APPENDICE

Senatore Carmela Pia

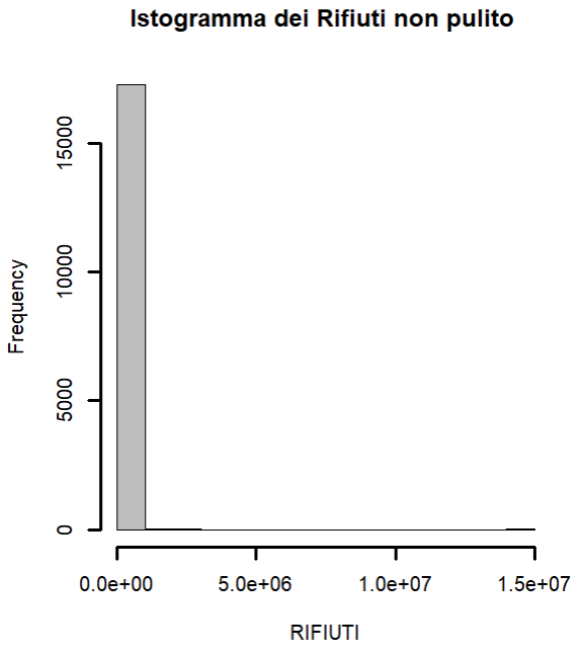
1) GRAFICI E STATISTICHE DURANTE E DOPO LA PULIZIA

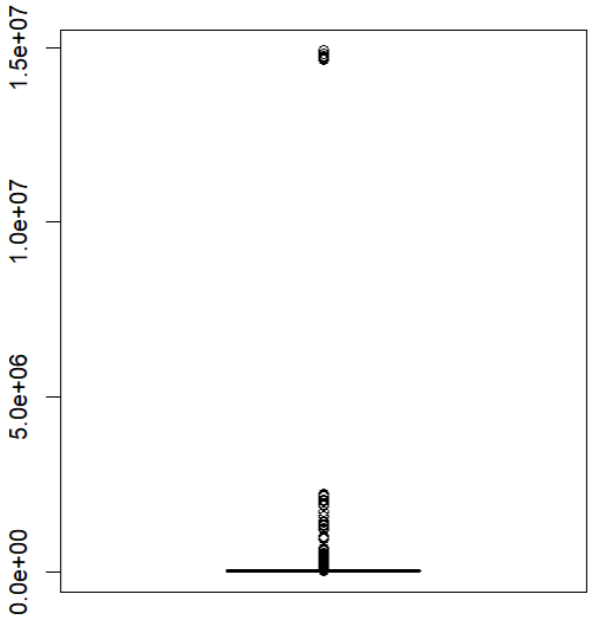
QUINTALI DEI RIFIUTI RACCOLTI

Statiche ottenute dalla summary dei quintali di rifiuti raccolti non pulito.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     30     900    2298   23936    7689 14901756
```

dalle misure riassuntive visualizziamo che media e mediana sono molto diverse tra di loro, e che, cosa più importante, il 3 quartile che raccoglie il 75 per cento delle osservazioni ha una differenza abbissale con il restante 25 per cento dei dati il cui massimo è 1 milione e mezzo.



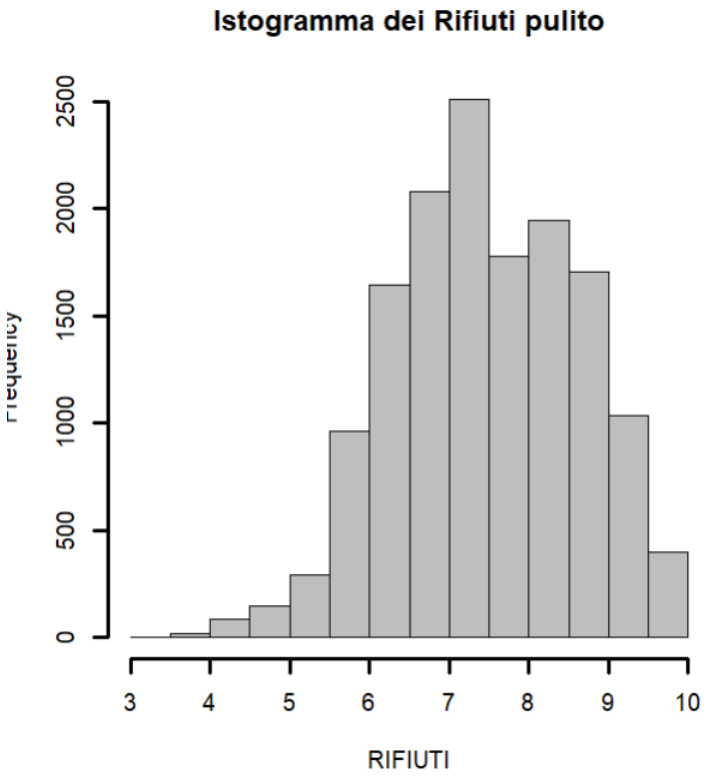


la rappresentazione grafica conferma i nostri sospetti potrebbero esserci outliers che influenzano notevolmente la nostra analisi. Applichiamo il metodo di Tukey Fences per individuare gli outliers.

Dopo i tukey fences:

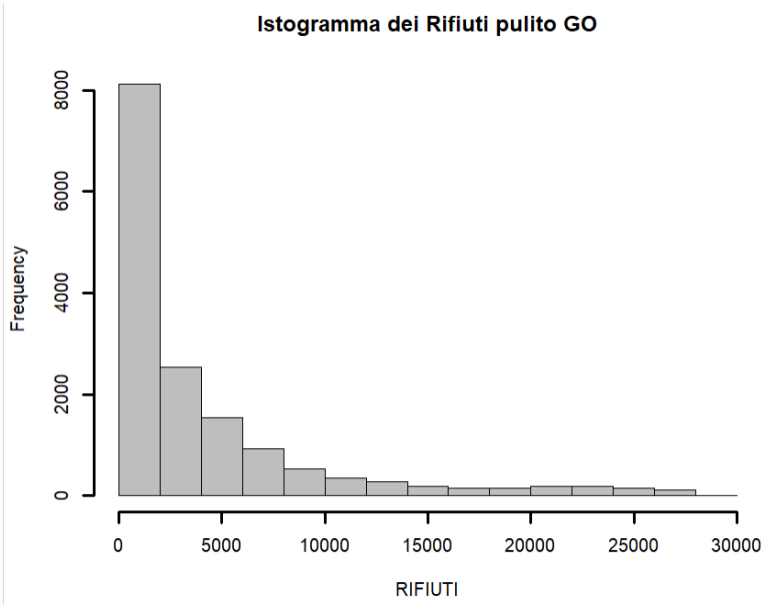
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
30.0   757.8  1650.0  3155.5  4330.9 17870.0
```

Guardiamo alla rappresentazione grafica dei logaritmi dei quintali dei rifiuti raccolti per visualizzare al meglio la distribuzione.



Considerando k=3 e individuando i gross outliers:

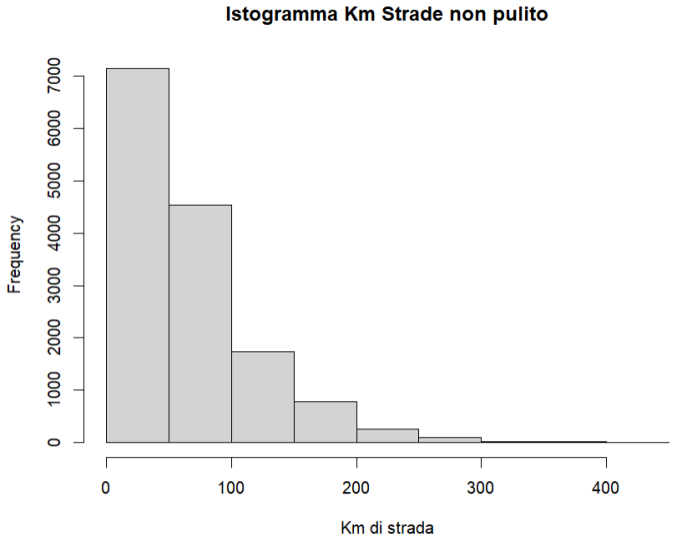
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 30.0   797.9  1797.1  4162.1  5000.0 28056.0
```

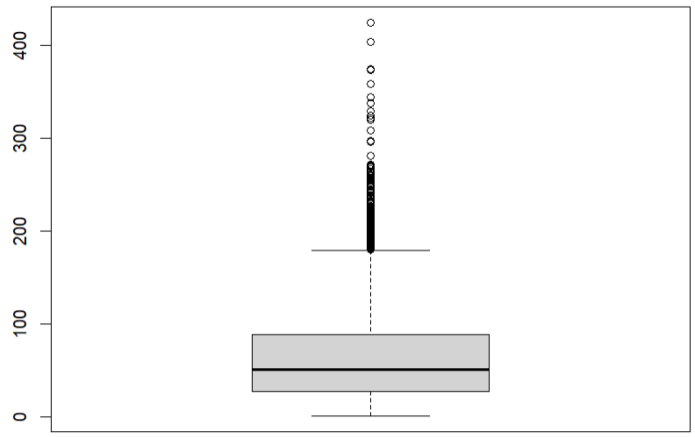


KMDISTRADE

Statiche ottenute dalla summary dei kilometri di strada non pulito.

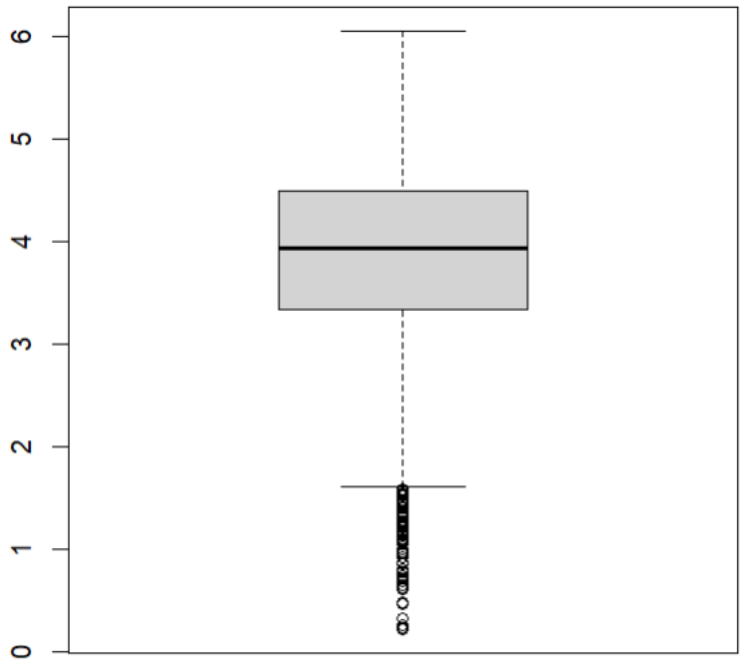
```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.244  27.977  51.017  65.943  89.000 424.380
```





Sembra essere caratterizzato fortemente da valori anomali.

Visualizziamo anche il boxplot dei logaritmi dei kilometri di strade, anche con il logaritmo il problema persiste:

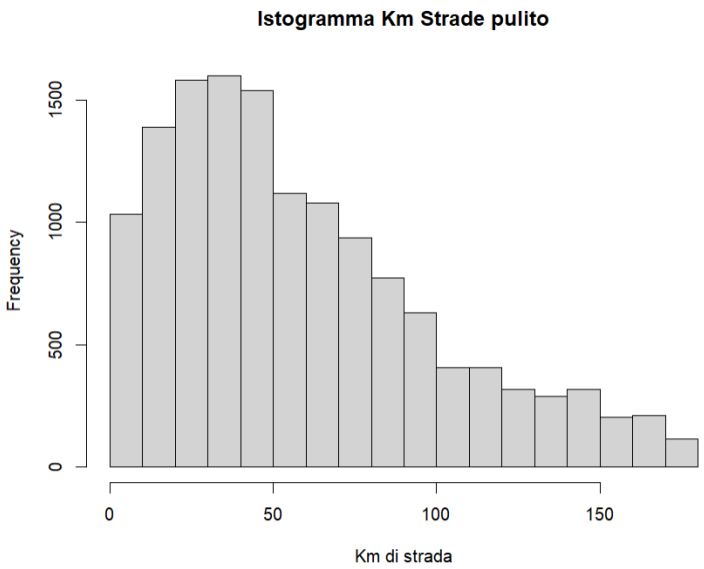


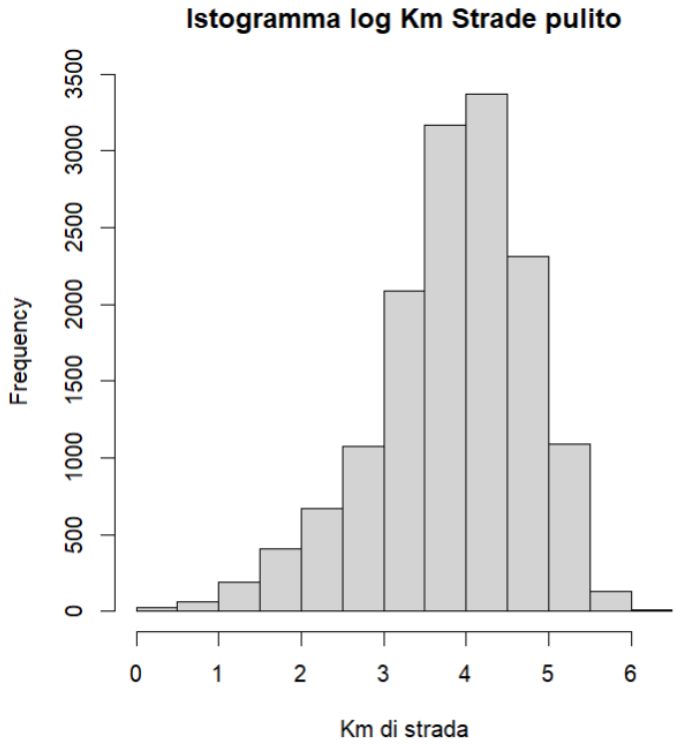
Dopo l'applicazione del metodo dei Tukey Fences:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.244  26.674  49.063  58.848  82.006 179.638
```

Summary del logaritmo:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.2183  3.2837  3.8931  3.7677  4.4068  5.1909
```



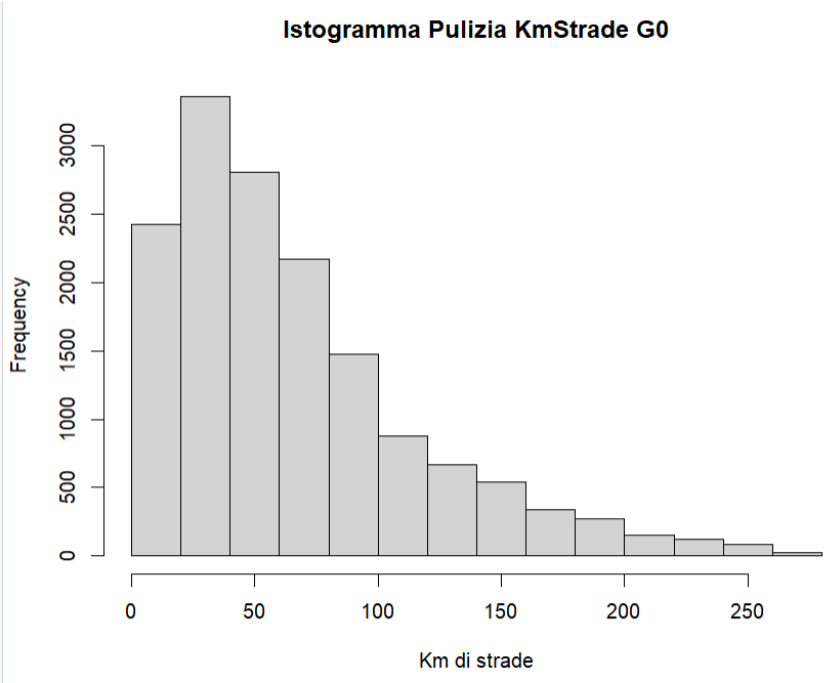


Considerando l'analisi con k=3 e individuano i gross-outliers il summary risulta essere:

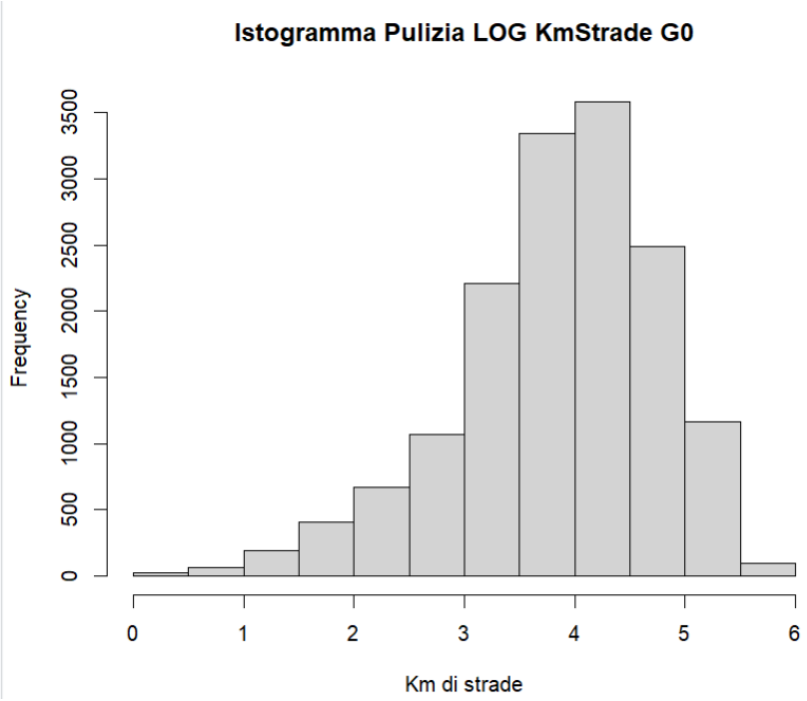
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.244  28.864  51.879  65.810  89.064 271.623
```

Verificando i suoi logaritmi:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.2183  3.3626  3.9489  3.8505  4.4894  5.6044
```



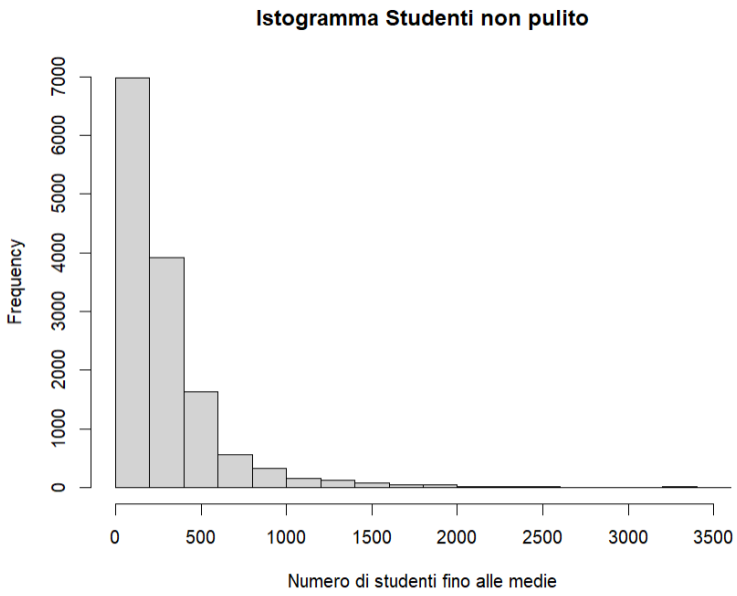
Visualizziamo l'istogramma dei logaritmi dei kilometri di strade: risulta essere piuttosto ripulito e soprattutto la distribuzione è unimodale.

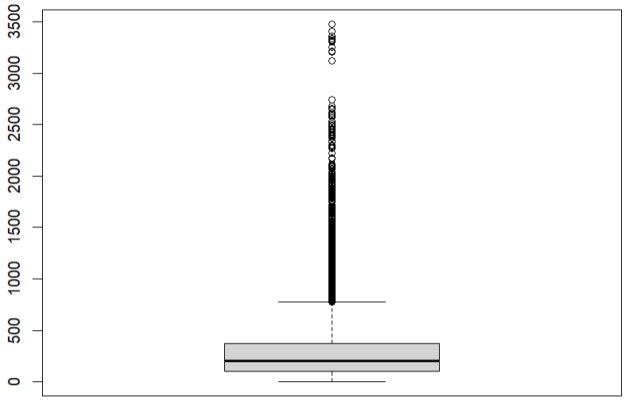


STUDENTI

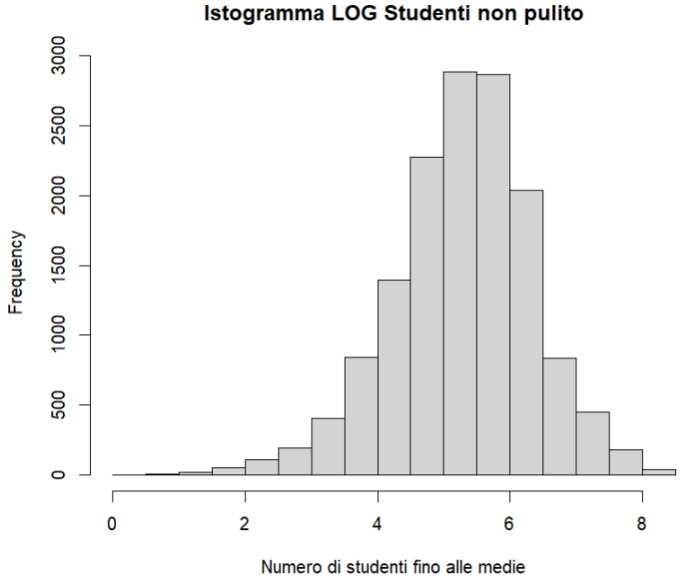
Numero di studenti. Analisi del summary pre operazione di pulizia:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.0   101.0   200.0   295.2   370.0  3477.0
```





Istogramma del numero di log di studenti non ripulito: la distribuzione è unimodale, tuttavia, presenta code di sinistra più lunghe.

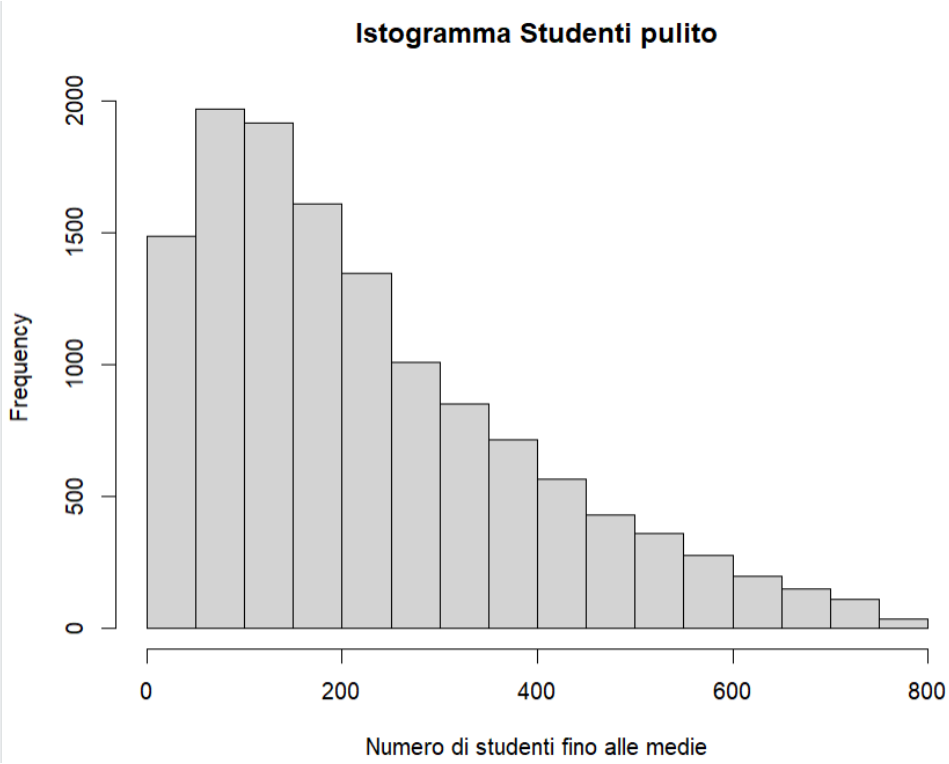


Summary della variabile post pulizia con i Tukey Fences:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.0    96.0   185.0   227.4   325.0   773.0
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   4.564   5.220   5.088   5.784   6.650
```



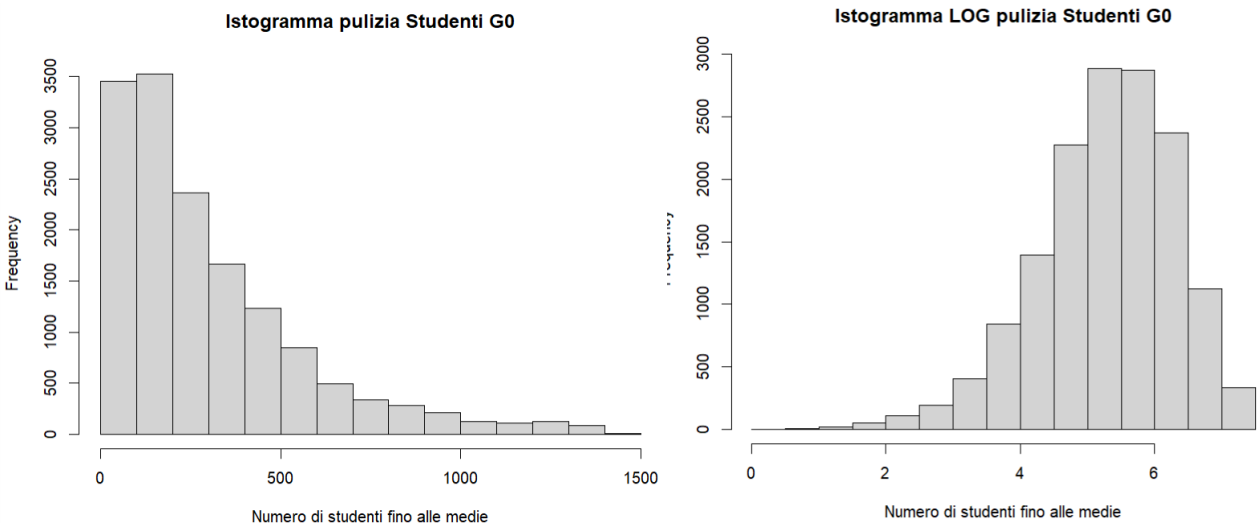
Considerando k=3 e individuando i gross-outliers, successivamente il summary della features è di questo tipo:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.0   107.0   216.0   299.9   411.0  1409.0
```

Verifichiamo i logaritmi:

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   4.673   5.375   5.281   6.019   7.251
```

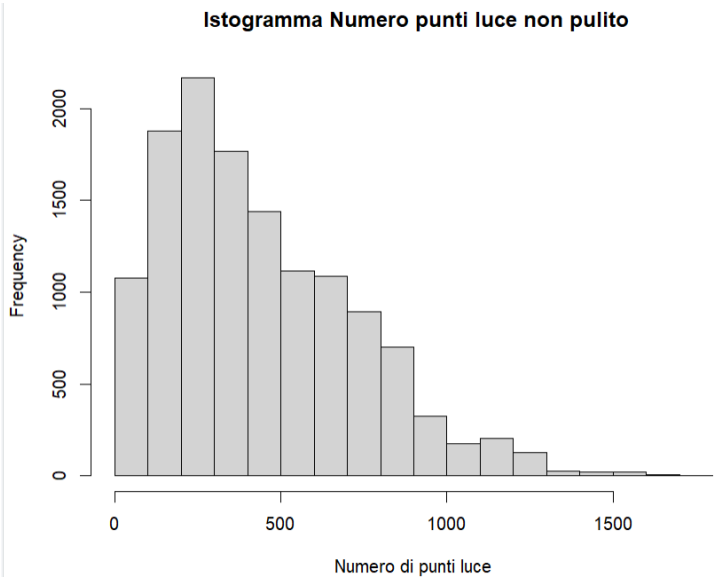
Visualizziamo i grafici post pulizia per la distribuzione di numero di studenti e del suo logaritmo:

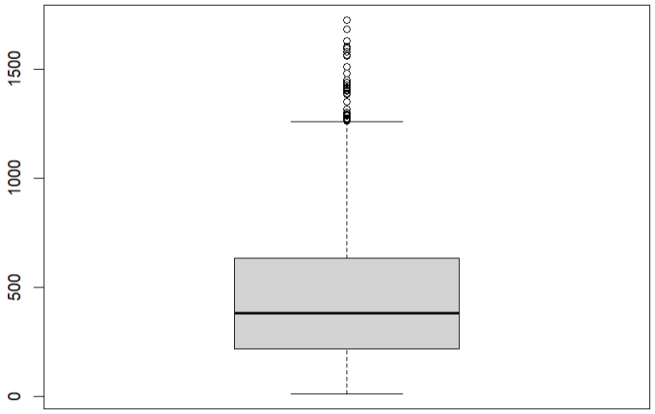


NR PUNTI LUCE

Statiche ottenute dalla summary numeri dei punti luce non pulito:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
11.0   216.0   380.0   442.8   634.0  1724.0
```





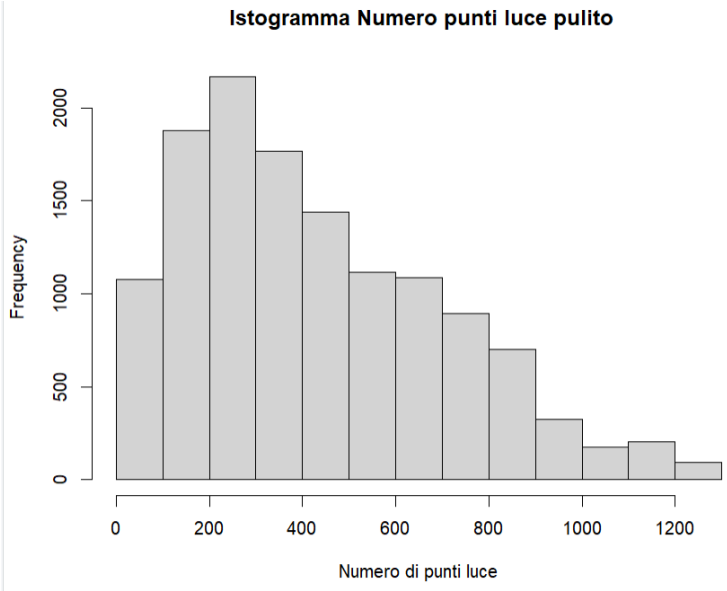
La coda di destra non sembra essere significativamente pesante, tuttavia ripuliamolo lo stesso. La differenza tra il 3 quartile che contiene il 75 per cento dei dati e il valore massimo è allarmante.

Summary della variabile dopo il metofo di Tukey fences:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11     215     375     435     628    1260
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.398   5.371   5.927   5.820   6.443   7.139
```



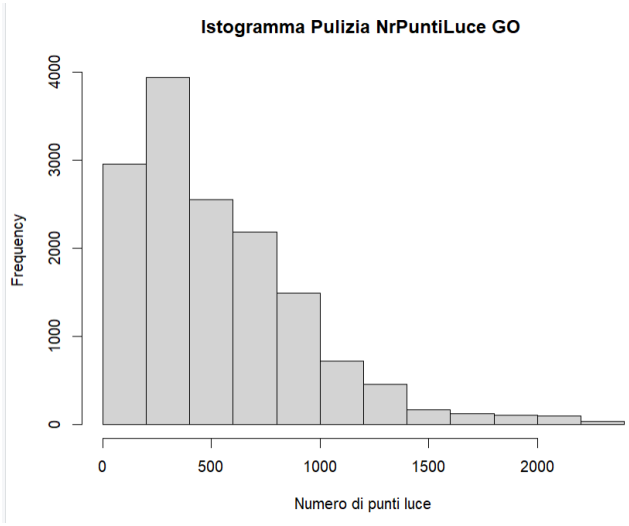
Considerando k=3 e individuando i gross-outliers, dopo l'applicazione del metodo il summary risuta:

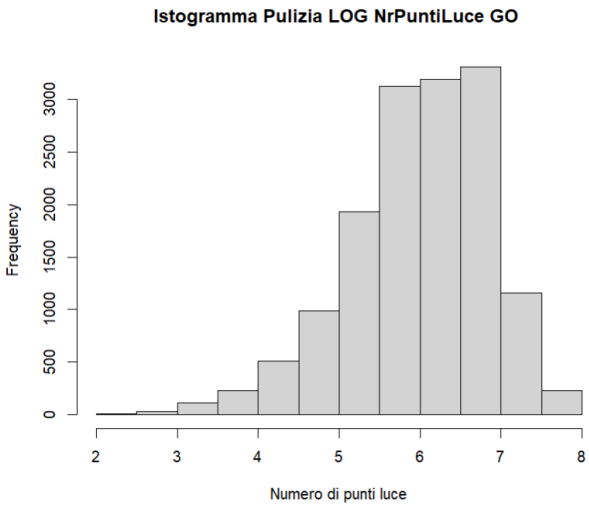
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 11.0   240.0   449.0   538.8   754.0  2304.0
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.398   5.481   6.107   5.981   6.625   7.742
```

Grafici post pulizia per la variabile di numero di punti luce e i suoi logaritmi:

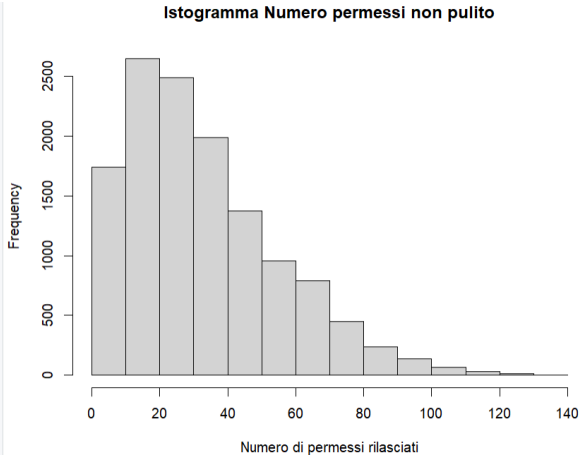


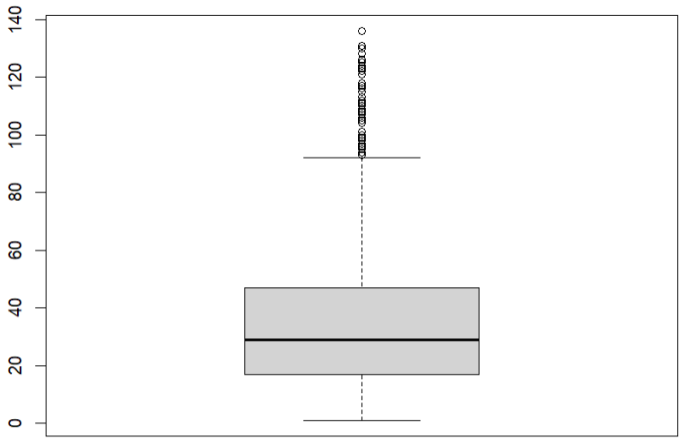


NR PERMESSI

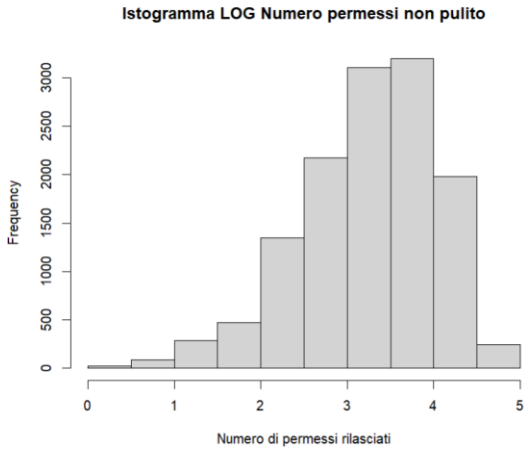
Statiche ottenute dalla summary numeri di permessi non pulito:

```
summary(dat_nuovo$NrPermessi)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   17.00   29.00   33.83   47.00  136.00
```





Non sembra avere una coda eccessivamente lunga a destra ma i valori esterni molto elevati potrebbero essere outliers. Visualizziamo i valori del logartimi della variabile:

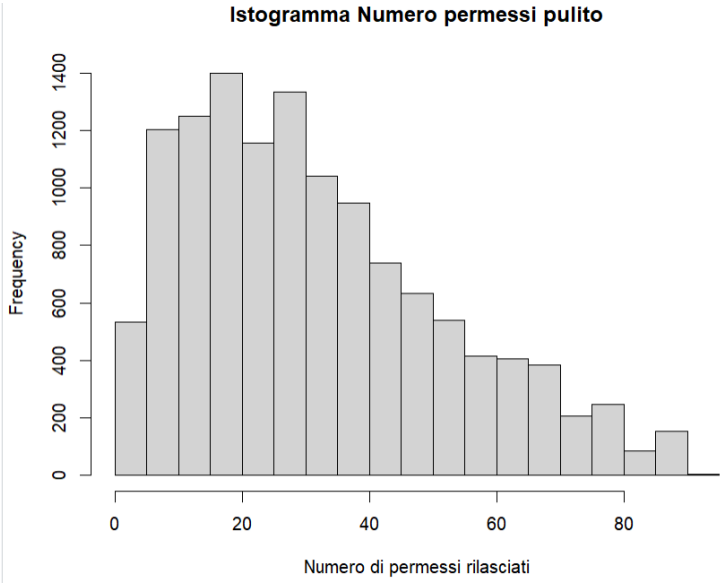


Summary della variabile dopo il metodo di Tukey fences:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00   17.00   29.00   32.49   45.00   91.00
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   2.833   3.367   3.240   3.807   4.511
```



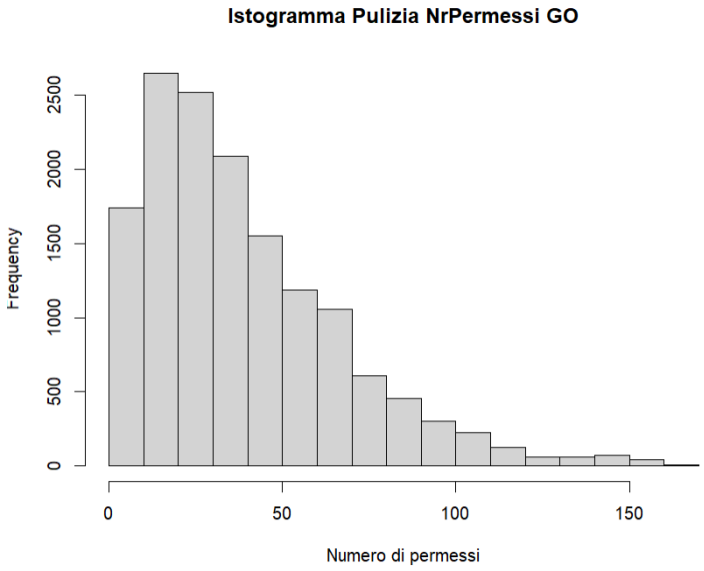
Considerando k=3 e individuando i gross-outliers, dopo l'applicazione del metodo il summery risuta:

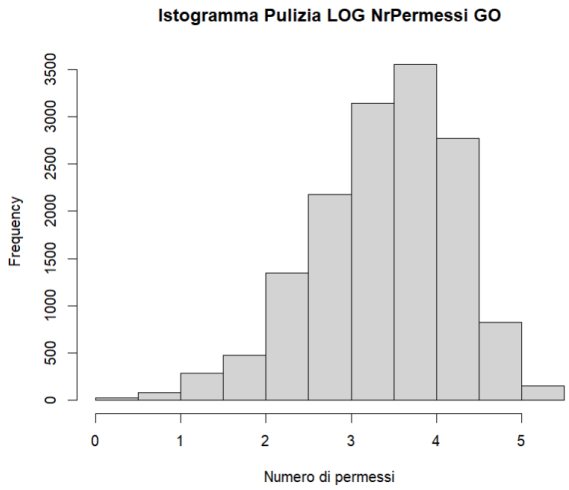
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   19.00   33.00   39.79   54.00  162.00
```

Verifichiamo i logaritmi:

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000   2.944   3.497   3.398   3.989   5.088
```

Grafici post pulizia per la variabile di numero di permessi e i suoi logaritmi:

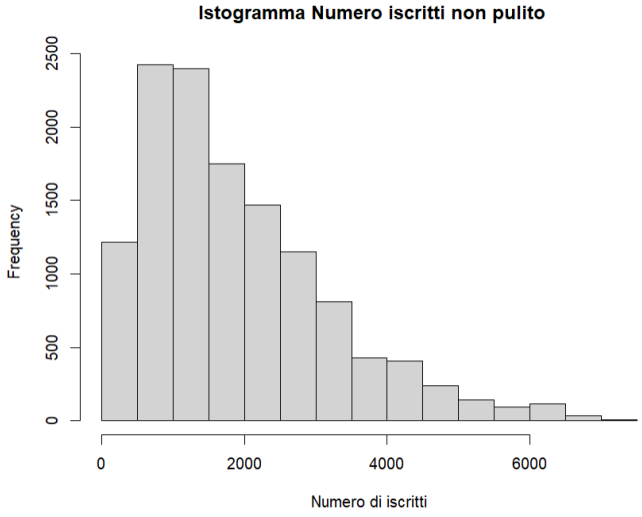


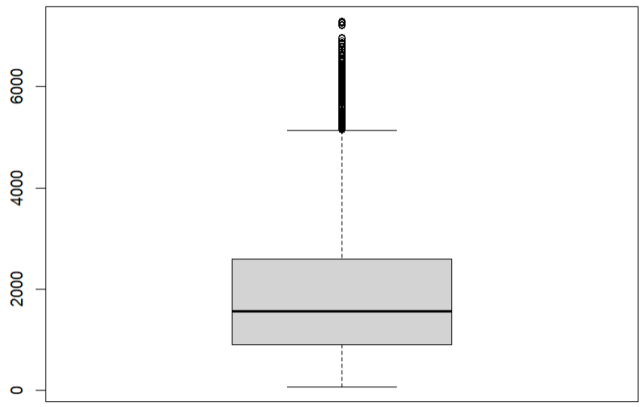


NR ISCRITTI

Statiche ottenute dalla summary numeri di iscritti alle liste elettorali non pulito:

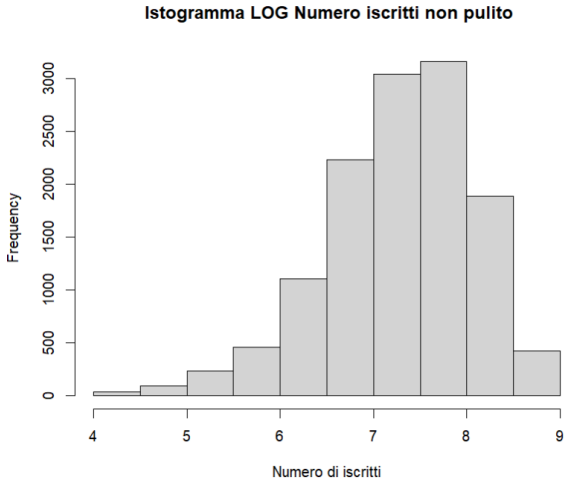
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
67.16  902.00 1565.64 1893.68 2599.39 7291.00
```





sembra avere una coda eccessivamente lunga a destra, per questo motivo ripeto TF.

Controlliamo i valori del logaritmo degli iscritti non pulito: coda di destra più pesante di quella di sinistra.

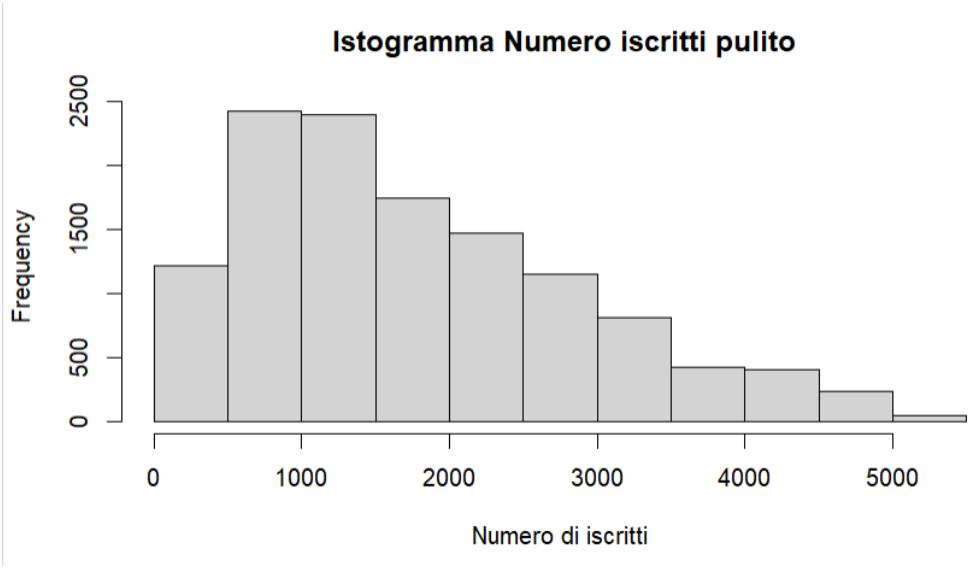


Summary della variabile dopo il metodo di Tukey fences:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
67.16  890.17 1528.95 1781.73 2498.63 5140.64
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
4.207   6.791   7.332   7.241   7.823   8.545
```



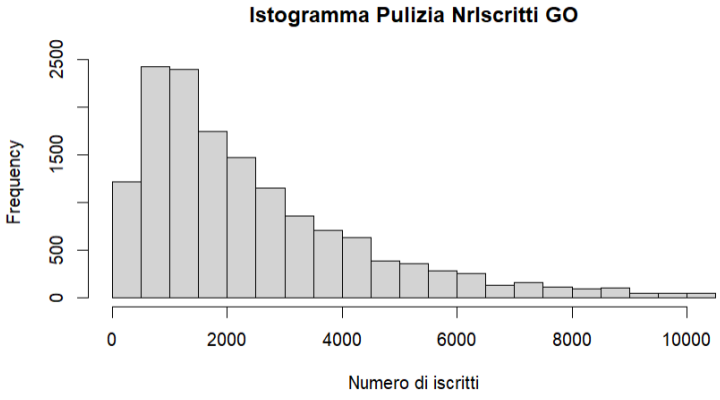
Considerando k=3 e individuando i gross-outliers, dopo l'applicazione del metodo il summery risuta:

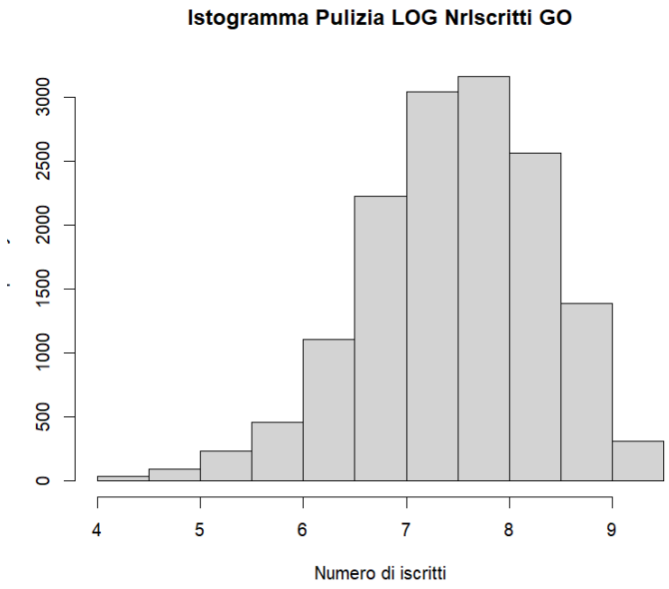
```
  Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 67.16   1003.39   1845.00   2436.51   3325.19 10482.00
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
4.207   6.911   7.520   7.463   8.109   9.257
```

Visualizziamo graficamente sia la distribuzione della variabile post pulizia che dei suoi logaritmi sembra essere distribuita in miglior modo:

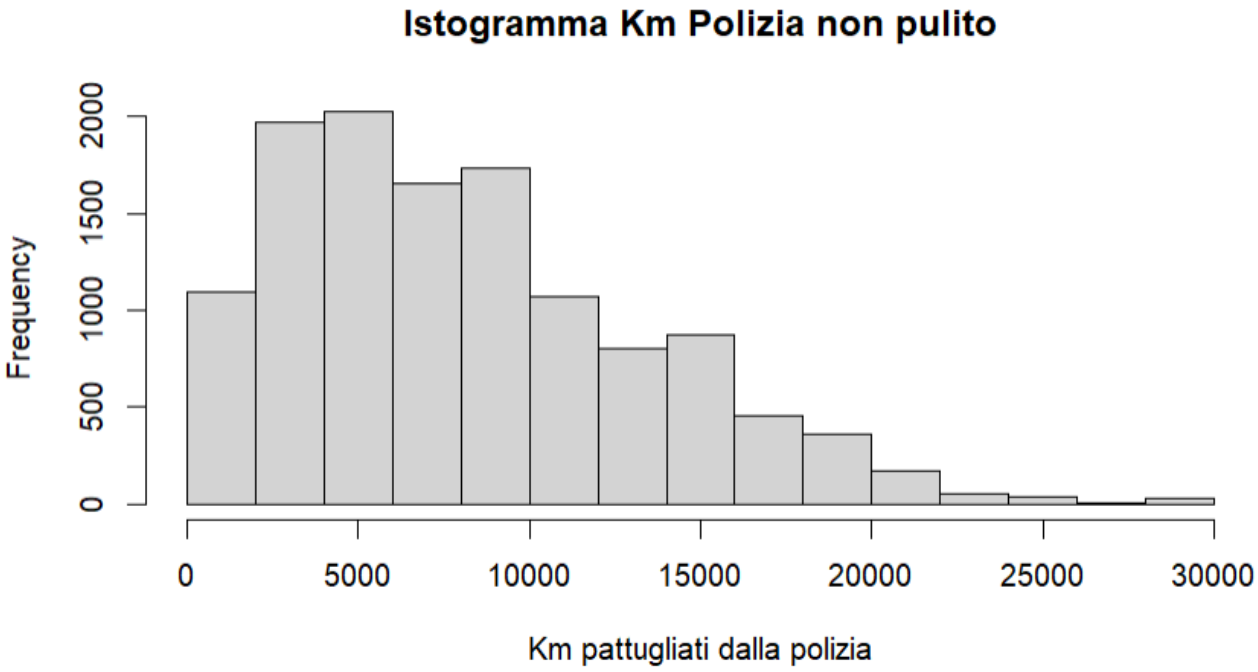


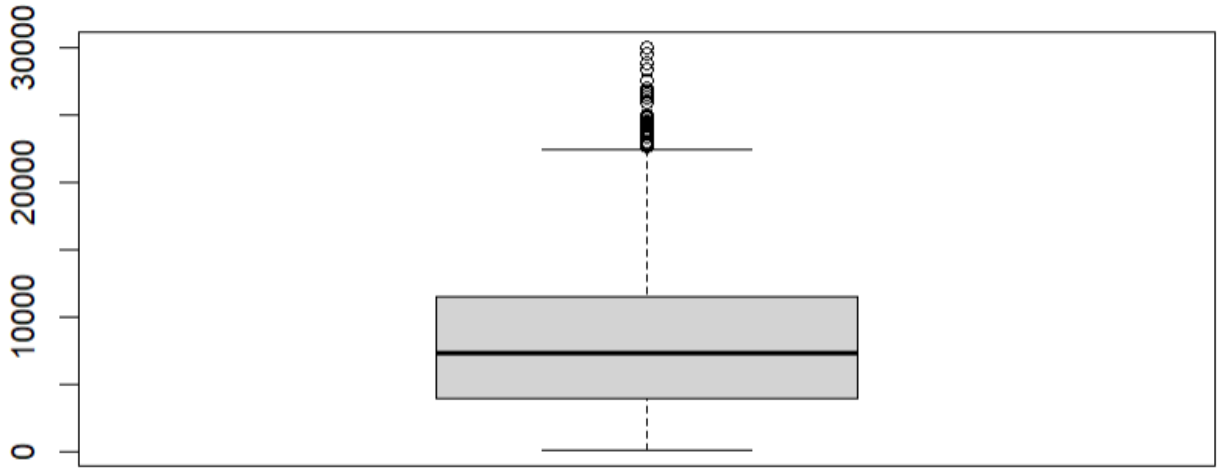


KM POLIZIA

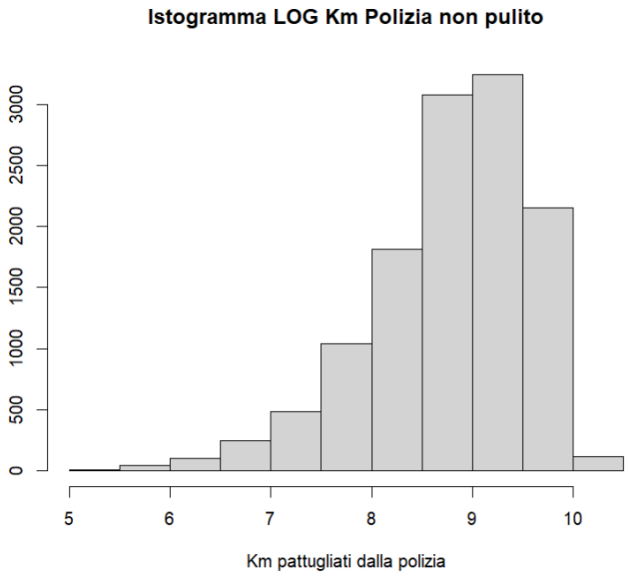
Statiche ottenute dalla summary km di pattugliamento effettuati dalla polizia municipalenon pulito:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  179    4037    7412    8297   11500   30000
```





Rappresentazione dei logaritmi della variabile non pulita:

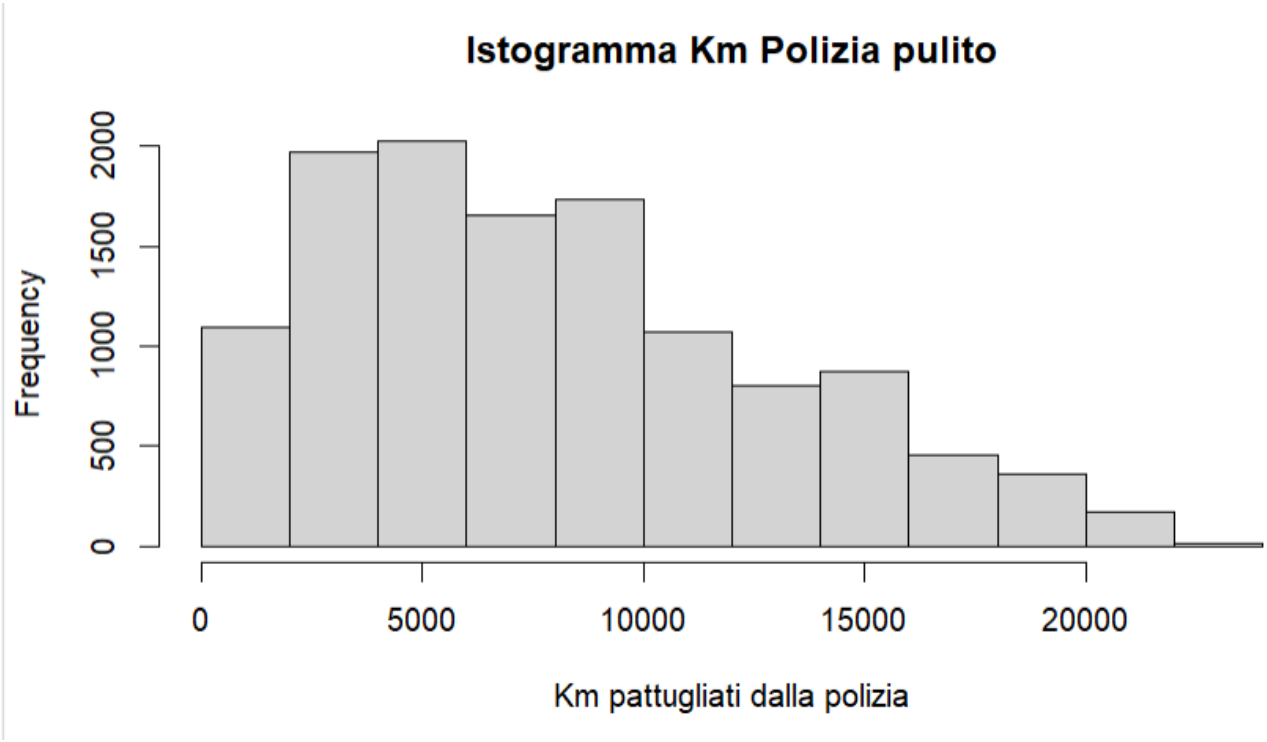


Summary della variabile dopo il metodo di Tukey fences:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 179    4000    7303    8145   11294   22500
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
5.187   8.294   8.896   8.759   9.332  10.021
```



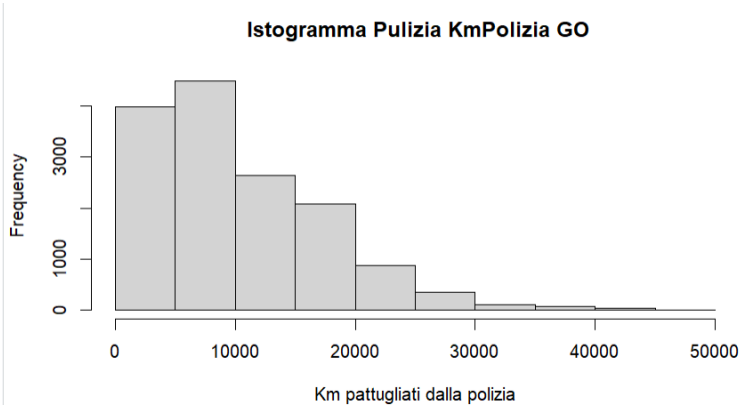
Considerando k=3 e individuando i gross-outliers, dopo l'applicazione del metodo il summary risuta:

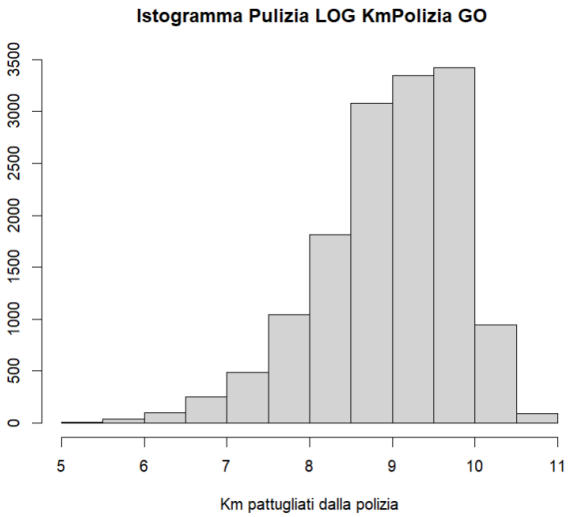
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  179    4842    8837   10390   15000   45400
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
5.187   8.485   9.087   8.955   9.616  10.723
```

Visualizziamo la distribuzione della variabile post pulizia dei gross outliers e i suoi logaritmi:

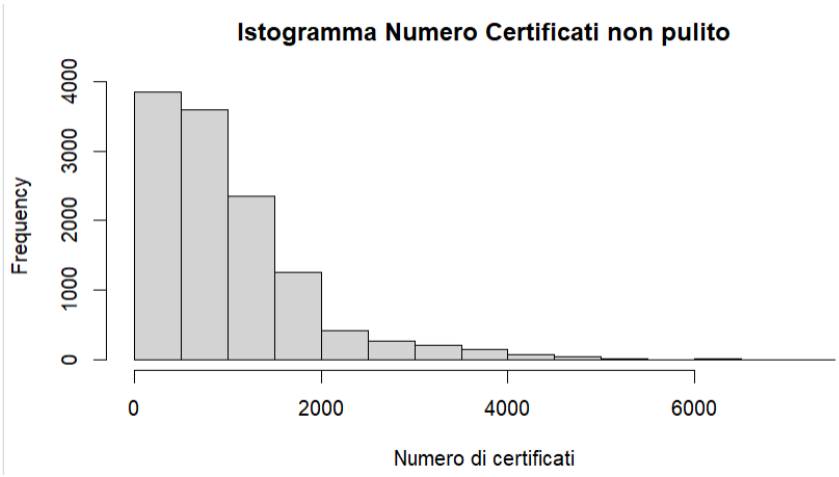


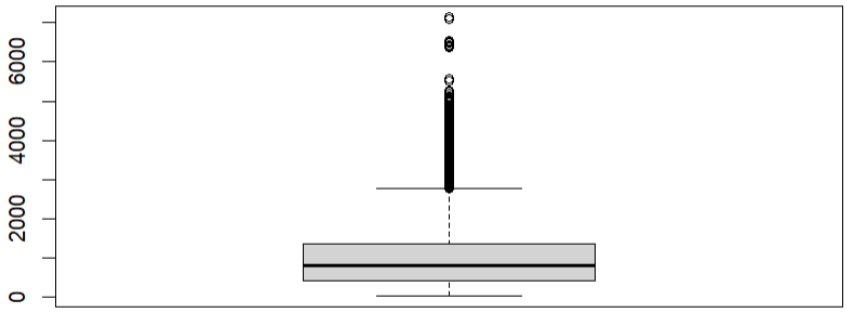


NR CERTIFICATI

Statiche ottenute dalla summary numero di certificati anagrafici rilasciati non pulito:

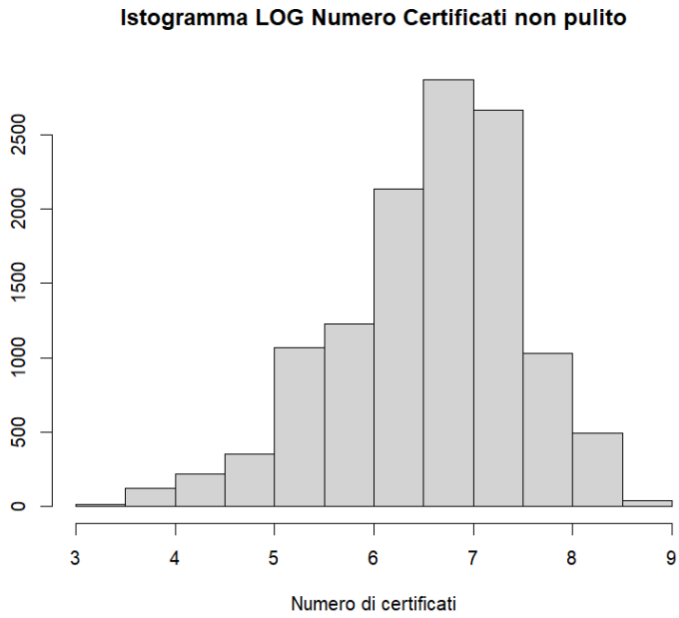
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25     413     800    1008    1353    7146
```





Coda significativamente pesante.

Visualizziamo la distribuzione dei suoi logaritmi: la coda di sinistra è più pesante di quelal di destra

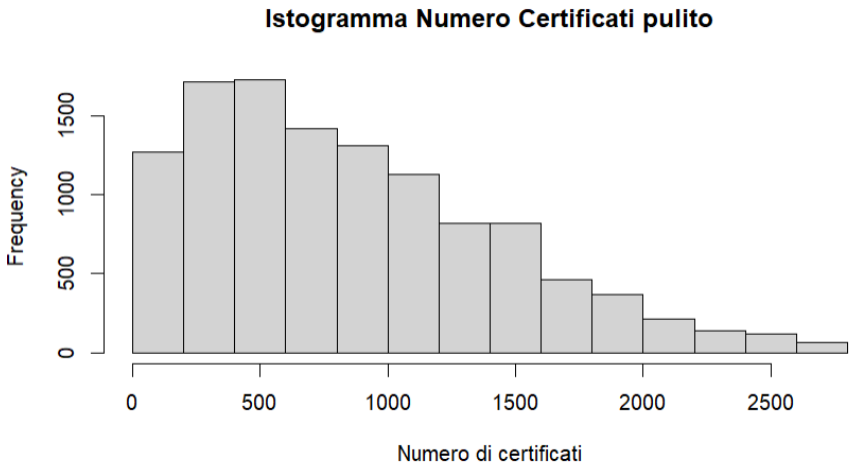


Summary della variabile dopo il metodo di Tukey fences:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
25.0   385.8   753.5   860.6  1232.2  2761.0
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3.219   5.955   6.625   6.460   7.117   7.923
```

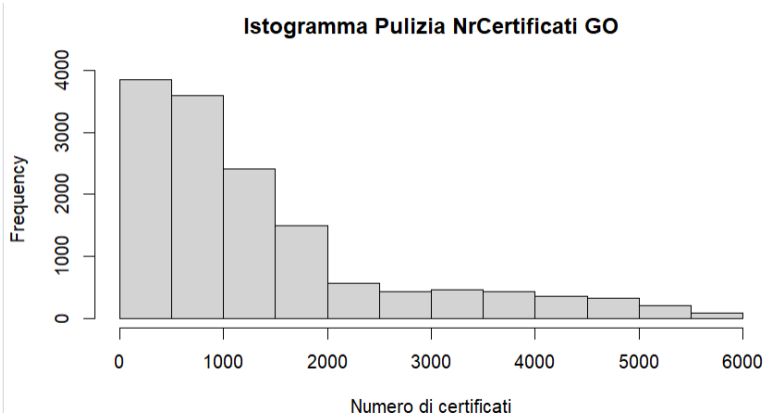


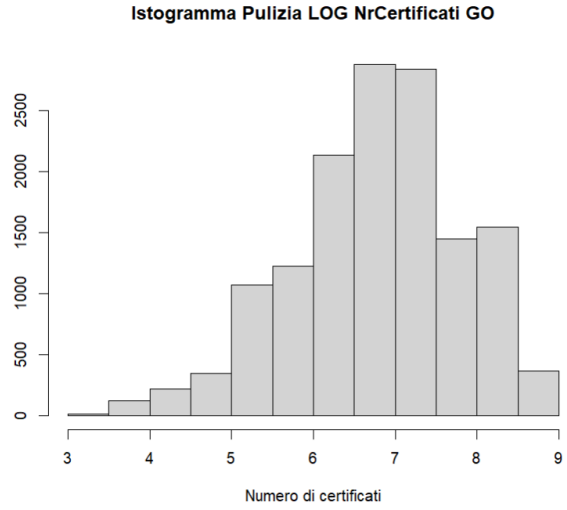
Considerando k=3 e individuando i gross-outliers, dopo l'applicazione del metodo il summary risuta:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   25     470     956    1370    1715    5955
```

Verifichiamo i logaritmi:

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.219   6.153   6.863   6.771   7.447   8.692
```

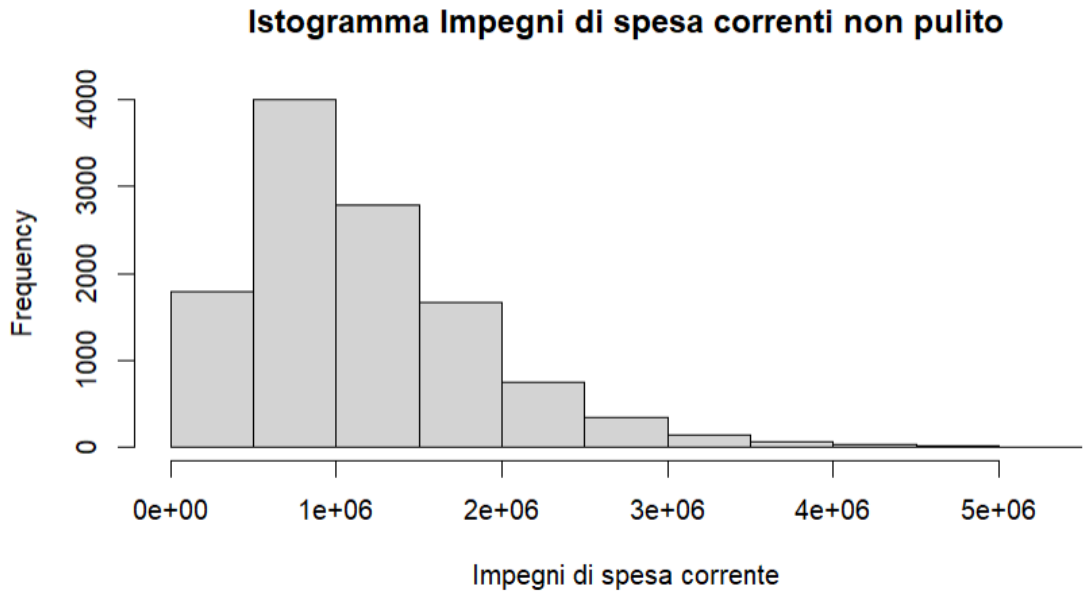


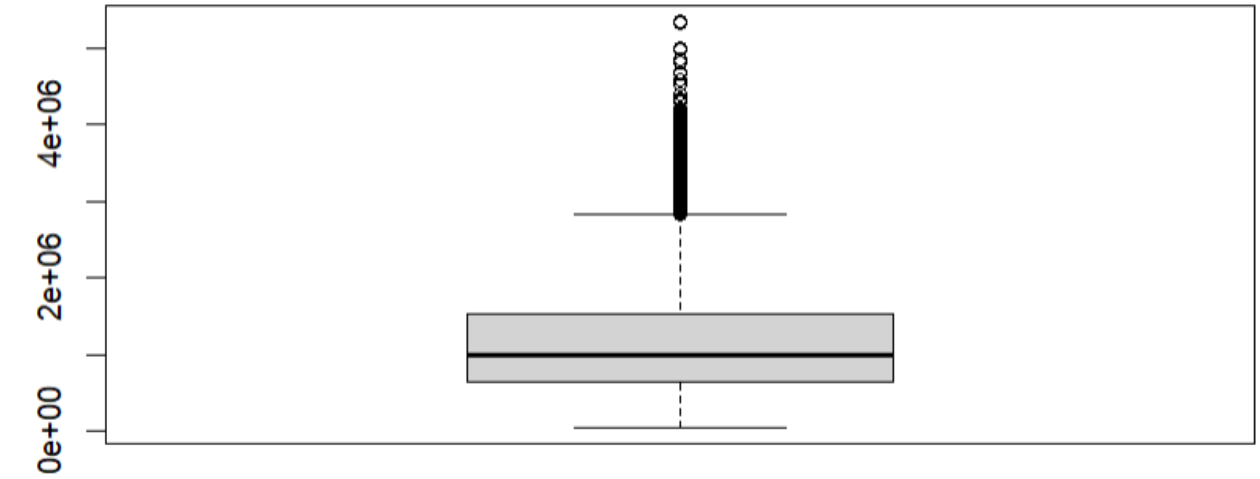


IMP TOT

Statiche ottenute dalla summary di impegni di spesa corrente totale non pulito:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
52796  647182  999169 1145797 1523127 5341500
```



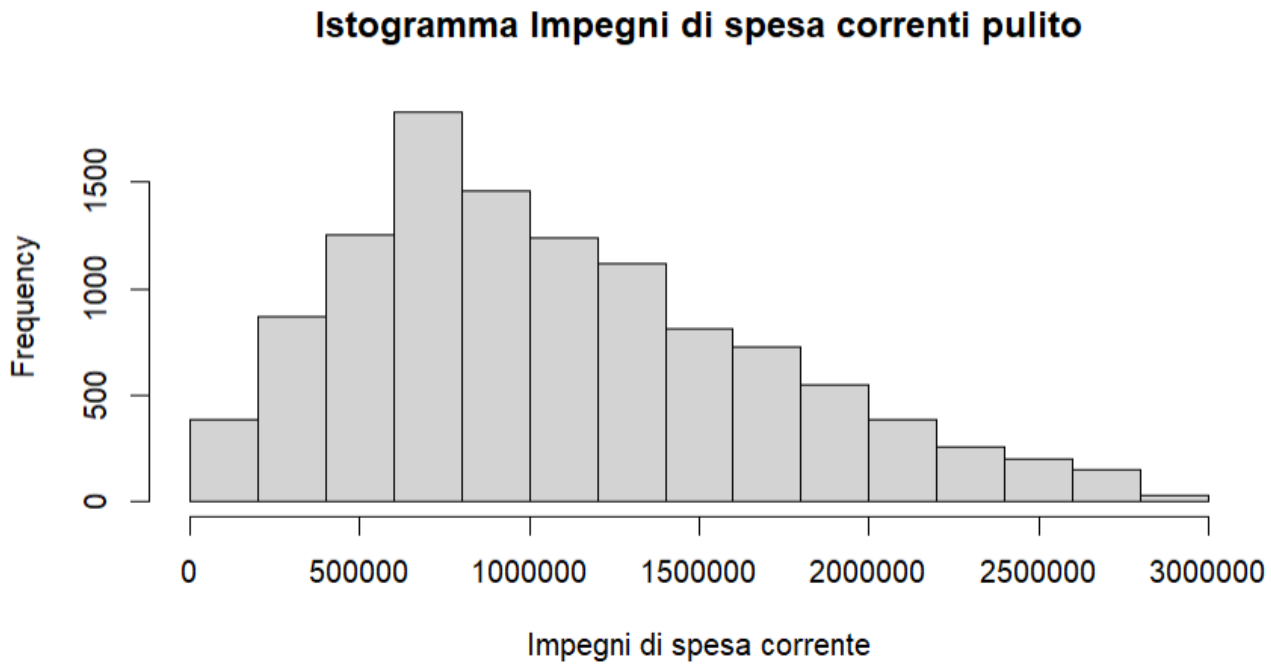


Summary della variabile dopo il metodo di Tukey fences:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
52796  638183  975439 1085208 1463778 2836491
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
10.87   13.37   13.79   13.71   14.20   14.86
```

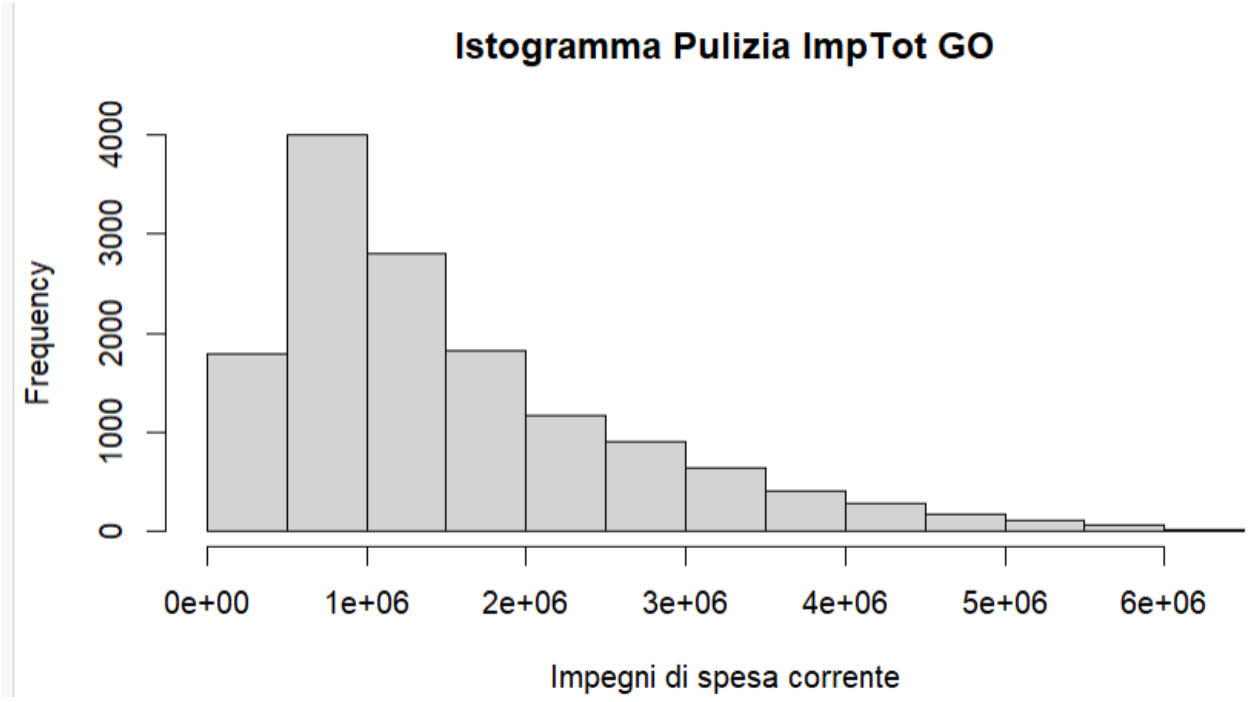


Considerando k=3 e individuando i gross-outliers, dopo l'applicazione del metodo il summary risuta:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
52796  716200 1205862 1535239 2058731 6151313
```

Verifichiamo i logaritmi:

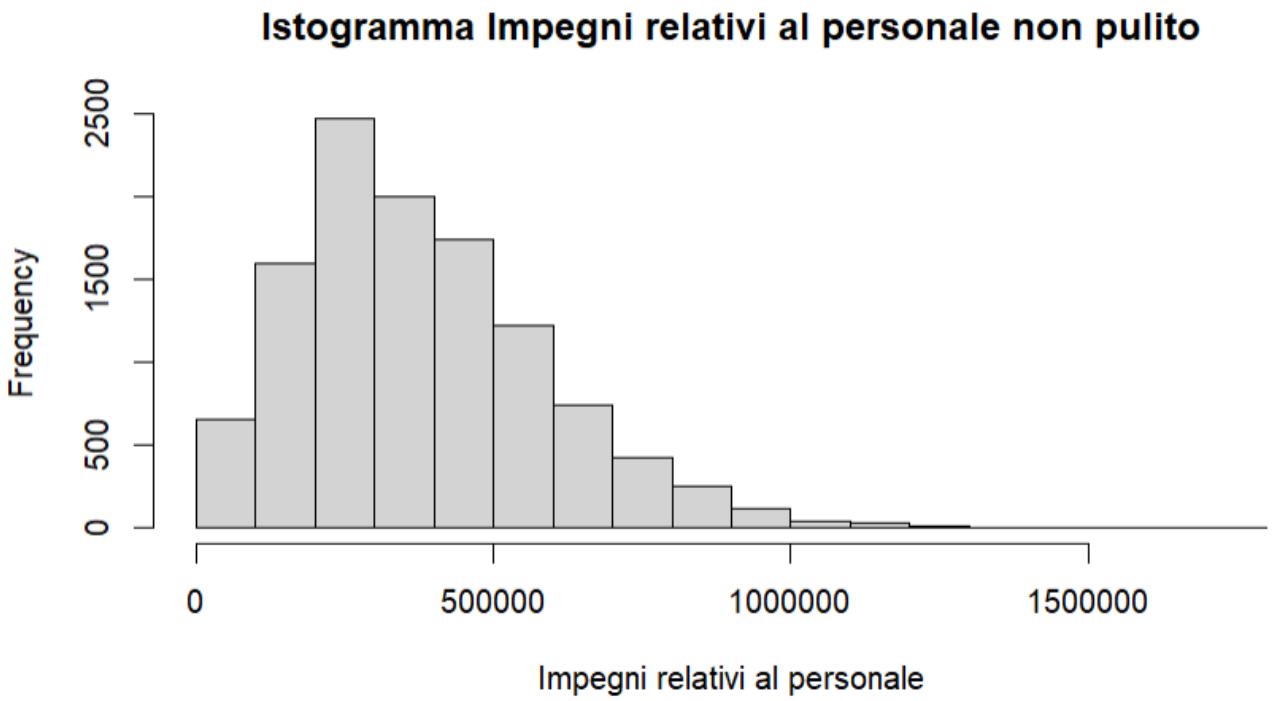
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
10.87   13.48   14.00   13.97   14.54   15.63
```



IMP PERS

Statiche ottenute dalla summary di impegni di spesa per personale non pulito:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
17166  224945  345845  377684  500571 1779358
```

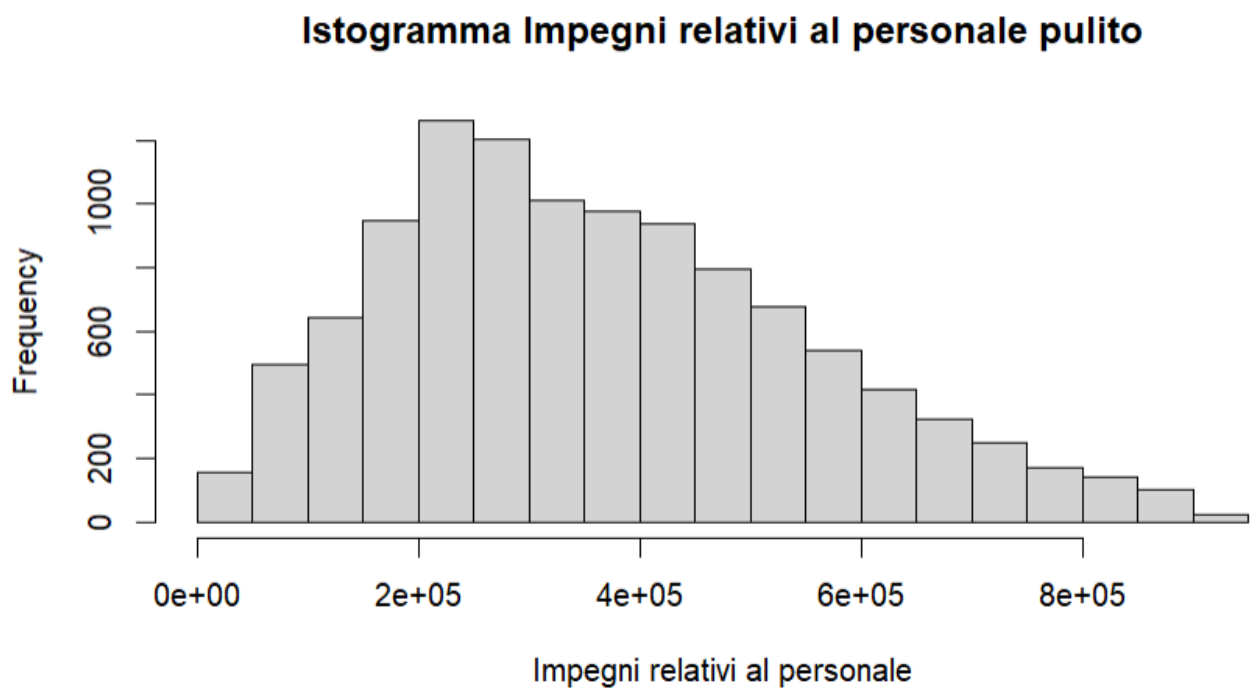


Summary della variabile dopo il metodo di Tukey fences:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
17166  223707  340981  366828  492317  913793
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
9.751  12.318  12.740  12.647  13.107  13.725
```

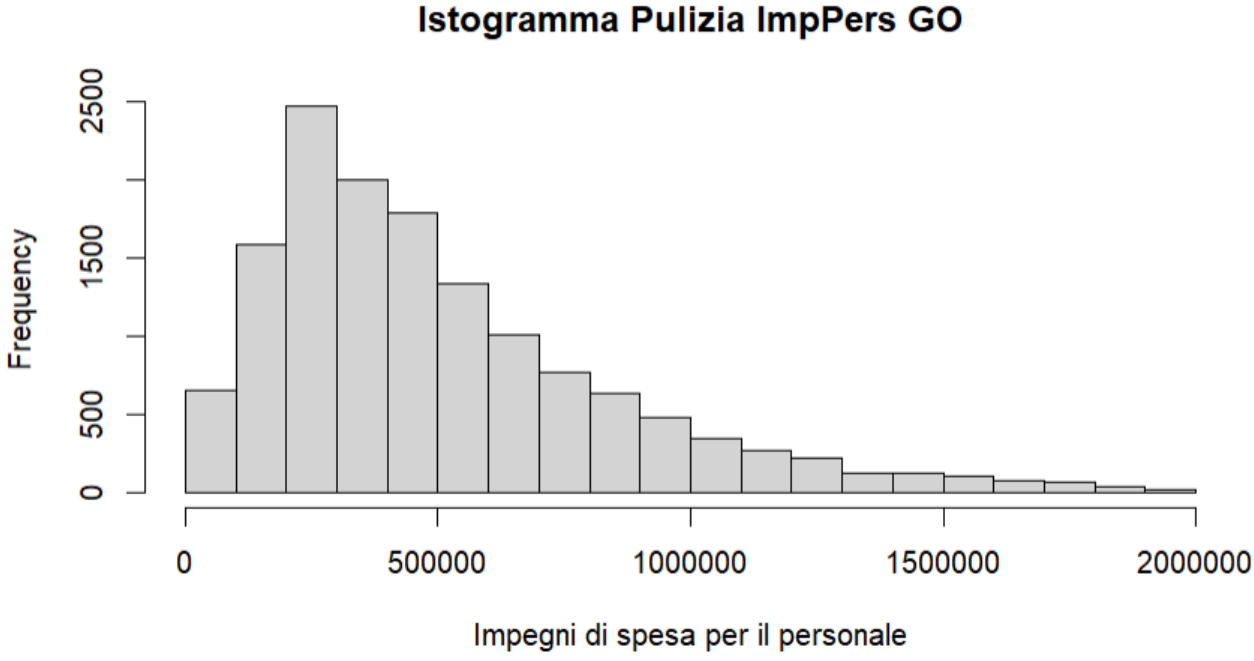


Considerando k=3 e individuando i gross-outliers, dopo l'applicazione del metodo il summary risuta:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
17166  251115  418646  507437  672714 1949183
```

Verifichiamo i logaritmi:

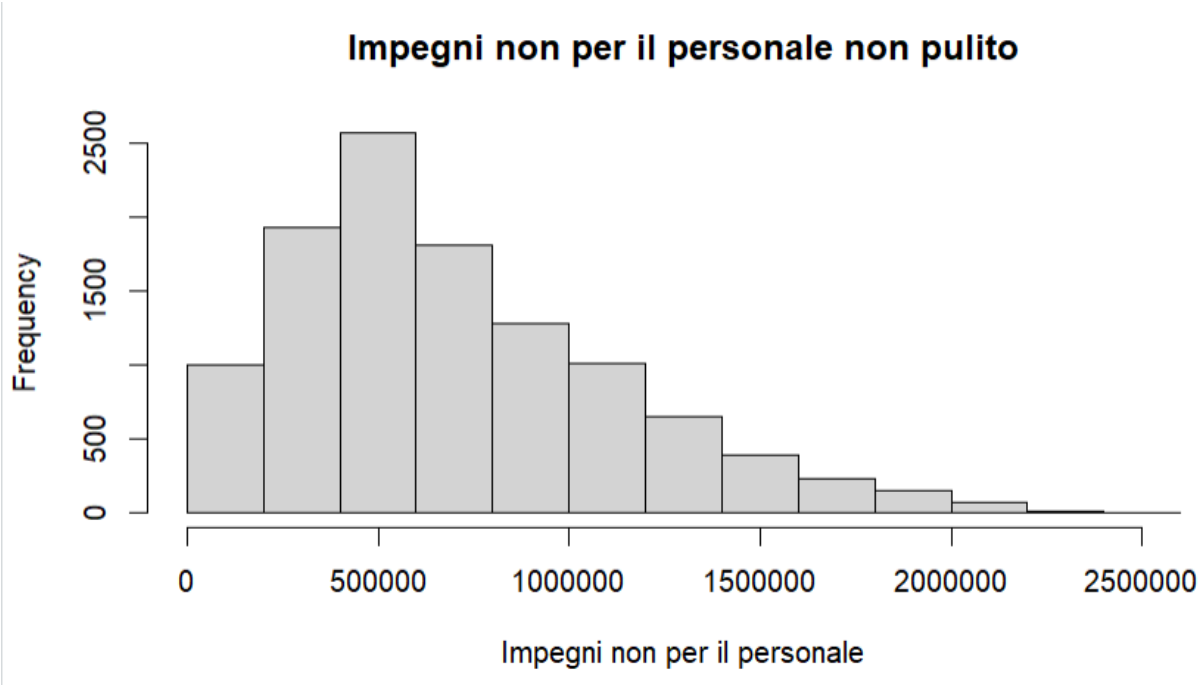
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
9.751  12.434  12.945  12.892  13.419  14.483
```



IMP NO PERS

Statiche ottenute dalla summary di impegni di spesa non per personale non pulito:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
7680  385868  603573  698867  955086 2433598
```

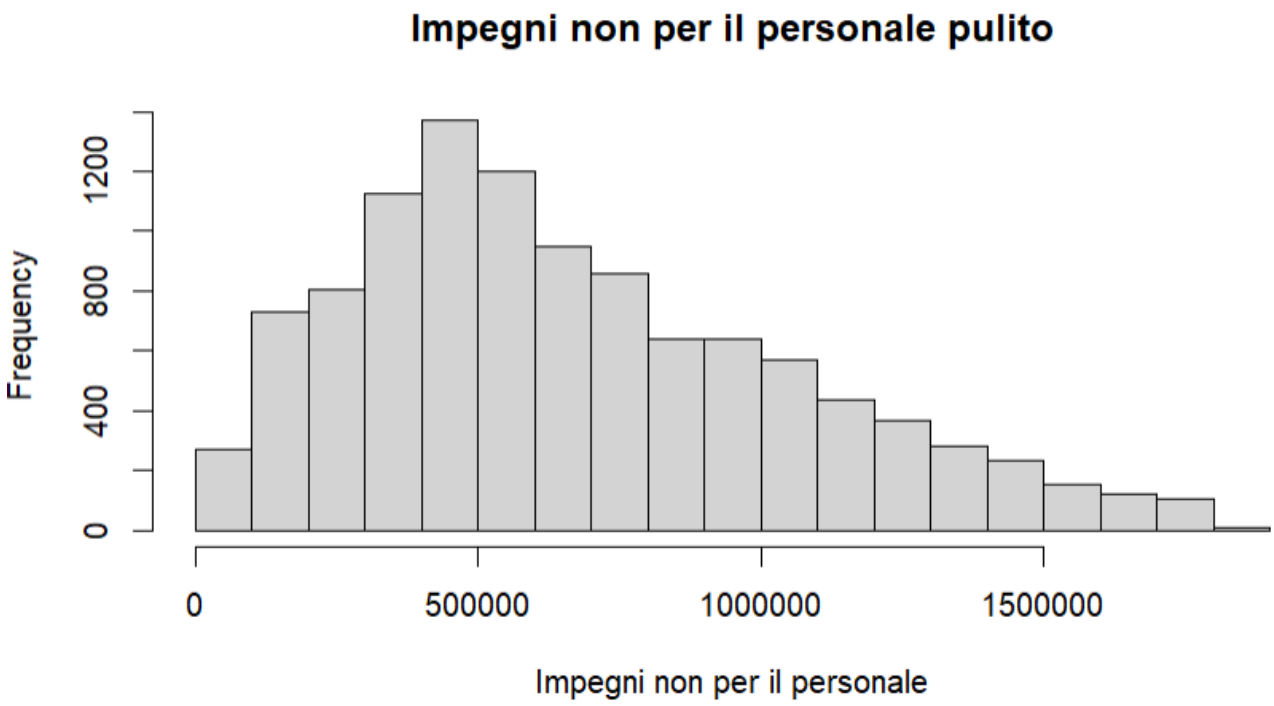


Summary della variabile dopo il metodo di Tukey fences:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
7680  381878  592691  674044  930949 1808396
```

Verifichiamo i logaritmi:

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
8.946  12.853  13.292  13.217  13.744  14.408
```

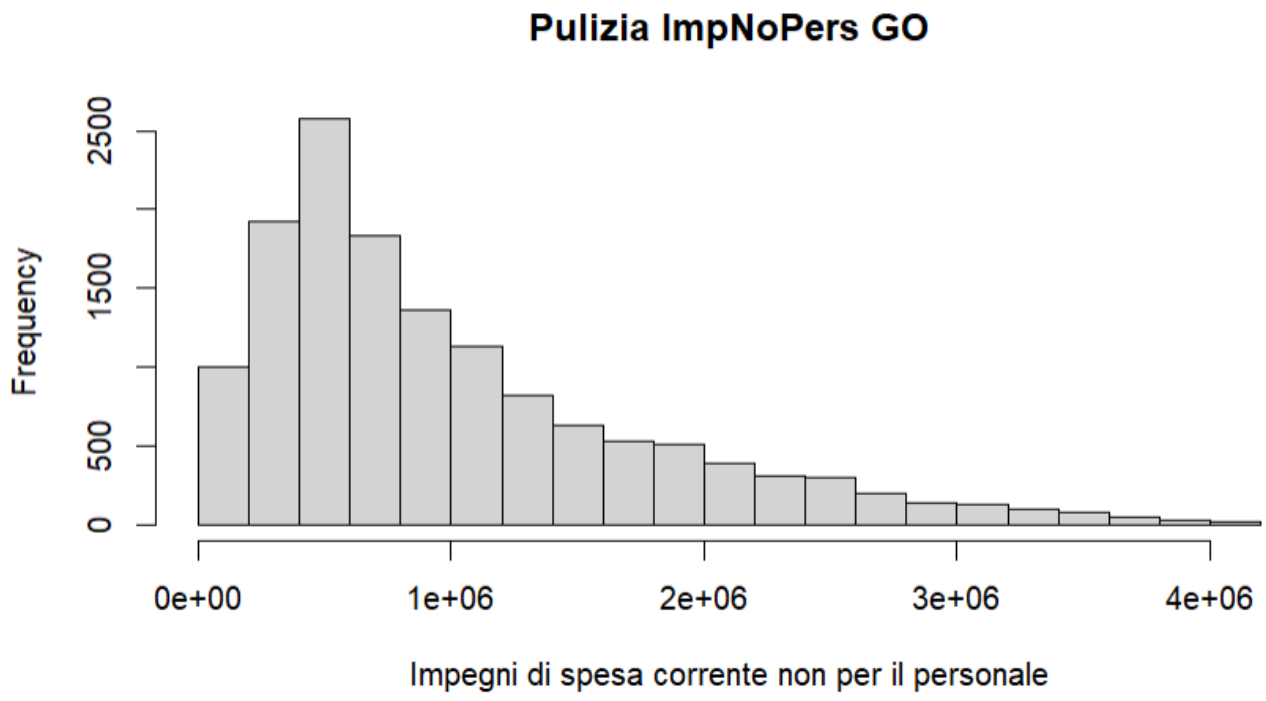


Considerando k=3 e individuando i gross-outliers, dopo l'applicazione del metodo il summary risuta:

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7680  445315  765196 1009663 1372633 4167561
```

Verifichiamo i logaritmi:

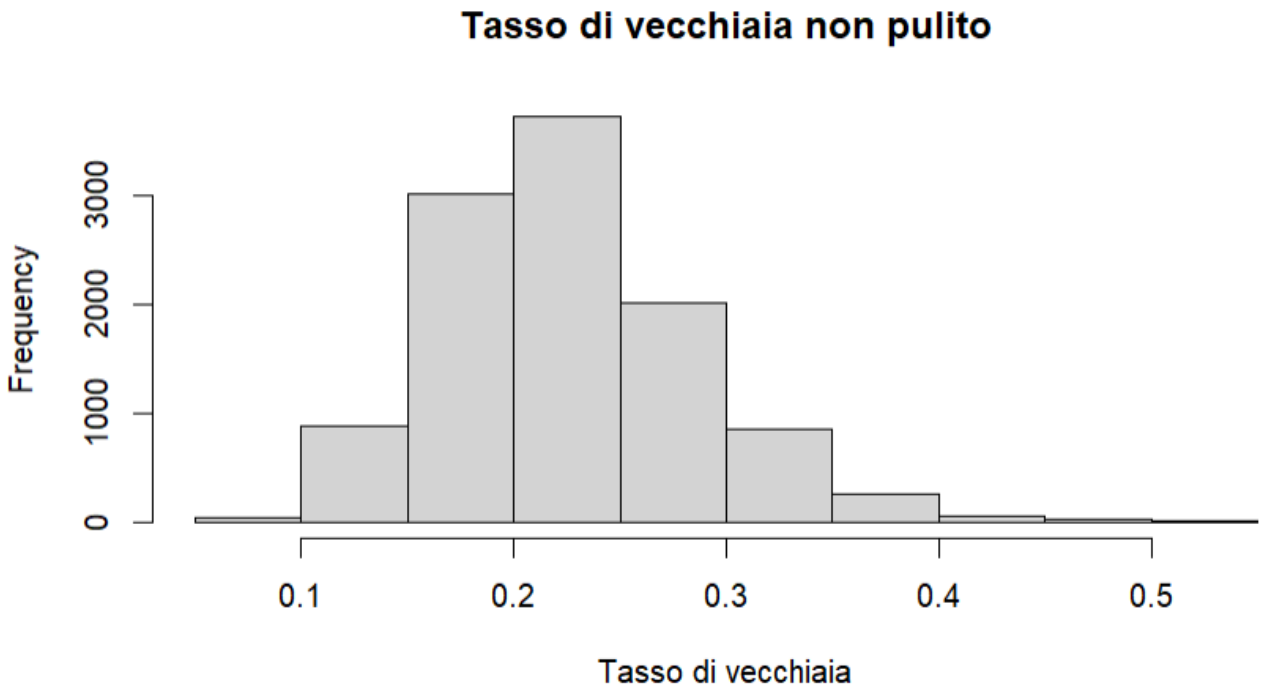
```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
8.946  13.007  13.548  13.515  14.132  15.243
```



VECCHIAIA

Statiche ottenute dalla summary del tasso di vecchiaia non pulito:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.06693 0.18231 0.21997 0.22511 0.25976 0.53009
```



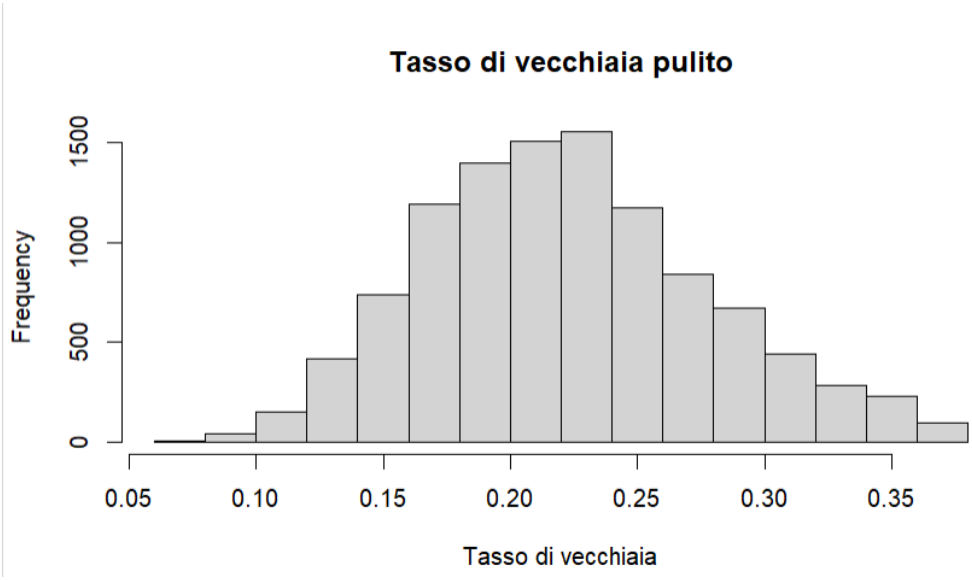
Sembra avere una coda più pesante a sinistra che a destra.

Summary della variabile dopo il metodo di Tukey fences:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.06693 0.18192 0.21889 0.22245 0.25764 0.37548
```

Verifichiamo i logaritmi:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.7041 -1.7042 -1.5192 -1.5353 -1.3562 -0.9796
```

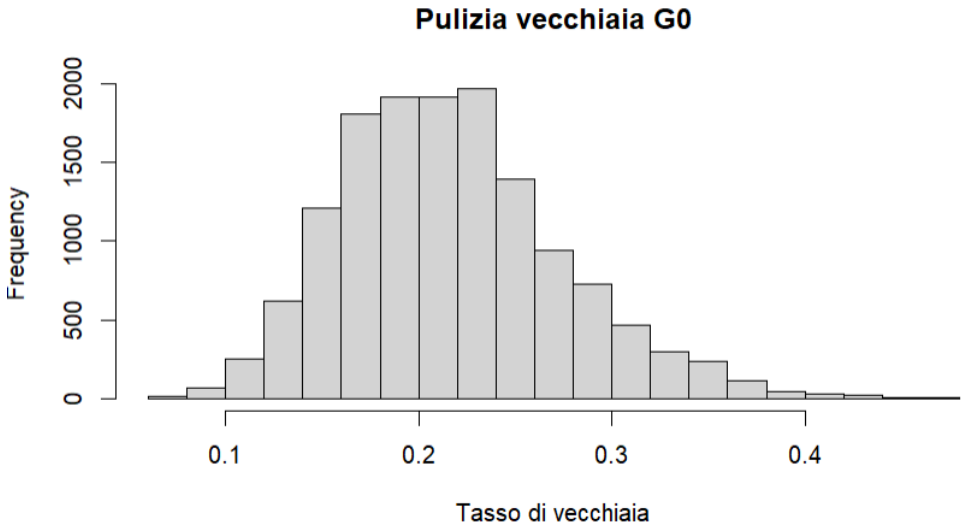


Considerando k=3 e individuando i gross-outliers, dopo l'applicazione del metodo il summary risuta:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.06052 0.17555 0.21168 0.21678 0.24962 0.46980
```

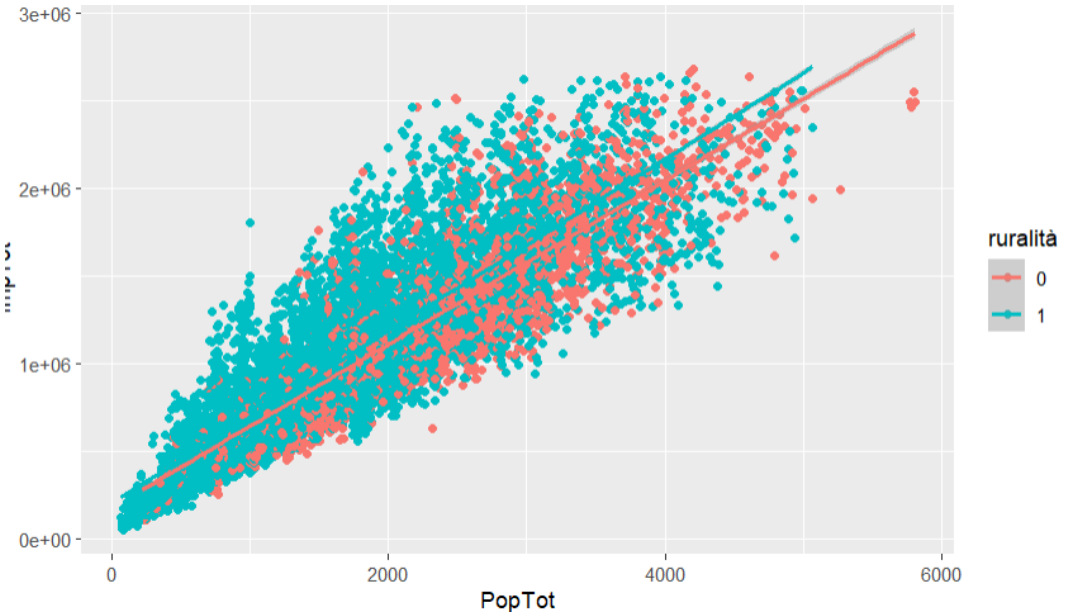
Verifichiamo i logaritmi:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.8048 -1.7398 -1.5527 -1.5641 -1.3878 -0.7555
```



BREVE ANALISI MULTIVARIATA DELLE FEATURES

I.Relazione tra IMPTOT E POPOLAZIONE TOTALE CONDIZIONATA SU RURALITA'



Dal grafico notiamo che in entrambi i gruppi è presente una relazione positiva tra i due caratteri, tuttavia, il gruppo 1 sembra avere una retta di regressione traslata verso l'alto rispetto al gruppo 0.

Covarianza elevata ma maggiore nel gruppo 1.

```
ruralita        Cov
<fct>         <dbl>
0         513248994.
1         451115492.
```

La correlazione è invece pressoché la stessa:

```
ruralità   Cov
<fct>    <dbl>
0        0.899
1        0.883
```

Verifichiamo di quanto , e se per davvero, il gruppo 1 ha una retta traslata verso l'alto, tramite il metodo di regressione lineare.

```
> rgr_1=lm(ImpTot~PopTot+ruralità, data=dat_nuovo)
> summary(rgr_1)

Call:
lm(formula = ImpTot ~ PopTot + ruralità, data = dat_nuovo)

Residuals:
    Min      1Q  Median      3Q     Max
-891978 -157819  -37816  121256 1260561

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.387e+05  7.195e+03   19.27   <2e-16 ***
PopTot      4.845e+02  2.435e+00  199.00   <2e-16 ***
ruralità1   7.692e+04  5.683e+03   13.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 247600 on 10721 degrees of freedom
Multiple R-squared:  0.7953,    Adjusted R-squared:  0.7952
F-statistic: 2.082e+04 on 2 and 10721 DF,  p-value: < 2.2e-16
```

I coefficienti risultano essere tutti significativi con un p-value prossimo allo 0. Ciò che notiamo è che a parità di condizioni avere ruralità=1 implica un'impegno di spesa totale incrementata di 7.69*10^4, per cui conferma che la sua retta di regressione è traslata verso l alta rispetto al gruppo 0. Il p-value è prossimo all'80 per cento, per cui queste variabili riescono a spiegare la variabilità di imptot all'80 per cento.

A) SLIDESET RIPULITO DA TUTTI I VALORI ANOMALI DI TUTTE LE VARIABILI

1) FDH 1994

```
Eff range         #     %
     F ==1     1255 70.43
  1< F =<1.1    209 11.73
1.1< F =<1.2    122  6.85
1.2< F =<1.3     76  4.26
1.3< F =<1.5     71  3.98
1.5< F =<  2     47  2.64
  2< F =<  5      2  0.11
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00    1.00    1.00    1.06    1.04    2.06
```

Implementando la fdh visualizziamo a schermo che il 70.43% delle osservazioni(ovvero 1255 comuni) sono efficienti. Il restante 30% circa sono inefficienti, ovvero 527. Il punteggio piu elevato è piu inefficiente ha un minimo in 1 e un massimo in 2.059, ovvero l'espansione degli output per arrivare alla frontiera.

2) DEA-V 1994

Come ci aspettavamo , considerata la DEA-V come un FDH con aggiunta di convessità, il numero di unità efficienti è calata in maniera drastica arrivando a ricoprire solo 185 comuni, ovvero il 10.4 per cento del totale. il restante 90 % di osservazioni(1597 comuni), da questo punto di vista,è inefficiente.

```
---
 Eff range         #    %
     F ==1     185 10.4
  1< F =<1.1   234 13.1
1.1< F =<1.2   303 17.0
1.2< F =<1.3   292 16.4
1.3< F =<1.5   382 21.4
1.5< F =<  2   309 17.3
  2< F =<  5    77  4.3
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00    1.11    1.26    1.33    1.46    3.19
```

3) DEA-C 1994

Ancora, ci rendiamo conto che con la DEA-C il numero di comuni effettivamente efficienti sono solo 92, ovvero il 5.2% del totale. I restanti 1690 comuni sono inefficienti sotto questo fronte.

```
  Eff range         #    %
       F ==1       92  5.2
    1< F =<1.1    150  8.4
  1.1< F =<1.2    237 13.3
  1.2< F =<1.3    267 15.0
  1.3< F =<1.5    470 26.4
  1.5< F =<  2    459 25.8
    2< F =<  5    107  6.0
   Min. 1st Qu.  Median    Mean
   1.00    1.19    1.35    1.42
3rd Qu.    Max.
   1.57    3.64
```

4) FDH 2003

Passiamo ora invece all'analisi dell'anno 2003, inteso come anno post-policy. analizziamolo nel particolare utilizzando le stesse tecniche di analisi non parametrica. Nel 2003, anno inteso come post-policy, il numero di comuni efficienti diminuisce rispetto al 1994. Infatti in questo anno i comuni efficienti sono 1031, ovvero il 58% del totale rispetto a 1255 comuni, ovvero il 70% del totale dell anno 1994.

```
---
  Eff range         #     %
       F ==1     1031 58.85
    1< F =<1.1    210 11.99
  1.1< F =<1.2    179 10.22
  1.2< F =<1.3    112  6.39
  1.3< F =<1.5    109  6.22
  1.5< F =<  2     98  5.59
    2< F =<  5     13  0.74
   Min. 1st Qu.  Median    Mean
   1.00    1.00    1.00    1.11
3rd Qu.    Max.
   1.13    2.66
```

5) DEA-V 2003

Come prima i comuni efficienti scendono rispetto al fdh a 148, ovvero l'8.4 per cento del totale. i punteggi hanno un minimo in 1 e un massimo in 3.99.

```
--
 Eff range         #    %
      F ==1      148  8.4
   1< F =<1.1    182 10.4
 1.1< F =<1.2    200 11.4
 1.2< F =<1.3    190 10.8
 1.3< F =<1.5    359 20.5
 1.5< F =<  2    478 27.3
   2< F =<  5    195 11.1
  Min. 1st Qu.  Median    Mean
  1.00    1.16    1.38    1.47
rd Qu.    Max.
  1.69    3.99
```

6) DEA-C 2003

Anche nella dea-c i paesi efficienti calano in maniera drastica, ora i paesi considerati efficienti sono 73, ovvero il 4.2% complessivo. I restanti 1679 comuni sono inefficienti sotto questo fronte.

```
  Eff range         #    %
       F ==1       73  4.2
    1< F =<1.1    113  6.4
  1.1< F =<1.2    156  8.9
  1.2< F =<1.3    167  9.5
  1.3< F =<1.5    349 19.9
  1.5< F =<  2    606 34.6
    2< F =<  5    288 16.4
   Min. 1st Qu.  Median    Mean
   1.00    1.25    1.51    1.58
3rd Qu.    Max.
   1.83    4.27
```

B) DATASET RIPULITO DELLA SOLA VARIABILE QUINTALI DI RIFIUTI

1) FDH 1994

Implementando la fdh per il dataset ripulito della sola variabile quintali di rifiuti raccolti, visualizziamo a schermo che il 71% delle osservazioni(ovvero 1394 comuni) sono efficienti rispetto al 70% del dataset completamente ripulito. Il restante 29% circa sono inefficienti, ovvero 569.

```
  Eff range         #    %
       F ==1     1394 71.0
    1< F =<1.1    227 11.6
  1.1< F =<1.2    126  6.4
  1.2< F =<1.3     85  4.3
  1.3< F =<1.5     82  4.2
  1.5< F =<  2     47  2.4
    2< F =<  5      2  0.1
   Min. 1st Qu.  Median    Mean
   1.00    1.00    1.00    1.06
3rd Qu.    Max.
   1.03    2.06
```

2) DEA-V 1994

Come ci aspettavamo, il numero di efficienti è calata in maniera decisiva arrivando a ricoprire solo 185 comuni, ovvero il 9.4 per cento del totale. Tuttavia passando alla dea-v il numero di comuni efficienti rispetto al dataset gross outliers è identica, anche nel caso precedente il numero di comuni efficienti è di 185.

```
  Eff range         #    %
       F ==1      185  9.4
    1< F =<1.1    242 12.3
  1.1< F =<1.2    311 15.8
  1.2< F =<1.3    333 17.0
  1.3< F =<1.5    455 23.2
  1.5< F =<  2    355 18.1
    2< F =<  5     82  4.2
   Min. 1st Qu.  Median    Mean
   1.00    1.12    1.27    1.34
3rd Qu.    Max.
```

3) DEA-C 1994

Ancora, ci rendiamo conto che con la DEA-C il numero di comuni effettivamente efficienti sono solo 88, ovvero il 4.5% del totale. In questo caso il numero è inferiore di 4 unità rispetto al dataset con cui stiamo facendo il confronto.

```
 Eff range         #    %
      F ==1       88  4.5
   1< F =<1.1    166  8.5
 1.1< F =<1.2    235 12.0
 1.2< F =<1.3    290 14.8
 1.3< F =<1.5    501 25.5
 1.5< F =<  2    551 28.1
   2< F =<  5    132  6.7
  Min. 1st Qu.   Median    Mean
  1.00    1.20    1.38    1.44
rd Qu.    Max.
  1.61    3.67
```

4) FDH 2003

```
Eff range          #     %
     F ==1      1151 60.36
  1< F =<1.1     227 11.90
1.1< F =<1.2     182  9.54
1.2< F =<1.3     120  6.29
1.3< F =<1.5     113  5.93
1.5< F =<  2     101  5.30
  2< F =<  5      13  0.68
 Min. 1st Qu.   Median    Mean
 1.00    1.00    1.00    1.10
d Qu.    Max.
 1.12    2.66
```

5) DEA-V 2003

Come prima i comuni efficienti scendono a 151, rispetto ai 185 del 1994, ovvero l'7.9 per cento del totale. i punteggi hanno un minimo in 1 e un massimo in 3.99.

```
 Eff range         #    %
      F ==1      151  7.9
   1< F =<1.1    190 10.0
 1.1< F =<1.2    216 11.3
 1.2< F =<1.3    203 10.6
 1.3< F =<1.5    413 21.7
 1.5< F =<  2    529 27.7
   2< F =<  5    205 10.7
  Min. 1st Qu.   Median    Mean
  1.00    1.16    1.38    1.47
3rd Qu.    Max.
  1.68    3.99
```

6) DEA-C 2003

Anche nella dea-c i paesi efficienti calano in maniera drastica, ora i paesi considerati efficienti sono 74, ovvero il 4.2% complessivo. I restanti comuni sono inefficienti sotto questo fronte. Il numero di comuni efficienti nel 1994 rispetto alla dea-c sono 92.

```
Eff range        #    %
     F ==1      74  3.9
   1< F =<1.1  109  5.7
 1.1< F =<1.2  162  8.5
 1.2< F =<1.3  181  9.5
 1.3< F =<1.5  360 18.9
 1.5< F =<  2  683 35.8
   2< F =<  5  338 17.7
  Min. 1st Qu.  Median    Mean
  1.00    1.27    1.53    1.60
3rd Qu.    Max.
  1.86    4.27
```

#SUMMARY PER L ANNO POST-POLICY: Quindi, in maniera sintetica, per l'anno post-policy: per la fdh e dea-v, il numero di comuni efficienti è maggiore nel caso del dataset ripulito da una sola variabile rispetto al dataset ripulito dai gross-outliers. Mentre per la dea-c il numero di comuni efficienti è quasi simile, si differenziano per un solo comune.

ANALISI DID:

Consideriamo la regressione in cui la variabile dipendente è l'efficienza e le variabili indipendenti sono dopo, doppio e did. Considerando un alpha pari a 0.05, tutti i p-value associati alle variabili sono prossimi allo zero per l'intercetta e la dummy dopo, leggermente più elevato per doppio e did, ma comunque tutti statisticamente significativi, per cui significa che hanno un impatto. Il concetto da fissare è che un valore dello stimatore negativo implica un miglioramento dell'efficienza. Per cui a parità di condizioni, il valore positivo per dopo implica che per l anno dopo il 1994 l'efficienza è diminuita. Al contrario, un valore negativo per il coefficiente doppio implica che per i paesi in cui è stato introdotto il doppio turno , quindi i comuni con piu di 15000 abitanti, l'efficienza è aumentata rispetto ai comuni in cui vige il singolo turno. In particolare, la variabile che ci interessa più di tutte è did, che è negativa, per cui tutti i comuni effettivamente trattati dopo il 1994 l'efficienza è aumentata.

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.4776 -0.2512 -0.0714  0.1599  2.5128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.34465    0.00811  165.70   <2e-16 ***
dopo         0.13297    0.01155   11.51   <2e-16 ***
doppio      -0.14361    0.04721   -3.04   0.0024 **
did         -0.16952    0.06969   -2.43   0.0150 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 0.354 on 3866 degrees of fr
eedom
Multiple R-squared:  0.0434,    Adjusted R-squared:
0.0426
F-statistic: 58.4 on 3 and 3866 DF,  p-value: <2e-16
```

ANALISI DID AUMENTATA

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.9671 -0.2156 -0.0542  0.1565  2.3117

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.42e+00   6.69e-02   21.26  < 2e-16 ***
vecchiaia    8.22e-01   1.57e-01    5.25  1.6e-07 ***
laurea      -1.16e+00   2.99e-01   -3.88  0.00011 ***
superficie  -2.03e-03   1.44e-04  -14.05  < 2e-16 ***
altitudine   1.06e-04   2.37e-05    4.45  8.7e-06 ***
litoraneità1 7.27e-02   2.64e-02    2.76  0.00586 **
gradourb2   -1.47e-01   1.57e-02   -9.39  < 2e-16 ***
gradourb3   -1.84e-01   2.56e-02   -7.21  6.8e-13 ***
ruralità1    7.70e-03   1.56e-02    0.49  0.62145
dopo         1.20e-01   1.07e-02   11.25  < 2e-16 ***
doppio       7.71e-02   4.55e-02    1.69  0.09053 .
did         -2.16e-01   6.42e-02   -3.36  0.00078 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.326 on 3858 degrees of freedom
Multiple R-squared:  0.192,     Adjusted R-squared:  0.19
F-statistic: 83.4 on 11 and 3858 DF,  p-value: <2e-16
```

Vediamo che did continua a rimanere significativo e negativo a un livello di alpha=0.05.Conferma la presenza di un miglioramento dell’efficienza in maniera generale. L'unico dei coefficienti che non è per nulla significativo è ruralità, mentre se fissiamo un livello di significatività 0.10 anche dopo rientra nelle variabili significative. Da un analisi generale di questo modello riusciamo a dedurre che: 1) i coefficienti di vecchiaia, altitudine, litoraneirà, grado di urbanizzazione basso, ruralità=1, dopo e doppio sono positivi il che significa , che nel quadro generale, sono legate a inefficenza.2) Mentre i coefficienti associati a laurea, superficie, grado di urbanizzazione alta e did sono negativi, per cui sono legati all'efficienza. Ho quindi arricchito le informazioni con i dati di panel:

```
Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
 -0.621  -0.259  -0.079   0.158   2.526

Coefficients:
              Estimate Std. Error t-value Pr(>|t|)
vecchiaia     2.93e-01   1.73e-01    1.69   0.0904 .
laurea        1.07e+00   3.31e-01    3.24   0.0012 **
superficie    6.25e-04   1.55e-04    4.04  5.5e-05 ***
altitudine    1.13e-05   2.61e-05    0.43   0.6652
litoraneità1  2.37e-02   2.87e-02    0.83   0.4079
gradourb2     1.74e-02   1.73e-02    1.01   0.3138
gradourb3     4.10e-02   2.80e-02    1.47   0.1424
ruralità1     9.63e-04   1.71e-02    0.06   0.9552
did          -2.56e-01   5.20e-02   -4.93  8.6e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    507
Residual Sum of Squares: 500
R-Squared:      0.0144
Adj. R-Squared: 0.0118
F-statistic: 6.24324 on 9 and 3859 DF, p-value: 8.44e-09
```

```
1994 2003
1.16 1.17
```

Impostando gli effetti fissi sul tempo ci rendiamo conto che entra in contrasto con ciò che abbiamo appena detto, per cui sembra che ci sia stato un lieve peggioramento dell' efficienza dopo l'applicazione del doppio turno.

STATISTICAL MATCHING

1) ANNO 1994

```
Sample Sizes:
          Control Treated
All          1842      43
Matched        43      43
Unmatched    1799       0
Discarded       0       0
```

```
Summary of Balance for Matched Data:
                  Means Treated Means Control
vecchiaia.1994            0.177         0.176
laurea.1994               0.140         0.138
superficie.1994          60.601        48.166
altitudine.1994         140.488       145.930
litoraneità.19940         0.674         0.674
litoraneità.19941         0.326         0.326
gradourb.19941            0.070         0.140
gradourb.19942            0.605         0.535
gradourb.19943            0.326         0.326
ruralità.19940            0.860         0.837
ruralità.19941            0.140         0.163
                  Std. Mean Diff. Var. Ratio
vecchiaia.1994               0.022      1.275
laurea.1994                  0.057      1.521
superficie.1994              0.278      0.900
altitudine.1994             -0.043      0.715
litoraneità.19940            0.000          .
litoraneità.19941            0.000          .
gradourb.19941              -0.274          .
gradourb.19942               0.143          .
gradourb.19943               0.000          .
ruralità.19940               0.067          .
ruralità.19941              -0.067          .
                  eCDF Mean eCDF Max
vecchiaia.1994        0.039    0.116
laurea.1994           0.047    0.163
superficie.1994       0.138    0.256
altitudine.1994       0.022    0.093
litoraneità.19940     0.000    0.000
litoraneità.19941     0.000    0.000
gradourb.19941        0.070    0.070
gradourb.19942        0.070    0.070
```

Le unità di controllo sono quelle del dataset mentre i trattati sono quelli effettivamente interessati dall'introduzione del doppio turno, ovvero i comuni con una popolazione totale >15000. Il numero di unità di controlli è decisamente maggiore rispetto ai trattati per cui l'implementazione è attendibile. Guardiamo al bilanciamento : con il matching andiamo a individuare i comuni gemelli tra trattati e non trattati che hanno tratti simili, come: tasso di vecchiaia simili, tasso di laurea simili, superficie simile, grado di urbanizzazione simile. Analizzando la media dei dati in generale notiamo una leggerenza differenza nel tasso di occupazione, superficie e ruralità.Forte differenza in grado di urbanizzazione e altitudine.Dopo il bilanciamento per i dati matchati riusciamo a comprendere che i dati sono pressocchè simili, quindi ha fatto un buon abbinamento.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.34686    0.00719   187.3  < 2e-16 ***
doppio.1994 -0.18092    0.04761    -3.8  0.00015 ***
---
```

Doppio continua a essere statisticamente significativo e negativo, quindi implica che i comuni trattati sono più efficienti.

```
mean in group 0 mean in group 1
          1.347           1.166
```

confermato anche dal test che individua una media di efficienza non trattati maggiore rispetto ai trattati. Un valore più alto di eff implica una maggiore inefficienza allora realmente i trattati sono più efficienti.

ANNO 2003

```
Sample Sizes:
          Control Treated
All          1836      49
Matched        49      49
Unmatched    1787       0
Discarded       0       0
```

```
Summary of Balance for Matched Data:
                  Means Treated
vecchiaia.1994            0.178
laurea.1994               0.139
superficie.1994          63.246
altitudine.1994         144.918
litoraneità.19940         0.714
litoraneità.19941         0.286
gradourb.19941            0.082
gradourb.19942            0.612
gradourb.19943            0.306
ruralità.19940            0.837
ruralità.19941            0.163
                  Means Control
vecchiaia.1994            0.178
laurea.1994               0.137
superficie.1994          48.623
altitudine.1994         146.184
litoraneità.19940         0.714
litoraneità.19941         0.286
gradourb.19941            0.143
gradourb.19942            0.551
gradourb.19943            0.306
ruralità.19940            0.816
ruralità.19941            0.184
                  Std. Mean Diff.
vecchiaia.1994             -0.005
laurea.1994                 0.076
superficie.1994             0.321
altitudine.1994            -0.009
litoraneità.19940           0.000
litoraneità.19941           0.000
gradourb.19941             -0.224
gradourb.19942              0.126
gradourb.19943              0.000
ruralità.19940              0.055
ruralità.19941             -0.055
```

Doppio continua a essere statisticamente significativo e negativo, il che implica che i comuni trattati dopo la policy sono più efficienti dei non trattati.

```
Coefficients:
            Estimate Std. Error t value
(Intercept)  1.47980    0.00926  159.84
doppio.2003 -0.31891    0.05742   -5.55
            Pr(>|t|)
(Intercept)  < 2e-16 ***
doppio.2003  3.2e-08 ***
---
```

come prima, viene individuata una media di eff dei non trattati maggiore della media dei trattati per cui continua a confermare che i trattati sono più efficienti.

(risultati relativi al t test e commenti nel file TESINA)

CONTROLLO DELLA ROBUSTEZZA

Dal modello di regressione visualizziamo che tutte e tre le variabili sono significative a un livello alpha pari a 0.10, did è negativa mentre dopo e doppio sono positive. Il che significa che la spesa per i comuni effettivamente trattati è diminuita leggermente. Addiritura eliminando l'intercetta doppio risulta essere statisticamente a livello alpha 0.05 per cento.

```
                  Estimate Std. Error t value Pr(>|t|)
log(QliRifiuti)    0.00348    0.00342    1.02    0.309
log(KmStrade)     -0.12760    0.00694  -18.39  < 2e-16 ***
log(Studenti)     -0.95534    0.01111  -86.01  < 2e-16 ***
log(NrPuntiLuce)  -0.02685    0.01523   -1.76    0.078 .
log(NrPermessi)   -0.38242    0.00898  -42.60  < 2e-16 ***
log(NrIscritti)    2.16194    0.02169   99.65  < 2e-16 ***
log(KmPolizia)     0.71993    0.01094   65.82  < 2e-16 ***
log(NrCertificati) -0.26906    0.00902  -29.82  < 2e-16 ***
dopo               0.21909    0.00859   25.51  < 2e-16 ***
doppio            -0.26929    0.04714   -5.71  1.1e-08 ***
did               -0.10706    0.05001   -2.14    0.032 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Did è non significativa con i dati panel , quindi non possiamo dare una spiegazione generale, è come se la politica non avesse avuto impatto.

```
                  Estimate Std. Error t-value Pr(>|t|)
log(QliRifiuti)   -0.02679    0.00214  -12.49  < 2e-16 ***
log(KmStrade)      0.03290    0.00444    7.41  1.3e-13 ***
log(Studenti)     -0.21142    0.00856  -24.69  < 2e-16 ***
log(NrPuntiLuce)   0.09126    0.00962    9.48  < 2e-16 ***
log(NrPermessi)   -0.10167    0.00588  -17.29  < 2e-16 ***
log(NrIscritti)    1.16207    0.01508   77.06  < 2e-16 ***
log(KmPolizia)    -0.01584    0.00842   -1.88    0.060 .
log(NrCertificati) 0.01096    0.00594    1.84    0.065 .
did               -0.00168    0.01178   -0.14    0.887
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#Svolgendo un subset della popolazione e restringendola a >8000, otteniamo che la variabile did è significativa al 10

#per cento, rimane negativa.

```
                  Estimate Std. Error t-value Pr(>|t|)
log(QliRifiuti)    -0.0011     0.0115   -0.10    0.924
log(KmStrade)      -0.0153     0.0132   -1.16    0.245
log(Studenti)      -0.6020     0.0303  -19.88  < 2e-16 ***
log(NrPuntiLuce)    0.1545     0.0321    4.81  1.7e-06 ***
log(NrPermessi)    -0.1177     0.0138   -8.54  < 2e-16 ***
log(NrIscritti)     1.4734     0.0541   27.22  < 2e-16 ***
log(KmPolizia)      0.2177     0.0308    7.06  2.6e-12 ***
log(NrCertificati) -0.0273     0.0217   -1.26    0.208
did                -0.0554     0.0332   -1.67    0.096 .
```

```
   1994    1999    2000    2001    2002    2003    2004    2005
-0.1034 -0.0741 -0.0503  0.0140  0.0326  0.0387  0.0668  0.0951
> fixef(panel_regr, effect="time")
1994 1999 2000 2001 2002 2003 2004 2005
3.77 3.79 3.82 3.88 3.90 3.91 3.94 3.96
```

Effetto sul tempo visualizziamo che con il crescere degli anni gli impegni di spesa totale crescono , anche se nel nostro modello il valore di did è negativo. Chiaramente non è piuttosto affidabile, considerato il p-value elevatissimo. Non possiamo affidarci ai risultati di panel per cui traiamo le conclusioni affidandoci ai risultati posti in essere dal passo due e dal comando lm.