Pilgrim Bank

Luca Bajardi e Francesca Collini

31/07/2020

Carichiamo i dati leggendo il file csv e settiamo il seme del generatore pseudo-casuale così da avere i risultati sempre uguali.

```
rm(list =ls())
Pilgrim = read.csv(file = "PilgrimABC.csv", header=T)
attach(Pilgrim)
```

Osserviamo che il nostro dataset contiene 31634 osservazioni e 11 variabili (le ultime due non sono interessanti ai fini della nostra analisi). I dati sono relativi a due anni, 1999 e 2000, per ciascuno abbiamo a disposizione un'informazione relativa al profitto (o alla perdita) di un determinato cliente in quell'anno e un'informazione che ci dice se quel cliente in quell'anno ha utilizzato l'opzione di online banking oppure no: la variabile Online quindi è una variabile binaria. Abbiamo poi a dispozione altre informazioni relative al cliente:

- AGE: variabile categorica che indica a che fascia di età appartiene il cliente(1 = less than 15 years; 2 = 15-24 years; 3 = 25-34 years; 4 = 35-44 years; 5 = 45-54 years; 6 = 55-64 years; 7 = 65 years and older.)
- INCOME: variabile categorica che indica il range del reddito di ciascun cliente (1 = less than $\$15,000$; 2 = $\$15,000 - \$19,999$; 3 = $\$20,000 - \$29,999$; 4 = $\$30,000 - \$39,999$; 5 = $\$40,000 - \$49,999$; 6 = $\$50,000 - \$74,999$; 7 = $\$75,000 - \$99,999$; 8 = $\$100,000 - \$124,999$; 9 = $\$125,000$ and more.)
- TENURE:
- DISTRICT: variabile categorica che indica uno delle tre regioni geografiche in cui si trova il cliente.

```
dim(Pilgrim)
```

```
## [1] 31634    11
```

```
names(Pilgrim[, 1:9])
```

```
## [1] "ID"         "X9Profit"   "X9Online"   "X9Age"      "X9Inc"
## [6] "X9Tenure"   "X9District" "X0Profit"   "X0Online"
```

Rinomico colonne per evitare problemi con i nomi

```
Profit=Pilgrim$X9Profit
Online=Pilgrim$X9Online
Age=Pilgrim$X9Age
Income=Pilgrim$X9Inc
Tenure=Pilgrim$X9Tenure
District=Pilgrim$X9District
```

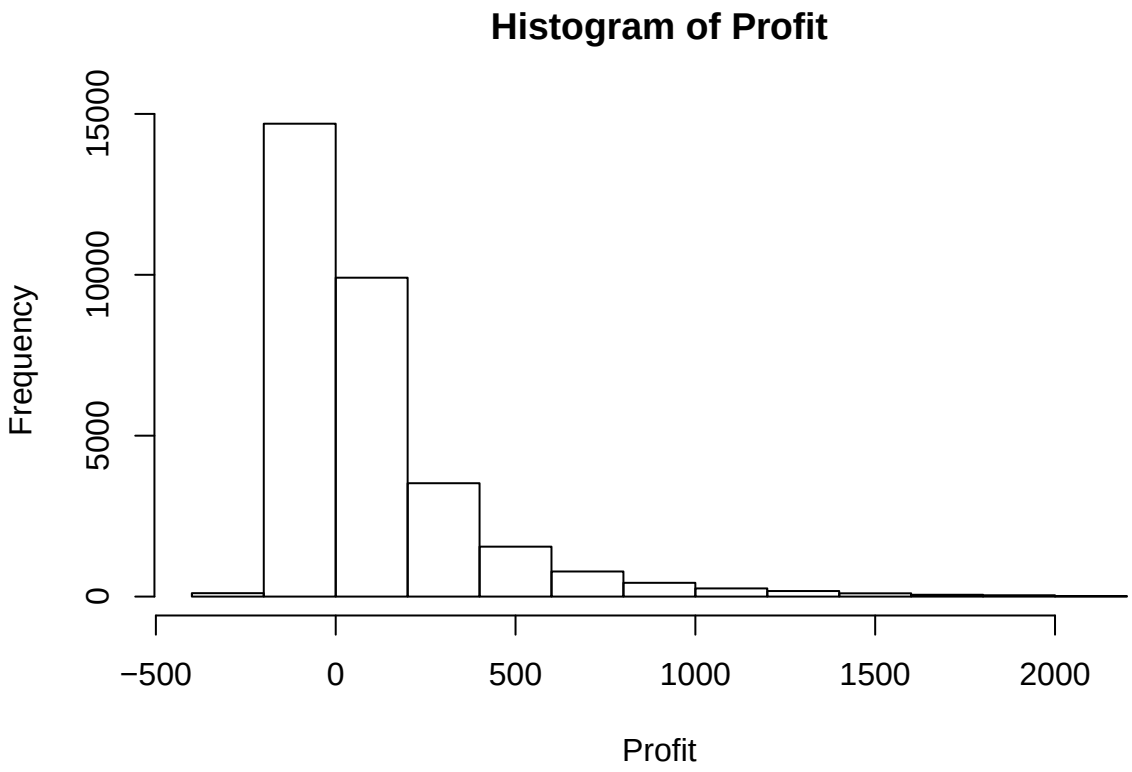
```
head(Pilgrim)
```

```
##   ID X9Profit X9Online X9Age X9Inc X9Tenure X9District X0Profit X0Online
## 1  1       21        0    NA    NA     6.33       1200       NA       NA
## 2  2       -6        0     6     3    29.50       1200      -32        0
## 3  3      -49        1     5     5    26.41       1100      -22        1
```

```
## 4  4       -4        0    NA    NA     2.25       1200       NA       NA
## 5  5      -61        0     2     9     9.91       1200       -4        0
## 6  6      -38        0    NA     3     2.33       1300       14        0
##   X9Billpay X0Billpay
## 1         0        NA
## 2         0         0
## 3         0         0
## 4         0        NA
## 5         0         0
## 6         0         0
```

Dall'istogramma sul profitto nel 1999 possiamo notare che c'è un discreto numero di clienti che mi fanno perdere e che l'istogramma è molto scodato.

```
hist(Profit)
```



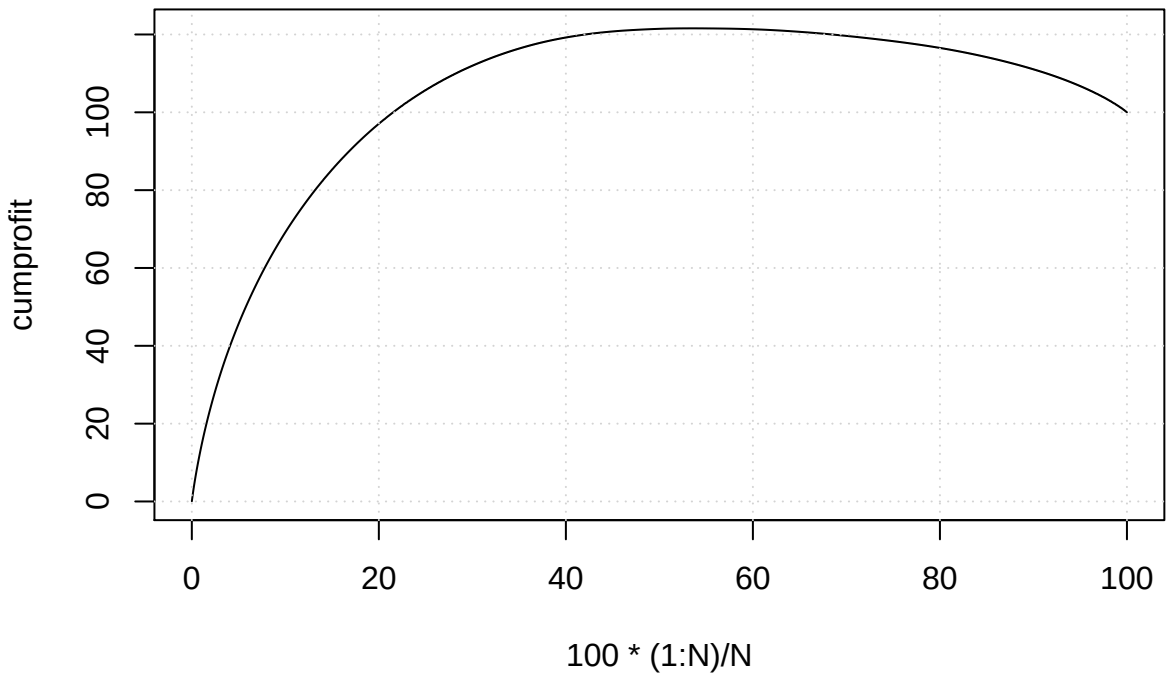
Possiamo notare che ci sono 16832 persone che generano profitto sulle 31634 presenti nel dataset.

```
N=length(Profit)
Nprofitable = sum(Profit>0)
cat('profitable = ', Nprofitable, ' out of ', N, '\n')
```

```
## profitable =  16832  out of  31634
```

Tramite la curva di Pareto possiamo notare che poche persone fanno guadagnare tanto, mentre la maggior parte delle persone fa guadagnare poco o addirittura fa perdere guadagno.

```
cumprofit=cumsum(sort(Profit,decreasing=TRUE))*100/sum(Profit)
plot(100*(1:N)/N,cumprofit,type='l')
grid()
```



Dall'analisi dei profitti medi possiamo notare che il profitto generato da chi utilizza i servizi online è maggiore rispetto a chi li utilizza offline.

```
cat('average profit ', mean(Profit), '\n')
```

```
## average profit  111.5027
```

```
ProfitOnline = Profit[Online==1]
cat('average profit ON', mean(ProfitOnline), '\n')
```

```
## average profit ON 116.6668
```

```
ProfitOffline = Profit[Online==0]
cat('average profit OFF', mean(ProfitOffline), '\n')
```

```
## average profit OFF 110.7862
```

Analizziamo l'intervallo di confidenza per vedere se i dati sono sufficienti o no, se intervallo grande avrei bisogno di più clienti

```
cat('Conf int profit ', t.test(Profit)$conf.int, '\n')
```

```
## Conf int profit  108.496 114.5094
```

```
cat('p-value difference ',t.test(ProfitOnline,ProfitOffline)$p.value, '\n')
```

```
## p-value difference  0.2254368
```

Eseguo il `t.test` su due popolazioni e vado a vedere il p-value, il p-value è maggiore di 0.05 e quindi vedo che non c'è una differenza significativa, quindi non rifiuto l'ipotesi nulla sulla differenza tra le medie.

Posso ottenere lo stesso risultato facendo:

```
mod = lm(Profit ~ Online)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ Online)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -337.67 -144.79 -101.79   52.21 1960.21
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  110.786      1.637  67.678   <2e-16 ***
## Online         5.881      4.690   1.254     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 272.8 on 31632 degrees of freedom
## Multiple R-squared:  4.97e-05,   Adjusted R-squared:  1.809e-05
## F-statistic: 1.572 on 1 and 31632 DF,  p-value: 0.2099
```

Infatti posso osservare che il valore dell'intercetta è la media del profitto di quelli che operano offline e il valore aggiunto di quelli che operano online è di 6 dollari come la differenza tra la media di quelli che operano online e la media di quelli che operano offline. Anche qua il p-value è superiore a 0.05 come il precedenza. Questo significa che non è significativa perhcé metto insieme clienti molto diversi tra loro, per questo controllo l'età.

La variabile `Age` è riconosciuta numerica anche se in realtà è categorica. Nell'analisi dovremo trasformarla in `factor` perché altrimenti vi è troppa influenza dell'ordinamento delle variabili categoriali.

```
Age1 = as.factor(Age)
summary(Age1)
```

```
##    1    2    3    4    5    6    7 NA's
##  710 3650 5390 5376 3236 2290 2693 8289
```

Possiamo notare che ci sono 8289 osservazioni in cui non è presente l'età, infatti nel `summary` sottostante possiamo notare che nonostante ci siano 32 mila osservazioni ci sono solo 23 mila gradi di libertà in quanto quelle con i missing values sono eliminate.

```
mod = lm(Profit ~ Online+Age1)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ Online + Age1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -404.52 -162.90  -84.62   68.80 1952.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.802     10.485  -0.172    0.864
## Online        27.246      5.519   4.937 8.01e-07 ***
```

```
## Age12         54.425     11.401   4.774 1.82e-06 ***
## Age13        112.699     11.098  10.155  < 2e-16 ***
## Age14        133.820     11.103  12.053  < 2e-16 ***
## Age15        144.986     11.531  12.574  < 2e-16 ***
## Age16        160.844     11.965  13.443  < 2e-16 ***
## Age17        193.072     11.757  16.422  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 277.9 on 23337 degrees of freedom
##   (8289 observations deleted due to missingness)
## Multiple R-squared:  0.02494,    Adjusted R-squared:  0.02464
## F-statistic: 85.26 on 7 and 23337 DF,  p-value: < 2.2e-16
```

Ora sulle fasce d'età vi è una significatività sia statistica che di business, infatti le diverse età potrebbero implicare diversi redditi, un giovane usa di più l'online banking ma ha un reddito minore quindi fa guadagnare di meno.

Posso notare che a prescindere dall'utilizzo dell'online o dell'offline i giovani sono meno profittevoli degli anziani:

```
lapply(split(Profit,as.factor(Age)),mean)
```

```
## $`1`
## [1] 3.45493
##
## $`2`
## [1] 58.48959
##
## $`3`
## [1] 115.1122
##
## $`4`
## [1] 135.6618
##
## $`5`
## [1] 145.7596
##
## $`6`
## [1] 160.41
##
## $`7`
## [1] 192.2614
```

Il 20% dei giovani usa l'online, questa percentuale scende per le fasce di età successive fino ad arrivare quasi allo 0.

```
lapply(split(Online,as.factor(Age)),mean)
```

```
## $`1`
## [1] 0.1929577
##
## $`2`
## [1] 0.2153425
##
## $`3`
## [1] 0.154731
```

```
##
## $`4`
## [1] 0.1337426
##
## $`5`
## [1] 0.09456119
##
## $`6`
## [1] 0.05021834
##
## $`7`
## [1] 0.03639064
```

Ci potrebbe essere un bias dovuto ai dati mancanti

```
AgeGiven = ifelse(is.na(Age),0,1) # 0 dove c'è NA, 1 se c'è l'età
mod = lm(Profit ~ AgeGiven)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ AgeGiven)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -346.19 -150.19  -90.96   50.81 1961.04
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   72.962      2.986   24.43   <2e-16 ***
## AgeGiven      52.224      3.476   15.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 271.9 on 31632 degrees of freedom
## Multiple R-squared:  0.007085,   Adjusted R-squared:  0.007054
## F-statistic: 225.7 on 1 and 31632 DF,  p-value: < 2.2e-16
```

C'è una differenza statisticamente significativa sul profitto tra dove c'è l'età e dove non c'è, quindi eliminando i dati dove manca l'età sto distorgendo l'analisi, infatti c'è una profittabilità media più alta tra chi mi ha dato l'età rispetto a chi non me l'ha data.

LUCA SI è FERMATO QUA

```
#Data Imputation = cerco di riempire i missing
#potremmo creare un modello di regressione per trovare i valori mancanti

# Replace missing with Zero
AgeZero = ifelse(is.na(Age),0,Age)
#dove c'è NA metto 0, ma essendo ordinali non ha senso, sembra che chi non dato l'età è un neonato
table(AgeZero)
```

```
## AgeZero
##    0    1    2    3    4    5    6    7
## 8289  710 3650 5390 5376 3236 2290 2693
```

```r
mod = lm(Profit ~ Online+AgeZero)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ Online + AgeZero)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -393.91 -147.07  -82.03   49.97 1976.97
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.0311     2.6014  21.923  < 2e-16 ***
## Online       13.7925     4.6487   2.967  0.00301 **
## AgeZero      17.6803     0.6697  26.402  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.9 on 31631 degrees of freedom
## Multiple R-squared:  0.02161,    Adjusted R-squared:  0.02155
## F-statistic: 349.3 on 2 and 31631 DF,  p-value: < 2.2e-16
```

```r
#ci è venuto questo valore di online perché ho scelto arbitrariamente di mettere 0

# Replace missing with mean
mm = mean(Age, na.rm=TRUE)
#na.rm=TRUE mi fa la media senza considerare i NA,
# anche se non ha molto senso perché è discretizzato, ma rimane comunque ordinato
AgeAverage = ifelse(is.na(Age),mm,Age)
table(AgeAverage)
```

```
## AgeAverage
##                1                2                3                4
##              710             3650             5390             5376
## 4.04604840436924                5                6                7
##             8289             3236             2290             2693
```

```r
mod = lm(Profit ~ Online+AgeAverage)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ Online + AgeAverage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -398.01 -144.99  -91.28   55.00 1981.04
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.911      4.776   1.028    0.304
## Online        22.005      4.699   4.683 2.84e-06 ***
## AgeAverage    25.682      1.090  23.572  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 270.5 on 31631 degrees of freedom
## Multiple R-squared:  0.01731,    Adjusted R-squared:  0.01725
## F-statistic: 278.6 on 2 and 31631 DF,  p-value: < 2.2e-16
```

```
#cambia il coefficiente di Online quindi a seconda di come tappo il buco esce un risultato diverso
# quindi non è così che dobbiamo procedere

# control for AgeGiven
mod = lm(Profit ~ Online+AgeZero+AgeGiven)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ Online + AgeZero + AgeGiven)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -409.34 -144.14  -82.69   52.07 1967.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   70.926      3.000  23.643  < 2e-16 ***
## Online        19.649      4.685   4.194 2.75e-05 ***
## AgeZero       25.603      1.086  23.582  < 2e-16 ***
## AgeGiven     -51.849      5.598  -9.263  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.5 on 31630 degrees of freedom
## Multiple R-squared:  0.02426,    Adjusted R-squared:  0.02417
## F-statistic: 262.1 on 3 and 31630 DF,  p-value: < 2.2e-16
```

```
#Online è venuto 19
#questo è un modo per considerare AgeZero come categorico

mod = lm(Profit ~ Online+AgeAverage+AgeGiven)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ Online + AgeAverage + AgeGiven)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -409.34 -144.14  -82.69   52.07 1967.12
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -32.663      5.377  -6.074 1.26e-09 ***
## Online        19.649      4.685   4.194 2.75e-05 ***
## AgeAverage    25.603      1.086  23.582  < 2e-16 ***
## AgeGiven      51.740      3.448  15.006  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 269.5 on 31630 degrees of freedom
## Multiple R-squared:  0.02426,    Adjusted R-squared:  0.02417
## F-statistic: 262.1 on 3 and 31630 DF,  p-value: < 2.2e-16
```

```
#Online viene 19 come prima
#ho tappato il buco in modo arbitrario per poter fare la regressione,
# ma poi la considero come categorica
#R^2 è miserevole quindi il modello spiega il 2.5% della variabilità

# Deal with missing income
#faccio lo stesso ragionamento di prima
IncomeZero = ifelse(is.na(Income),0,Income)
IncomeGiven = ifelse(is.na(Income),0,1)
mod = lm(Profit ~ Online+AgeAverage+AgeGiven+IncomeZero+IncomeGiven)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ Online + AgeAverage + AgeGiven + IncomeZero +
##     IncomeGiven)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -459.03 -144.61  -74.61   50.17 1963.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -38.191      5.336  -7.158 8.38e-13 ***
## Online        11.867      4.650   2.552   0.0107 *
## AgeAverage    26.891      1.078  24.941  < 2e-16 ***
## AgeGiven      14.490      8.272   1.752   0.0799 .
## IncomeZero    18.771      0.748  25.094  < 2e-16 ***
## IncomeGiven  -63.553      9.047  -7.024 2.20e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 266.8 on 31628 degrees of freedom
## Multiple R-squared:  0.04365,    Adjusted R-squared:  0.0435
## F-statistic: 288.7 on 5 and 31628 DF,  p-value: < 2.2e-16
```

```
#Online scende un po', R^2 è cresciuto, è raddoppiato ma rimane sempre piccolo

# Control for Tenure and district
any(is.na(District))
```

```
## [1] FALSE
```

```
table(District) #ha tre valori e nessun NA, ma sono considerati numerici
```

```
## District
##  1100  1200  1300
##  3142 24342  4150
```

```
# quindi devo creare varibili categoriche
any(is.na(Tenure))
```

```
## [1] FALSE
```

```
District1100 = ifelse(District==1100,1,0)
District1200 = ifelse(District==1200,1,0)
mod = lm(Profit ~ Online+AgeAverage+AgeGiven+IncomeZero+IncomeGiven+Tenure+District1100+District1200)
summary(mod)
```

```
##
## Call:
## lm(formula = Profit ~ Online + AgeAverage + AgeGiven + IncomeZero +
##     IncomeGiven + Tenure + District1100 + District1200)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -487.17 -141.21  -65.88   48.87 1993.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -44.2382     6.5180  -6.787 1.16e-11 ***
## Online        13.8233     4.6091   2.999  0.00271 **
## AgeAverage    16.6701     1.1482  14.519  < 2e-16 ***
## AgeGiven       4.3913     8.2017   0.535  0.59237
## IncomeZero    16.8530     0.7554  22.310  < 2e-16 ***
## IncomeGiven  -57.1191     8.9956  -6.350 2.19e-10 ***
## Tenure         4.7464     0.1918  24.742  < 2e-16 ***
## District1100  -7.9955     6.2582  -1.278  0.20140
## District1200  13.1986     4.4734   2.950  0.00318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 264.2 on 31625 degrees of freedom
## Multiple R-squared:  0.06234,    Adjusted R-squared:  0.0621
## F-statistic: 262.8 on 8 and 31625 DF,  p-value: < 2.2e-16
```

```
#qualcosa di R^2 lo grattiamo ma poco
#Online è circa 13

#potremmo considerare Age come categorico
#potremmo considerare il rapporto tra Online e Age

#un valore di soglia dell'R^2 non c'è, dipende se migliora la mia prestazione economica
```

```
sum(is.na(Pilgrim$XOProfit))
```

```
## [1] 5238
```

```
detach(Pilgrim)
```