Nursing Home Utilization

Data Analytics

Cancian Piero Castagnotto Alessandro Majer William

30-06-2022

Prime analisi esplorative

Dati

```
head(WNH)
```

```
##   hospID CRYEAR      TPY NUMBED SQRFOOT MSA URBAN PRO TAXEXEMPT SELFFUNDINS
## 1    101   2000 16.48087     18  10.861   0     0   0         1           0
## 2    103   2000 59.24590     63  19.782   0     0   0         0           1
## 3    105   2000 49.63661     54  26.868   1     1   1         0           1
## 4    107   2000 51.87432     60  26.319   0     0   0         1           1
## 5    108   2000 94.56011    104  30.700  10     1   1         0           0
## 6    109   2000 69.70492     79  24.270  11     1   1         0           0
##   MCERT ORGSTR
## 1     0      2
## 2     1      3
## 3     1      1
## 4     1      2
## 5     1      1
## 6     1      1
```

Dati

```
str(WNH)
```

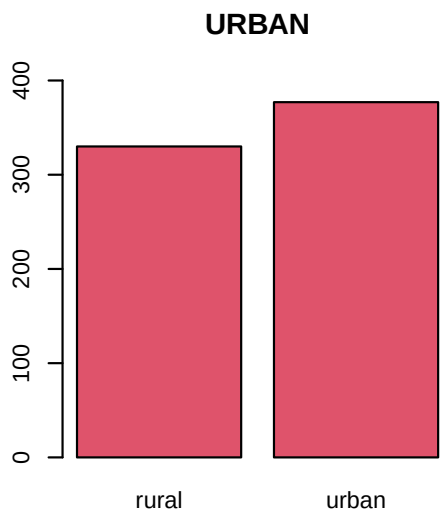
```
## 'data.frame':    717 obs. of  12 variables:
##  $ hospID     : int  101 103 105 107 108 109 110 111 112 113 ...
##  $ CRYEAR     : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ TPY        : num  16.5 59.2 49.6 51.9 94.6 ...
##  $ NUMBED     : num  18 63 54 60 104 79 105 129 200 184 ...
##  $ SQRFOOT    : num  10.9 19.8 26.9 26.3 30.7 ...
##  $ MSA        : int  0 0 1 0 10 11 11 6 10 6 ...
##  $ URBAN      : int  0 0 1 0 1 1 1 1 1 1 ...
##  $ PRO        : int  0 0 1 0 1 1 1 1 0 0 ...
##  $ TAXEXEMPT  : int  1 0 0 1 0 0 0 0 1 1 ...
##  $ SELFFUNDINS: int  0 1 1 1 0 0 0 1 1 0 ...
##  $ MCERT      : int  0 1 1 1 1 1 1 1 1 1 ...
##  $ ORGSTR     : int  2 3 1 2 1 1 1 1 2 2 ...
```

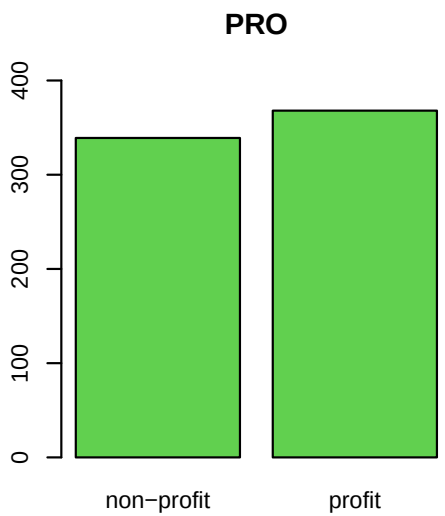
Fattorizzazione

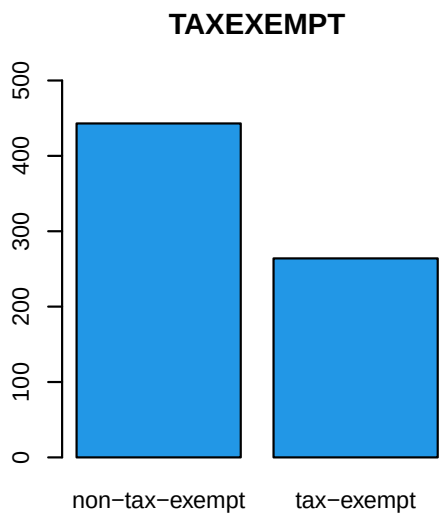
```
WNH <- WNH[,-1]
WNH <- na.omit(WNH)

WNH$CRYEAR <- factor(WNH$CRYEAR)
WNH$URBAN <- factor(WNH$URBAN)
levels(WNH$URBAN) <- c("rural","urban")
WNH$PRO <- factor(WNH$PRO)
levels(WNH$PRO) <- c("non-profit","profit")
WNH$TAXEXEMPT <- factor(WNH$TAXEXEMPT)
levels(WNH$TAXEXEMPT) <- c("non-tax-exempt", "tax-exempt")
WNH$SELFFUNDINS <- factor(WNH$SELFFUNDINS)
levels(WNH$SELFFUNDINS) <- c("non-self-funded", "self-funded")
WNH$MCERT <- factor(WNH$MCERT)
levels(WNH$MCERT) <- c("non-Medicare certified", "Medicare certified")
WNH$ORGSTR <- factor(WNH$ORGSTR)
levels(WNH$ORGSTR) <- c("profit", "tax-exempt", "governmental unit")
```

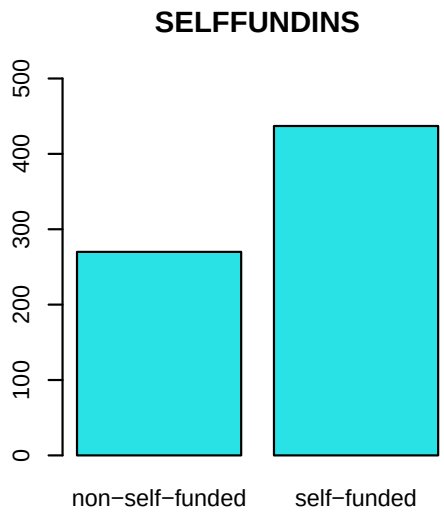
Variabili qualitative

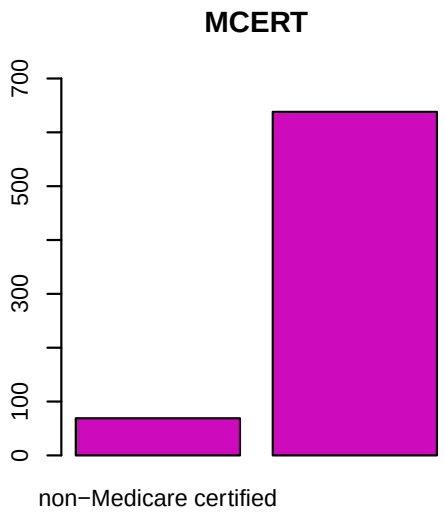


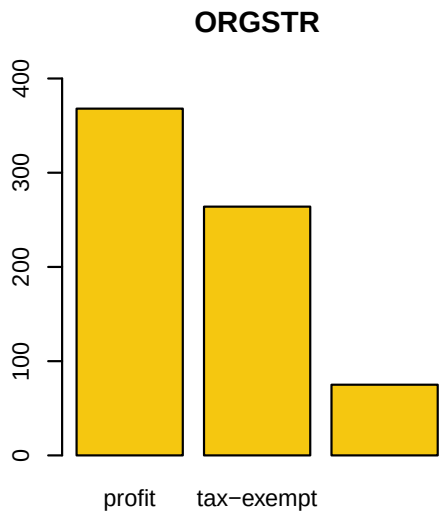




Variabili qualitative







Variabili qualitative

```
table(WNH$CRYEAR)
```

```
##
## 2000 2001
##  357  350
```

```
prop.table(table(WNH$CRYEAR))
```

```
##
##      2000      2001
## 0.5049505 0.4950495
```

Variabili qualitative

```
table(WNH$URBAN)
```

```
##
## rural urban
##   330   377
```

```
prop.table(table(WNH$URBAN))
```

```
##
##    rural    urban
## 0.466761 0.533239
```

Variabili qualitative

```
table(WNH$PRO)
```

```
##
## non-profit     profit
##        339        368
```

```
prop.table(table(WNH$PRO))
```

```
##
## non-profit     profit
##  0.4794908  0.5205092
```

Variabili qualitative

```
table(WNH$TAXEXEMPT)
```

```
##
## non-tax-exempt     tax-exempt
##            443            264
```

```
prop.table(table(WNH$TAXEXEMPT))
```

```
##
## non-tax-exempt     tax-exempt
##      0.6265912      0.3734088
```

Variabili qualitative

```
table(WNH$SELFFUNDINS)
```

```
##
## non-self-funded     self-funded
##             270             437
```

```
prop.table(table(WNH$SELFFUNDINS))
```

```
##
## non-self-funded     self-funded
##       0.3818953       0.6181047
```

Variabili qualitative

```
table(WNH$MCERT)
```

```
##
## non-Medicare certified     Medicare certified
##                     69                    638
```

```
prop.table(table(WNH$MCERT))
```

```
##
## non-Medicare certified     Medicare certified
##             0.09759547             0.90240453
```

Variabili qualitative

```
table(WNH$ORGSTR)
```

```
##
##            profit       tax-exempt governmental unit
##               368               264                75
```

```
prop.table(table(WNH$ORGSTR))
```

```
##
##            profit       tax-exempt governmental unit
##         0.5205092        0.3734088         0.1060820
```

Variabili qualitative

```
tab1 <- table(WNH$ORGSTR,WNH$URBAN)
tab1
```

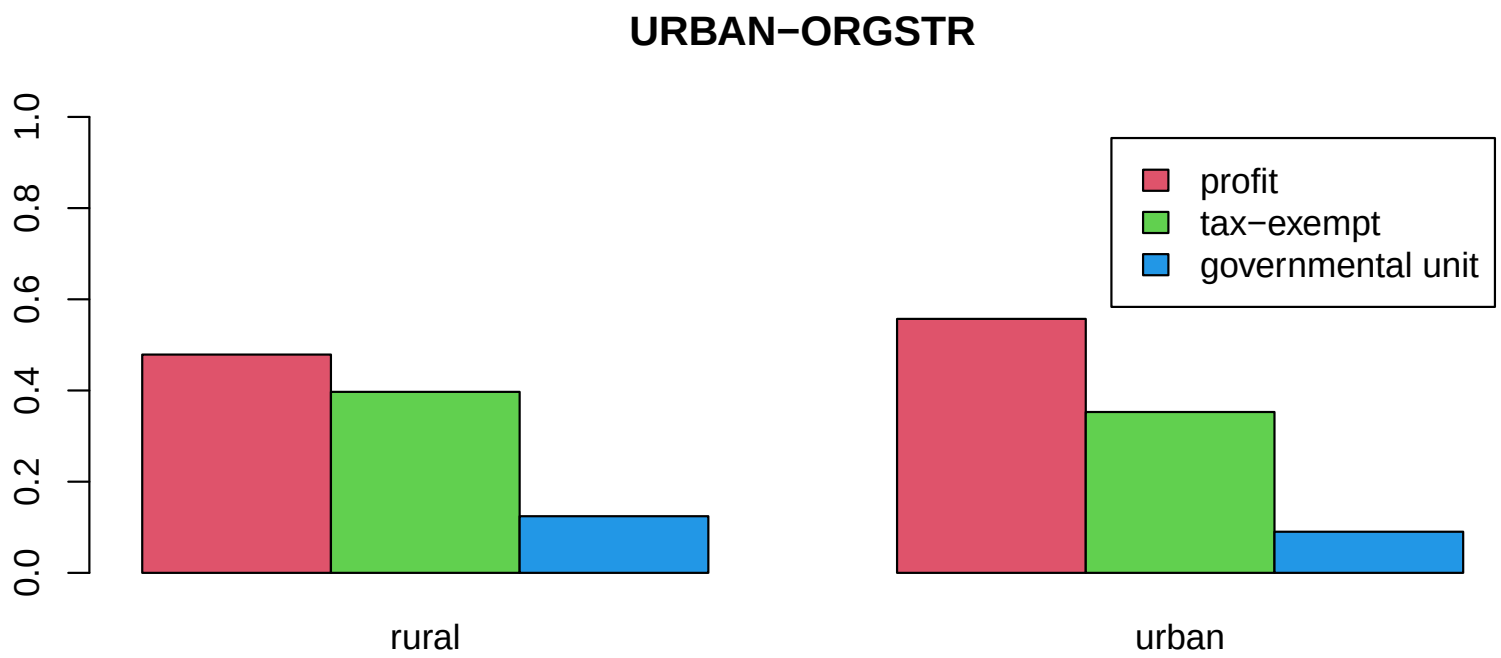
```
##
##                     rural urban
##   profit              158   210
##   tax-exempt          131   133
##   governmental unit    41    34
```

```
ptab1 <- prop.table(tab1,2)
ptab1
```

```
##
##                         rural      urban
##   profit            0.47878788 0.55702918
##   tax-exempt        0.39696970 0.35278515
##   governmental unit 0.12424242 0.09018568
```

Variabili qualitative

```
barplot(ptab1,legend=T,beside=TRUE, main="URBAN-ORGSTR", ylim=c(0,1), col=(2:4))
```



Variabili qualitative

```
tab2 <- table(WNH$PRO,WNH$URBAN)
tab2
```

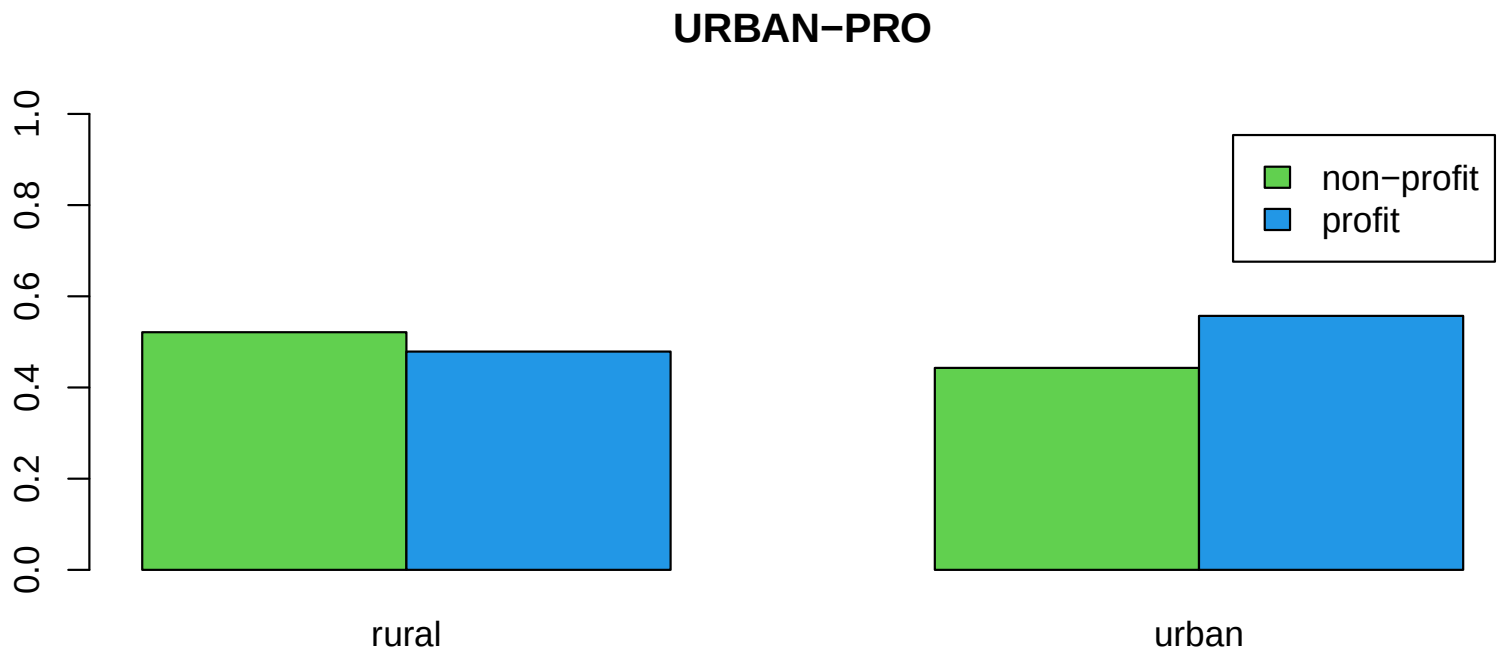
```
##
##              rural urban
##   non-profit   172   167
##   profit       158   210
```

```
ptab2 <- prop.table(tab2,2)
ptab2
```

```
##
##                  rural     urban
##   non-profit 0.5212121 0.4429708
##   profit     0.4787879 0.5570292
```

Variabili qualitative

```
barplot(ptab2,legend=T,beside=TRUE, main="URBAN-PRO", ylim=c(0,1), col=(3:4))
```



Variabili qualitative

```
tab3 <- table(WNH$PRO,WNH$MCERT)
tab3
```

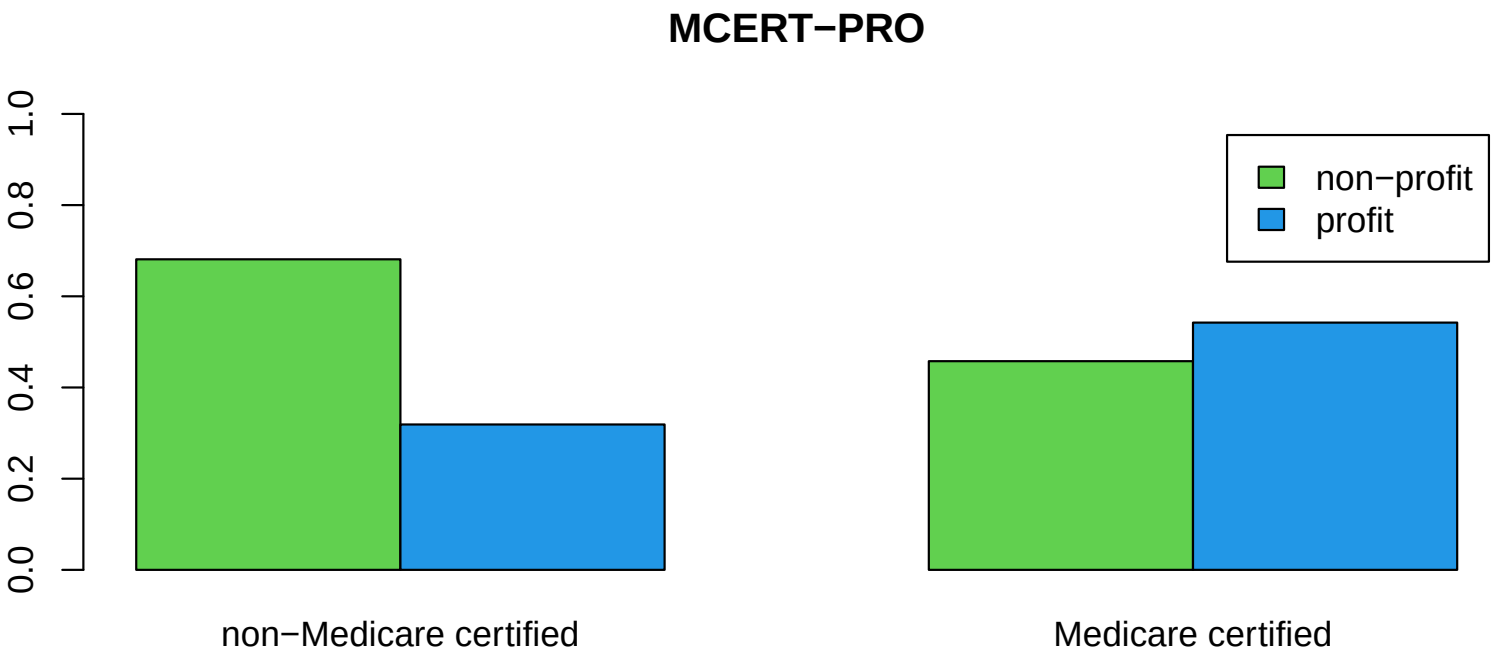
```
##
##              non-Medicare certified Medicare certified
##   non-profit                     47                292
##   profit                         22                346
```

```
ptab3 <- prop.table(tab3,2)
ptab3
```

```
##
##              non-Medicare certified Medicare certified
##   non-profit              0.6811594          0.4576803
##   profit                  0.3188406          0.5423197
```

Variabili qualitative

```
barplot(ptab3,legend=T,beside=TRUE, main="MCERT-PRO", ylim=c(0,1), col=(3:4))
```



Variabili qualitative

```
tab4 <- table(WNH$SELFFUNDINS,WNH$MCERT)
tab4
```

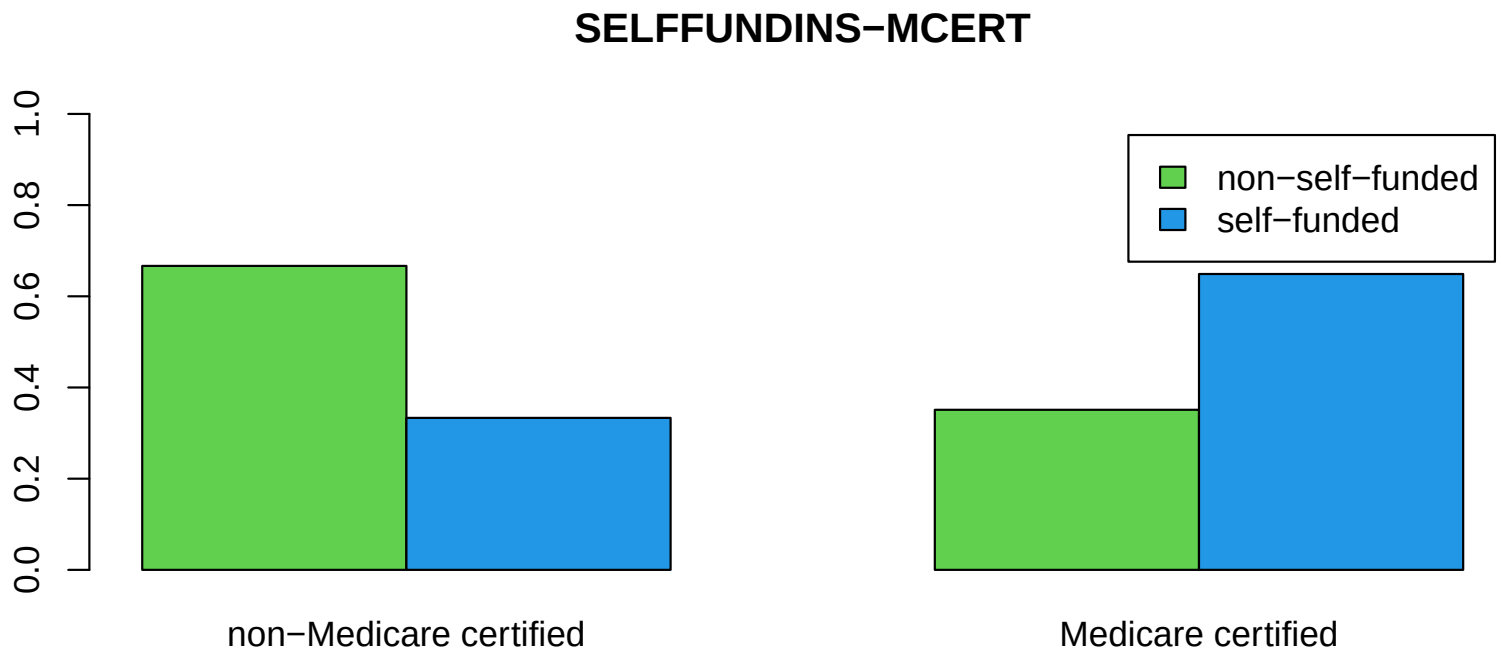
```
##
##                   non-Medicare certified Medicare certified
##   non-self-funded                     46                 224
##   self-funded                         23                 414
```

```
ptab4 <- prop.table(tab4,2)
ptab4
```

```
##
##                   non-Medicare certified Medicare certified
##   non-self-funded              0.6666667          0.3510972
##   self-funded                  0.3333333          0.6489028
```

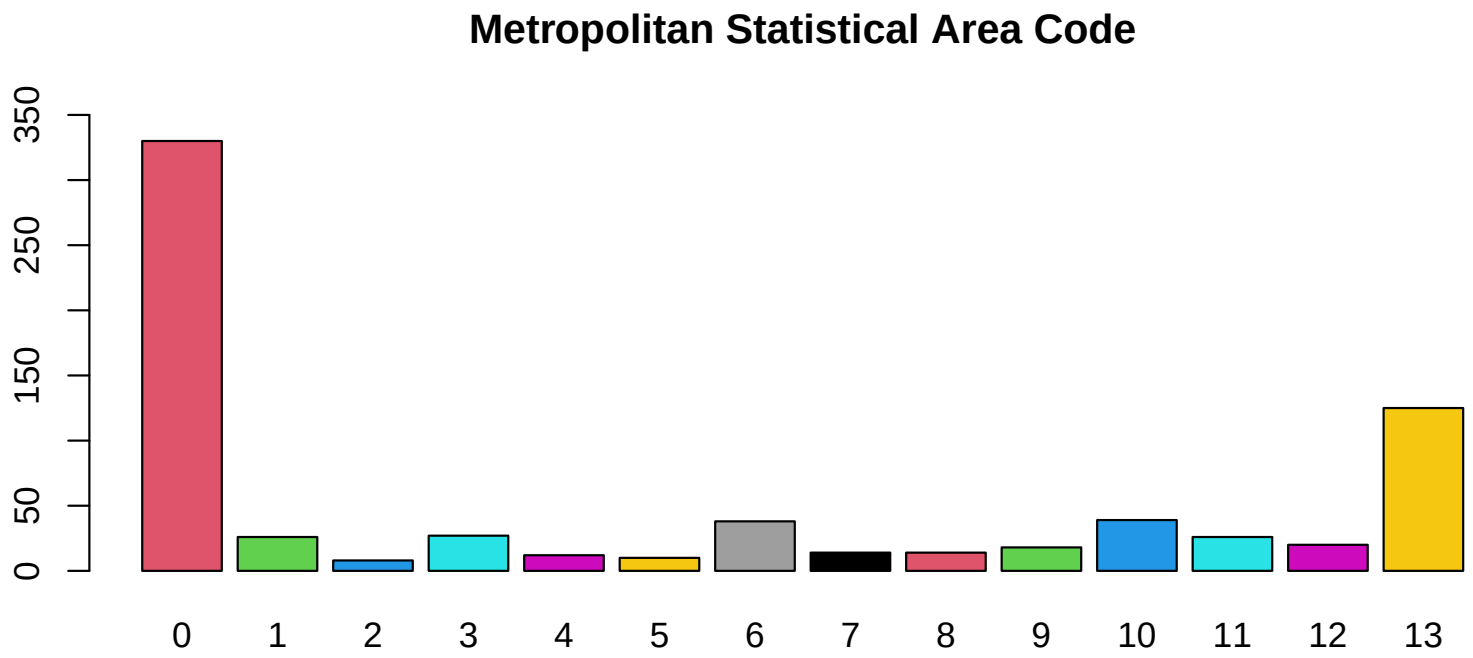
Variabili qualitative

```
barplot(ptab4,legend=T,beside=TRUE, main="SELFFUNDINS-MCERT", ylim=c(0,1), col=(3:4))
```

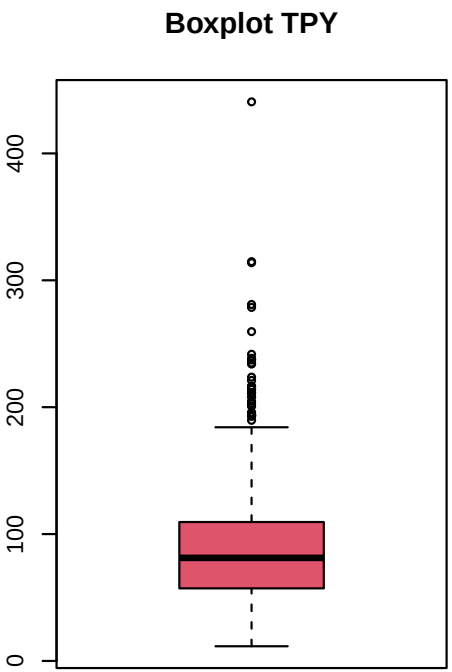


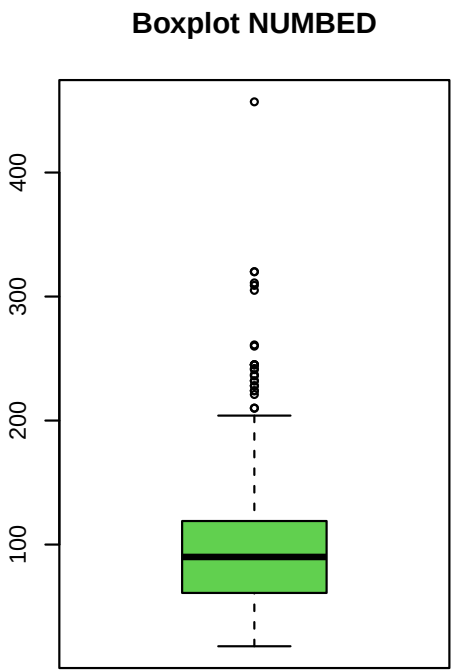
Variabili qualitative

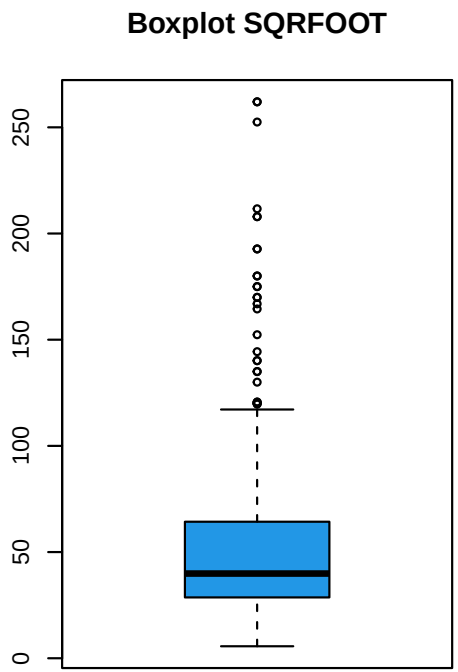
```
barplot(table(WNH$MSA), main="Metropolitan Statistical Area Code", ylim=c(0,350), col=2:15)
```



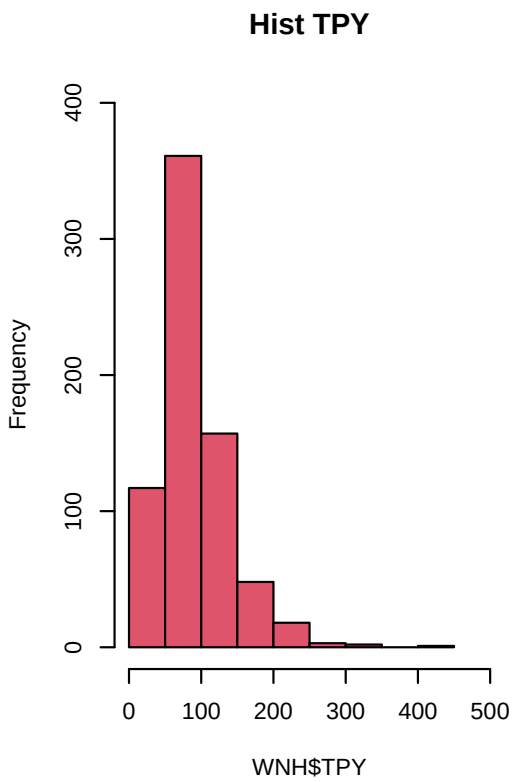
Variabili quantitative

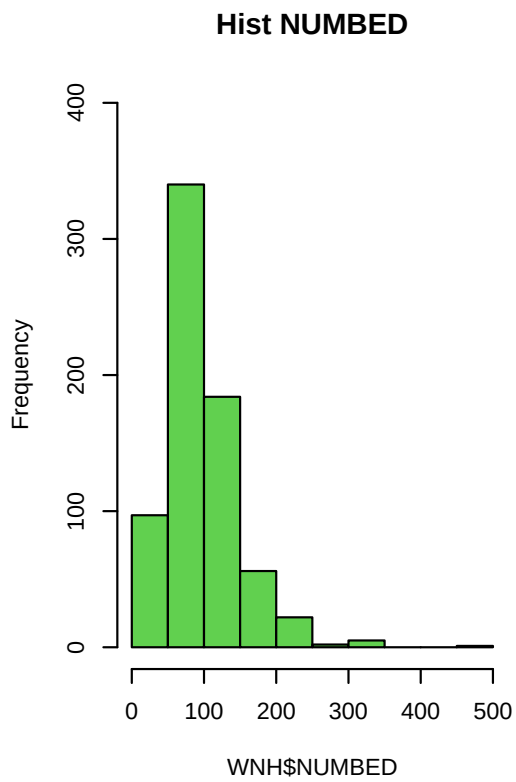


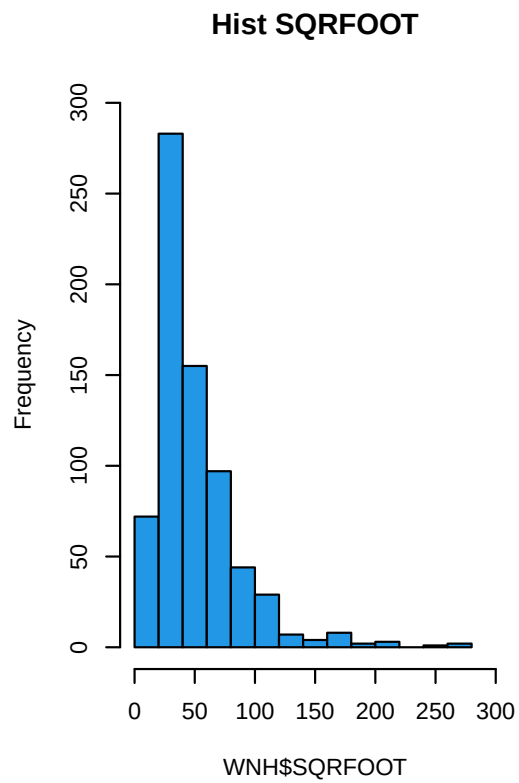




Variabili quantitative







Variabili quantitative

```
skewness(WNH$TPY)
```

```
## [1] 1.779206
```

```
skewness(WNH$NUMBED)
```

```
## [1] 1.745066
```

```
skewness(WNH$SQRFOOT)
```

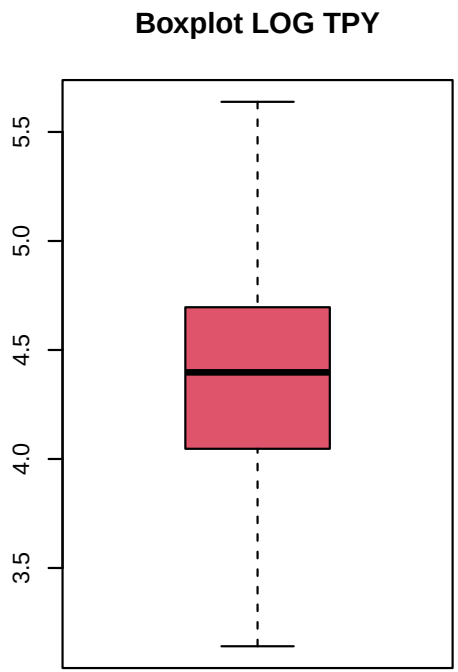
```
## [1] 2.301746
```

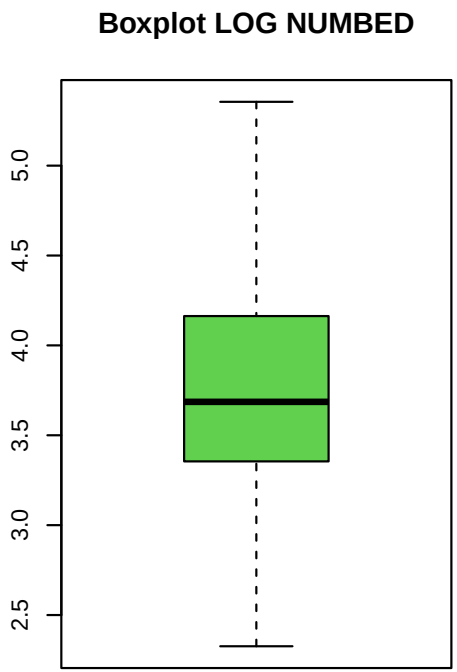
Variabili quantitative

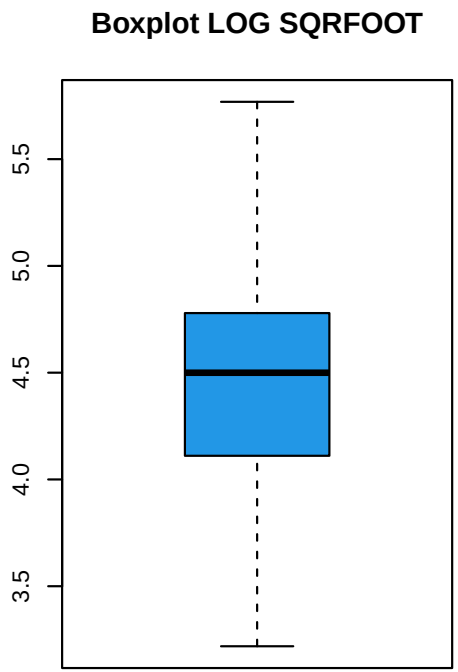
```
apply(WNH[, c("TPY","NUMBED", "SQRFOOT")], 2, summary, na.rm=TRUE)
```

```
##                TPY    NUMBED   SQRFOOT
## Min.      11.56831  18.00000   5.64400
## 1st Qu.   57.21233  61.00000  28.63800
## Median    81.23497  90.00000  39.88300
## Mean      89.72828  97.71867  50.25744
## 3rd Qu.  109.50683 119.00000  64.28050
## Max.     440.66575 457.00000 262.00000
```

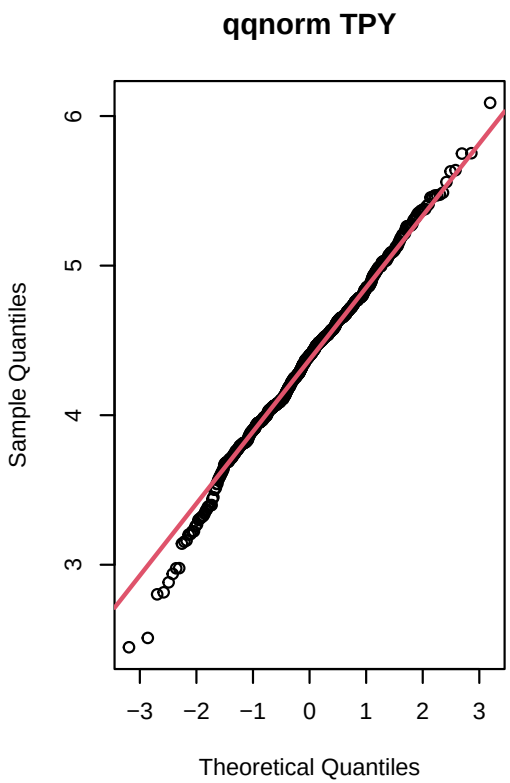
Variabili quantitative

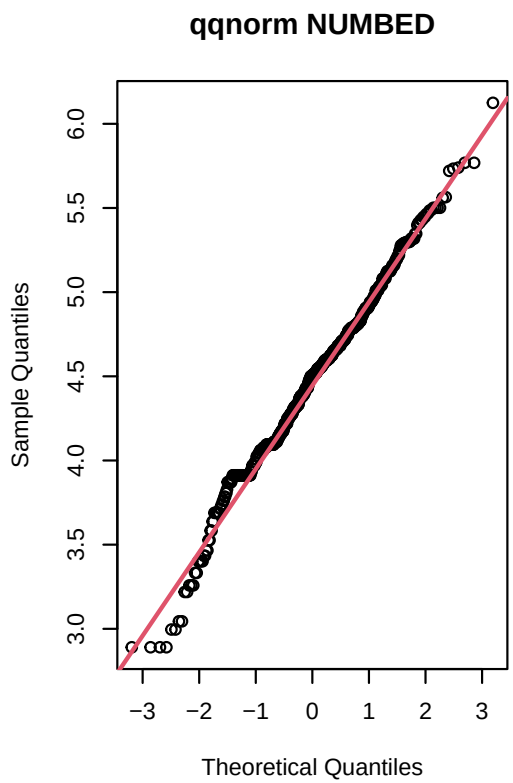


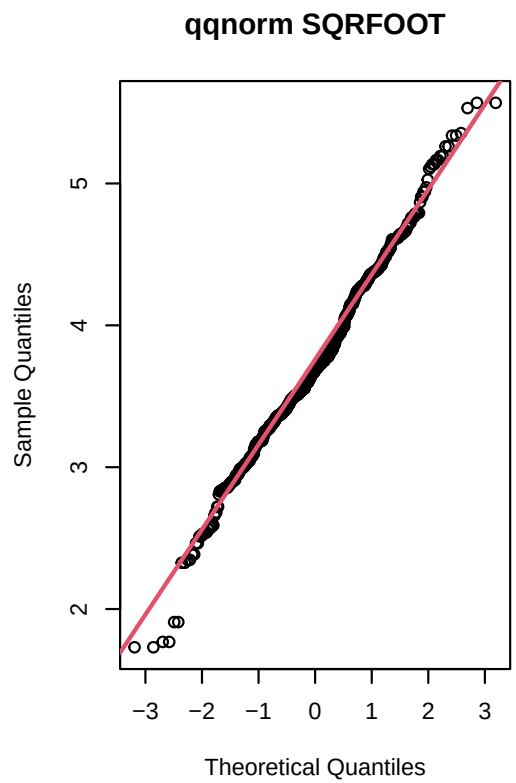




Confronto con distribuzione Normale







Regressione

Misure di relazione lineare

- Covarianza

```
cov(WNH$SQRFOOT, WNH$TPY)
```

```
## [1] 1370.001
```

```
cov(WNH$NUMBED, WNH$TPY)
```

```
## [1] 2365.443
```

- Correlazione

```
cor(WNH$SQRFOOT, WNH$TPY)
```

```
## [1] 0.8219443
```

```
cor(WNH$NUMBED, WNH$TPY)
```

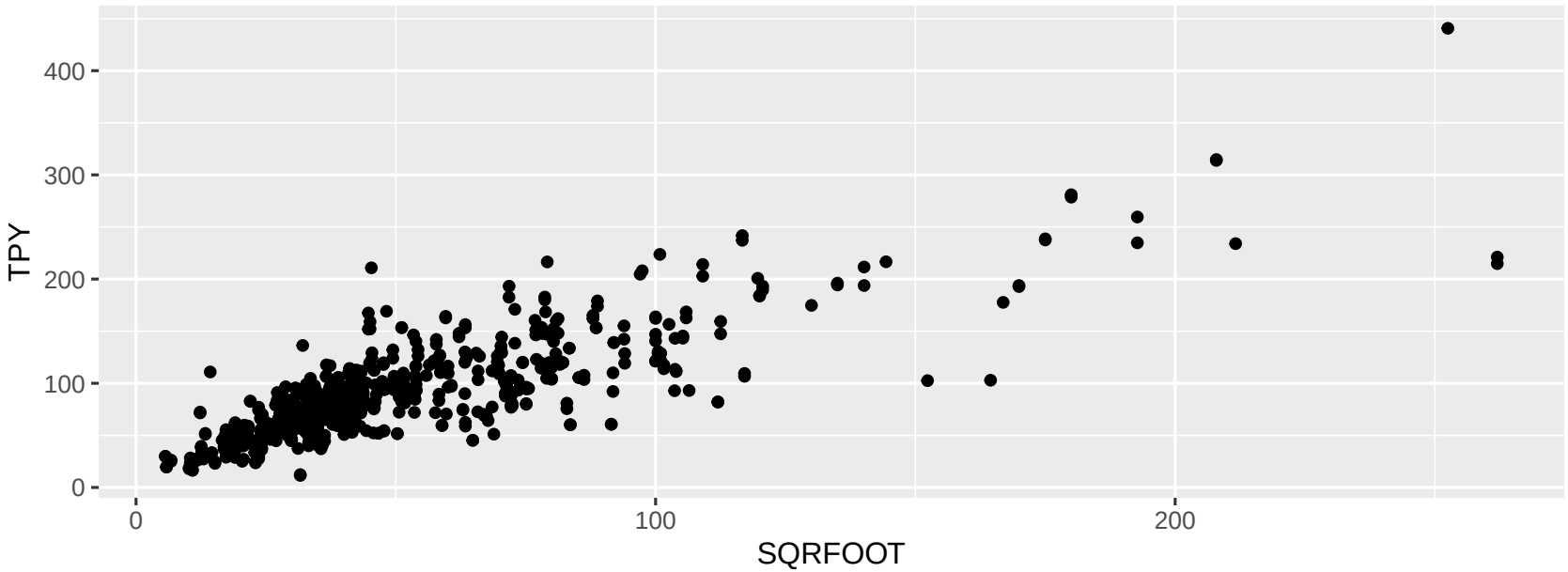
```
## [1] 0.9836241
```

Regressione lineare semplice

- Modello 1:
  - Y = Total patient years (`TPY`)
  - X = Square footage of the nursing home (`SQRFOOT`)
- Modello 2:
  - Y = Total patient years (`TPY`)
  - X = Number of beds (`NUMBED`)

Modello 1 - Grafico dispersione

```
ggplot(data = WNH )+geom_point(aes(x=SQRFOOT,y=TPY))
```



Modello 1

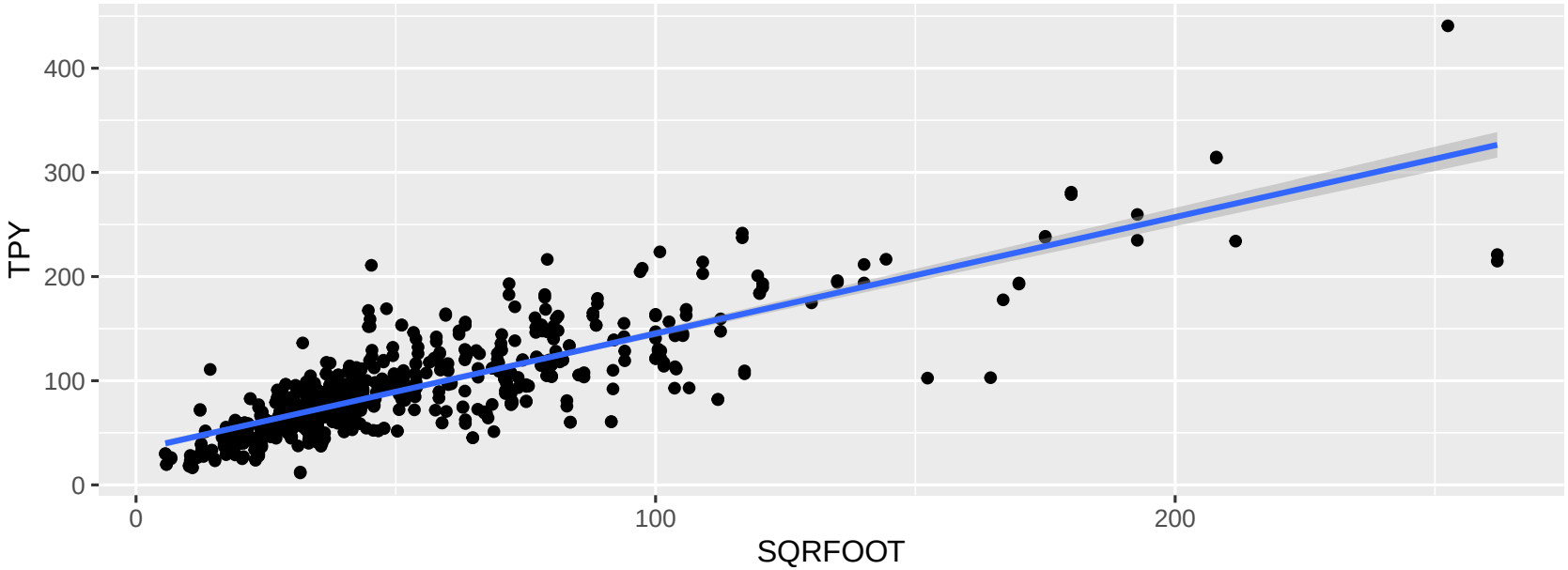
```
mod1<-lm(TPY ~ SQRFOOT, data=WNH)
summary(mod1)
```

```
##
## Call:
## lm(formula = TPY ~ SQRFOOT, data = WNH)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -114.501  -15.391   -2.426   15.615  126.599
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.54754    1.78645   18.78   <2e-16 ***
## SQRFOOT      1.11786    0.02917   38.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.14 on 705 degrees of freedom
## Multiple R-squared:  0.6756, Adjusted R-squared:  0.6751
## F-statistic:  1468 on 1 and 705 DF,  p-value: < 2.2e-16
```

Modello 1 - Retta di regressione

```
ggplot(data=WNH) + geom_point(aes(x=SQRFOOT, y=TPY))+geom_smooth(aes(x=SQRFOOT, y=TPY), method='lm')
```

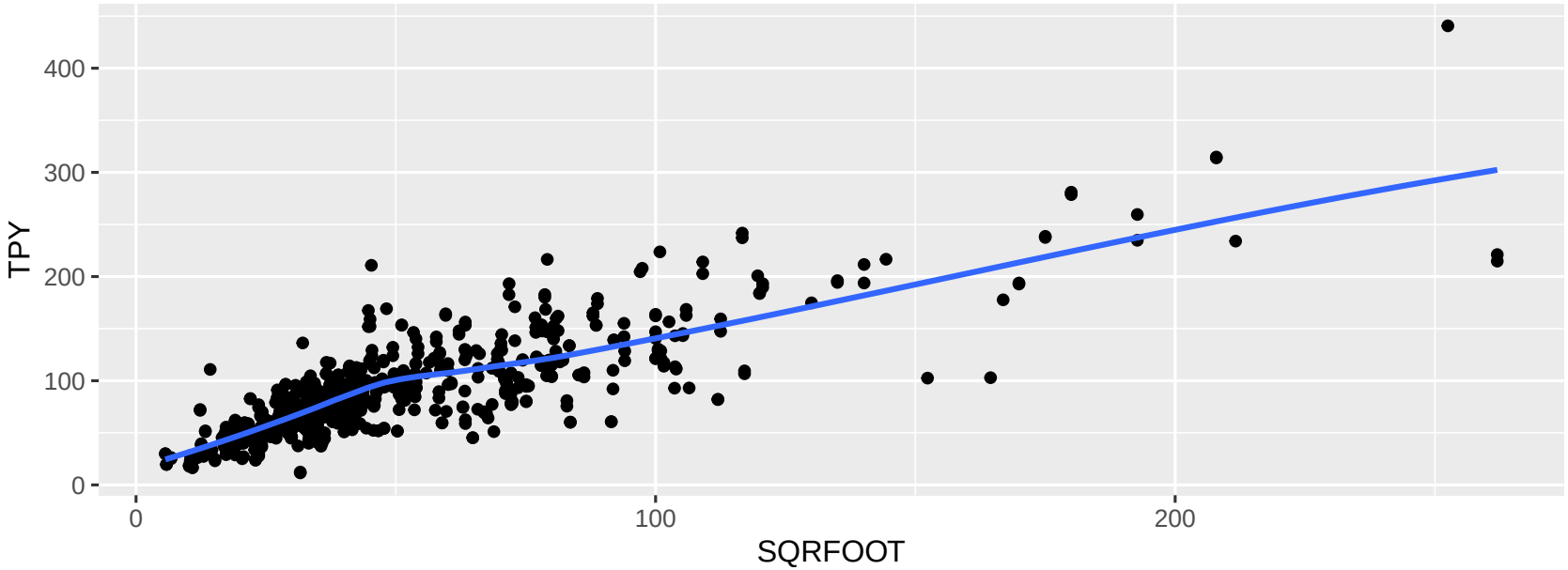
```
## `geom_smooth()` using formula 'y ~ x'
```



Modello 1 - Curva di regressione (metodo Lowess)

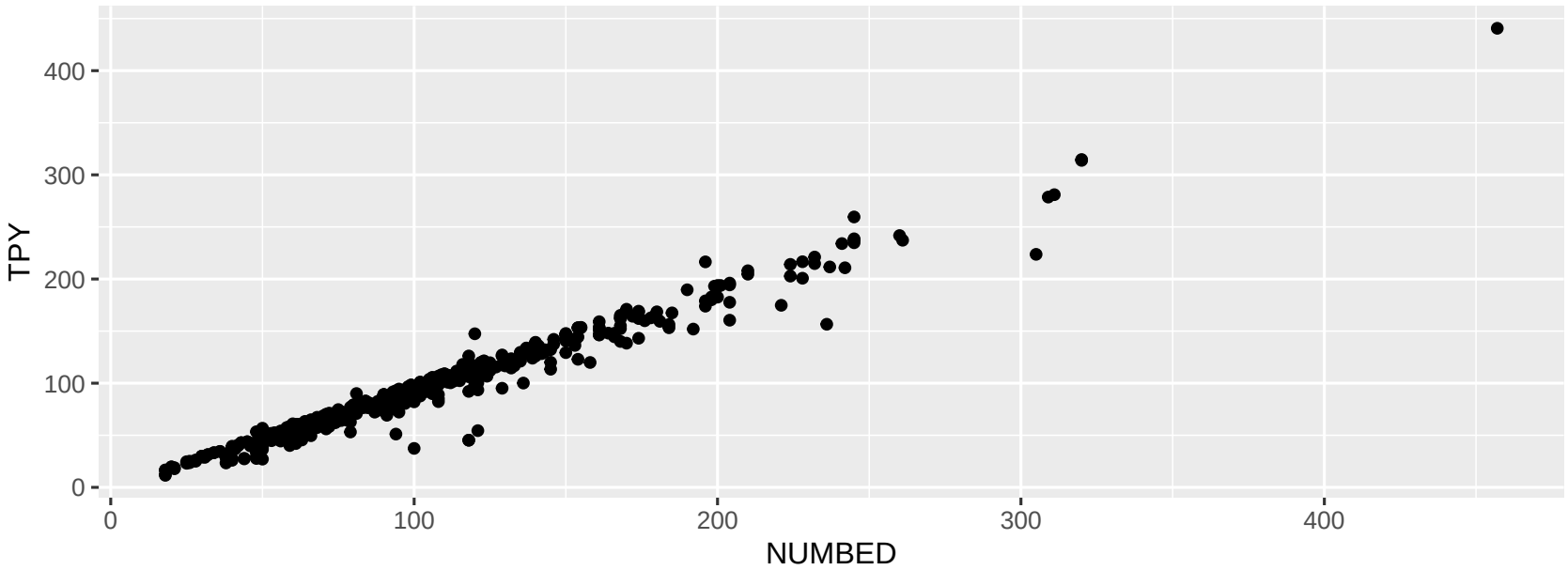
```
ggplot(mapping = aes(x=SQRFOOT, y=TPY), data = WNH)+geom_point()+geom_smooth(aes(), se=F)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Modello 2 - Grafico dispersione

```
ggplot(data = WNH )+geom_point(aes(x=NUMBED,y=TPY))
```



Modello 2

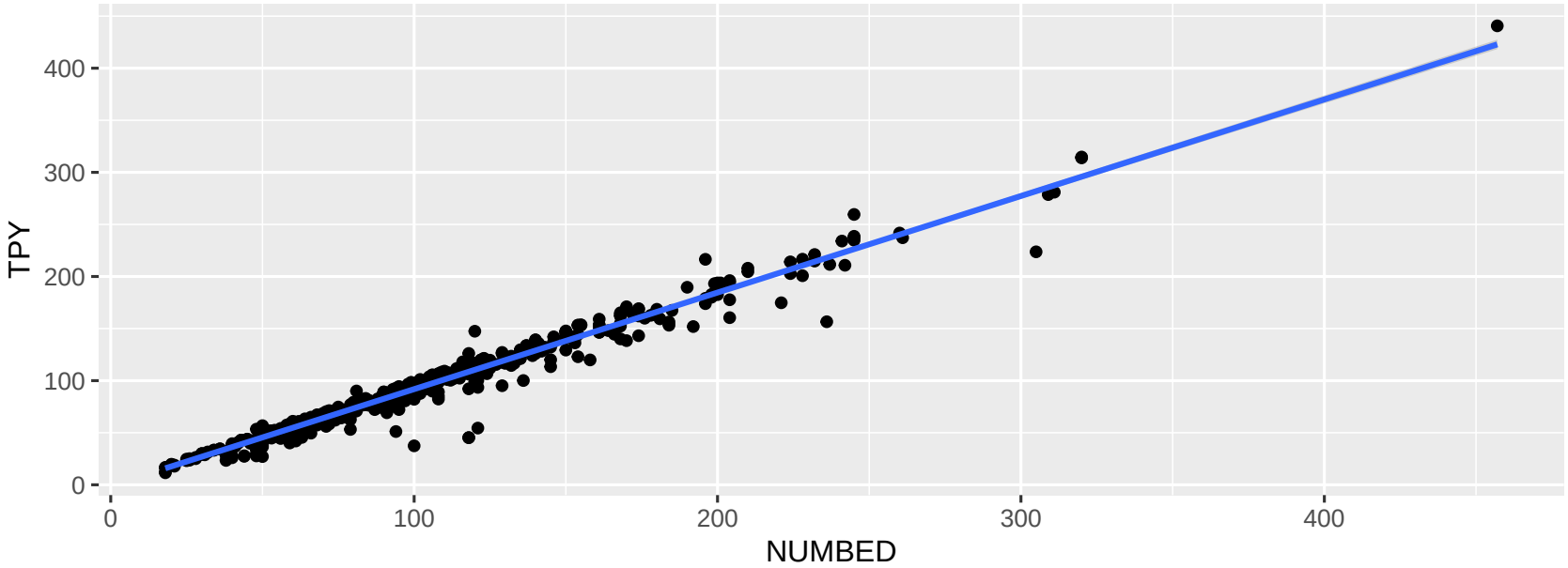
```
mod2<-lm(TPY ~ NUMBED, data=WNH)
summary(mod2)
```

```
##
## Call:
## lm(formula = TPY ~ NUMBED, data = WNH)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.251  -2.075   0.882   3.983  37.062
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.875548   0.703734  -1.244    0.214
## NUMBED       0.927191   0.006398 144.908   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.587 on 705 degrees of freedom
## Multiple R-squared:  0.9675, Adjusted R-squared:  0.9675
## F-statistic: 2.1e+04 on 1 and 705 DF,  p-value: < 2.2e-16
```

Modello 2 - Retta di regressione

```
ggplot(data=WNH) + geom_point(aes(x=NUMBED, y=TPY))+geom_smooth(aes(x=NUMBED, y=TPY), method='lm')
```

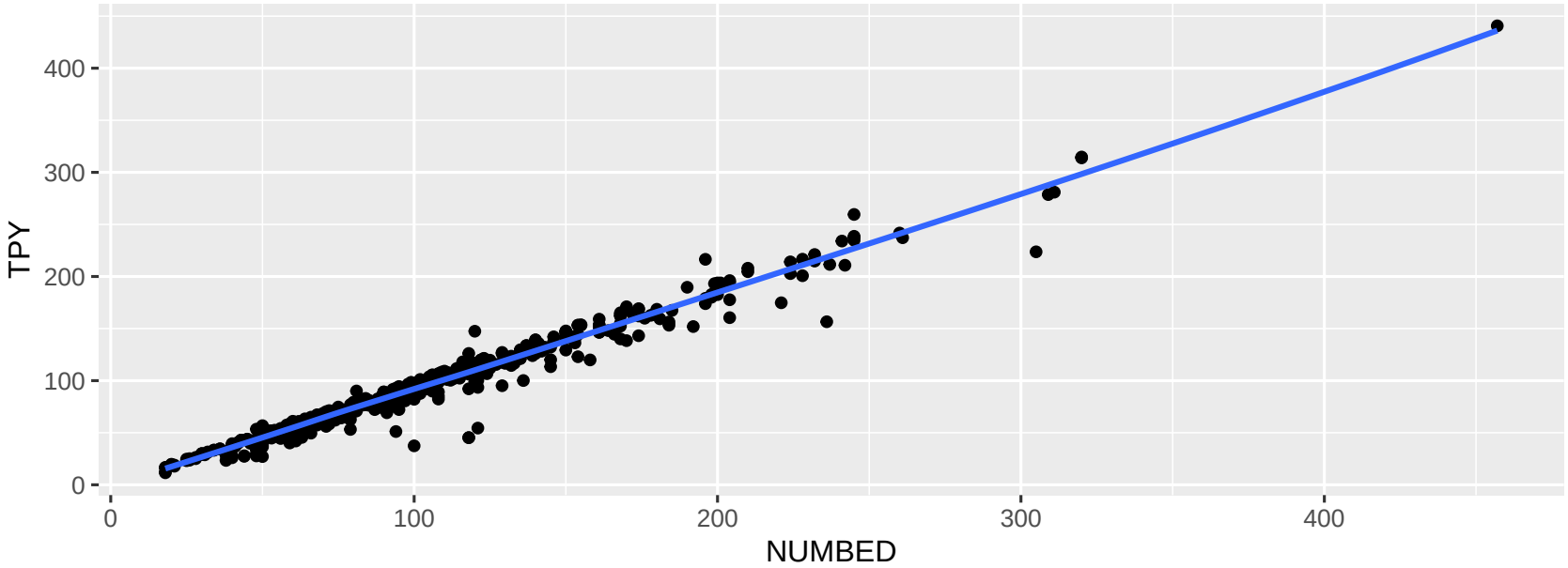
```
## `geom_smooth()` using formula 'y ~ x'
```



Modello 2 - Curva di regressione (metodo Lowess)

```
ggplot(mapping = aes(x=NUMBED, y=TPY), data = WNH)+geom_point()+geom_smooth(aes(), se=F)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Regressione multipla

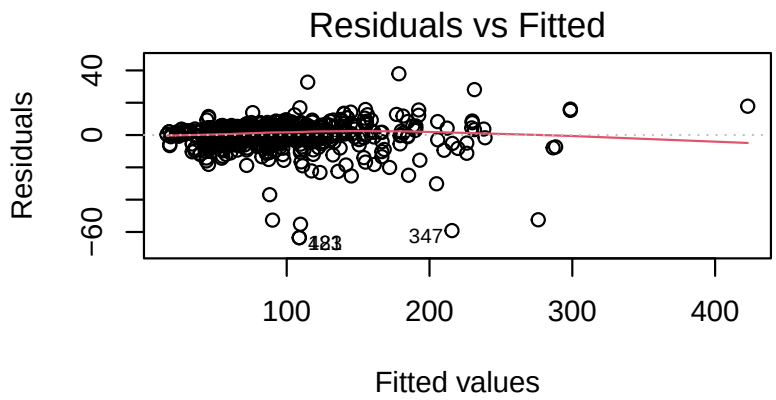
- Modello 3:
  - Y = Total patient years (`TPY`)
  - X = Number of beds (`NUMBED`)
  - Z = Square footage of the nursing home (`SQRFOOT`)

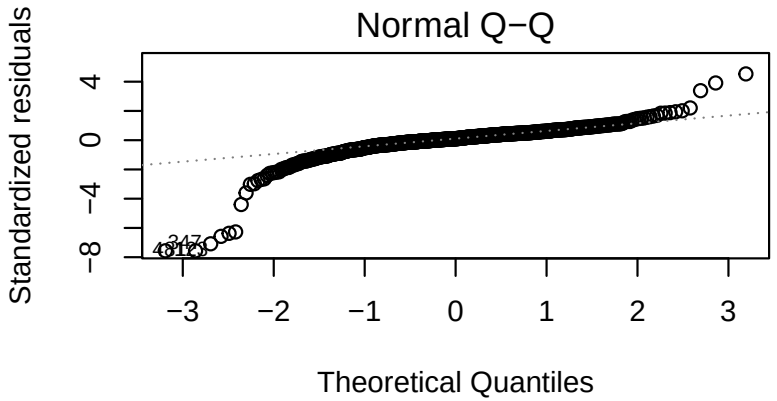
Modello 3

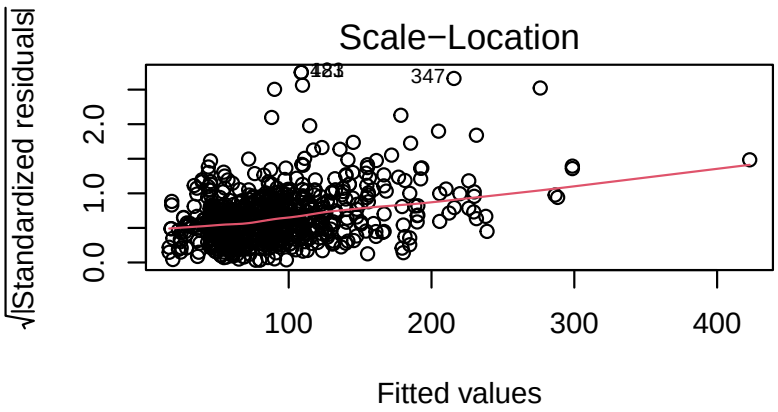
```
mod3<-lm(TPY ~ NUMBED+SQRFOOT, data=WNH)
summary(mod3)
```

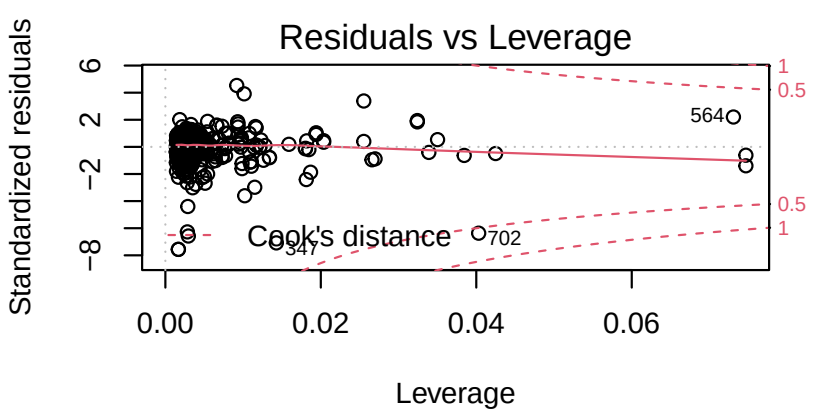
```
##
## Call:
## lm(formula = TPY ~ NUMBED + SQRFOOT, data = WNH)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.523  -2.073   0.868   3.875  37.927
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.45331    0.69315  -0.654    0.513
## NUMBED       0.87808    0.01078  81.487  < 2e-16 ***
## SQRFOOT      0.08709    0.01555   5.602 3.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.408 on 704 degrees of freedom
## Multiple R-squared:  0.9689, Adjusted R-squared:  0.9688
## F-statistic: 1.097e+04 on 2 and 704 DF,  p-value: < 2.2e-16
```

Modello 3 - Residui







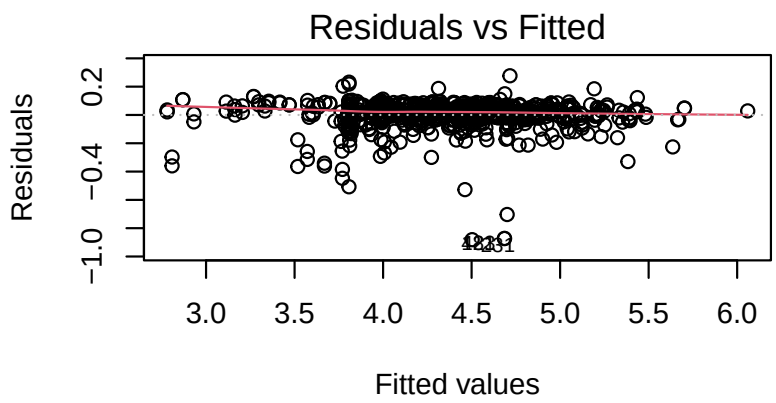


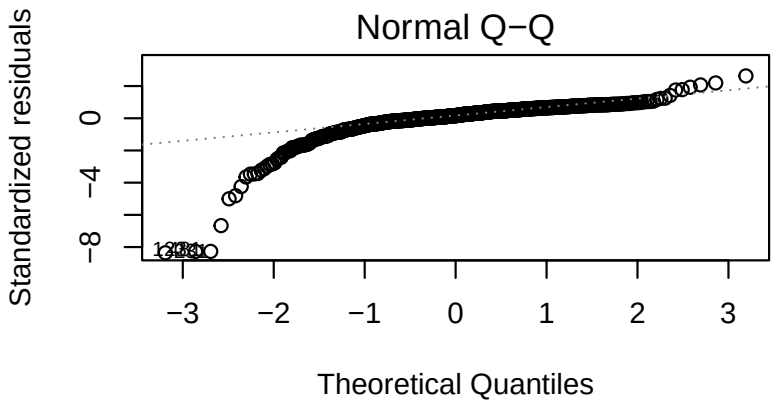
Modello 3 log

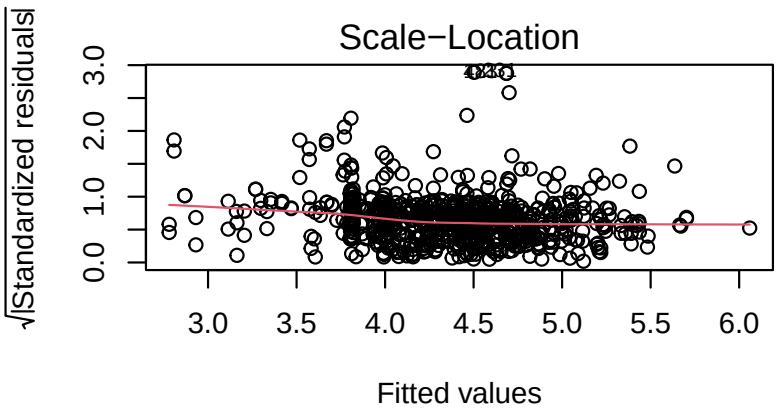
```
mod3l<-lm(log(TPY)~ log(NUMBED)+log(SQRFOOT), data=WNH)
summary(mod3l)
```

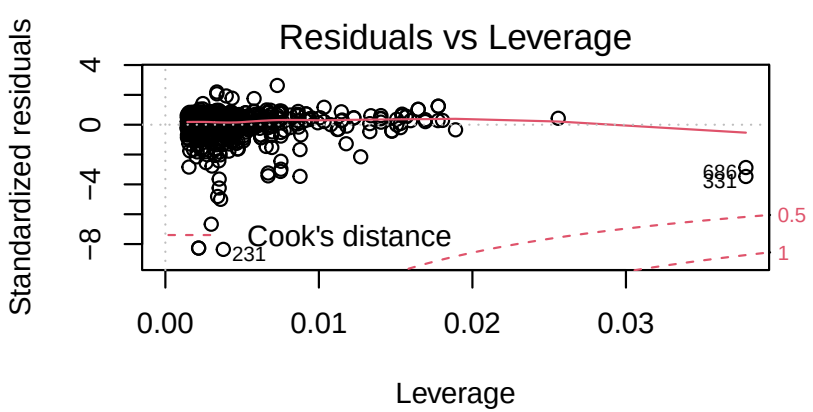
```
##
## Call:
## lm(formula = log(TPY) ~ log(NUMBED) + log(SQRFOOT), data = WNH)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88114 -0.01930  0.01720  0.05531  0.27691
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.13917    0.03779  -3.683 0.000249 ***
## log(NUMBED)   0.98877    0.01442  68.554  < 2e-16 ***
## log(SQRFOOT)  0.02583    0.01157   2.234 0.025825 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1057 on 704 degrees of freedom
## Multiple R-squared:  0.9571, Adjusted R-squared:  0.9569
## F-statistic:  7846 on 2 and 704 DF,  p-value: < 2.2e-16
```

Modello 3 log - Residui









Test ANOVA Modello 2 - Modello 3

```
anova(mod2,mod3)
```

```
## Analysis of Variance Table
##
## Model 1: TPY ~ NUMBED
## Model 2: TPY ~ NUMBED + SQRFOOT
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1    705 51987
## 2    704 49768  1    2218.4 31.38 3.044e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod3bis<-lm(TPY ~ NUMBED+SQRFOOT+PRO+URBAN+TAXEXEMPT+SELFFUNDINS+MCERT, data=WNH)
summary(mod3bis)
```

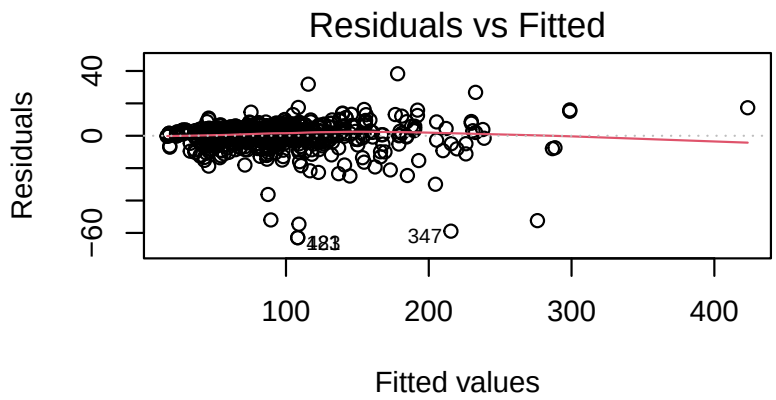
```
##
## Call:
## lm(formula = TPY ~ NUMBED + SQRFOOT + PRO + URBAN + TAXEXEMPT +
##     SELFFUNDINS + MCERT, data = WNH)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.613  -2.182   1.049   3.841  38.891
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.11232    1.53426  -0.725   0.4687
## NUMBED                    0.88475    0.01122  78.879  < 2e-16 ***
## SQRFOOT                   0.07949    0.01671   4.756  2.4e-06 ***
## PROprofit                 0.05346    1.13293   0.047   0.9624
## URBANurban               -1.34140    0.66356  -2.022   0.0436 *
## TAXEXEMPTtax-exempt       1.44303    1.12020   1.288   0.1981
## SELFFUNDINSself-funded    0.46701    0.66752   0.700   0.4844
## MCERTMedicare certified   0.27550    1.12463   0.245   0.8066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.382 on 699 degrees of freedom
## Multiple R-squared:  0.9693, Adjusted R-squared:  0.969
## F-statistic:  3154 on 7 and 699 DF,  p-value: < 2.2e-16
```

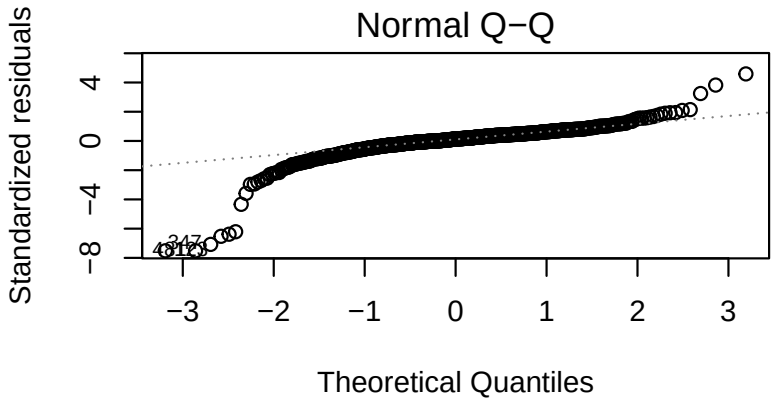
Modello 3 bis

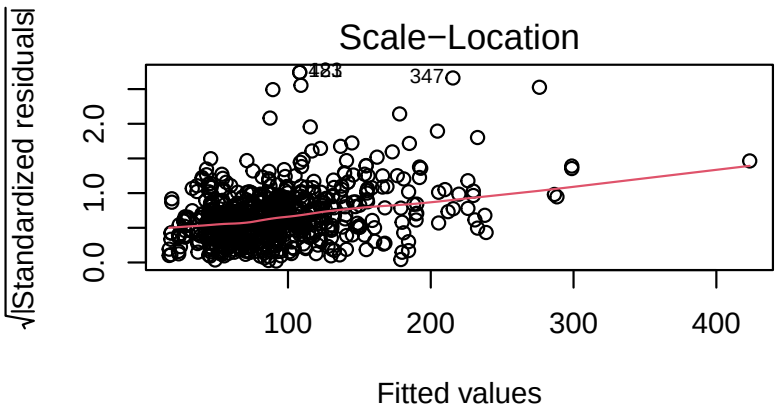
```
mod3bis<-lm(TPY ~ NUMBED+SQRFOOT+URBAN, data=WNH)
summary(mod3bis)
```

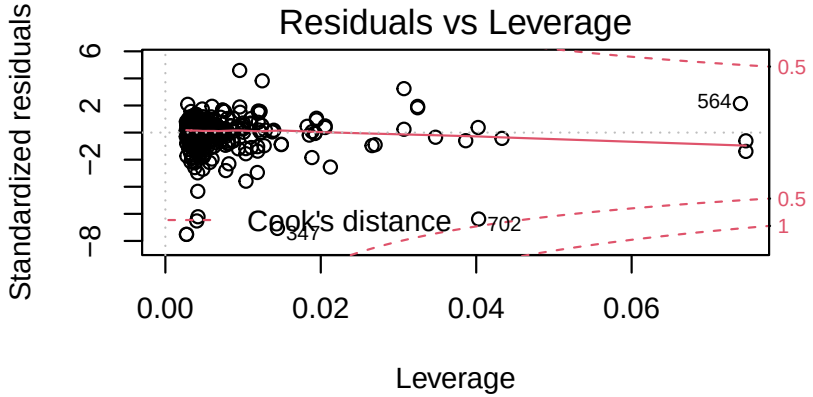
```
##
## Call:
## lm(formula = TPY ~ NUMBED + SQRFOOT + URBAN, data = WNH)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.933  -2.146   1.120   3.897  38.282
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02342    0.71910  -0.033   0.9740
## NUMBED       0.88090    0.01083  81.373   <2e-16 ***
## SQRFOOT      0.08808    0.01551   5.678    2e-08 ***
## URBANurban  -1.41560    0.65167  -2.172   0.0302 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.386 on 703 degrees of freedom
## Multiple R-squared:  0.9691, Adjusted R-squared:  0.969
## F-statistic:  7352 on 3 and 703 DF,  p-value: < 2.2e-16
```

Modello 3 bis - Residui







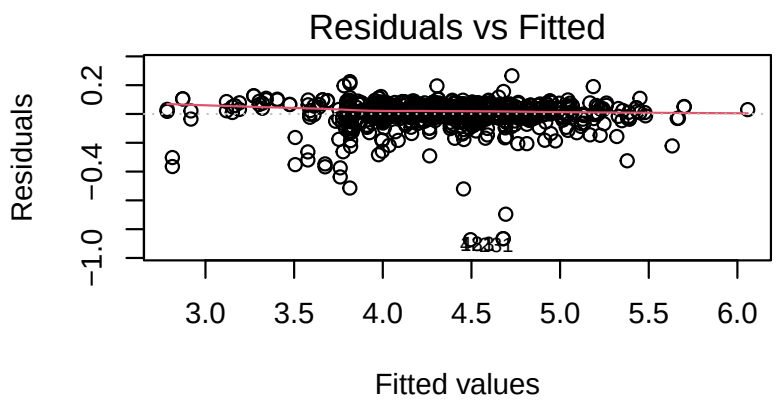


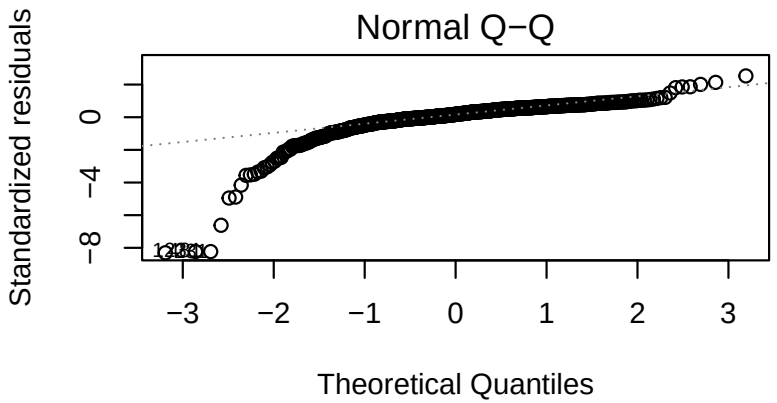
Modello 3 bisl

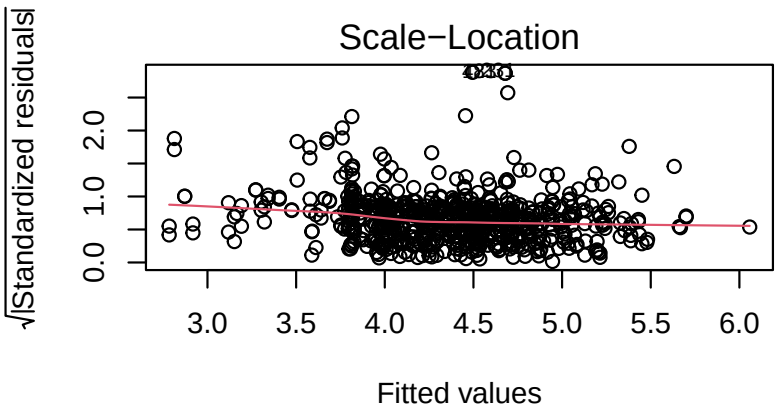
```
mod3bisl<-lm(log(TPY) ~ log(NUMBED)+log(SQRFOOT)+URBAN, data=WNH)
summary(mod3bisl)
```

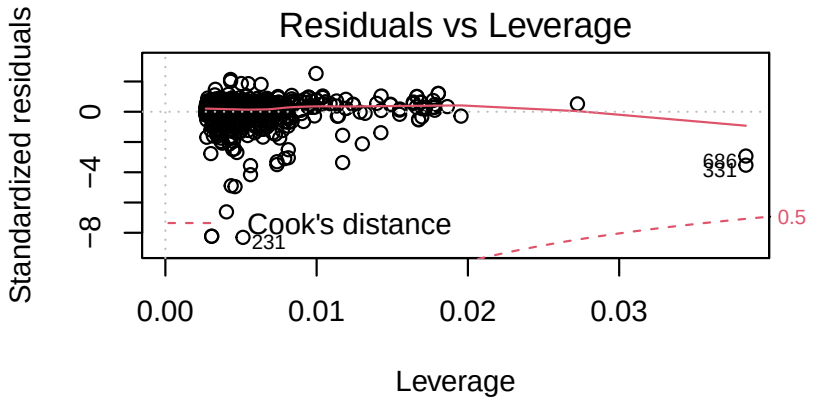
```
##
## Call:
## lm(formula = log(TPY) ~ log(NUMBED) + log(SQRFOOT) + URBAN, data = WNH)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87290 -0.02260  0.01799  0.05686  0.26534
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.145836   0.037827  -3.855 0.000126 ***
## log(NUMBED)   0.990985   0.014425  68.698  < 2e-16 ***
## log(SQRFOOT)  0.027449   0.011562   2.374 0.017865 *
## URBANurban   -0.017335   0.008162  -2.124 0.034030 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1054 on 703 degrees of freedom
## Multiple R-squared:  0.9573, Adjusted R-squared:  0.9572
## F-statistic:  5258 on 3 and 703 DF,  p-value: < 2.2e-16
```

Modello 3 bisl - Residui









Predict

Divisione per anno

```
a20 <- WNH[WNH$CRYEAR == "2000",]
a21 <- WNH[WNH$CRYEAR == "2001",]
```

```
modP0 <- lm(TPY ~ NUMBED + SQRFOOT , data=a20)
modP1 <- lm(TPY ~ NUMBED + SQRFOOT, data=a21)
```

```
new_obs = data.frame(NUMBED=39, SQRFOOT=25.845)
predict(modP0, newdata = new_obs)
```

```
##        1
## 36.14731
```

```
predict(modP1, newdata = new_obs,)
```

```
##        1
## 36.00048
```

Predict - Modello 1

```
mod1
```

```
##
## Call:
## lm(formula = TPY ~ SQRFOOT, data = WNH)
##
## Coefficients:
## (Intercept)      SQRFOOT
##      33.548        1.118
```

```
new_obs = data.frame(SQRFOOT=c(52, 89))
predict(mod1, newdata = new_obs, interval = 'confidence')
```

```
##         fit       lwr       upr
## 1  91.67623  89.66996  93.68249
## 2 133.03702 130.04711 136.02693
```

Predict - Modello 2

```
mod2
```

```
##
## Call:
## lm(formula = TPY ~ NUMBED, data = WNH)
##
## Coefficients:
## (Intercept)       NUMBED
##     -0.8755       0.9272
```

```
new_obs2 = data.frame(NUMBED=86)
predict(mod2, newdata = new_obs2)
```

```
##        1
## 78.86284
```

Predict - Modello 3

```
mod3
```

```
##
## Call:
## lm(formula = TPY ~ NUMBED + SQRFOOT, data = WNH)
##
## Coefficients:
## (Intercept)       NUMBED      SQRFOOT
##    -0.45331      0.87808      0.08709
```

```
new_obs3 = data.frame(SQRFOOT=93.522, NUMBED = 65)
predict(mod3, newdata = new_obs3)
```

```
##        1
## 64.76676
```

Predict - Modello 3 bis

```
mod3bis
```

```
##
## Call:
## lm(formula = TPY ~ NUMBED + SQRFOOT + URBAN, data = WNH)
##
## Coefficients:
## (Intercept)       NUMBED      SQRFOOT   URBANurban
##    -0.02342      0.88090      0.08808     -1.41560
```

```
new_obs3 = data.frame(SQRFOOT=81.222, NUMBED = 56, URBAN = 'urban')
predict(mod3bis, newdata = new_obs3)
```

```
##        1
## 55.04502
```

Predict

```
s1 <- sample(1:length(WNH$TPY ),floor(length(WNH$TPY)/3))
test <- WNH[s1,]
train <- WNH[-s1,]
finmod1 <- lm(TPY~SQRFOOT + NUMBED, data=train)
pred1 <- predict(finmod1,newdata=test)
```

Predict

```
RMSD1 <- sqrt(sum((pred1 - log(test$TPY))^2)/length(test$TPY))
head(pred1)
```

```
##       584       678       300       179       574        62
## 156.27979  44.81110  87.66712  97.42565  88.66785 127.57716
```

```
RMSD1
```

```
## [1] 92.68712
```

Analisi

Anova

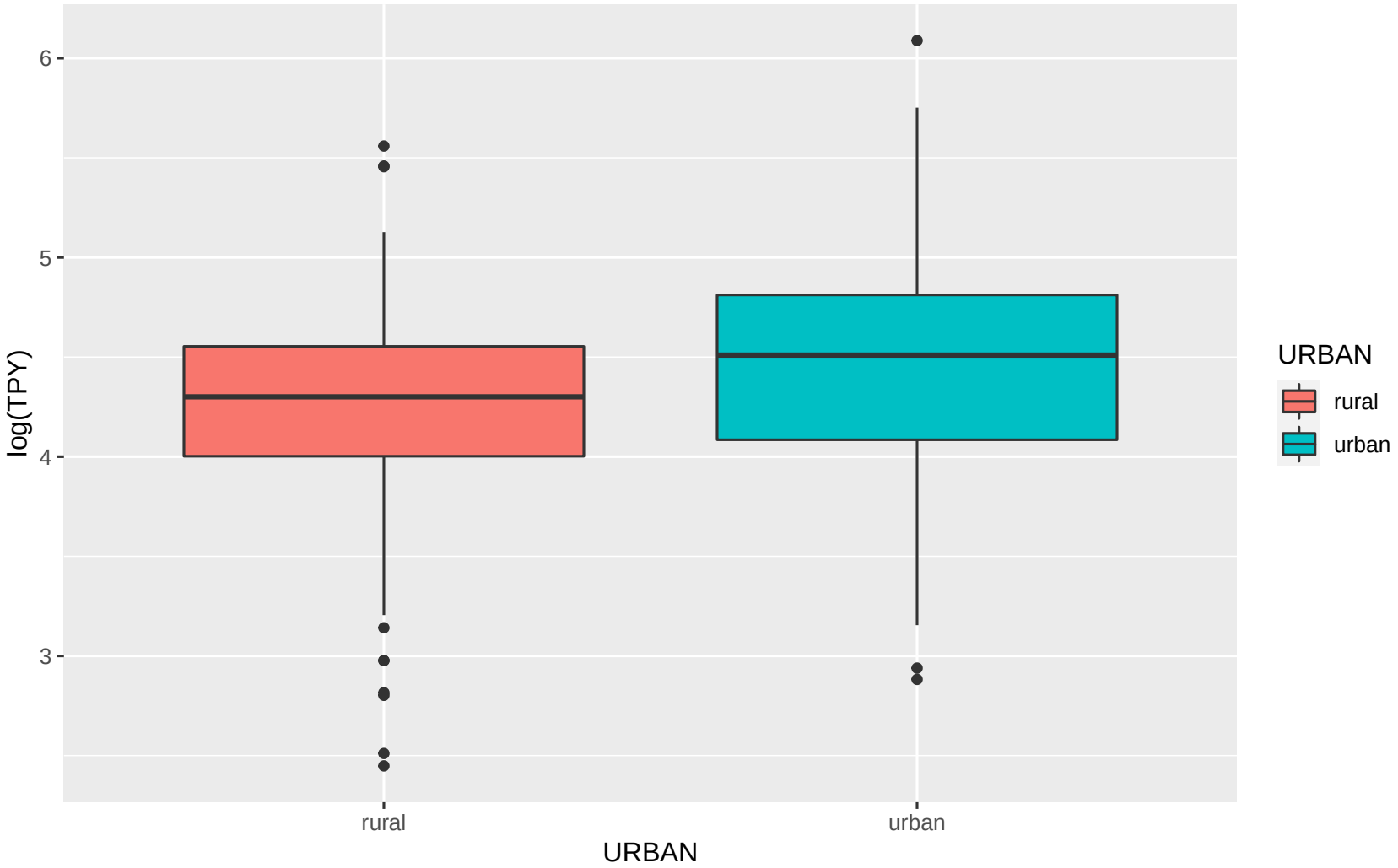
```
anova1<-aov(TPY~URBAN, data=WNH)
summary(anova1)
```

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## URBAN         1   80516   80516   37.35 1.64e-09 ***
## Residuals   705 1519881    2156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

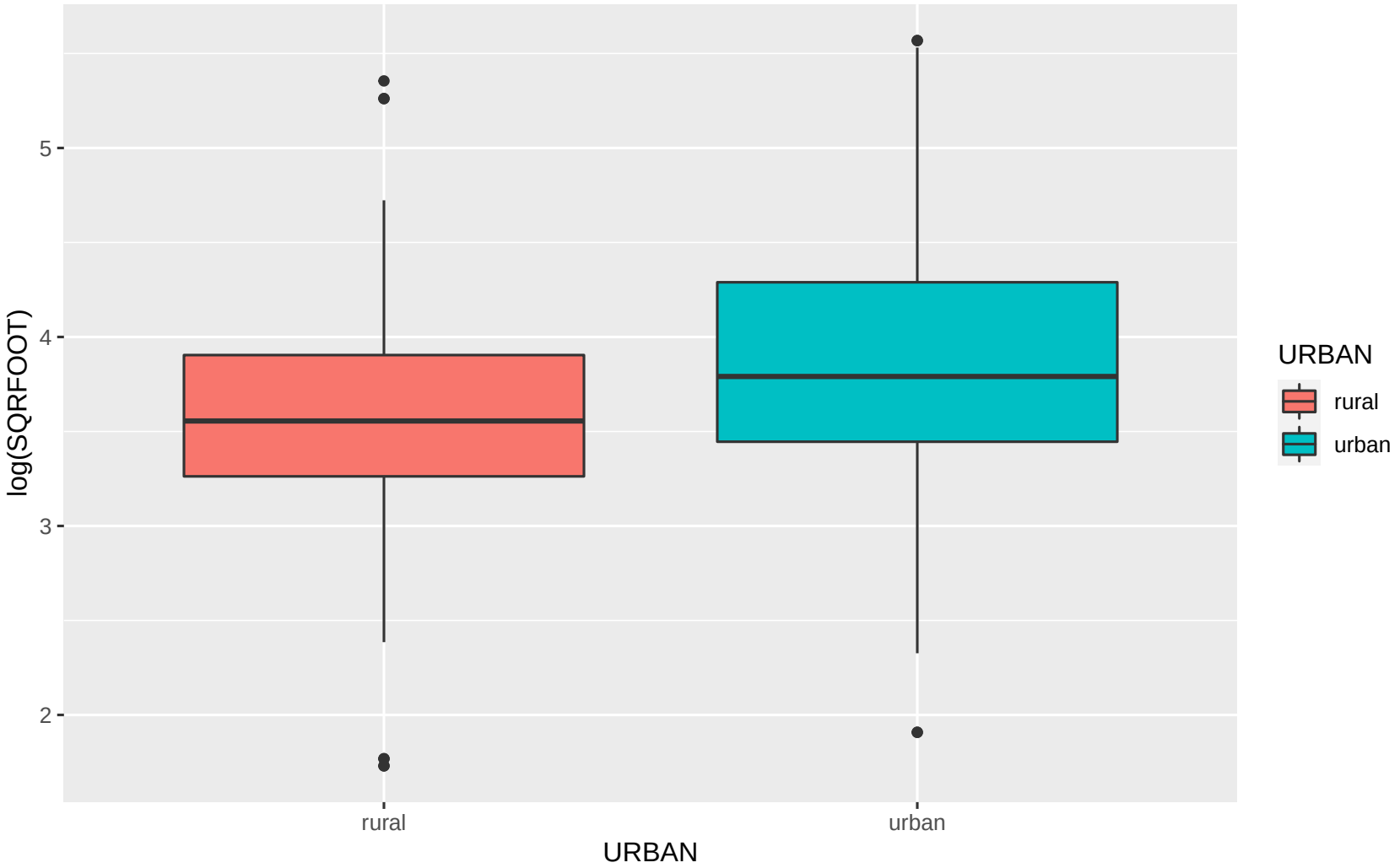
```
anova2<-aov(SQRFOOT~URBAN, data=WNH)
summary(anova2)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## URBAN         1  39151   39151   33.41 1.12e-08 ***
## Residuals   705 826092    1172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

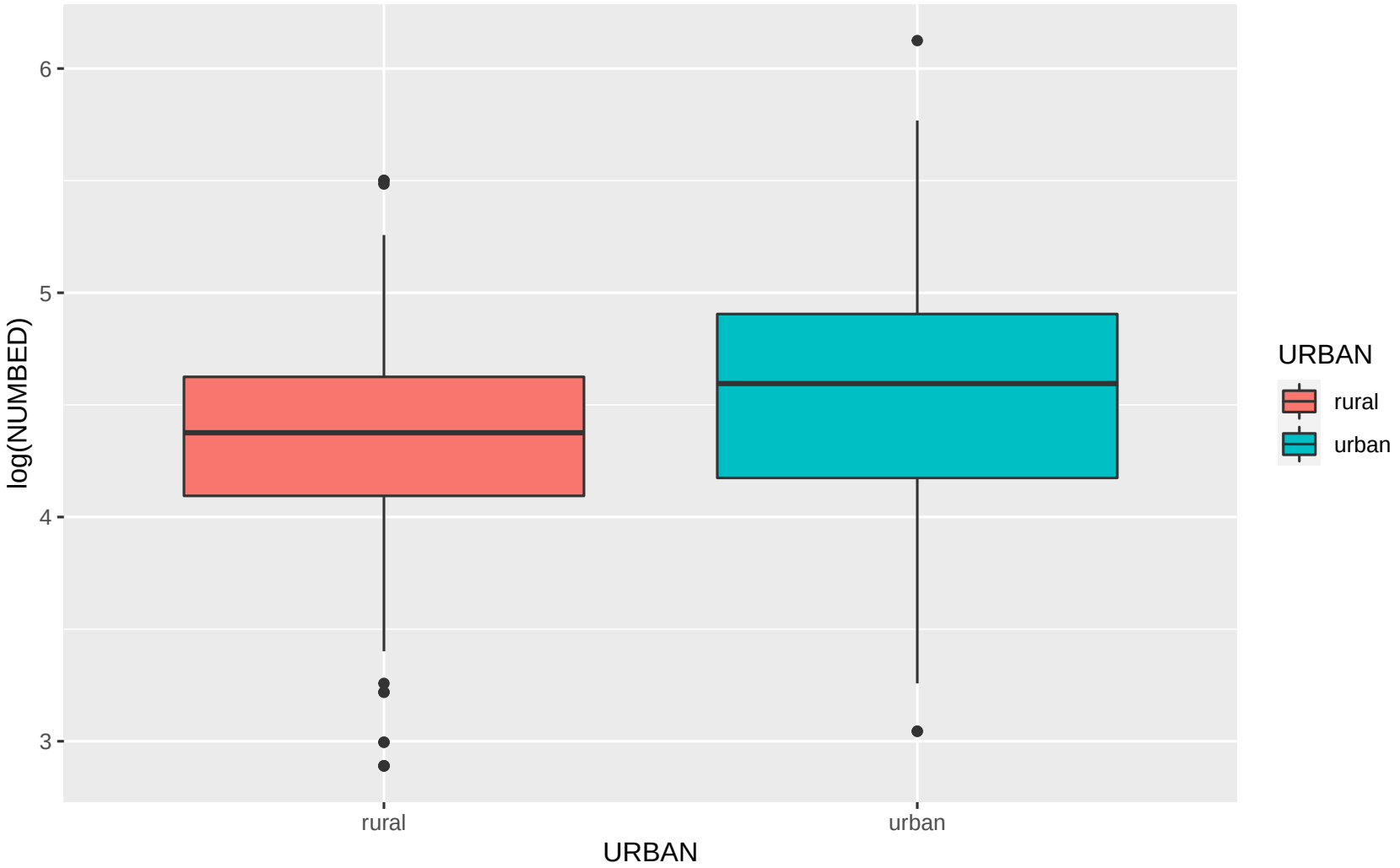
```
ggplot(data = WNH, aes(x=URBAN, y=log(TPY), fill=URBAN)) + geom_boxplot()
```



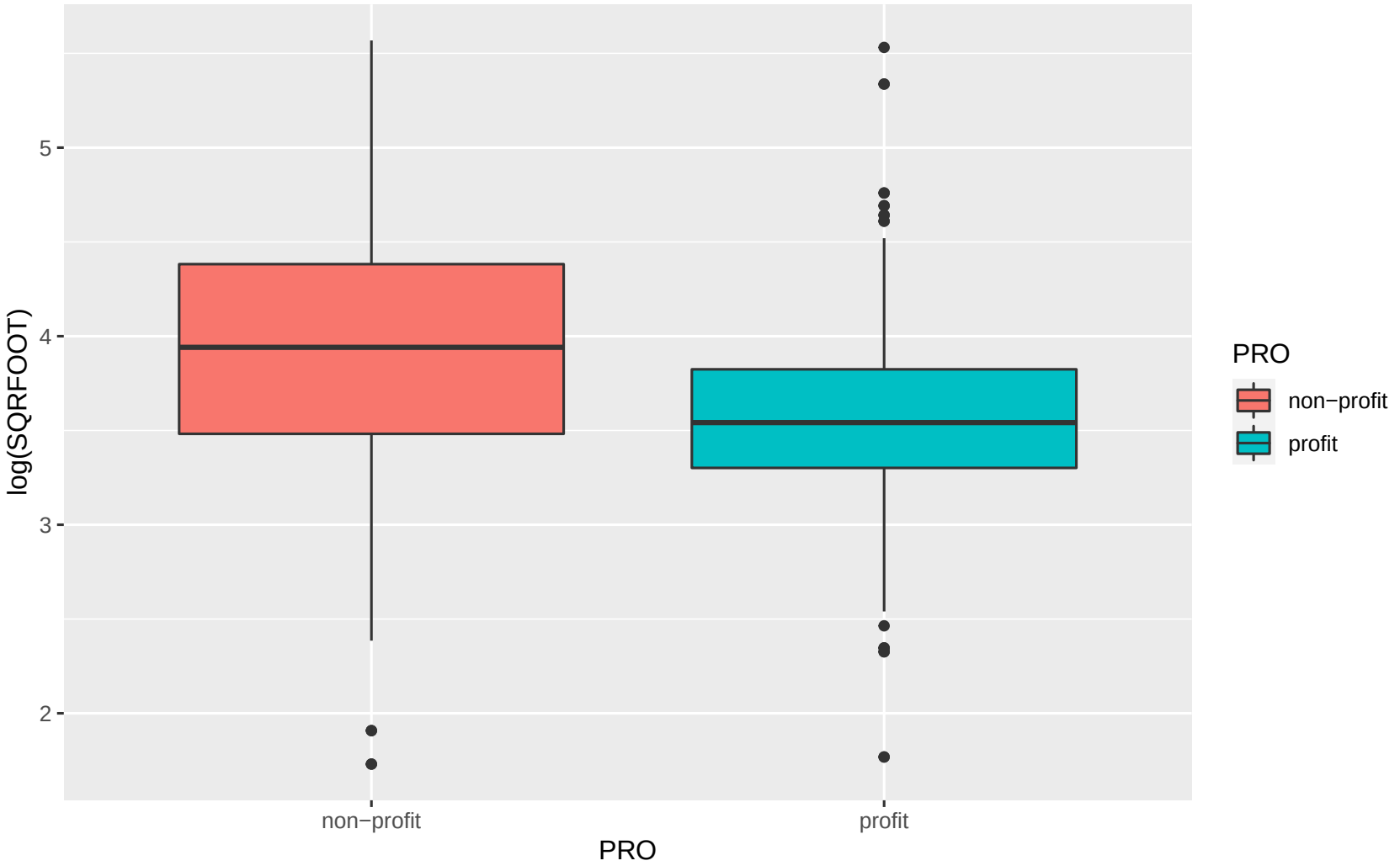
```
ggplot(data = WNH, aes(x=URBAN, y=log(SQRFOOT), fill=URBAN)) + geom_boxplot()
```



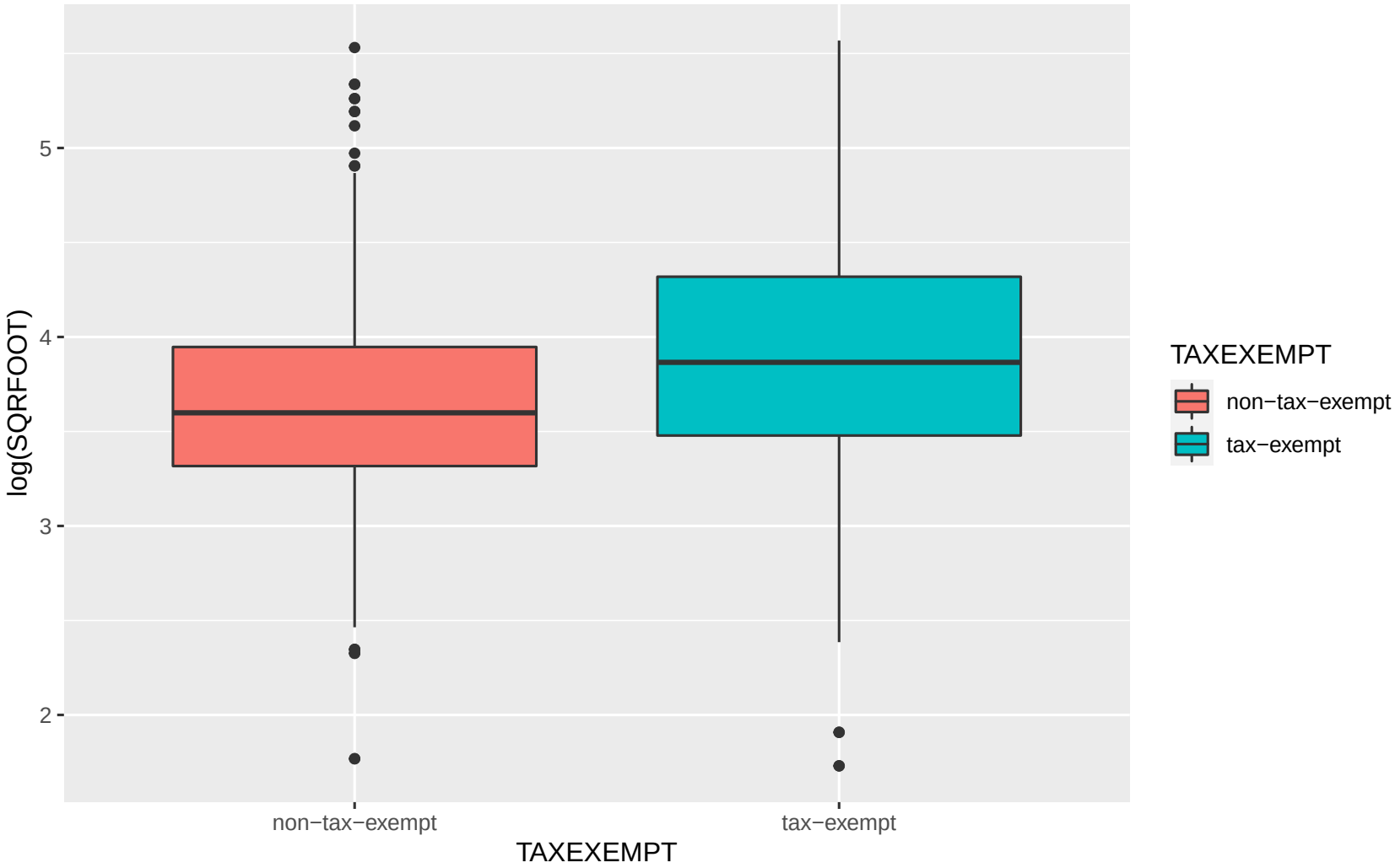
```
ggplot(data = WNH, aes(x=URBAN, y=log(NUMBED), fill=URBAN)) + geom_boxplot()
```



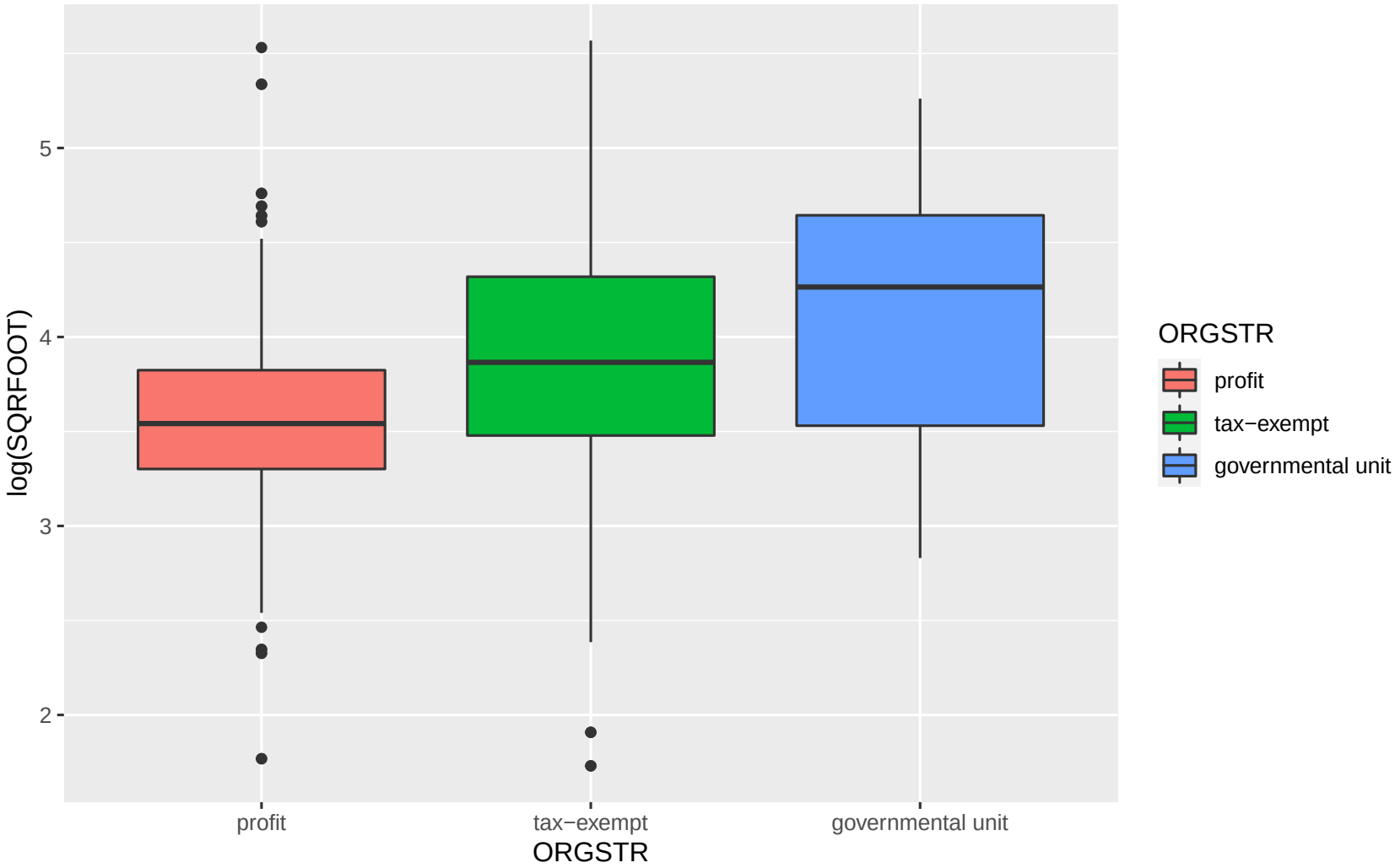
```
ggplot(data = WNH, aes(x=PRO, y=log(SQRFOOT), fill=PRO)) + geom_boxplot()
```



```
ggplot(data = WNH, aes(x=TAXEXEMPT, y=log(SQRFOOT), fill=TAXEXEMPT)) + geom_boxplot()
```



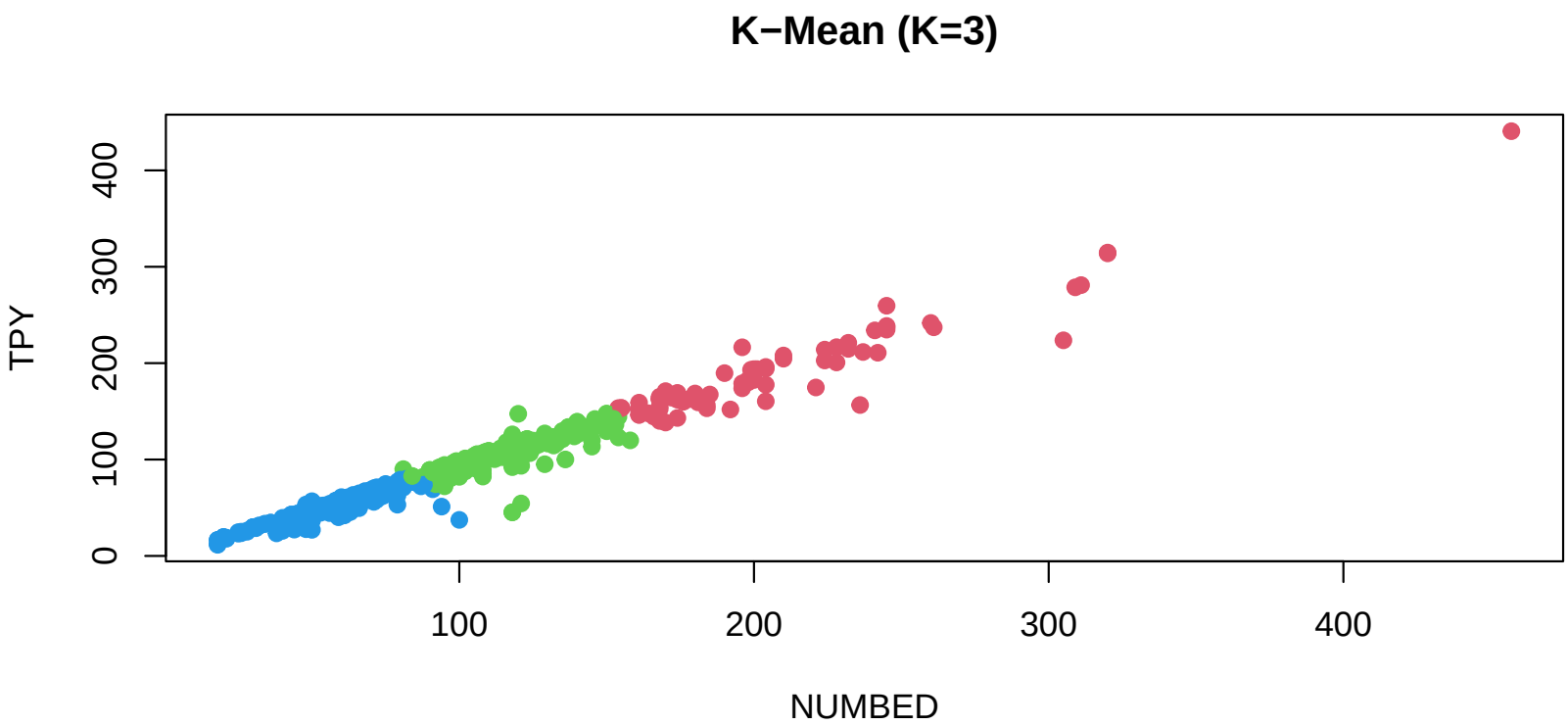
```
ggplot(data = WNH, aes(x=ORGSTR, y=log(SQRFOOT), fill=ORGSTR)) + geom_boxplot()
```



Clustering

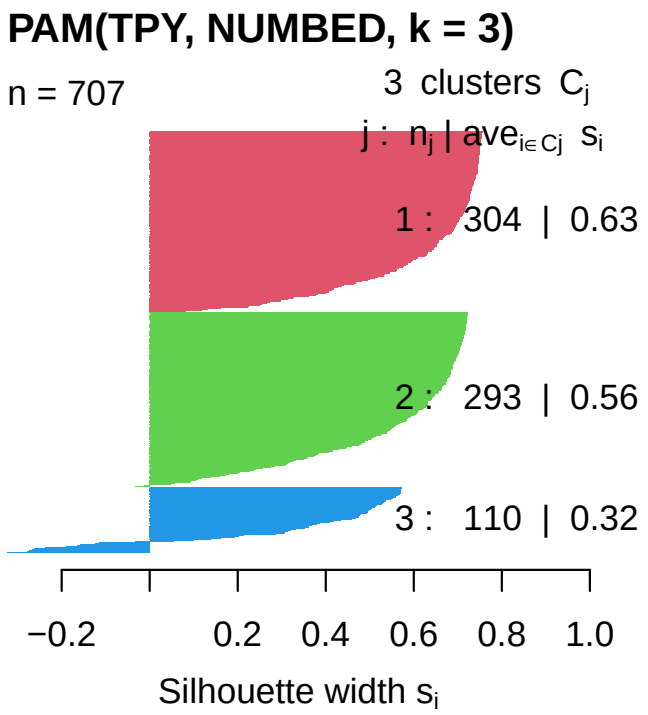
K-Means

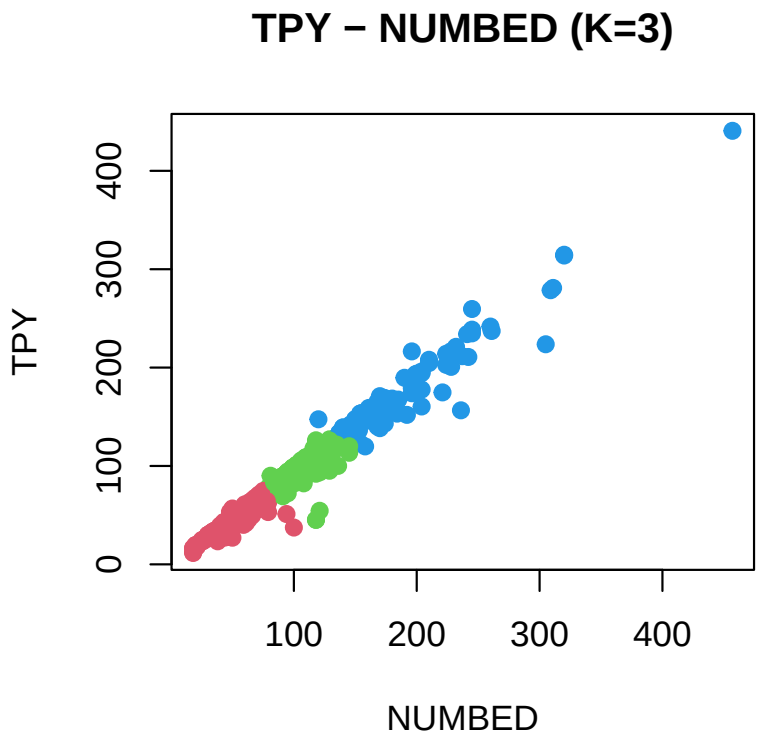
```
k<-kmeans(WNH[,c(3,2)], 3, nstart=5)
plot(WNH$NUMBED, WNH$TPY, col =(k$cluster+1),main="K-Mean (K=3)", xlab ="NUMBED", ylab="TPY", pch=19)
```



PAM

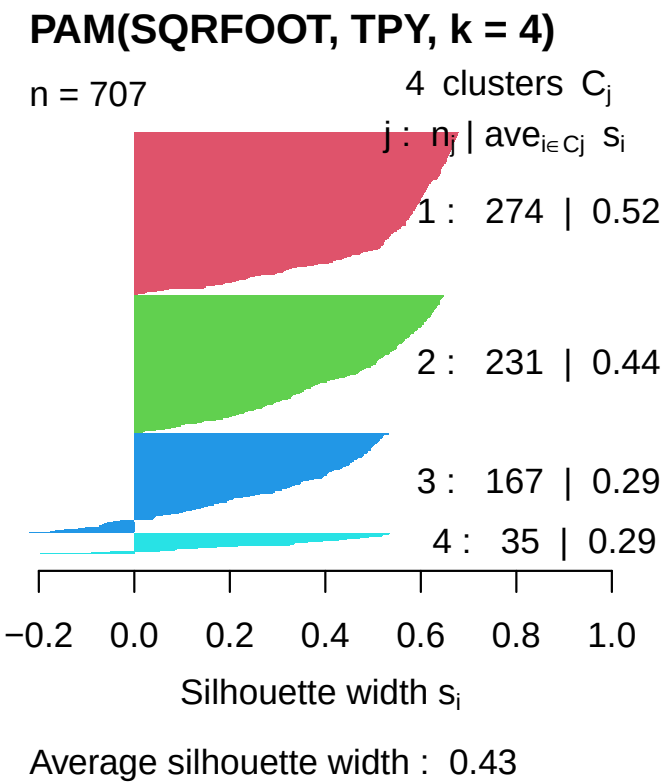
```
pam.out1 <-pam(WNH[,c(3,2)], 3, metric="euclidean", stand=TRUE)
```

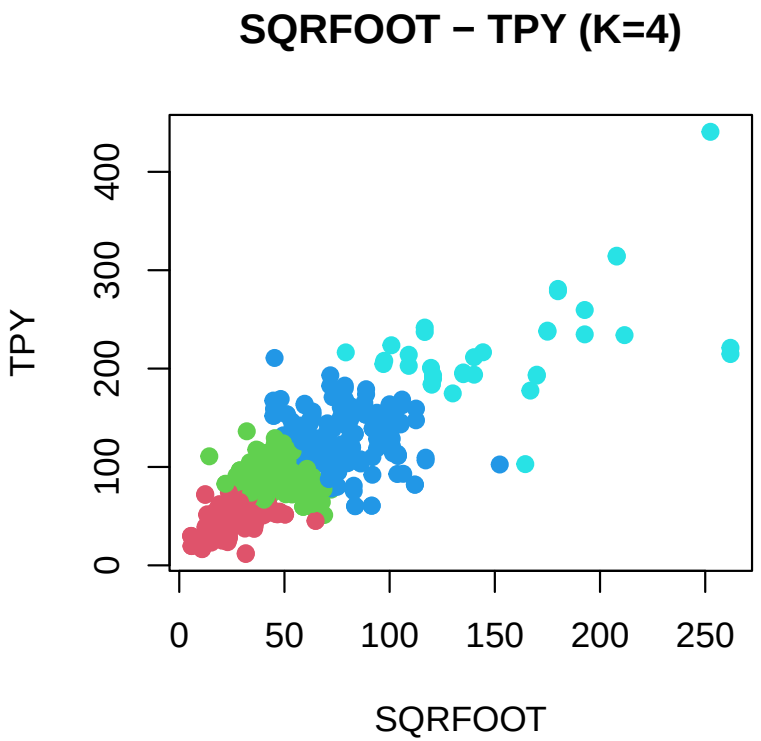




PAM

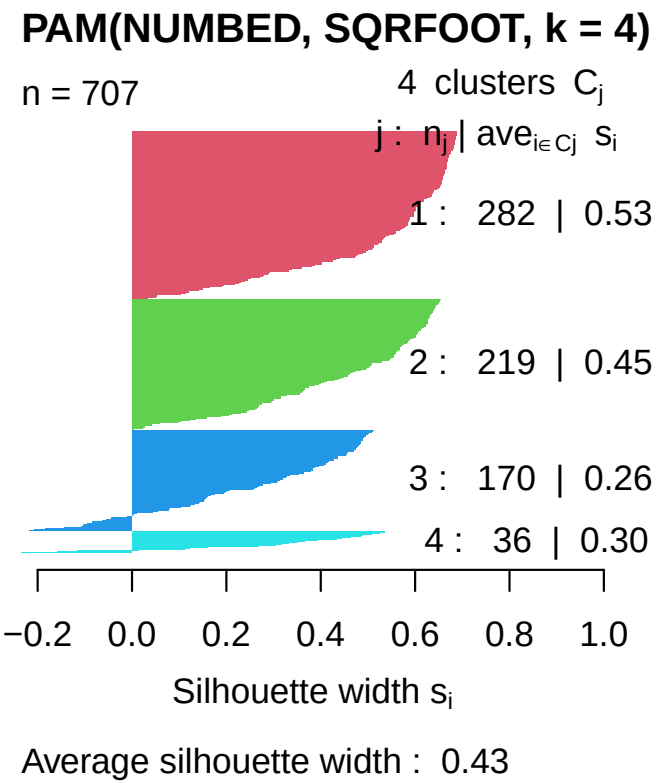
```
pam.out2 <-pam(WNH[,c(4,2)], 4, metric="euclidean", stand=TRUE)
```

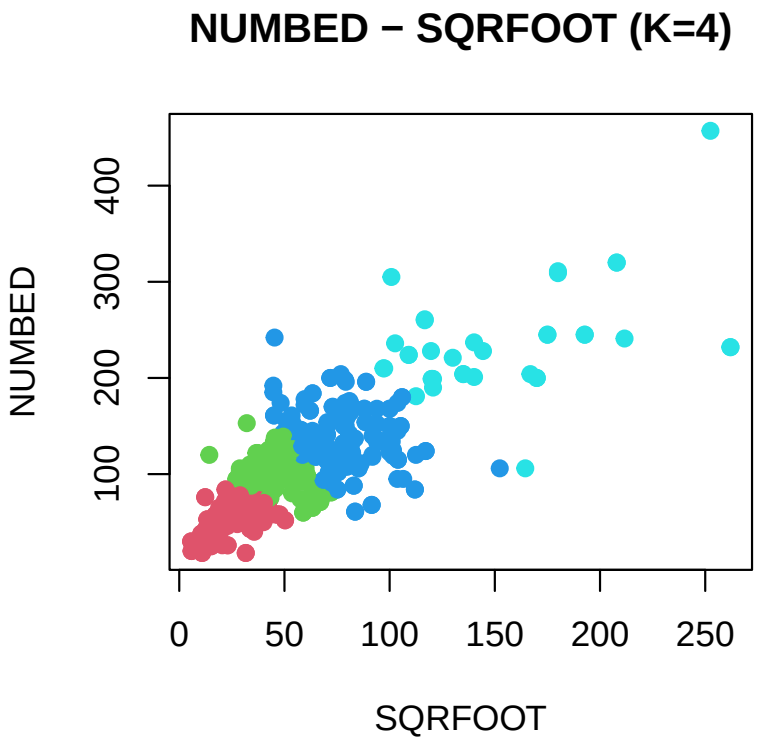




PAM

```
pam.out3 <-pam(WNH[,c(4,3)], 4, metric="euclidean", stand=TRUE)
```





Conclusione

Sitografia

- https://instruction.bus.wisc.edu/jfrees/jfreesbooks/regression%20modeling/bookwebdec2010/CSVData/WiscNursingHome.csv
- https://istruction.bus.wisc.edu/jfrees/jfreesbooks/regression%20modeling/bookwebdec2010/DataDescriptions.pdf

Progetto

- **Prime analisi esplorative** a cura di CANCIAN PIERO
- **Regressione** a cura di MAJER WILLIAM
- **Predict**,**Clustering** a cura di CASTAGNOTTO ALESSANDRO
- **Slides** a cura di MAJER WILLIAM
- **Presentazione** a cura di:
  - CANCIAN PIERO
  - CASTAGNOTTO ALESSANDRO
  - MAJER WILLIAM