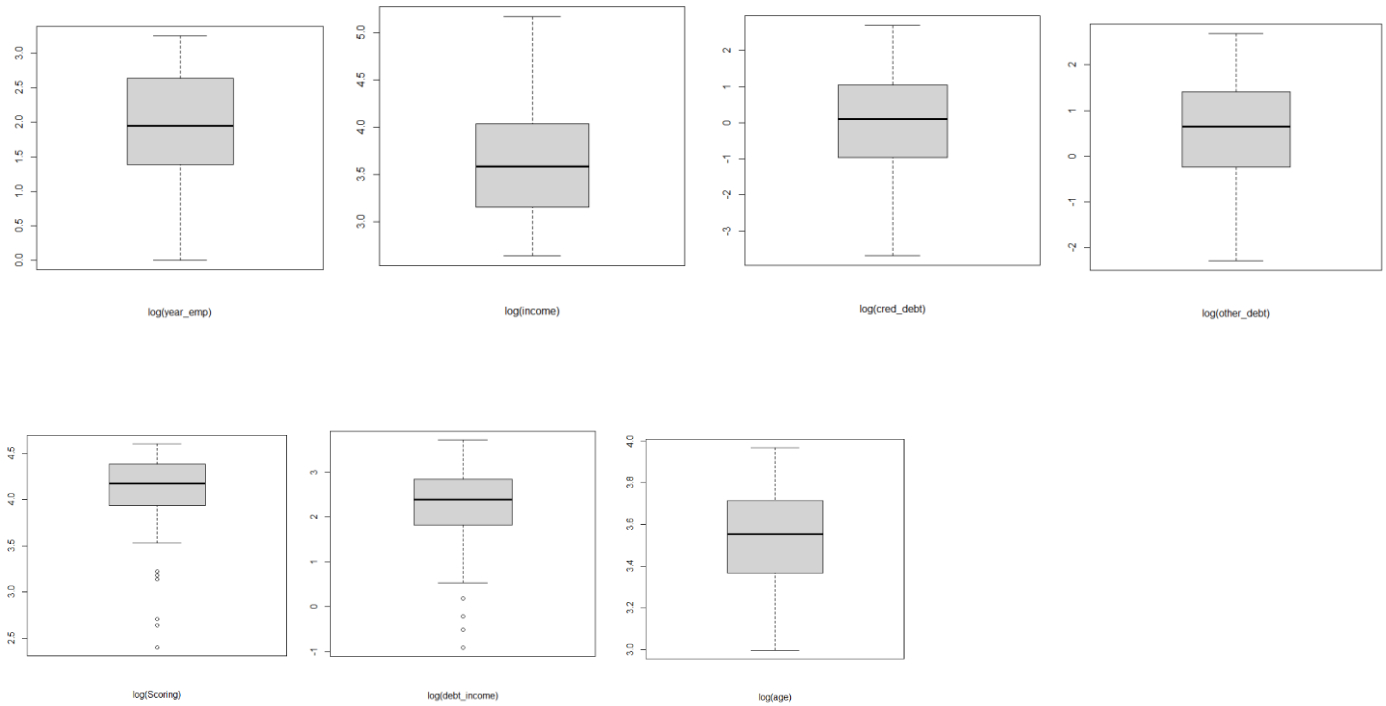
Θηκογράμματα (box plots) για ποσοτικές μεταβλητές

Η ασυμμετρία φαίνεται και εδώ μαζί με τις έκτοπες τιμές (outliers)



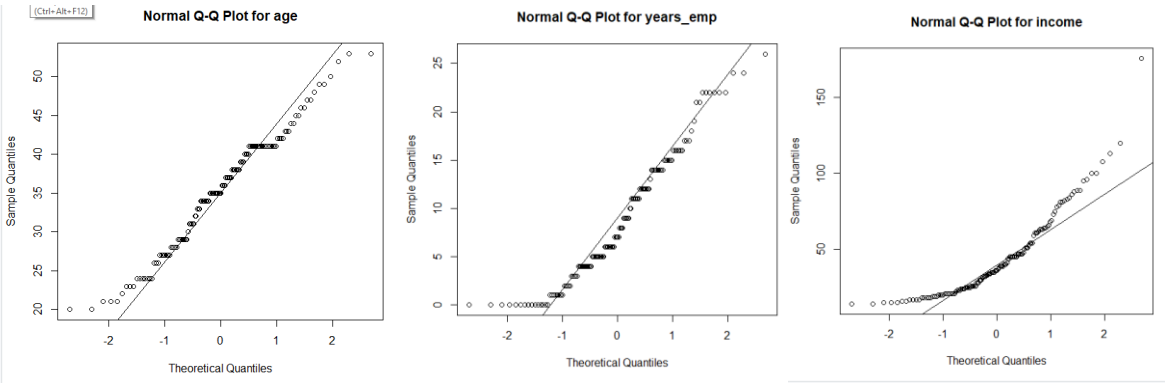
Θηκογράμματα (box plots) για ποσοτικές μεταβλητές (λογαριθμικός μετασχηματισμός)

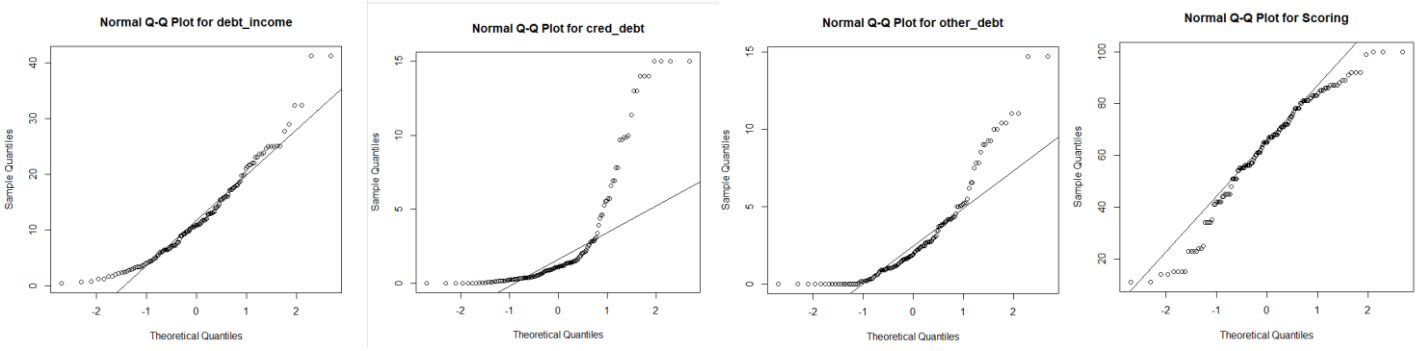
Ο λογαριθμικός μετασχηματισμός εξαφανίζει τα έκτοπα σημεία και κάνει πιο συμμετρική την κατανομή



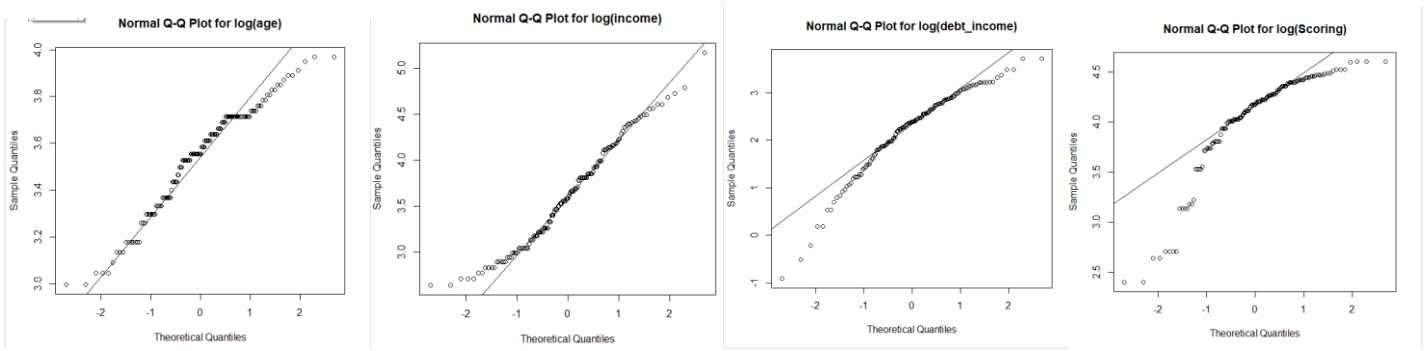
Έλεγχος κανονικότητας

Αν εξαιρέσουμε τις age και year_emp οι υπόλοιπες μεταβλητές είναι μακριά από την κανονική κατανομή.



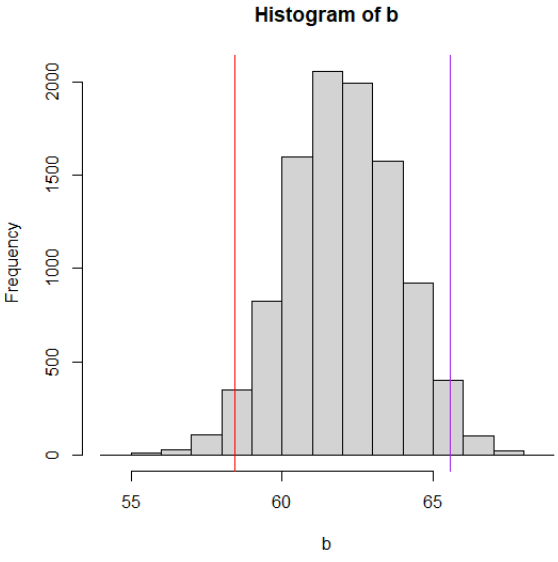


Έλεγχος κανονικότητας για λογαριθμικούς μετασχηματισμούς



Bootstrap διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού του scoring

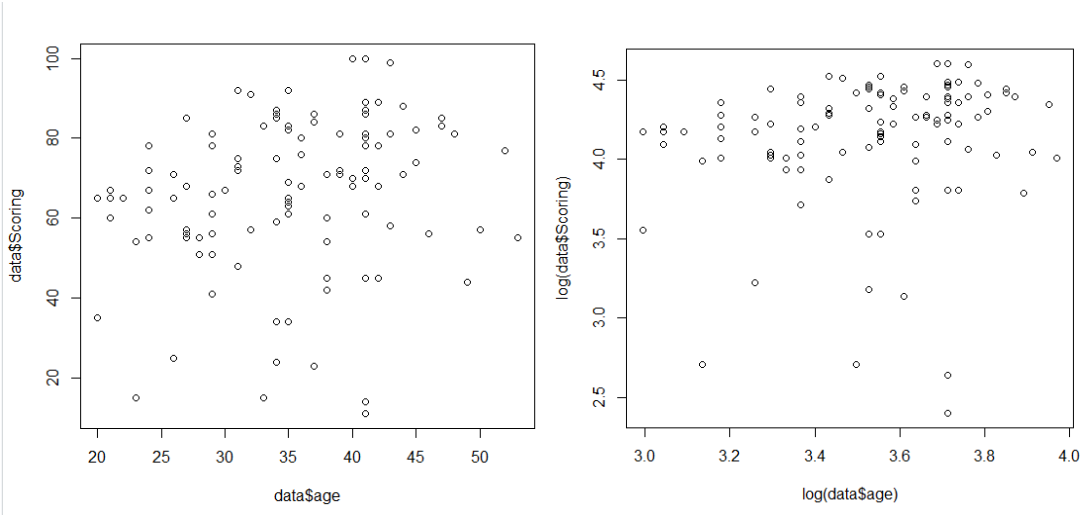
Το bootstrap 95% δ.ε. για τη μέση τιμή του πληθυσμού απ' όπου προέρχεται το δείγμα είναι [58.4, 65.5]



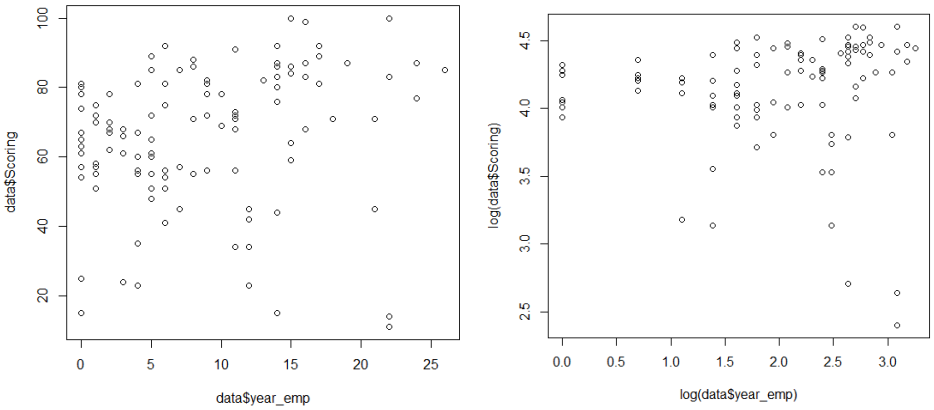

```
> b<-numeric(10000)
> for (i in 1:10000) {b[i]<-
+   mean(sample(data$Scoring,length(data$Scoring),replace=T))}
> hist(b)
> quantile(b,c(0.025,0.975))
    2.5%    97.5%
58.42839 65.58571
> abline(v=quantile(b,0.025),col="red")
> abline(v=quantile(b,0.975),col="purple")
>
```

Σχέση Scoring με

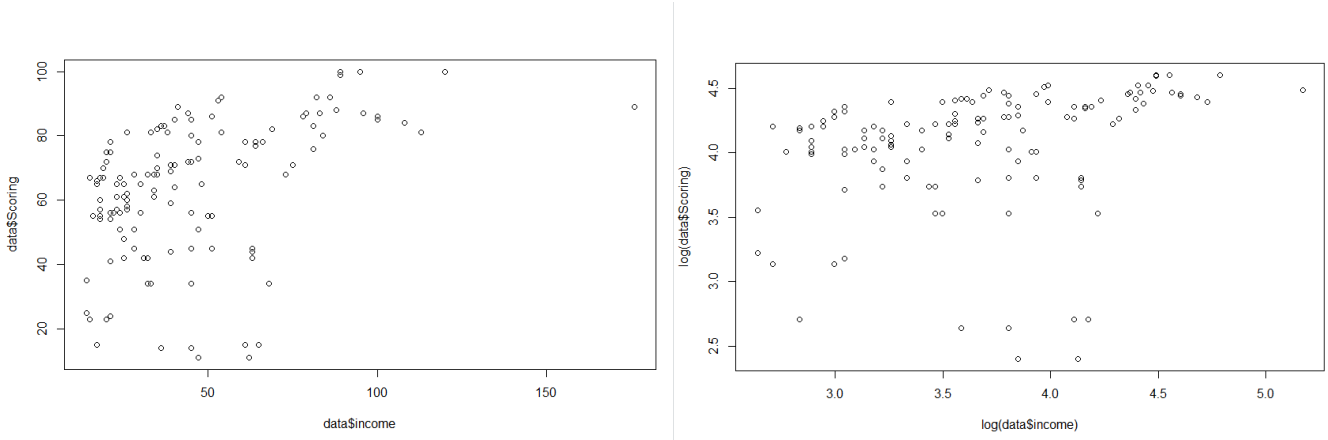
age: Δεν φαίνεται κάποια συσχέτιση



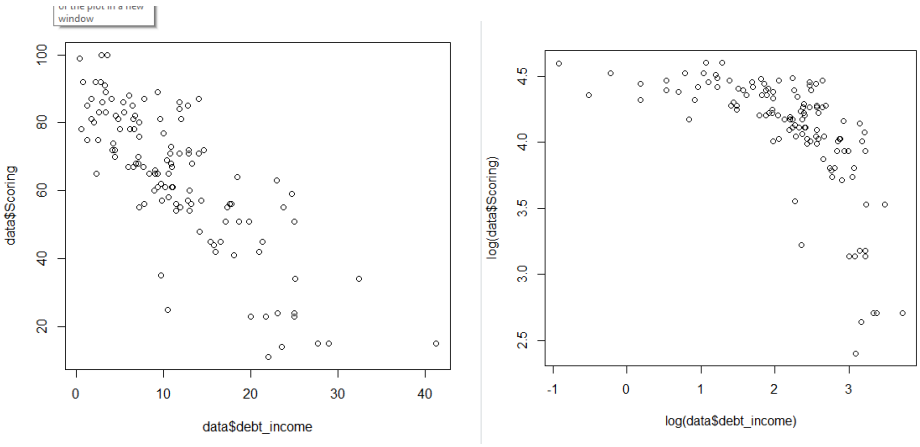
years_emp: Δεν φαίνεται κάποια συσχέτιση



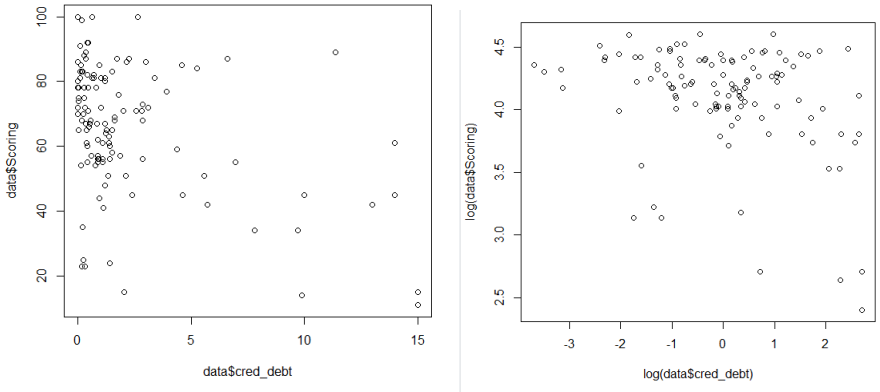
Income: Φαίνεται ισχυρή θετική συσχέτιση



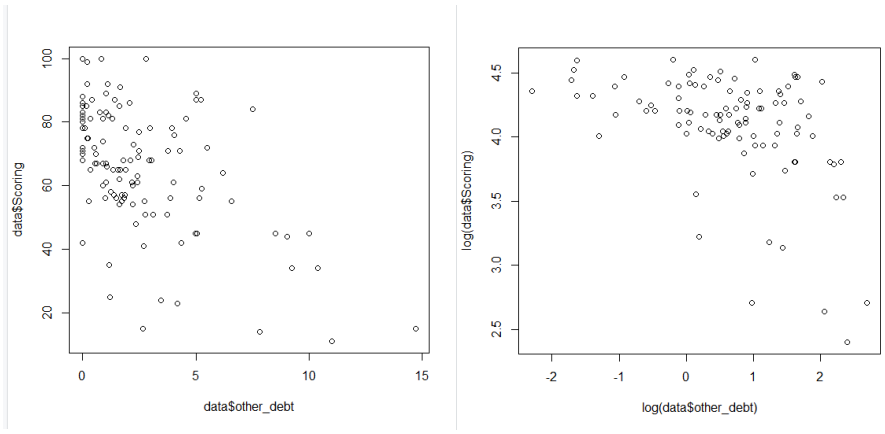
debt_income: Φαίνεται ισχυρή αρνητική συσχέτιση



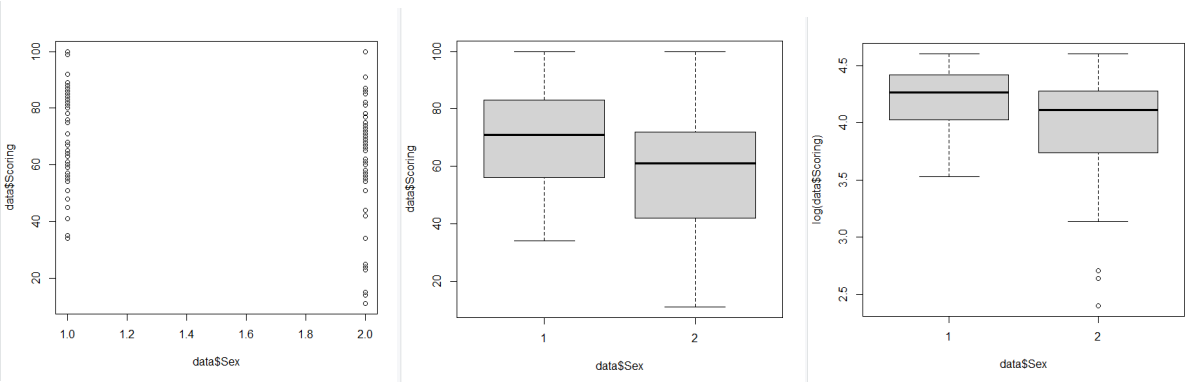
cred_debt: Φαίνεται ισχυρή αρνητική συσχέτιση



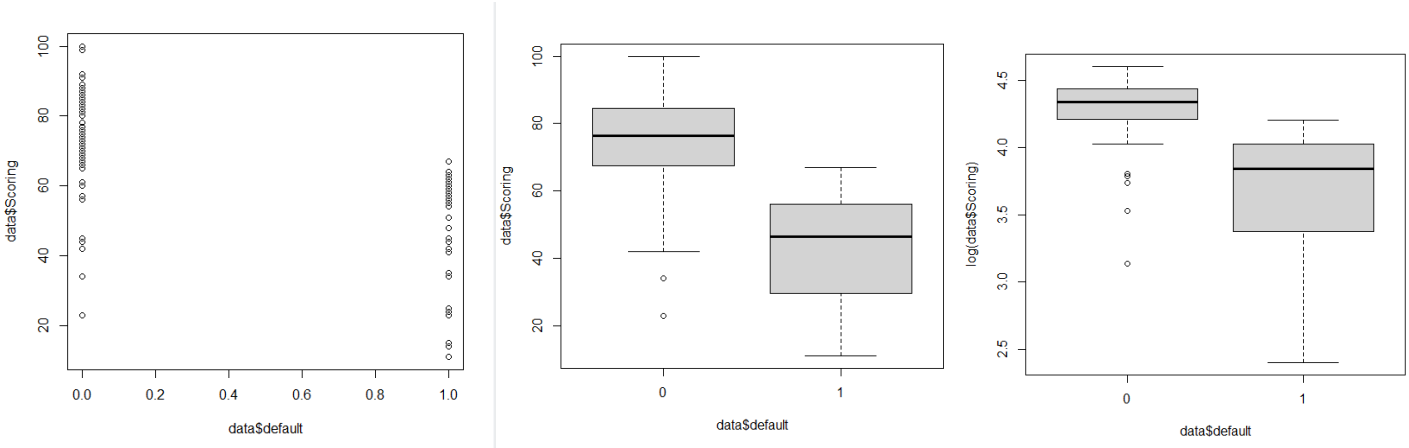
other_debt: Φαίνεται ισχυρή αρνητική συσχέτιση



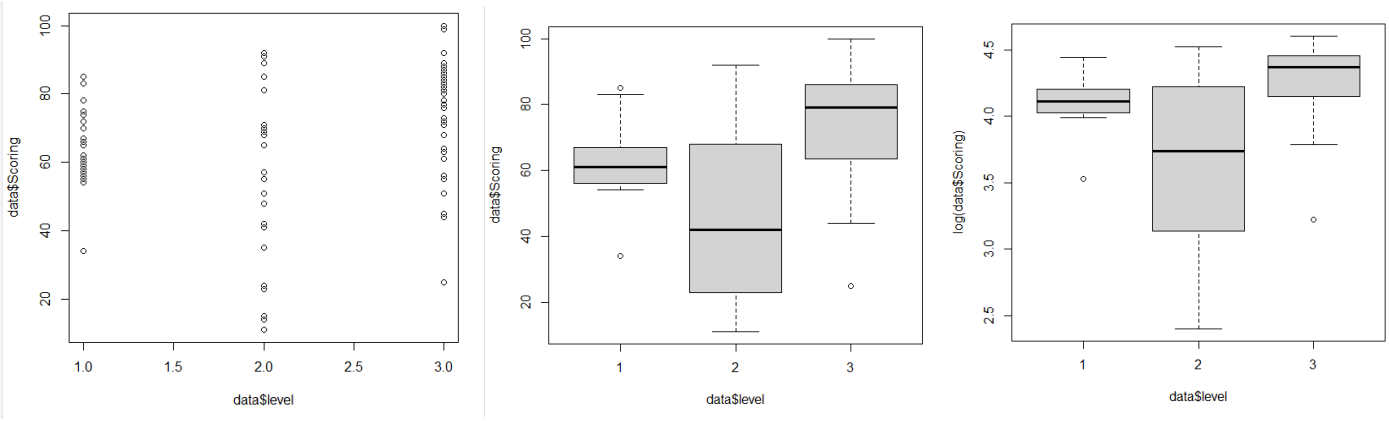
Sex: Δεν φαίνονται έντονες διαφορές στα επίπεδα



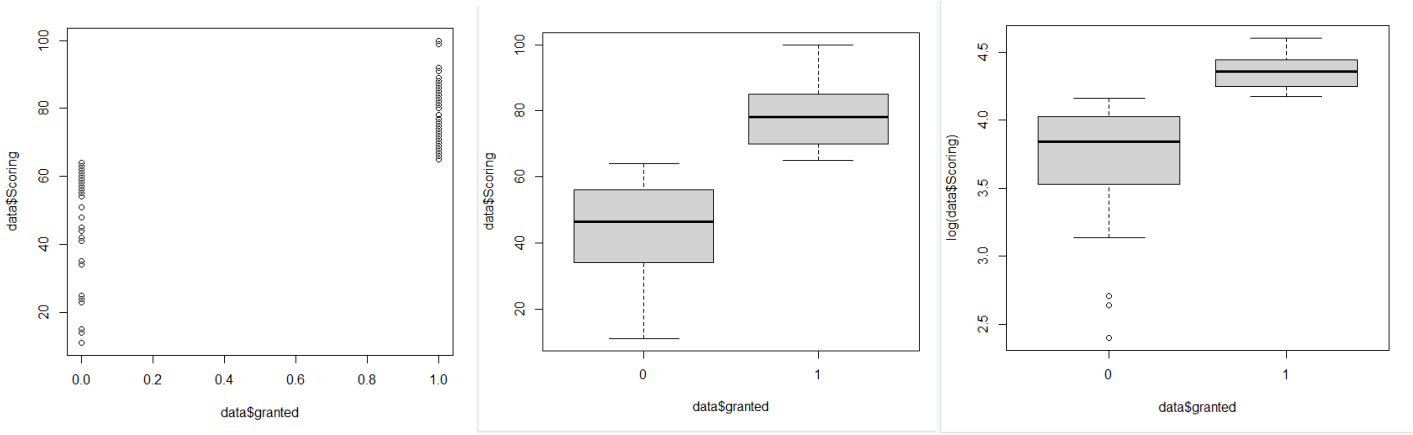
Default: Φαίνεται έντονη διαφορά στα επίπεδα



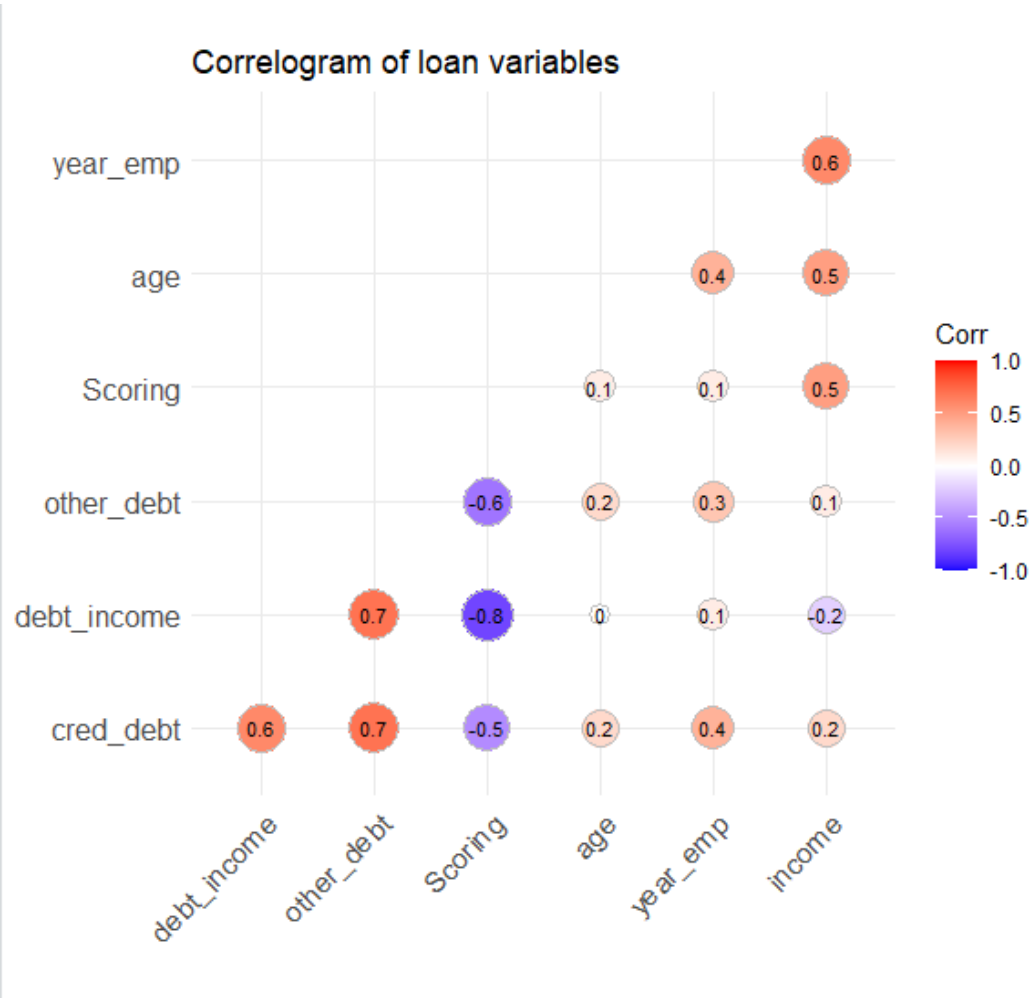
Level: Δεν φαίνονται έντονες διαφορές στα επίπεδα



Granted: Φαίνεται έντονη διαφορά



Διάγραμμα συσχέτισης του Scoring με όλες τις υπόλοιπες ποσοτικές μεταβλητές



Παρατηρούμε διάφορες συσχετίσεις μεταξύ των μεταβλητών αλλά στην περίπτωση μας ενδιαφερόμαστε για τις συσχετίσεις της Scoring όπου βλέπουμε ισχυρές αρνητικές συσχετίσεις με τις other_debt, debt_income, cred_income και ισχυρή θετική συσχέτιση με την income. Αυτά με την μέθοδο Pearson. Θετική συσχέτιση σημαίνει αύξηση της μιας μεταβλητής οδηγεί σε αύξηση και της άλλης, ενώ αρνητική συσχέτιση σημαίνει αύξηση της μιας μεταβλητής οδηγεί σε μείωση της άλλης και όσο μεγαλύτερη η απόλυτη τιμή τόσο σημαντικότερη η συσχέτιση.

Στη συνέχεια κάνουμε ξεχωριστά correlation test για τη Scoring με την εκάστοτε μεταβλητή αξιοποιώντας και την μέθοδο spearman και τους λογαριθμικούς μετασχηματισμούς για να επιβεβαιώσουμε αυτά που προαναφέραμε. Την μέθοδο spearman την χρησιμοποιούμαι κυρίως για να βρούμε τα p-values.

Correlation Scoring and age[στατιστικά σημαντική συσχέτιση(p<0.05)]

```
> cor(data$Scoring,data$age)
[1] 0.143218
> cor(log(data$Scoring),log(data$age))
[1] 0.0906487
> cor(data$Scoring,data$age, method="spearman")
[1] 0.1934475
> cor(log(data$Scoring),log(data$age), method="spearman")
[1] 0.1934475
> cor.test(data$Scoring,data$age, method = "spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  data$Scoring and data$age
S = 368845, p-value = 0.02202
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1934475

> |
```

Correlation Scoring and year_emp[στατιστικά σημαντική συσχέτιση(p<0.05)]

```
> cor(data$Scoring,data$year_emp)
[1] 0.1075456
> cor(log(data$Scoring),log(data$year_emp))
[1] NaN
> cor(data$Scoring,data$year_emp, method="spearman")
[1] 0.2035065
> cor(log(data$Scoring),log(data$year_emp), method="spearman")
[1] 0.2035065
> cor.test(data$Scoring,data$year_emp, method = "spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  data$Scoring and data$year_emp
S = 364244, p-value = 0.01588
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2035065

> |
```

Correlation Scoring and income[στατιστικά σημαντική συσχέτιση(p<0.05)]

```
> cor(data$Scoring,data$income)
[1] 0.4638873
> cor(log(data$Scoring),log(data$income))
[1] 0.3333387
> cor(data$Scoring,data$income, method="spearman")
[1] 0.5025965
> cor(log(data$Scoring),log(data$income), method="spearman")
[1] 0.5025965
> cor.test(data$Scoring,data$income, method = "spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  data$Scoring and data$income
S = 227468, p-value = 2.49e-10
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5025965

> |
```

Correlation Scoring and debt_income[στατιστικά σημαντική συσχέτιση(p<0.05)]

```
> cor(data$Scoring,data$debt_income)
[1] -0.8220649
> cor(log(data$Scoring),log(data$debt_income))
[1] -0.6644383
> cor(data$Scoring,data$debt_income, method="spearman")
[1] -0.8206416
> cor(log(data$Scoring),log(data$debt_income), method="spearman")
[1] -0.8206416
> cor.test(data$Scoring,data$debt_income, method = "spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  data$Scoring and data$debt_income
S = 832598, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.8206416

> |
```

Correlation Scoring and cred_debt[στατιστικά σημαντική συσχέτιση(p<0.05)]

```
> cor(data$Scoring,data$cred_debt)
[1] -0.4716147
> cor(log(data$Scoring),log(data$cred_debt))
[1] NaN
> cor(data$Scoring,data$cred_debt, method="spearman")
[1] -0.3868011
> cor(log(data$Scoring),log(data$cred_debt), method="spearman")
[1] -0.3868011
> cor.test(data$Scoring,data$cred_debt, method = "spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  data$Scoring and data$cred_debt
S = 634198, p-value = 2.35e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.3868011

> |
```

Correlation Scoring and other_debt[στατιστικά σημαντική συσχέτιση(p<0.05)]

```
> cor(data$Scoring,data$other_debt)
[1] -0.6006245
> cor(log(data$Scoring),log(data$other_debt))
[1] NaN
> cor(data$Scoring,data$other_debt, method="spearman")
[1] -0.5626778
> cor(log(data$Scoring),log(data$other_debt), method="spearman")
[1] -0.5626778
> cor.test(data$Scoring,data$other_debt, method = "spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  data$Scoring and data$other_debt
S = 714628, p-value = 4.643e-13
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.5626778

> |
```

Για τις συσχετίσεις τις Scoring με τις κατηγορικές μεταβλητές κάνουμε ένα correlation test και ανάλυση διακύμανσης με τη χρήση της ANOVA.

Correlation Scoring and Sex :

Αρνητική συσχέτιση και η ανάλυση διακύμανσης δείχνει στατιστικά σημαντική επίδραση του παράγοντα Sex στο Scoring (p<0.05)

```
> cor.test(data$Scoring,data$Sex, method = "spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  data$Scoring and data$Sex
S = 581904, p-value = 0.001128
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.2724507

> summary(aov(data$Scoring~data$Sex))
             Df Sum Sq Mean Sq F value   Pr(>F)
data$Sex      1   5865    5865   13.56 0.000331 ***
Residuals   138  59693     433
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(log(data$Scoring)~data$Sex))
             Df Sum Sq Mean Sq F value   Pr(>F)
data$Sex      1  3.168   3.168   14.81 0.000181 ***
Residuals   138 29.518   0.214
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Correlation Scoring and default :

Ισχυρή αρνητική συσχέτιση και η ανάλυση διακύμανσης δείχνει στατιστικά σημαντική επίδραση του παράγοντα default στο Scoring (p<0.05)

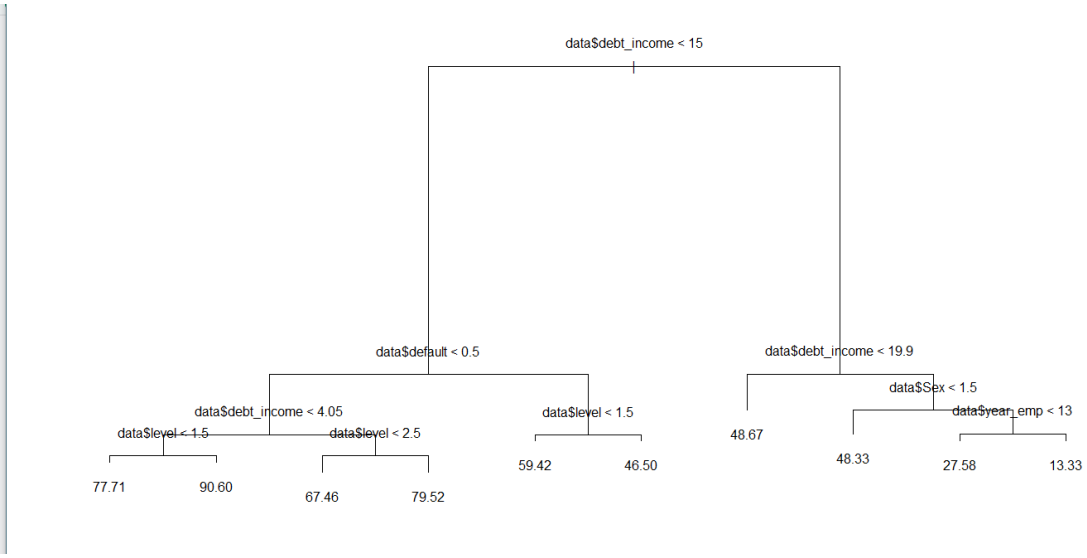
```
> cor.test(data$Scoring,data$default, method = "spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  data$Scoring and data$default
S = 799701, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.7487057

> summary(aov(data$Scoring~data$default))
              Df Sum Sq Mean Sq F value Pr(>F)
data$default   1  33201   33201   141.6 <2e-16 ***
Residuals    138  32357     234
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(log(data$Scoring)~data$default))
              Df Sum Sq Mean Sq F value Pr(>F)
data$default   1  13.45  13.448   96.47 <2e-16 ***
Residuals    138  19.24   0.139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Correlation Scoring and level :

Θετική συσχέτιση και η ανάλυση διακύμανσης δείχνει στατιστικά σημαντική επίδραση του παράγοντα level στο Scoring (p<0.05)

```
> cor.test(data$Scoring,data$level, method = "spearman",exact=FALSE)

        Spearman's rank correlation rho

data:  data$Scoring and data$level
S = 303212, p-value = 4.681e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3369663

> summary(aov(data$Scoring~data$level))
              Df Sum Sq Mean Sq F value  Pr(>F)
data$level     1   4606    4606   10.43 0.00155 **
Residuals    138  60952     442
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(aov(log(data$Scoring)~data$level))
              Df Sum Sq Mean Sq F value Pr(>F)
data$level     1   1.04  1.0448   4.557 0.0346 *
Residuals    138  31.64  0.2293
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Δενδρόγραμμα



Χρησιμοποιούμαι τη μεταβλητή debt_income αντί για τις επιμέρους για απλότητα.

Παρατηρούμαι ότι ο λόγος του χρέους προς το εισόδημα είναι και ο σημαντικότερος παράγοντας.

Βλέπουμε πως αν ο λόγος αυτός είναι μεγαλύτερος από 15 τότε σχεδόν βέβαια η αίτηση δανείου απορρίπτεται, από την άλλη αν είναι μικρότερος από 15 το επόμενο σημαντικό στάδιο είναι αν του έχει απορριφθεί προηγούμενο δάνειο, αν όχι η αίτηση του γίνεται αποδεκτή. Για το level παρατηρείται πως όσο μικρότερο debt_income με default 0 τότε καλύτερα να είναι μεγαλύτερο ενώ αν default 1 καλύτερα να είναι μικρότερο

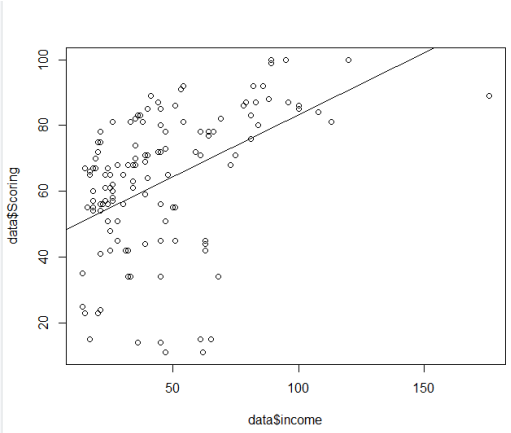
Σε κάθε περίπτωση χαμηλότερο debt_income και default συνδέεται με υψηλότερο scoring.

ΜΟΝΤΕΛΑ

Κάποια απλά μοντέλα γραμμικής παλινδρόμησης στα οποία υπάρχει στατιστικά σημαντική σχέση μεταξύ των μεταβλητών είναι τα εξής:

Scoring and income (Το μοντέλο εξηγεί το 21.5% της μεταβλητότητας του Scoring

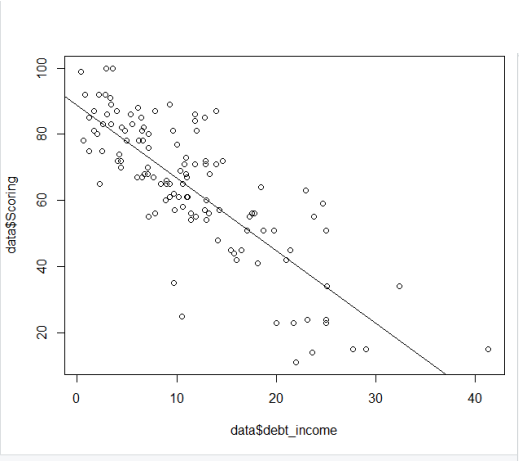
Scoring =45.5+0.37* income)




```
> plot(data$income,data$Scoring)
> model3<-lm(data$Scoring~data$income)
> abline(model3)
> summary(model3)

Call:
lm(formula = data$Scoring ~ data$income)

Residuals:
    Min      1Q  Median      3Q     Max
-57.832 -11.325   4.501  12.627  28.071

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.49957    3.13782  14.500  < 2e-16 ***
data$income  0.37633    0.06118   6.151 7.81e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.31 on 138 degrees of freedom
Multiple R-squared:  0.2152,    Adjusted R-squared:  0.2095
F-statistic: 37.84 on 1 and 138 DF,  p-value: 7.813e-09

>
```

Scoring and debt_income (Το μοντέλο εξηγεί το 67.5% της μεταβλητότητας του Scoring

Scoring =88.76-2.19* debt_income)




```
> plot(data$debt_income,data$Scoring)
> model3<-lm(data$Scoring~data$debt_income)
> abline(model3)
> summary(model3)

Call:
lm(formula = data$Scoring ~ data$debt_income)

Residuals:
    Min      1Q  Median      3Q     Max
-40.717  -8.099   0.082   7.977  28.962

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      88.7553     1.8951   46.83   <2e-16 ***
data$debt_income -2.1941     0.1294  -16.96   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.41 on 138 degrees of freedom
Multiple R-squared:  0.6758,    Adjusted R-squared:  0.6734
F-statistic: 287.7 on 1 and 138 DF,  p-value: < 2.2e-16

> |
```

Υπάρχουν όλοι οι συνδυασμοί στον κώδικα της R.

Θα φτιάξουμε ένα μοντέλο χωρίς αλληλεπιδράσεις βάζουμε μέσα όλες τις μεταβλητές που είδαμε προηγουμένως ότι είχαν στατιστικά σημαντική επίδραση στην scoring εκτός από την debt_income εφόσον θα αξιοποιήσουμε την πληροφορία από τις μεταβλητές που εκφράζει και την granted που είναι binary response της scoring.

```
> model5<-lm(data$Scoring~data$income+data$cred_debt+data$other_debt+data$age+data$Sex+data$default+data$level+data$year_emp)
> summary(model5)

Call:
lm(formula = data$Scoring ~ data$income + data$cred_debt + data$other_debt +
    data$age + data$Sex + data$default + data$level + data$year_emp)

Residuals:
    Min      1Q  Median      3Q     Max
-35.346  -7.602   1.505   7.408  27.446

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     66.71229    5.80447  11.493  < 2e-16 ***
data$income      0.35056    0.05638   6.218 6.25e-09 ***
data$cred_debt  -1.05995    0.38269  -2.770  0.00643 **
data$other_debt -2.41223    0.45727  -5.275 5.33e-07 ***
data$age         0.18625    0.15003   1.241  0.21668
data$Sex        -4.14107    2.01037  -2.060  0.04139 *
data$default   -14.43161    2.53027  -5.704 7.44e-08 ***
data$level      -1.55600    1.43903  -1.081  0.28156
data$year_emp   -0.16011    0.20227  -0.792  0.43004
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.27 on 131 degrees of freedom
Multiple R-squared:  0.7463,    Adjusted R-squared:  0.7308
F-statistic: 48.18 on 8 and 131 DF,  p-value: < 2.2e-16
```

Και όπως βλέπουμε οι μεταβλητές age, level και year_emp δεν είναι στατιστικά σημαντικές οπότε μπορούμε να τις απαλείψουμε και καταλήγουμε:

```
> model5<-lm(data$Scoring~data$income+data$cred_debt+data$other_debt+data$Sex+data$default)
> summary(model5)

Call:
lm(formula = data$Scoring ~ data$income + data$cred_debt + data$other_debt +
    data$Sex + data$default)

Residuals:
    Min      1Q  Median      3Q     Max
-34.981  -7.192   1.486   7.293  29.926

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     69.23439    4.06679  17.024  < 2e-16 ***
data$income      0.32242    0.04356   7.402 1.33e-11 ***
data$cred_debt  -1.04295    0.37198  -2.804   0.0058 **
data$other_debt -2.43033    0.45124  -5.386 3.14e-07 ***
data$Sex        -3.69727    1.98736  -1.860   0.0650 .
data$default   -14.50416    2.50701  -5.785 4.87e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.27 on 134 degrees of freedom
Multiple R-squared:  0.7405,    Adjusted R-squared:  0.7308
F-statistic: 76.47 on 5 and 134 DF,  p-value: < 2.2e-16
```

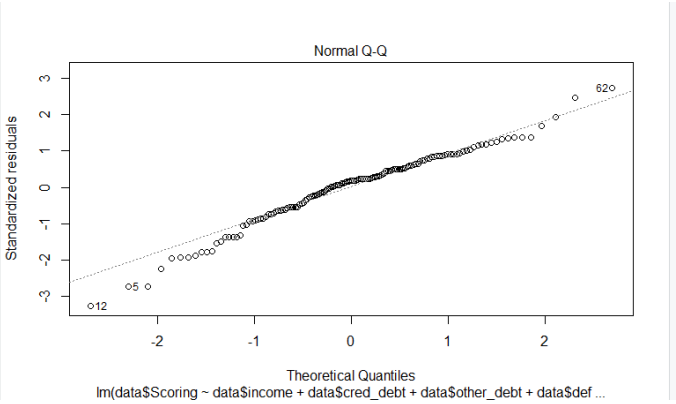
Παρατηρούμαι ότι ούτε η sex δεν είναι στατιστικά σημαντική οπότε την αφαιρούμε και τελικώς έχουμε:

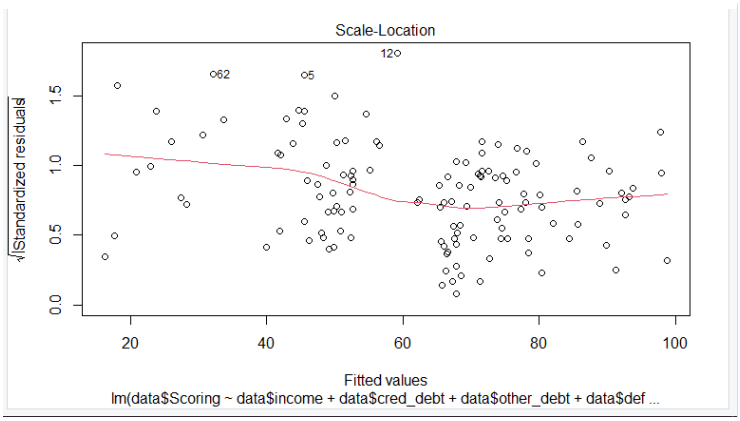
```
> model5<-lm(data$Scoring~data$income+data$cred_debt+data$other_debt+data$default)
> summary(model5)

Call:
lm(formula = data$Scoring ~ data$income + data$cred_debt + data$other_debt +
    data$default)

Residuals:
    Min      1Q  Median      3Q     Max
-36.288  -6.498   1.970   7.201  28.857

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     63.18462    2.46429  25.640  < 2e-16 ***
data$income      0.33821    0.04311   7.845 1.18e-12 ***
data$cred_debt  -1.06318    0.37519  -2.834  0.00531 **
data$other_debt -2.51441    0.45304  -5.550 1.46e-07 ***
data$default   -14.55492    2.52960  -5.754 5.59e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.37 on 135 degrees of freedom
Multiple R-squared:  0.7338,    Adjusted R-squared:  0.7259
F-statistic: 93.03 on 4 and 135 DF,  p-value: < 2.2e-16
```

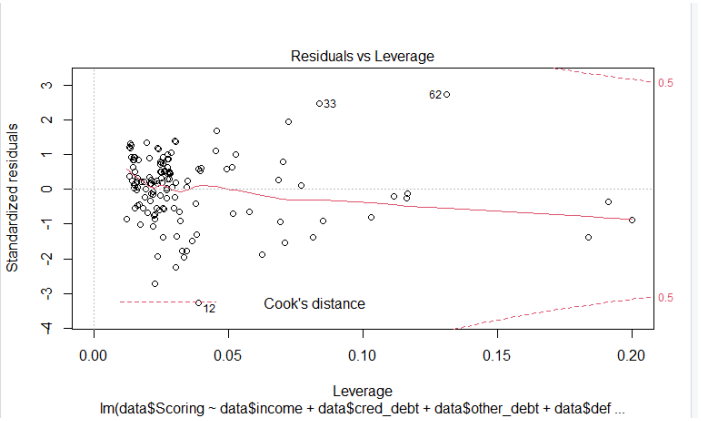
Το μοντέλο εξηγεί το 73.3% της μεταβλητότητας του Scoring

Scoring =63.18+0.33* income-1.06*cred_debt -2.51*other_debt-14.55*default

Από τα αποτελέσματα βλέπουμε ότι η απόκλιση των καταλοίπων είναι σε πολύ καλά επίπεδα σε σχέση με τους βαθμούς ελευθερίας.
Επίσης από τα σχήματα παρακάτω βλέπουμε ότι η διακύμανση των καταλοίπων φαίνεται να είναι σταθερή και τα κατάλοιπα να ακολουθούν κανονική κατανομή.
Λέγοντας όλα αυτά μπορούμε να πούμε ότι το μοντέλο αυτό είναι αξιόπιστο







Θα κάνουμε και ένα δεύτερο μοντέλο αντίστοιχο αλλά με αλληλεπιδράσεις ωστόσο επειδή οι συντελεστές γίνονται πάρα πολλοί θα χρησιμοποιήσουμε τη συνάρτηση step για να αφαιρέσουμε τους περισσότερους μη σημαντικούς. Το αποτέλεσμα που παίρνουμε είναι ένα ιδανικό μοντέλο που εξηγεί το 100% της μεταβλητότητας του Scoring.

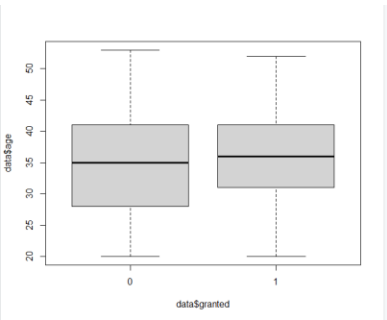
```
data$cred_debt:data$other_debt:data$default:data$level   -2.626e+01  1.375e+01  -1.909 0.092690 .
data$income:data$age:data$default:data$level              8.648e+01  1.240e+01   6.975 0.000115 ***
data$cred_debt:data$age:data$default:data$level          -2.691e+02  4.907e+01  -5.485 0.000584 ***
data$other_debt:data$age:data$default:data$level         -1.875e+02  2.074e+01  -9.039 1.80e-05 ***
data$income:data$Sex:data$default:data$level             -2.745e+02  5.198e+01  -5.282 0.000745 ***
data$cred_debt:data$Sex:data$default:data$level          -1.103e+04  9.131e+02 -12.085 2.03e-06 ***
data$other_debt:data$Sex:data$default:data$level          7.174e+03  7.829e+02   9.163 1.62e-05 ***
data$income:data$cred_debt:data$other_debt:data$year_emp  8.092e-02  5.274e-03  15.342 3.23e-07 ***
data$income:data$cred_debt:data$age:data$year_emp        -1.383e-01  5.918e-03 -23.378 1.19e-08 ***
data$income:data$other_debt:data$age:data$year_emp         1.755e-01  6.954e-03  25.236 6.51e-09 ***
data$cred_debt:data$other_debt:data$age:data$year_emp     -4.695e+00  3.613e-01 -12.995 1.17e-06 ***
data$income:data$cred_debt:data$Sex:data$year_emp         -1.412e+00  1.384e-01 -10.199 7.32e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5 on 8 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:  0.9995
F-statistic:  2002 on 131 and 8 DF,  p-value: 7.257e-13
```

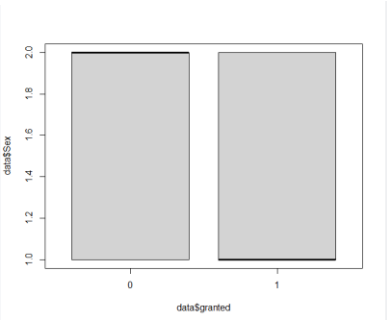
Θα εξετάσουμε την granted ως binary response(εξαρτημένη μεταβλητή).

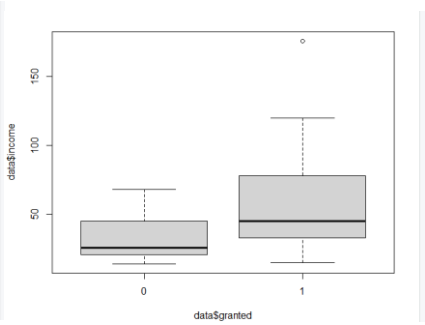
Είδαμε προηγουμένως την κατανομή των δύο τιμών της granted όπως και τα ποσοστά.

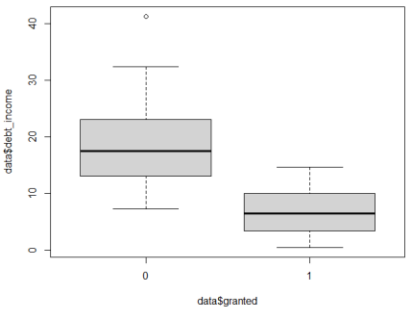
```
> tapply(data$age,data$granted,mean)
       0        1
34.72727 35.74324
> tapply(data$Sex,data$granted,mean)
       0        1
1.606061 1.472973
> tapply(data$year_emp,data$granted,mean)
       0        1
7.606061 9.608108
> tapply(data$income,data$granted,mean)
       0        1
32.34848 54.02703
> tapply(data$debt_income,data$granted,mean)
       0        1
18.360606  6.706757
> tapply(data$cred_debt,data$granted,mean)
       0        1
4.300866 1.264894
> tapply(data$other_debt,data$granted,mean)
       0        1
4.546236 1.457780
> tapply(data$default,data$granted,mean)
         0          1
0.81818182 0.02702703
> tapply(data$level,data$granted,mean)
       0        1
1.863636 2.283784
> tapply(data$Scoring,data$granted,mean)
       0        1
43.51515 78.45946
>
```

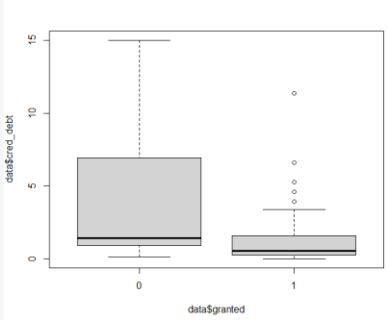
Κοιτώντας τις μέσες τιμές των ανεξάρτητων μεταβλητών ως προς την εξαρτημένη βλέπουμε πάνω κάτω τα αναμενόμενα δηλαδή η granted φαίνεται να επηρεάζεται θετικά από το υψηλό εισόδημα και αρνητικά από τα χρέη. Μικρή θετική επιρροή έχει το year_emp ενώ σημαντικά αρνητική το default. Το φύλο και η ηλικία δεν φαίνεται να επηρεάζουν. Τα συμπεράσματα αυτά φαίνονται και στα ακόλουθα boxplots.

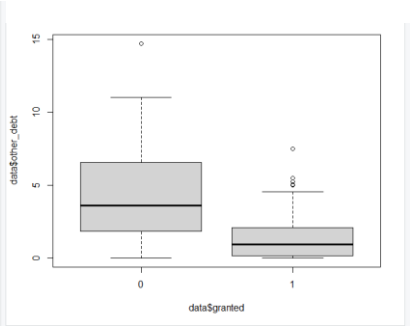


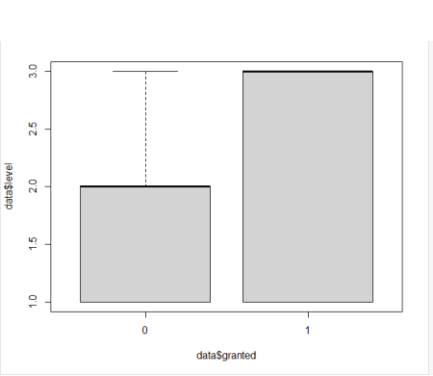


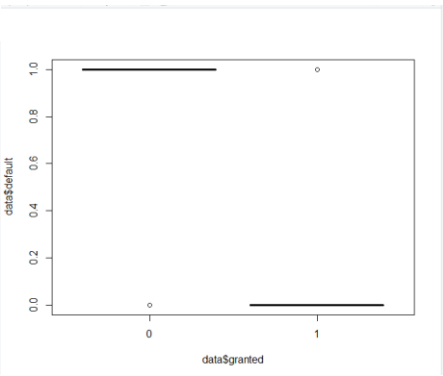


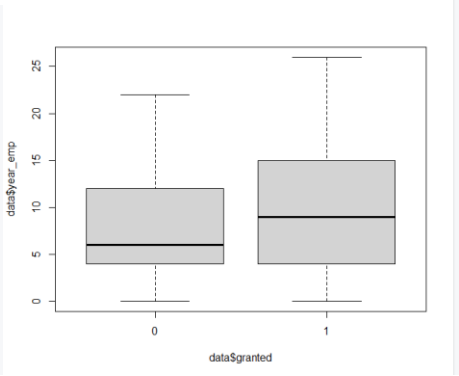












Επειδή η μεταβλητή granted είναι κατά κάποιον τρόπο ordinal (0 ή 1) μπορούμε να χρησιμοποιήσουμε τον συντελεστή συσχέτισης του Kendall.

```
> cor.test(data$age,data$granted,method="kendall")

        Kendall's rank correlation tau

data:  data$age and data$granted
z = 1.1509, p-value = 0.2498
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
0.08188152

> cor.test(data$Sex,data$granted,method="kendall")

        Kendall's rank correlation tau

data:  data$Sex and data$granted
z = -1.5705, p-value = 0.1163
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.1332104

> cor.test(data$income,data$granted,method="kendall")

        Kendall's rank correlation tau

data:  data$income and data$granted
z = 4.6082, p-value = 4.061e-06
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.3230381
```

```
> cor.test(data$debt_income,data$granted,method="kendall")

        Kendall's rank correlation tau

data:  data$debt_income and data$granted
z = -9.0257, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.628999

> cor.test(data$cred_debt,data$granted,method="kendall")

        Kendall's rank correlation tau

data:  data$cred_debt and data$granted
z = -4.6798, p-value = 2.871e-06
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.3259011

> cor.test(data$other_debt,data$granted,method="kendall")

        Kendall's rank correlation tau

data:  data$other_debt and data$granted
z = -6.3866, p-value = 1.696e-10
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.4478687
```

```
> cor.test(data$level,data$granted,method="kendall")

        Kendall's rank correlation tau

data:  data$level and data$granted
z = 3.0408, p-value = 0.002359
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.2434712

> cor.test(data$default,data$granted,method="kendall")

        Kendall's rank correlation tau

data:  data$default and data$granted
z = -9.5044, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.8061496

> cor.test(data$year_emp,data$granted,method="kendall")

        Kendall's rank correlation tau

data:  data$year_emp and data$granted
z = 1.6901, p-value = 0.09101
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.1204806
```

Όπως σωστά είπαμε προηγουμένως δεν φαίνεται κάποια στατιστικά σημαντική συσχέτιση της granted με την sex, year_emp και την age(p>0.05). Στατιστικά σημαντική θετική συσχέτιση με την income και level και στατιστικά σημαντική αρνητική συσχέτιση με τις debt_income,cred_debt,other_debt, default.

Τώρα θα εξετάσουμε τις διαφορές των ανεξάρτητων μεταβλητών ως προς την εξαρτημένη με το Wilcoxon Test.

```
> wilcox.test(data$age~data$granted)

        Wilcoxon rank sum test with continuity correction

data:  data$age by data$granted
W = 2167, p-value = 0.2506
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(data$Sex~data$granted)

        Wilcoxon rank sum test with continuity correction

data:  data$Sex by data$granted
W = 2767, p-value = 0.1169
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(data$income~data$granted)

        Wilcoxon rank sum test with continuity correction

data:  data$income by data$granted
W = 1338.5, p-value = 4.102e-06
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(data$debt_income~data$granted)

        Wilcoxon rank sum test with continuity correction

data:  data$debt_income by data$granted
W = 4604, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

```
> wilcox.test(data$cred_debt~data$granted)

        Wilcoxon rank sum test with continuity correction

data:  data$cred_debt by data$granted
W = 3563, p-value = 2.901e-06
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(data$other_debt~data$granted)

        Wilcoxon rank sum test with continuity correction

data:  data$other_debt by data$granted
W = 3970, p-value = 1.719e-10
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(data$level~data$granted)

        Wilcoxon rank sum test with continuity correction

data:  data$level by data$granted
W = 1758, p-value = 0.002377
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(data$default~data$granted)

        Wilcoxon rank sum test with continuity correction

data:  data$default by data$granted
W = 4374, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(data$year_emp~data$granted)

        Wilcoxon rank sum test with continuity correction

data:  data$year_emp by data$granted
W = 2038, p-value = 0.09141
```

Στατιστικά σημαντική διαφορά στην κατανομή των τιμών τους ως προς τις δύο τιμές (0 και 1) της εξαρτημένης μεταβλητής βλέπουμε στις income, debt_income, cred_debt, other_debt, default και level.

Για μεταβλητές απόκρισης που έχουν τιμές δυαδικές (0 ή 1) χρησιμοποιούμε τα γενικευμένα γραμμικά μοντέλα (generalized linear models – glm)
Και εδώ έχουμε μια πολλαπλή παλινδρόμηση που λόγω της μορφής της ανήκει στην λογιστική παλινδρόμηση. Στα μοντέλα αυτά χρησιμοποιείται η κατανομή binomial.

Δοκίμασα το εξής μοντέλο με αλληλεπιδράσεις

mod7 = glm(data$granted ~ data$income*data$cred_debt*data$other_debt*data$default*data$level, family = "binomial")

καμία μεταβλητή δεν φαινόταν σημαντική, ύστερα έκανα το ίδιο χωρίς αλληλεπιδράσεις

mod8 = glm(data$granted ~ data$income+data$cred_debt+data$other_debt+data$default+data$level, family = "binomial")

τα σύγκρινα και πήρα ότι η αφαίρεση των αρχικών αλληλεπιδράσεων δεν είναι στατιστικά σημαντική

```
> anova (mod7,mod8,test="Chi")
Analysis of Deviance Table

Model 1: data$granted ~ data$income * data$cred_debt * data$other_debt *
    data$default * data$level
Model 2: data$granted ~ data$income + data$cred_debt + data$other_debt +
    data$default + data$level
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1       108      0.000
2       134     37.359 -26  -37.359  0.06936 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

αφαιρούμε και την level γιατί δεν έχει στατιστικά σημαντική επίδραση και στη συνέχεια εξετάζουμε πόσο καλά προβλέπει το τελικό μοντέλο μας.

Αρχικά υπολογίζουμε για κάθε πελάτη την προβλεπόμενη από το μοντέλο πιθανότητα να πάρει το δάνειο.

Στη συνέχεια αν η πιθανότητα είναι >0.5 κατατάσσουμε το πελάτη ως επιτυχημένο (pos) και αν είναι <0.5 όχι αποτυχημένο (neg).

Βλέπουμε ότι από τους πραγματικά αποτυχημένους πελάτες (0) προβλέπει σωστά (neg) το 94% .

Από τους πραγματικά επιτυχημένους πελάτες (1) προβλέπει σωστά (pos) το 97% περίπου.

```
> probabilities<-predict(mod9,type="response")
> predicted.classes <- ifelse(probabilities > 0.5, "pos", "neg")
> table(data$granted,predicted.classes)
   predicted.classes
    neg pos
  0  62   4
  1   2  72
> prop.table(table(data$granted,predicted.classes),margin=1)
   predicted.classes
             neg        pos
  0 0.93939394 0.06060606
  1 0.02702703 0.97297297
>
```