Untitled

Americo

29 aprile 2016

Introduzione

Questo capitolo tratta lo svolgimento di una analisi dati relativa al **direct marketing** bancario, utilizzando un dataset disponibile sul web a questo indirizzo: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing). Questo dataset contiene le informazioni raccolte da una banca portoghese che dal 2008 al 2013 ha effettuato chiamate in outbound a un campione del suo portafoglio per proporre la sottoscrizione di un ulteriore prodotto, un deposito a termine. Avendo a disposizione alcune informazioni sui clienti e sapendo quali hanno sottoscritto e quali no, l'obiettivo dell'analisi sarà quello di costruire un modello di regressione logistica che sia in grado di discriminare i clienti, tra quelli mai chiamati, che sottoscriveranno il prodotto se contattati da quelli che non lo soottoscriveranno. Una tale analisi predittiva avrebbe enormi benefici in termini di efficacia dell'attività di vendita: senza il supporto di modelli infatti ogni n clienti contattati si avrà una penetrazione del prodotto identica, che sarà molto prossimo a quello del campione analizzato (11,7% circa); con un modello a disposizione invece si potrà individuare un segmento di clienti entro la quale la **penetrazione** sarà maggiore, e si potranno collocare più prodotti a parità di chiamate.

Metodologia

Strumenti

Questa analisi viene svolta utilizzando il linguaggio di analisi statistica R, tramite il modulo di literate statical programming **R markdown**, che permette il rispetto dei principi della ricerca riproducibile. Per ogni risultato dell'analisi viene riportato il codice che lo ha generato. Di seguito i package utilizzati per l'analisi:

```
require(gains)
require(PRROC)
require(broom)
require(ResourceSelection)
require(MKmisc)
require(perturb)
require(caret)
require(DAAG)
library(pROC)
library(ROCR)
require(MASS)
library(devtools)
require(GGally)
library(woe)
require(ggplot2)
require(dplyr)
require(tidyr)
require(knitr)
require(e1071)
select <- dplyr::select
```

Approccio statistico

I modelli che verranno testati rientrano tutti nella famiglia della regressione logistica, di cui si è trattato nel capitolo primo. L'approccio si concretizzerà in una prima fase di **exploratory data analysis**, soprattutto in formato grafico, seguita da un confronto tra diverse ipotesi di modello guidate dalla prima fase. Per ogni variabile verrà eseguita una analisi univariata e una della sua relazione con la variabile target `y`, inoltre in alcuni casi verrà approfondita la relazione trivariata tra `y` e due variabili indipendenti.

Analisi esplorativa

Caricamento del dataset

```
getwd()
setwd("/Users/Americo/Documents/Education/Unitelma/tesi/data_analysis/dataset")
bank0_df <- read.csv(file = "bank_full.csv", sep = ";")
bank0 <- tbl_df(read.csv(file = "bank_full.csv", sep = ";"))
bank0_sm <- read.csv(file = "bank.csv", sep = ";")
```

`bank0` è il nome dato al dataset che utilizzeremo.

Informazioni sul dataset

```
dim(bank0)
```

```
## [1] 45211    17
```

Il dataset è composto da circa 45000 osservazioni e da 17 variabili. Vediamo meglio quali sono le variabili:

```
str(bank0)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    45211 obs. of  17 variables:
##  $ age       : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job       : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 .
..
##  $ marital   : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
##  $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
##  $ default   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ balance   : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing   : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
##  $ loan      : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
##  $ contact   : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ day       : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month     : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays     : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ y         : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Al link sopra riportato vi è la descrizione delle 17 variabili presenti nel dataset, che noi approfondiremo singolarmente nella nostra analisi. Per ora basti la seguente sintesi:

Variabili legate al profilo del cliente

- age: età del cliente
- job: professione svolta dal cliente
- marital : stato coniugale del cliente
- education: titolo di studio del cliente
- default: presenza di crediti in default
- balance: saldo medio annuale del conto
- housing: presenza di mutuo per la casa
- loan: presenza di prestiti

Variabili legate all'ultimo contatto dell'attuale campagna di marketing

- contact: modalità di comunicazione per l'ultimo contatto avvenuto
- day: giorno del mese dell'ultimo contatto avvenuto
- month: mese dell'ultimo contatto avvenuto
- duration: durata (in secondi) dell'ultimo contatto avvenuto

Variabili legate all'attuale o a precedente campagna di marketing

- campaign: totale di contatti avvenuti durante l'attuale campagna di marketing per ogni cliente
- pdays: numero di giorni trascorsi prima che il cliente fosse contattato per questa campagna dopo la fine della campagna precedente
- previous: numerodi contatti avvenuti prima di questa campagna
- poutcome: esito della precedente campagna di marketing

Variabile target: sottoscrizione o meno del deposito

Verifichiamo che non ci siano valori mancanti

```
bank0[!complete.cases(bank0),]
```

```
## Source: local data frame [0 x 17]
##
## Variables not shown: age (int), job (fctr), marital (fctr), education
##   (fctr), default (fctr), balance (int), housing (fctr), loan (fctr),
##   contact (fctr), day (int), month (fctr), duration (int), campaign (int),
##   pdays (int), previous (int), poutcome (fctr), y (fctr)
```

Nessuna riga contiene valori mancanti.

Esplorazione variabili: age

Da indicazioni del dizionario dati, `age` è una variabile numerica che misura l'età del cliente, anche se non sappiamo in quale momento (assumiamo quello attuale di processamento del modello).

```
class(bank0$age)
```

```
## [1] "integer"
```

```
summary(bank0$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   33.00   39.00   40.94   48.00   95.00
```

Il range è dai 18 ai 95, vediamo in forma tabellare e grafica la distribuzione delle varie età.

```
t_age <- bank0 %>%
        group_by(age) %>%
        summarise(frequenza = n()) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza))
kable(t_age, digits = 4, format = "markdown")
```

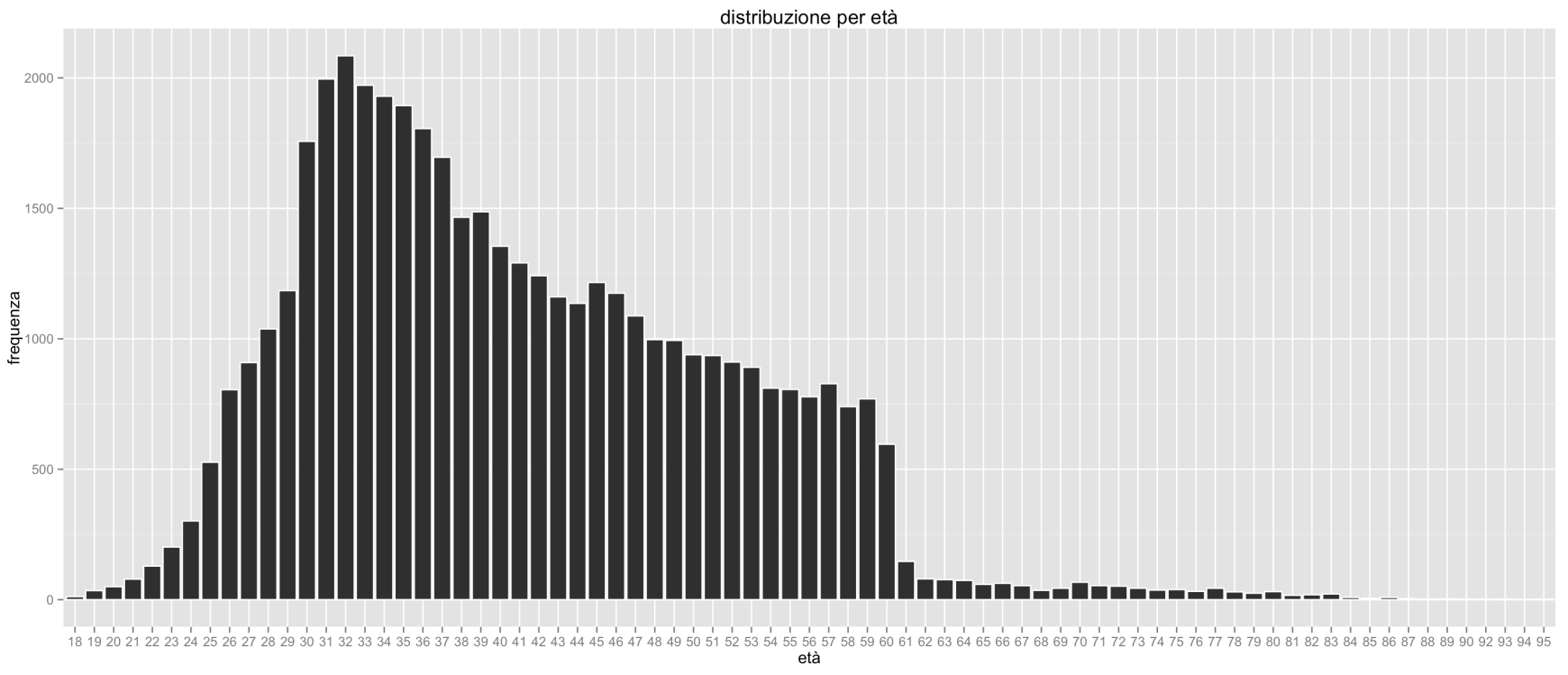
age	frequenza	frequenza_relativa
18	12	0.0003
19	35	0.0008

20	50	0.0011
21	79	0.0017
22	129	0.0029
23	202	0.0045
24	302	0.0067
25	527	0.0117
26	805	0.0178
27	909	0.0201
28	1038	0.0230
29	1185	0.0262
30	1757	0.0389
31	1996	0.0441
32	2085	0.0461
33	1972	0.0436
34	1930	0.0427
35	1894	0.0419
36	1806	0.0399
37	1696	0.0375
38	1466	0.0324
39	1487	0.0329
40	1355	0.0300
41	1291	0.0286
42	1242	0.0275
43	1161	0.0257
44	1136	0.0251
45	1216	0.0269
46	1175	0.0260
47	1088	0.0241
48	997	0.0221
49	994	0.0220
50	939	0.0208
51	936	0.0207
52	911	0.0201

53	891	0.0197
54	811	0.0179
55	806	0.0178
56	778	0.0172
57	828	0.0183
58	740	0.0164
59	770	0.0170
60	596	0.0132
61	147	0.0033
62	80	0.0018
63	77	0.0017
64	74	0.0016
65	59	0.0013
66	63	0.0014
67	54	0.0012
68	36	0.0008
69	44	0.0010
70	67	0.0015
71	54	0.0012
72	52	0.0012
73	44	0.0010
74	37	0.0008
75	39	0.0009
76	32	0.0007
77	44	0.0010
78	30	0.0007
79	25	0.0006
80	31	0.0007
81	17	0.0004
82	19	0.0004
83	22	0.0005
84	9	0.0002
85	5	0.0001
86	9	0.0002

87	4	0.0001
88	2	0.0000
89	3	0.0001
90	2	0.0000
92	2	0.0000
93	2	0.0000
94	1	0.0000
95	2	0.0000

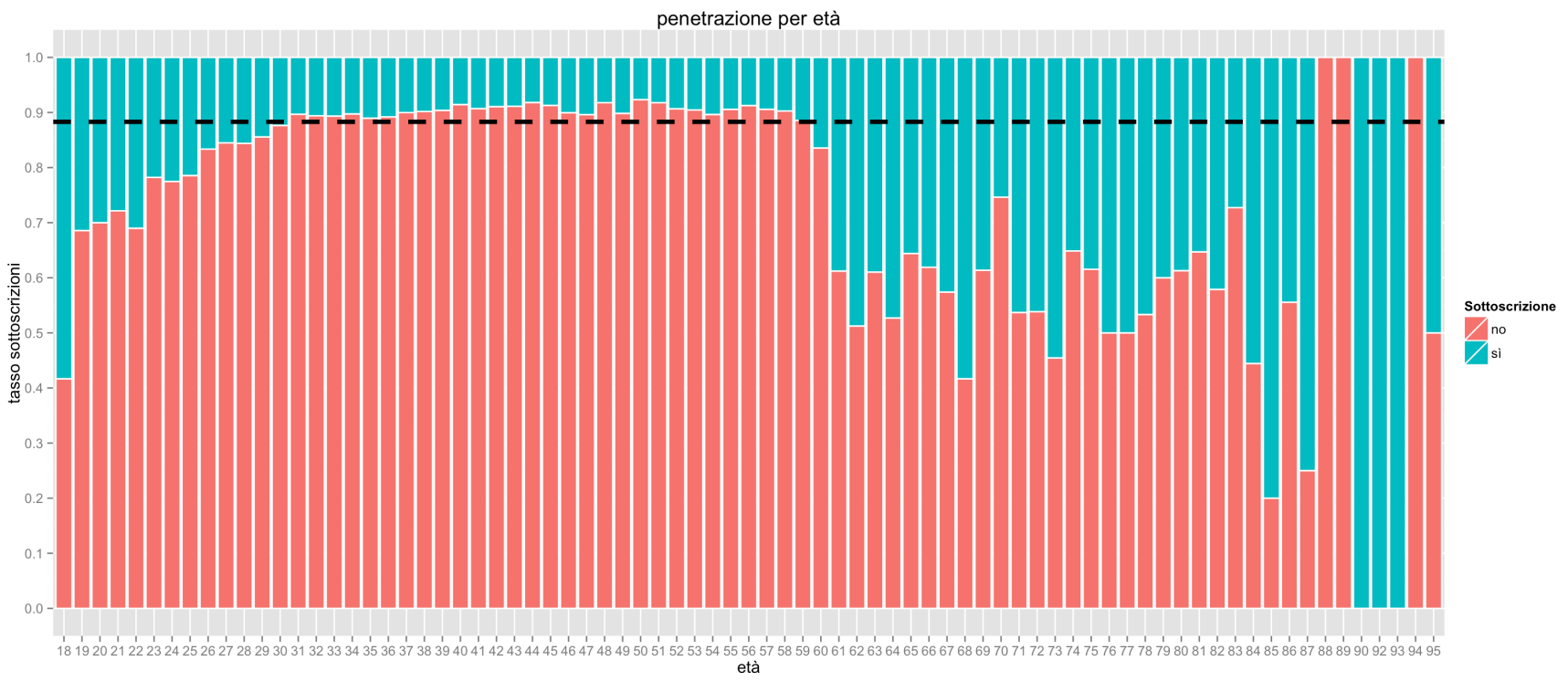
```
g_age <- ggplot(bank0, aes(x = factor(age))) + geom_bar(col = "white") +
        ggtitle("distribuzione per età") +
        xlab("età") +
        ylab("frequenza")
g_age
```



Una distribuzione quasi normale, con maggioritaria presenza di clienti dai 30 ai 60 anni, pochissimi diciotenni e ultra-ottantenni.

Ora vediamo come la penetrazione del prodotto si distribuisce all'interno delle varie età (in questo e in tutti gli altri grafici la linea orizzontale trattegiata rappresenta la percentuale di sottoscrizioni complessiva del campione, 0.1169848.

```
g_age_y <- ggplot(bank0, aes(x = factor(age), fill = y)) +
        geom_bar(col = "white", position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), size = 1.5, col = "black", linetyp
e = 2) +
        ggtitle("penetrazione per età") +
        xlab("età") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì"))
g_age_y
```




```
t_age_y <- bank0 %>%
        group_by (age) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y == "yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(age, frequenza, frequenza_relativa, tasso_sottoscrizioni)
kable(t_age_y, digits = 4, format = "markdown")
```

age	frequenza	frequenza_relativa	tasso_sottoscrizioni
18	12	0.0003	0.5833
19	35	0.0008	0.3143
20	50	0.0011	0.3000
21	79	0.0017	0.2785
22	129	0.0029	0.3101
23	202	0.0045	0.2178
24	302	0.0067	0.2252
25	527	0.0117	0.2144
26	805	0.0178	0.1665

27	909	0.0201	0.1551
28	1038	0.0230	0.1561
29	1185	0.0262	0.1443
30	1757	0.0389	0.1235
31	1996	0.0441	0.1032
32	2085	0.0461	0.1060
33	1972	0.0436	0.1065
34	1930	0.0427	0.1026
35	1894	0.0419	0.1103
36	1806	0.0399	0.1080
37	1696	0.0375	0.1002
38	1466	0.0324	0.0982
39	1487	0.0329	0.0962
40	1355	0.0300	0.0856
41	1291	0.0286	0.0930
42	1242	0.0275	0.0894
43	1161	0.0257	0.0887
44	1136	0.0251	0.0819
45	1216	0.0269	0.0872
46	1175	0.0260	0.1004
47	1088	0.0241	0.1039
48	997	0.0221	0.0822
49	994	0.0220	0.1016
50	939	0.0208	0.0767
51	936	0.0207	0.0823
52	911	0.0201	0.0933
53	891	0.0197	0.0954
54	811	0.0179	0.1036
55	806	0.0178	0.0943
56	778	0.0172	0.0874
57	828	0.0183	0.0942
58	740	0.0164	0.0973
59	770	0.0170	0.1143
60	596	0.0132	0.1644

61	147	0.0033	0.3878
62	80	0.0018	0.4875
63	77	0.0017	0.3896
64	74	0.0016	0.4730
65	59	0.0013	0.3559
66	63	0.0014	0.3810
67	54	0.0012	0.4259
68	36	0.0008	0.5833
69	44	0.0010	0.3864
70	67	0.0015	0.2537
71	54	0.0012	0.4630
72	52	0.0012	0.4615
73	44	0.0010	0.5455
74	37	0.0008	0.3514
75	39	0.0009	0.3846
76	32	0.0007	0.5000
77	44	0.0010	0.5000
78	30	0.0007	0.4667
79	25	0.0006	0.4000
80	31	0.0007	0.3871
81	17	0.0004	0.3529
82	19	0.0004	0.4211
83	22	0.0005	0.2727
84	9	0.0002	0.5556
85	5	0.0001	0.8000
86	9	0.0002	0.4444
87	4	0.0001	0.7500
88	2	0.0000	0.0000
89	3	0.0001	0.0000
90	2	0.0000	1.0000
92	2	0.0000	1.0000
93	2	0.0000	1.0000
94	1	0.0000	0.0000

95	2	0.0000	0.5000

Vediamo che dai 30 ai 59 anni l'incidenza di sottoscrizioni è prossima a quella complessiva, e rappresenta la parte più numerosa del campione. La penetrazione del prodotto comincia a essere superiore a quella complessiva, di portafoglio diciamo, per fasce di età molto giovani o ultra-sessantenni, dove però la numerosità (e quindi la significatività) è inferiore.

Potrebbe essere interessante raggruppare le età per fasce dal comportamento simile: [18, 30], [31, 59], [60, 95]

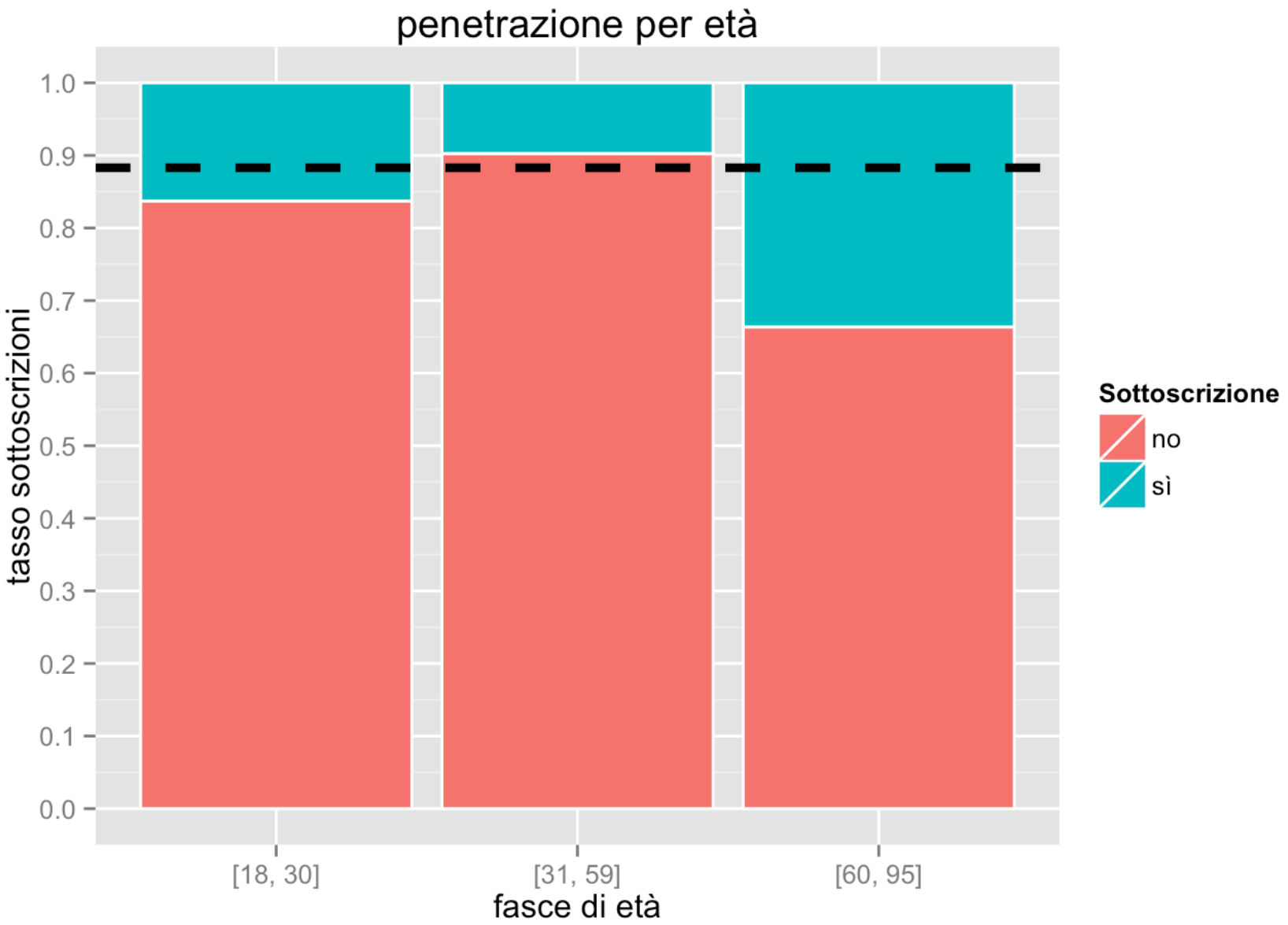
```
bank0$age <- as.numeric(bank0$age)
bank0$age_class <- cut(bank0$age, breaks = c(min(bank0$age)-1, 30, 59, max(bank0$age)),
labels = c("[18, 30]", "[31, 59]", "[60, 95]"))
summary(bank0$age_class)
```

```
## [18, 30] [31, 59] [60, 95]
##     7030    36397     1784
```

```
t_age_class <- bank0 %>%
        select(age_class, y) %>%
        group_by(age_class) %>%
        summarise(frequenza = n(), tasso_sottoscrizioni = mean(y =="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(age_class, frequenza, frequenza_relativa, tasso_sottoscrizioni)
kable(t_age_class, digits = 4, format = "markdown")
```

age_class	frequenza	frequenza_relativa	tasso_sottoscrizioni
[18, 30]	7030	0.1555	0.1629
[31, 59]	36397	0.8050	0.0974
[60, 95]	1784	0.0395	0.3363

```
g_age_class_y <- ggplot(bank0, aes(x = age_class, fill = y)) +
        geom_bar(col = "white", position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), size = 1.5, col = "black", linetyp
e = 2) +
        ggtitle("penetrazione per età") +
        xlab("fasce di età") +
        ylab("tasso sottoscrizioni") +
        scale_x_discrete(labels = c("[18, 30]", "[31, 59]", "[60, 95]")) +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì"))
g_age_class_y
```



è evidente che giovani e anziani sottoscrivono più del campione nel complesso, che nel comportamento è guidato dalla fascia media, gli adulti. Inoltre l'alta penetrazione della fascia anziana va pesato per la scarsa rilevanza numerica nel campione, il 4% circa, ma comunque di dimensione significativa.

Ha senso questa distribuzione? Beh, ci si può aspettare che un deposito a lungo termine venga sottoscritto di più da giovani che vogliono investire i loro risparmi (magari regali dei loro cari), mentre la fascia adulta avendo già molte spese (mantenimento della famiglia, mutui, presiti - in questi ultimi due casi potremo verificare l'ipotesi di una qualche interazione) abbia meno possibilità di farlo. Quanto agli anziani, per saperlo dovremmo conoscere meglio le caratteristiche del prodotto, magari lo hanno sottoscritto ma per farne beneficiare gli eredi.

Come per ogni variabile che analizzeremo, l'analisi termina con il calcolo dell'**information value**, una misura del potere predittivo della variabile per la cui trattazione teorica rimandiamo al capitolo secondo.

```
age_class_woe <- bank0 %>%
        select(age_class, y) %>%
        group_by(age_class) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
age_class_woe$woe <- log(age_class_woe$perc_no / age_class_woe$perc_y)
age_class_IV <- sum((age_class_woe$perc_no - age_class_woe$perc_y) * age_class_woe$woe)
age_class_IV
```

```
## [1] 0.1703601
```

0.1703601 , confrontato con la griglia del capitolo secondo, ci fa dire che l'età, raggruppata in tre classi, è una variabile mediamente predittiva della sottoscrizione.

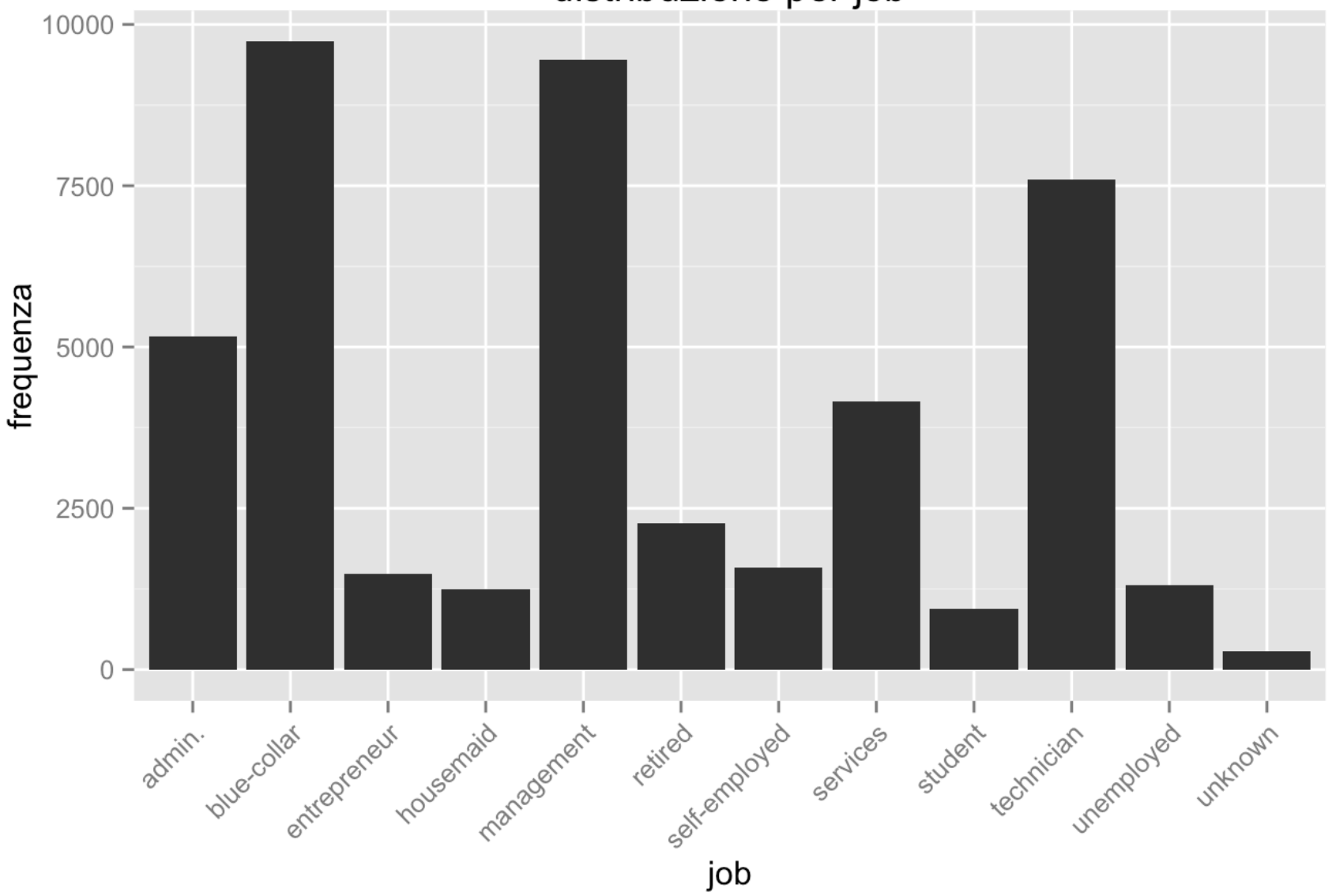
Job

La professione svolta dal cliente. Iniziamo con tabella e grafico della distribuzione:

```
t_job <- bank0 %>%
        group_by (job) %>%
        summarise (frequenza = n()) %>%
        mutate (frequenza_relativa = frequenza / sum(frequenza)) %>%
        arrange(desc(frequenza_relativa))
kable(t_job, digits = 4, format = "markdown")
```

job	frequenza	frequenza_relativa
blue-collar	9732	0.2153
management	9458	0.2092
technician	7597	0.1680
admin.	5171	0.1144
services	4154	0.0919
retired	2264	0.0501
self-employed	1579	0.0349
entrepreneur	1487	0.0329
unemployed	1303	0.0288
housemaid	1240	0.0274
student	938	0.0207
unknown	288	0.0064

```
g_job  <- ggplot(bank0, aes(x = job)) +
        geom_bar() +
        ggtitle("distribuzione per job") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_job
```



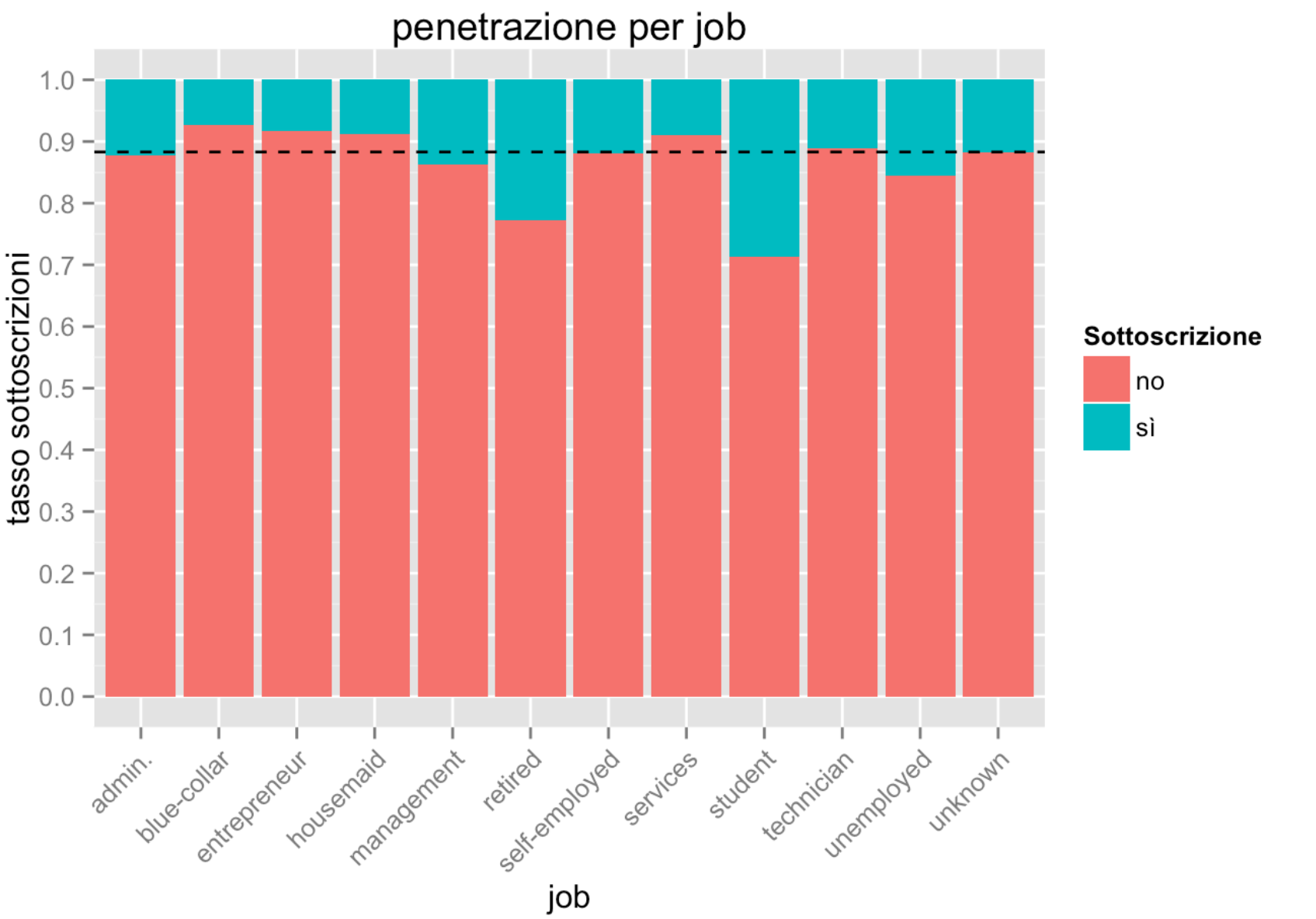
Vediamo adesso come, all'interno dei vari livelli della professione, si distribuiscono i sottoscrittori:

```
t_job_y <- bank0 %>%
        group_by (job) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(job, frequenza, frequenza_relativa, tasso_sottoscrizioni) %>%
        arrange(desc(tasso_sottoscrizioni))
kable(t_job_y, digits = 4, format = "markdown")
```

job	frequenza	frequenza_relativa	tasso_sottoscrizioni
student	938	0.0207	0.2868
retired	2264	0.0501	0.2279
unemployed	1303	0.0288	0.1550
management	9458	0.2092	0.1376
admin.	5171	0.1144	0.1220
self-employed	1579	0.0349	0.1184
unknown	288	0.0064	0.1181
technician	7597	0.1680	0.1106
services	4154	0.0919	0.0888

housemaid	1240	0.0274	0.0879
entrepreneur	1487	0.0329	0.0827
blue-collar	9732	0.2153	0.0727

```
g_job_y  <- ggplot(bank0, aes(x = job, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per job") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_job_y
```



Le professioni con il maggior tasso di sottoscrittori sono le meno frequenti, studenti e pensionati. Sarà interessante vedere - e lo faremo nella sezione delle analisi trivariate - se c'è un legame da età e professione tale per cui una delle due variabili può spiegare parzialmente la relazione con la sottoscrizione.

Non facile da interpretare invece il dato sui disoccupati: perché mai chi non ha reddito dovrebbe tendere più della media a sottoscrivere depositi a lungo termine? Forse vi rientrano giovani non categorizzati come studenti il cui deposito viene finanziato dai genitori. Anche qui l'analisi trivariata ci potrà svelare se

c’è un legame.

Da Ultimo calcoliamo l’information value della variabile `job`:

```
job_woe <- bank0 %>%
        select(job, y) %>%
        group_by(job) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
job_woe$woe <- log(job_woe$perc_no / job_woe$perc_y)
job_IV <- sum((job_woe$perc_no - job_woe$perc_y) * job_woe$woe)
job_IV
```

```
## [1] 0.1556973
```

Anche `job` ha un potere predittivo di media entità.

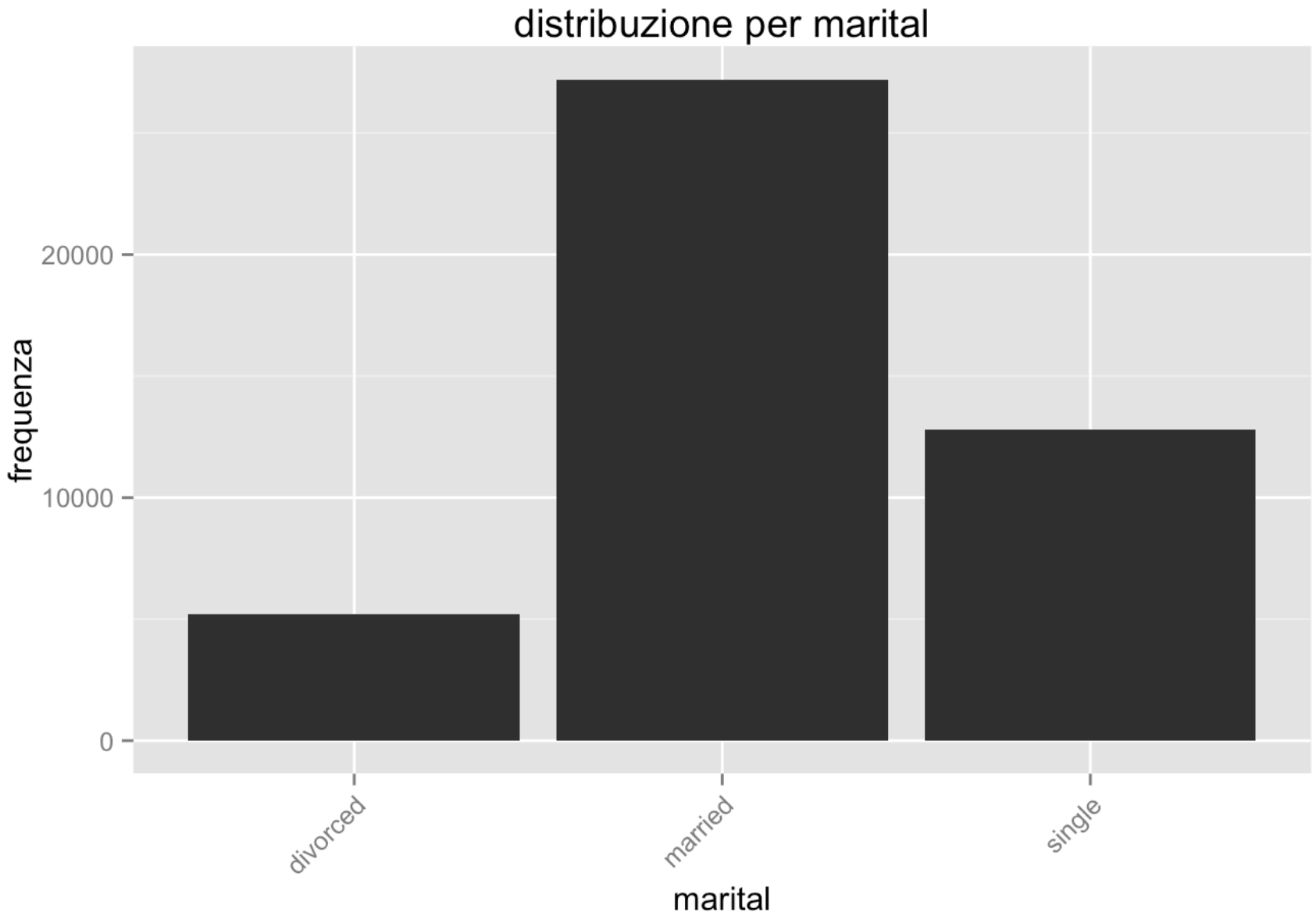
Marital

La variabile descrive lo stato coniugale. Dal dizionario dati sappiamo che il valore “divorced” include anche i vedovi.

```
t_marital <- bank0 %>%
        group_by(marital) %>%
        summarise(frequenza = n()) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        arrange(desc(frequenza))
kable(t_marital, digits = 4, format = "markdown")
```

marital	frequenza	frequenza_relativa
married	27214	0.6019
single	12790	0.2829
divorced	5207	0.1152

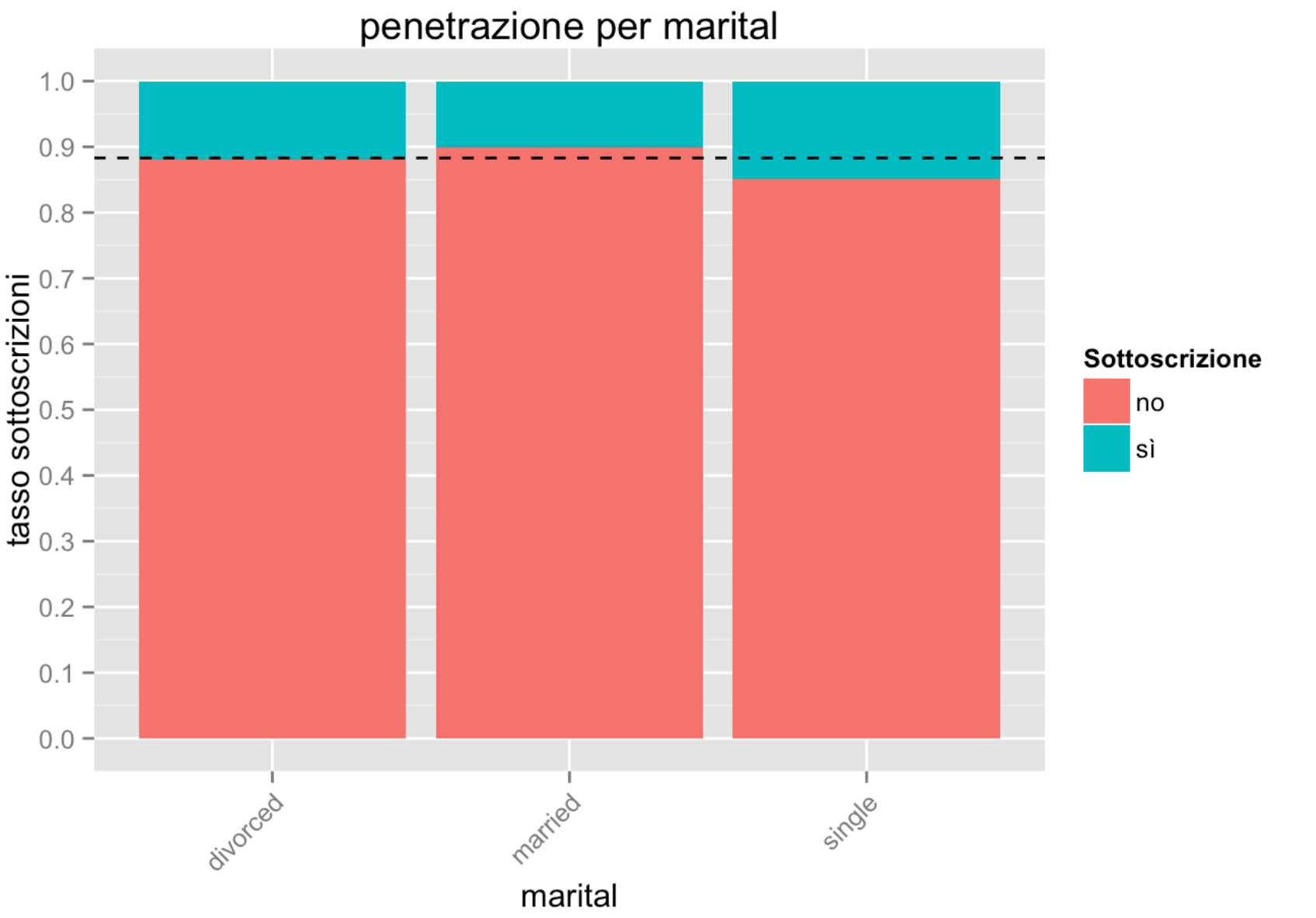
```
g_marital  <- ggplot(bank0, aes(x = marital)) +
        geom_bar() +
        ggtitle("distribuzione per marital") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_marital
```




```
t_marital_y <- bank0 %>%
        group_by (marital) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(marital, frequenza, frequenza_relativa, tasso_sottoscrizioni) %>%
        arrange(desc(tasso_sottoscrizioni))
kable(t_marital_y, digits = 4, format = "markdown")
```

marital	frequenza	frequenza_relativa	tasso_sottoscrizioni
single	12790	0.2829	0.1495
divorced	5207	0.1152	0.1195
married	27214	0.6019	0.1012

```
g_marital_y  <- ggplot(bank0, aes(x = marital, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per marital") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_marital_y
```



I single tendono a sottoscrivere un po' di più. Anche qui andrà indagata la relazione con l'età (la maggior parte dei single sono giovani e quindi per questo i single sottoscrivono di più?) e la professione (specialmente studentesca).

L'information value di `marital`:

```
marital_woe <- bank0 %>%
        select(marital, y) %>%
        group_by(marital) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
marital_woe$woe <- log(marital_woe$perc_no / marital_woe$perc_y)
marital_IV <- sum((marital_woe$perc_no - marital_woe$perc_y) * marital_woe$woe)
marital_IV
```

```
## [1] 0.04012659
```

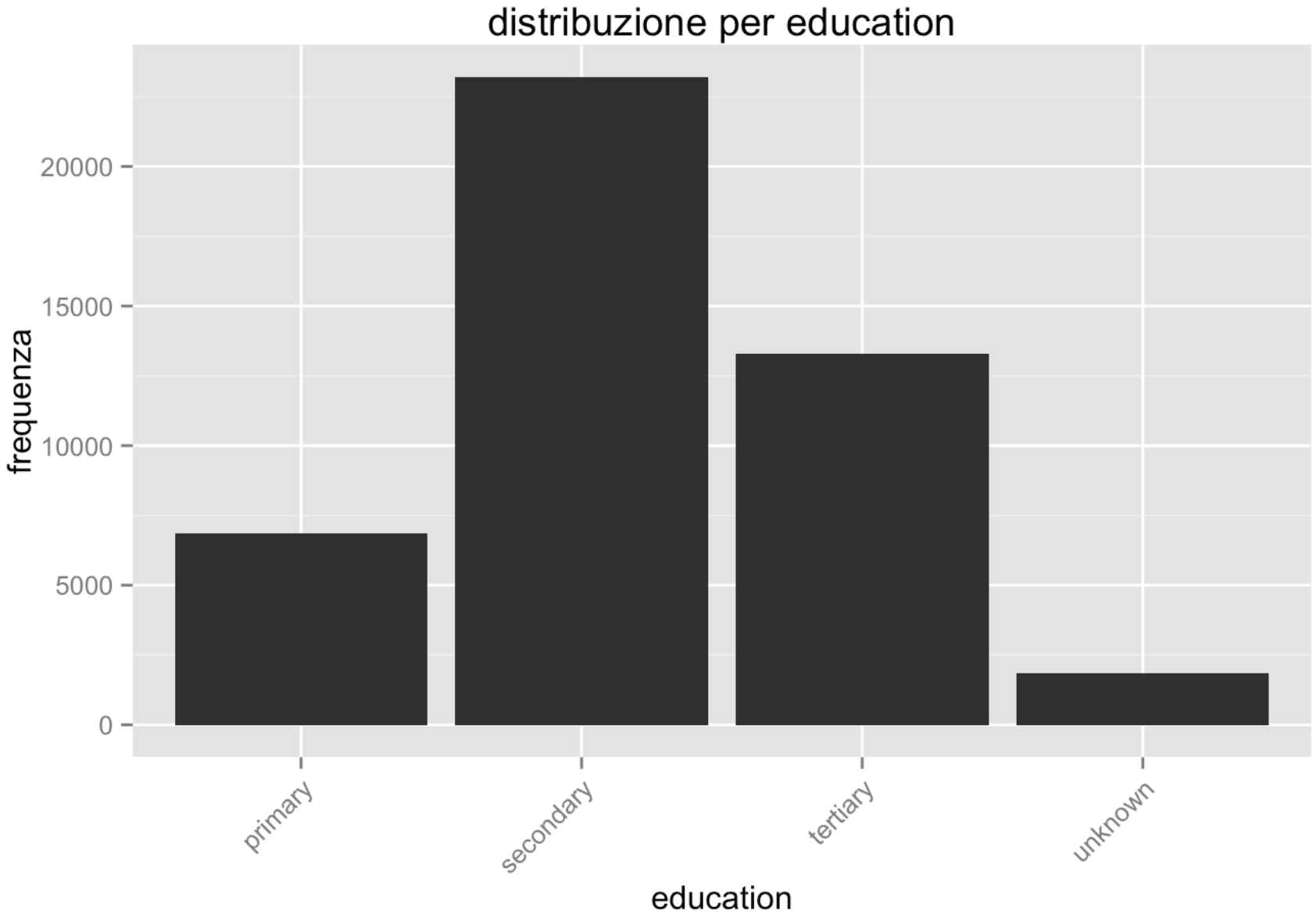
Ha una predittività molto debole.

Education

```
t_education  <- bank0 %>%
        group_by(education) %>%
        summarise(frequenza = n()) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        arrange(desc(frequenza))
kable(t_education, digits = 4, format = "markdown")
```

education	frequenza	frequenza_relativa
secondary	23202	0.5132
tertiary	13301	0.2942
primary	6851	0.1515
unknown	1857	0.0411

```
g_education  <- ggplot(bank0, aes(x = education)) +
        geom_bar() +
        ggtitle("distribuzione per education") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_education
```

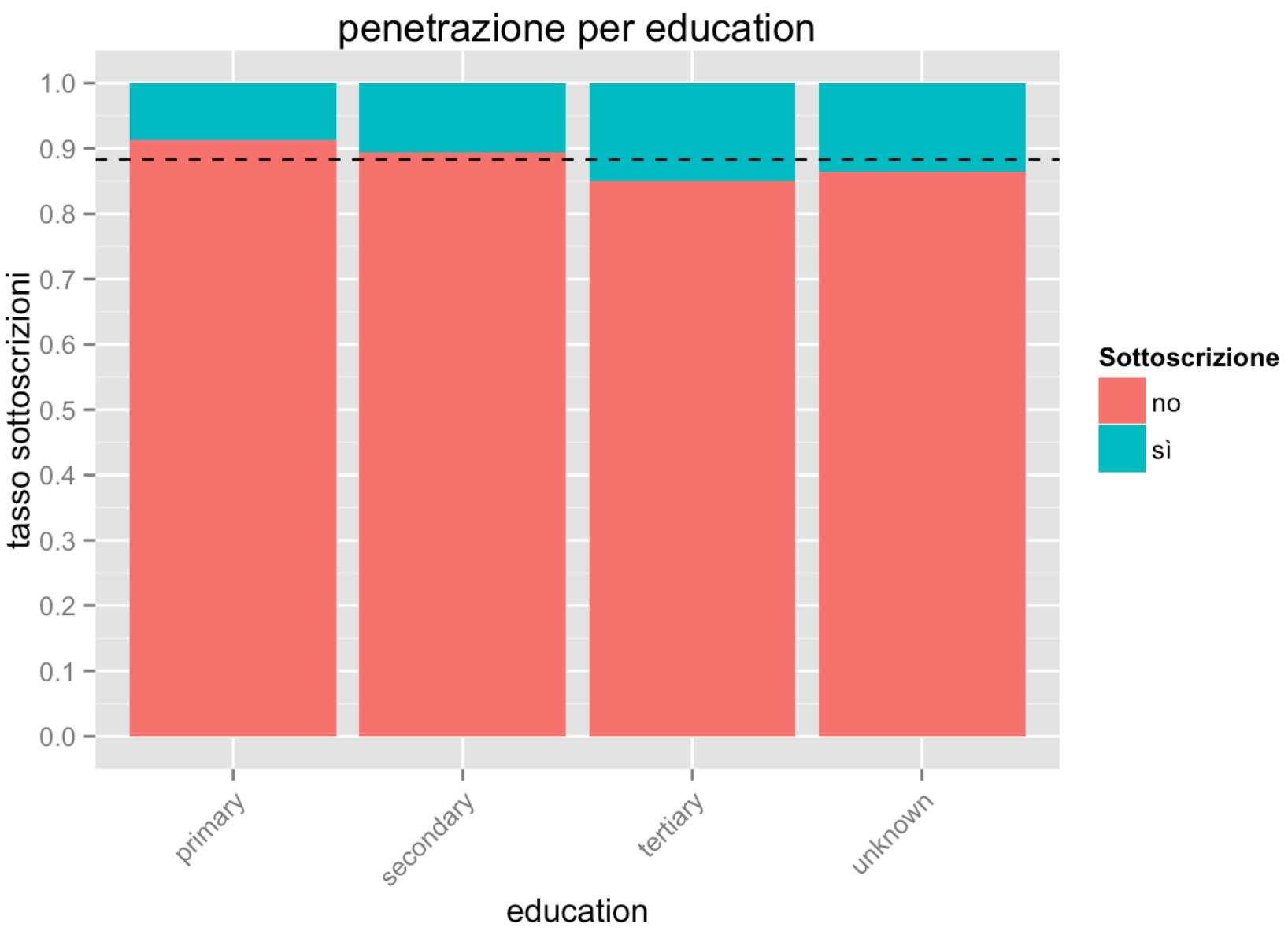


Analizziamo come la penetrazione si distribuisce tra i titoli di studio:

```
t_education_y <- bank0 %>%
        group_by (education) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(education, frequenza, frequenza_relativa, tasso_sottoscrizioni) %>%
        arrange(desc(tasso_sottoscrizioni))
kable(t_education_y, digits = 4, format = "markdown")
```

education	frequenza	frequenza_relativa	tasso_sottoscrizioni
tertiary	13301	0.2942	0.1501
unknown	1857	0.0411	0.1357
secondary	23202	0.5132	0.1056
primary	6851	0.1515	0.0863

```
g_education_y  <- ggplot(bank0, aes(x = education, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per education") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_education_y
```



Sembra che (ignorando il livello “unknown”) a maggior livello culturale segua maggiore propensione a sottoscrivere depositi, e questa appare come una informazione aggiuntiva rispetto a quanto abbiamo scoperto sinora. Anche qui dovremo capire se l’educazione terziaria è maggiormente presente negli under 30, che per condizioni socio-economiche mutevoli hanno studiato in proporzione di più degli adulti.

L’information value di `èducation`:

```
education_woe <- bank0 %>%
        select(education, y) %>%
        group_by(education) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
education_woe$woe <- log(education_woe$perc_no / education_woe$perc_y)
education_IV <- sum((education_woe$perc_no - education_woe$perc_y) * education_woe$woe)
education_IV
```

```
## [1] 0.05011195
```

Predittività debole ma non assente.

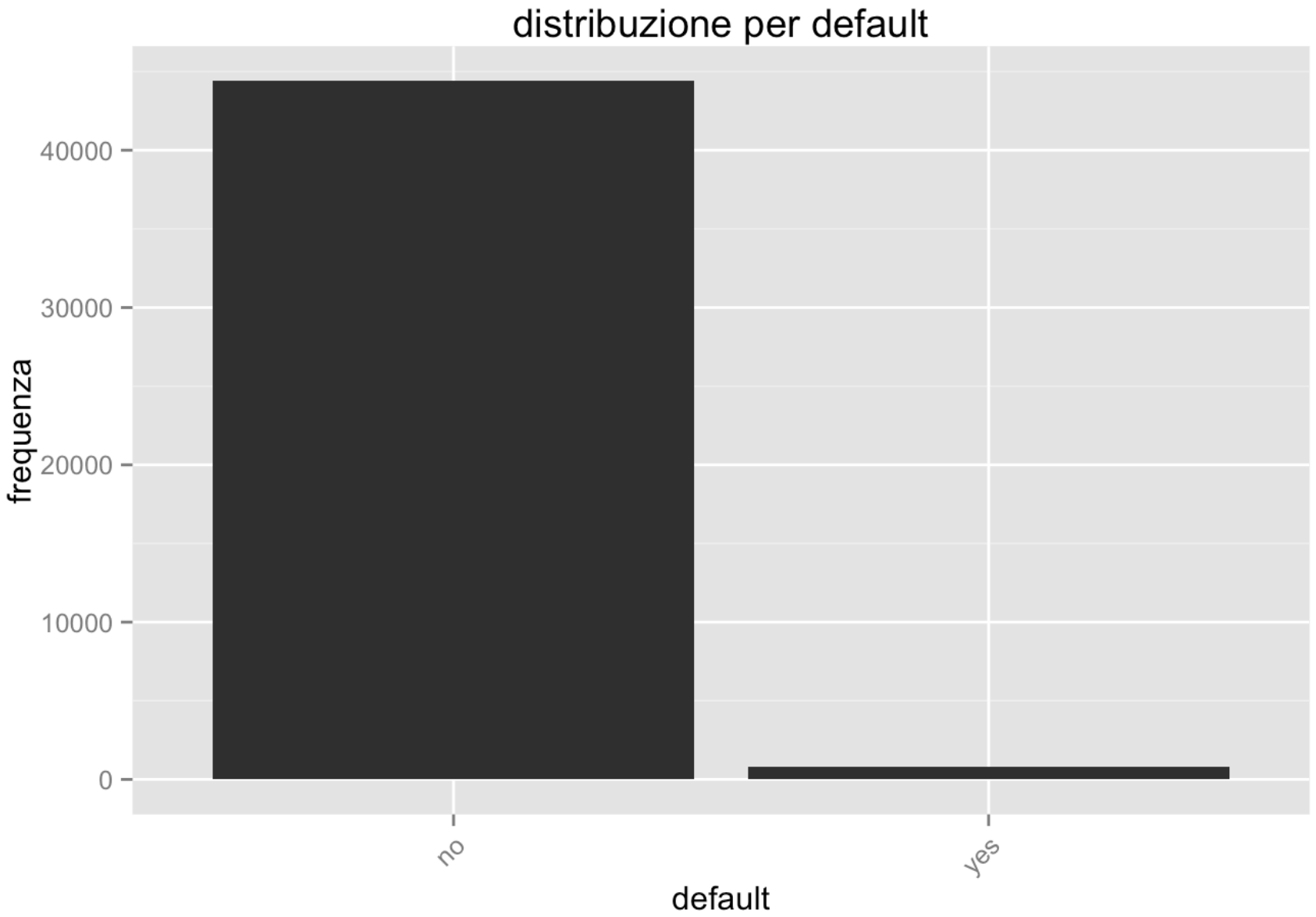
Default

Variabile dicotomica che quando assume valore 1 indica la presenza di crediti in defualt.

```
t_default  <- bank0 %>%
        group_by(default) %>%
        summarise(frequenza = n()) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        arrange(desc(frequenza))
kable(t_default, digits = 4, format = "markdown")
```

default	frequenza	frequenza_relativa
no	44396	0.982
yes	815	0.018

```
g_default  <- ggplot(bank0, aes(x = default)) +
        geom_bar() +
        ggtitle("distribuzione per default") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_default
```

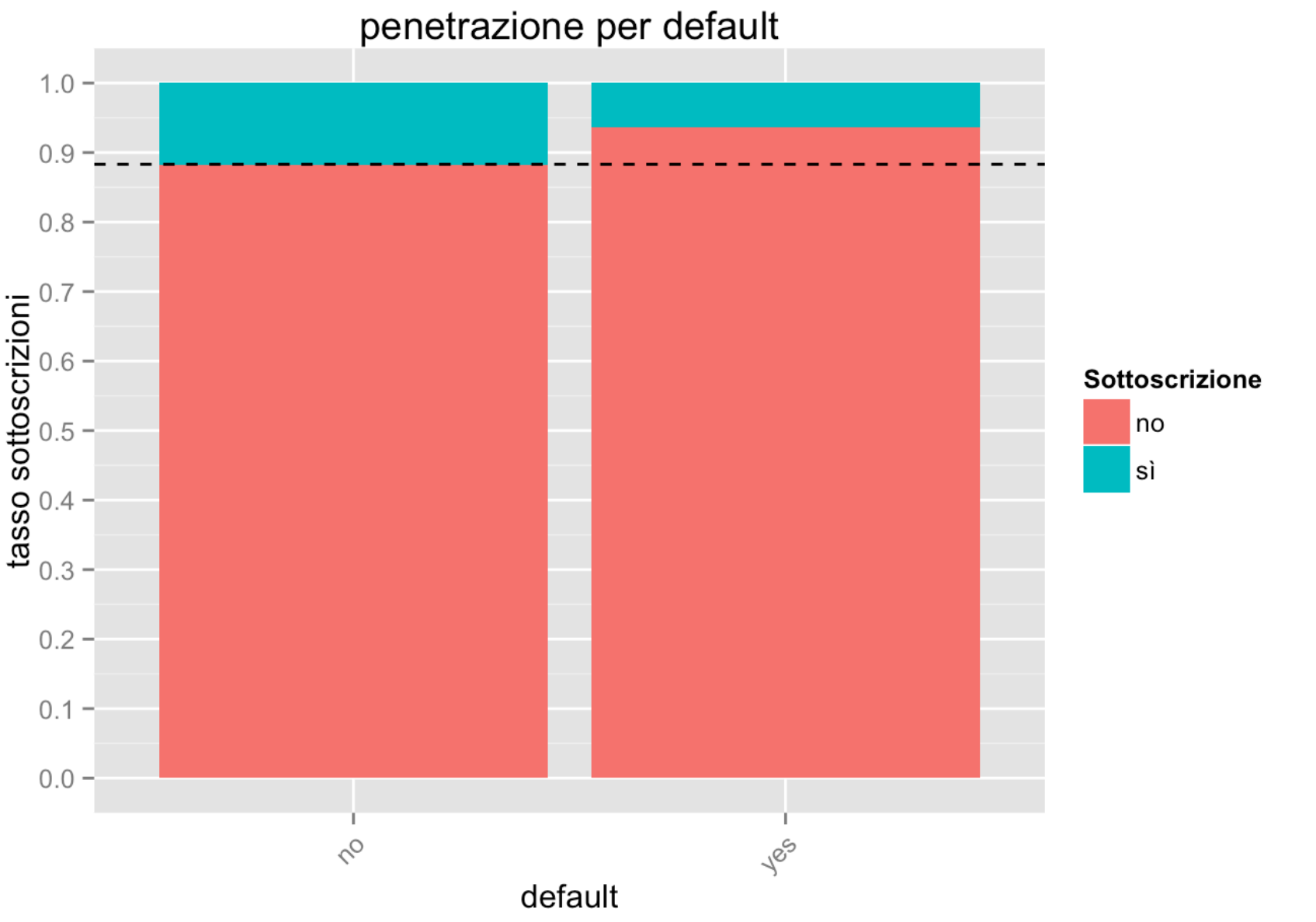


Pochissimi in default, forse troppo pochi per trovare significatività in questa variabile.

```
t_default_y <- bank0 %>%
        group_by (default) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(default, frequenza, frequenza_relativa, tasso_sottoscrizioni) %>%
        arrange(desc(tasso_sottoscrizioni))
kable(t_default_y, digits = 4, format = "markdown")
```

default	frequenza	frequenza_relativa	tasso_sottoscrizioni
no	44396	0.982	0.1180
yes	815	0.018	0.0638

```
g_default_y  <- ggplot(bank0, aes(x = default, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per default") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_default_y
```



I clienti non in default si comportano, in termini di sottoscrizioni, assolutamente nella media, ma d'altronde rappresentano pressoché tutto il campione. Il 0.0180266 che è in default sottoscrive (prevedibilmente) molto meno.

Passiamo all'information value

```
default_woe <- bank0 %>%
        select(default, y) %>%
        group_by(default) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
default_woe$woe <- log(default_woe$perc_no / default_woe$perc_y)
default_IV <- sum((default_woe$perc_no - default_woe$perc_y) * default_woe$woe)
default_IV
```

```
## [1] 0.006256319
```

Predittività assente.

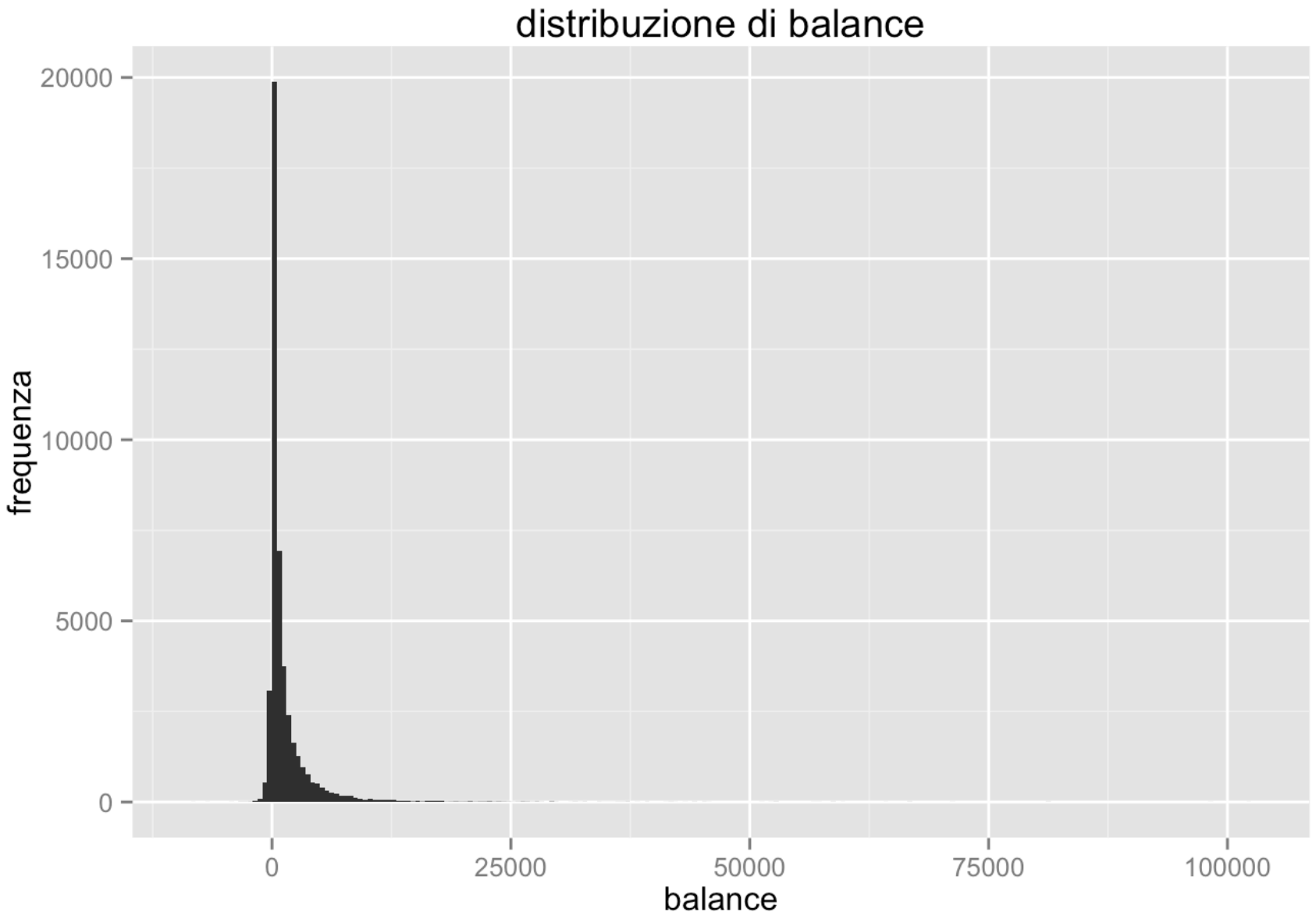
Balance

Il saldo medio sul conto corrente del cliente; in questo caso adottiamo un istogramma, data la natura quantitativa della variabile:

```
balance_quantile <- quantile(bank0$balance, c(seq(0.1, 1, 0.05)))
balance_quantile
```

```
##    10%    15%    20%    25%    30%    35%    40%    45%    50%    55%
##      0      0     22     72    131    198    272    352    448    563
##    60%    65%    70%    75%    80%    85%    90%    95%   100%
##    701    883   1126   1428   1859   2539   3574   5768 102127
```

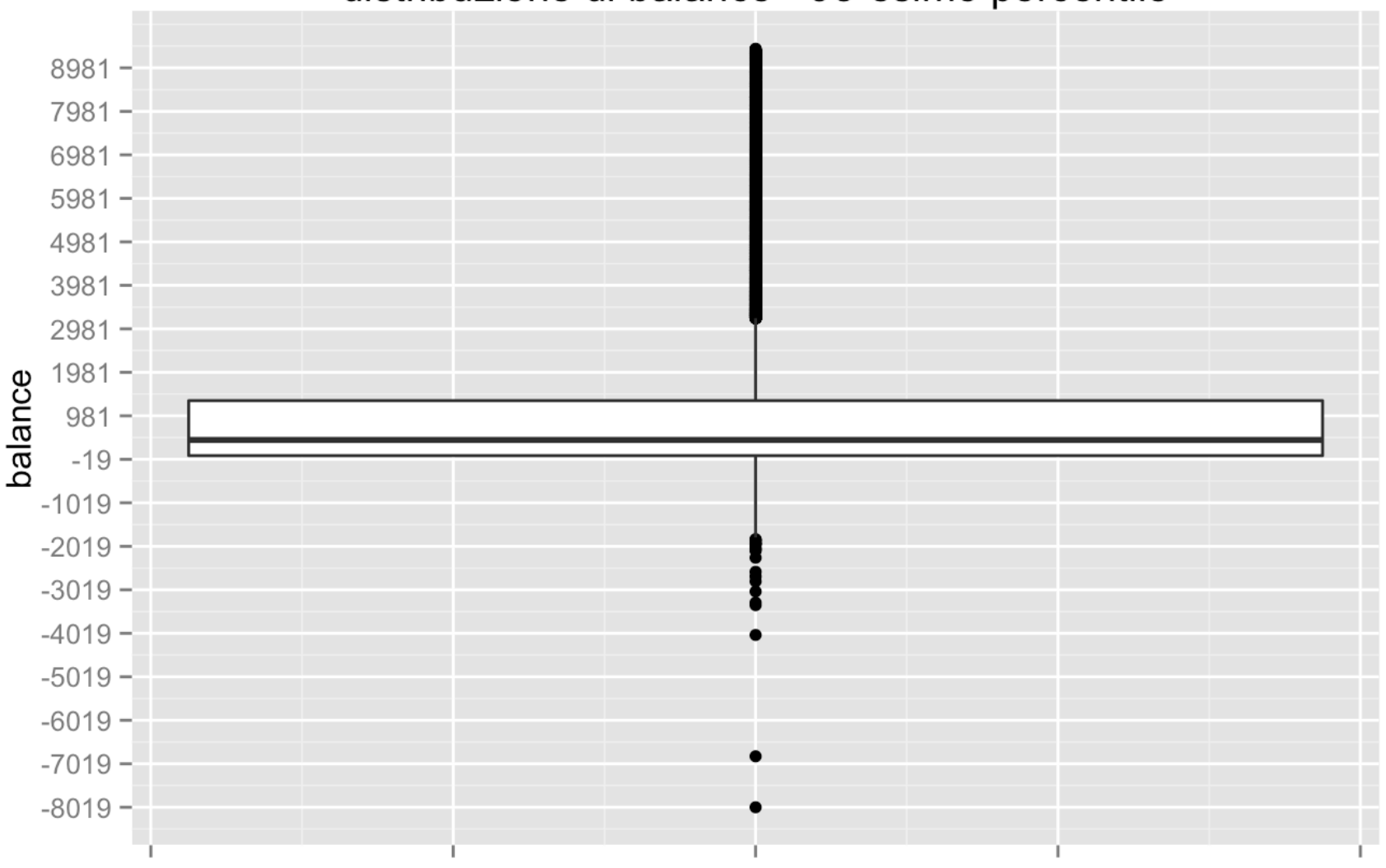
```
g_balance <- ggplot(bank0, aes(x = balance)) +
        geom_histogram(binwidth = 500) +
        ggtitle("distribuzione di balance") +
        ylab("frequenza")
g_balance
```



Come sia il grafico che la distribuzione dei percentili ci mostrano, `balance` presenta dei valori anomali che rendono l'intervallo dei valori incredibilmente ampio; per così dire, i pochi ricchi clienti della banca, mentre il 79% è tra 0 e 1000.

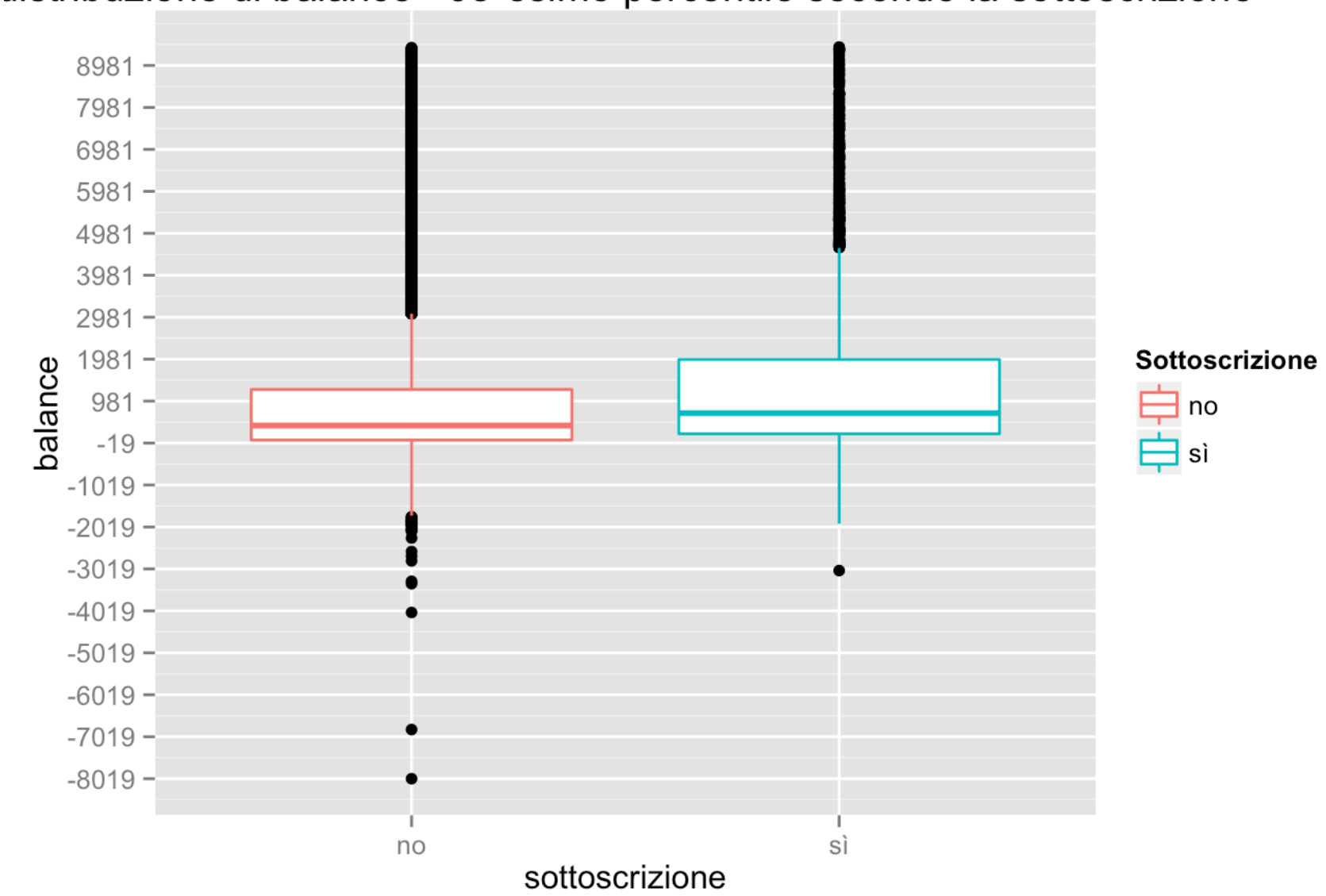
Vediamo che tipo di distribuzione prende forma se escludiamo il 2% dei valori più alti.

```
balance_p_98 <- quantile(bank0$balance, 0.98)
bank0_p98 <- bank0 %>%
        filter(balance <= balance_p_98)
g_balance_p98 <- ggplot(bank0_p98, aes(x = 1, y = balance)) +
        geom_boxplot() +
        ggtitle("distribuzione di balance - 98-esimo percentile") +
        xlab("") +
        scale_x_continuous(labels = c("", "", "", "", "")) +
        scale_y_continuous(breaks = seq(min(bank0$balance), balance_p_98, 1000))
g_balance_p98
```



Il 50% dei valori è tra 0 e 1000 euro, ora la distribuzione di `balance` è più chiara. Vediamo cosa succede condizionando la distribuzione alla sottoscrizione o meno

```
g_balance_y_p98 <- ggplot(bank0_p98, aes(x = y, y = balance)) +
        geom_boxplot(aes(col = y)) +
        ggtitle("distribuzione di balance - 98-esimo percentile secondo la sottoscrizion
e") +
        xlab("sottoscrizione") +
        scale_x_discrete(labels = c("no", "sì")) +
        scale_color_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        scale_y_continuous(breaks = seq(min(bank0$balance), balance_p_98, 1000))
g_balance_y_p98
```



Ora comincia a essere evidente che chi sottoscrive depositi tende a avere un saldo medio del conto più alto; vale per la mediana del gruppo dei sottoscrittori, ma anche per il terzo quartile. Quindi per saldi che crescono si tende a sottoscrivere più depositi, il che ha senso. Più soldi hai, più ne puoi depositare.

Ora, anche ai fini del calcolo dell'information value, raggruppiamo `balance` in classi e analizziamo la distribuzione dei sottoscrittori all'interno dei vari livelli.

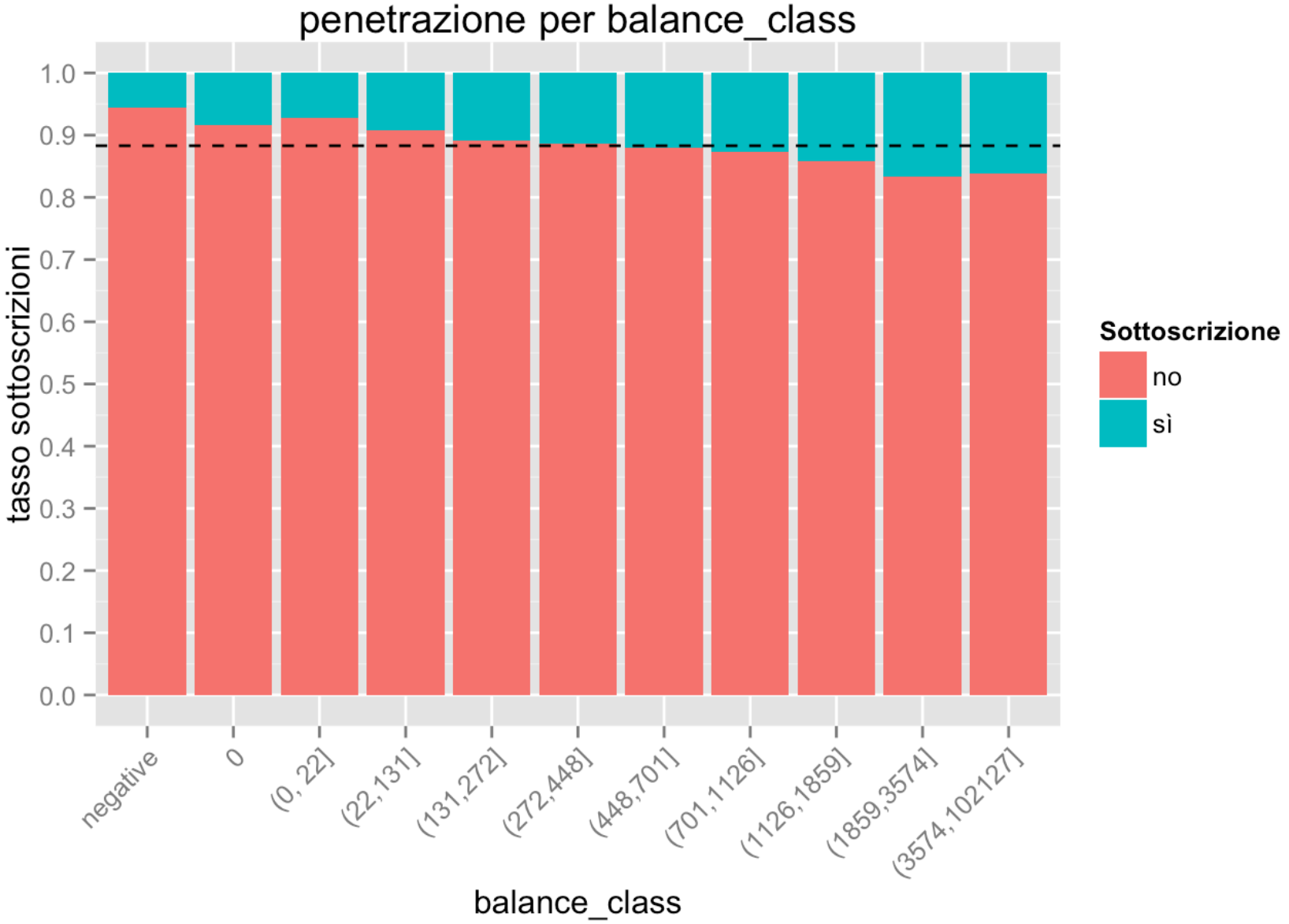
```
bank0$balance <- as.numeric(bank0$balance)
balance_decile <- quantile(bank0$balance, probs = seq(0.1,1,0.1))
bank0$balance_class <- cut(bank0$balance, breaks = c(min(bank0$balance)-1, -1, balance_d
ecile), right = TRUE, labels = c("negative", "0", "(0, 22]", "(22,131]", "(131,272]", "(
272,448]", "(448,701]", "(701,1126]", "(1126,1859]", "(1859,3574]", "(3574,102127]"))
summary(bank0$balance_class)
```

```
##      negative             0       (0, 22]      (22,131]     (131,272]
##          3766          3514          1773          4544          4516
##     (272,448]     (448,701]    (701,1126]   (1126,1859]   (1859,3574]
##          4495          4522          4526          4513          4521
## (3574,102127]
##          4521
```

```
t_balance_class_y <- bank0 %>%
        group_by (balance_class) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(balance_class, frequenza, frequenza_relativa, tasso_sottoscrizioni)
kable(t_balance_class_y, digits = 4, format = "markdown")
```

balance_class	frequenza	frequenza_relativa	tasso_sottoscrizioni
negative	3766	0.0833	0.0558
0	3514	0.0777	0.0831
(0, 22]	1773	0.0392	0.0722
(22,131]	4544	0.1005	0.0918
(131,272]	4516	0.0999	0.1083
(272,448]	4495	0.0994	0.1141
(448,701]	4522	0.1000	0.1201
(701,1126]	4526	0.1001	0.1268
(1126,1859]	4513	0.0998	0.1425
(1859,3574]	4521	0.1000	0.1661
(3574,102127]	4521	0.1000	0.1612

```
g_balance_class_y  <- ggplot(bank0, aes(x = balance_class, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per balance_class") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_balance_class_y
```



Evidente come spostandosi verso gruppi di saldo maggiore (gruppi di numerosità simile, ricordiamolo) cresce la probabilità di sottoscrivere un deposito a medio termine. E l'information value?

```
balance_class_woe <- bank0 %>%
        select(balance_class, y) %>%
        group_by(balance_class) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
balance_class_woe$woe <- log(balance_class_woe$perc_no / balance_class_woe$perc_y)
balance_class_IV <- sum((balance_class_woe$perc_no - balance_class_woe$perc_y) * balance
_class_woe$woe)
balance_class_IV
```

```
## [1] 0.1079715
```

Il valore di 0.1079715 rientra tra i predittori di media rilevanza.

Housing

La variabile di dice se il cliente ha o meno sottoscritto un mutuo ipotecario per l'acquisto di un immobile

```
t_housing  <- bank0 %>%
        group_by(housing) %>%
        summarise(frequenza = n()) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        arrange(desc(frequenza))
kable(t_housing, digits = 4, format = "markdown")
```

housing	frequenza	frequenza_relativa
yes	25130	0.5558
no	20081	0.4442

```
g_housing  <- ggplot(bank0, aes(x = housing)) +
        geom_bar() +
        ggtitle("distribuzione per housing") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_housing
```

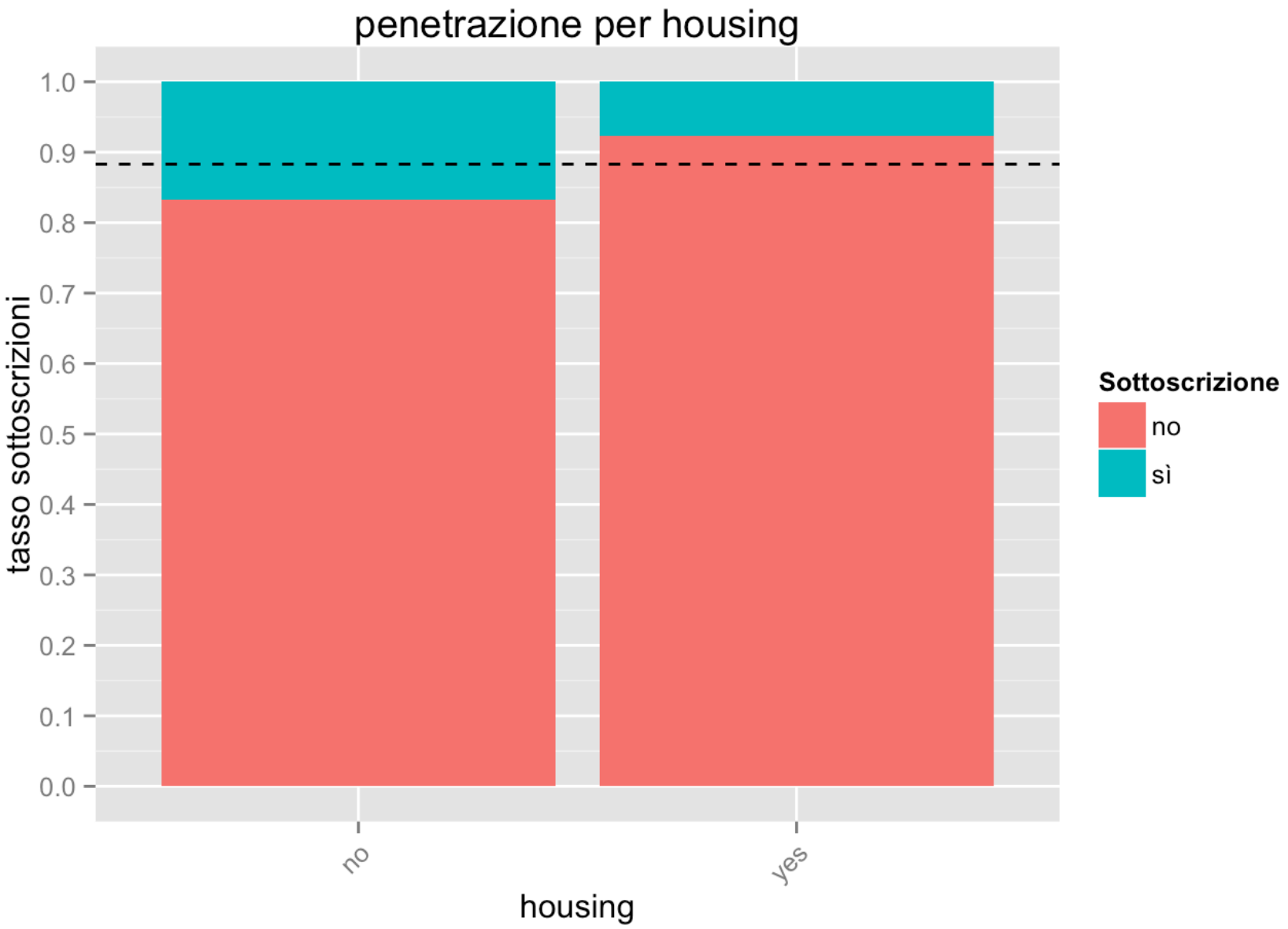
Vediamo ora il comportamento dei sottoscrittori.

Poco più della metà dei clienti ha il mutuo, l’altra no. Entrambe le classi sono estremanente significative, quindi.

```
t_housing_y <- bank0 %>%
        group_by (housing) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(housing, frequenza, frequenza_relativa, tasso_sottoscrizioni) %>%
        arrange(desc(tasso_sottoscrizioni))
kable(t_housing_y, digits = 4, format = "markdown")
```

housing	frequenza	frequenza_relativa	tasso_sottoscrizioni
no	20081	0.4442	0.167
yes	25130	0.5558	0.077

```
g_housing_y  <- ggplot(bank0, aes(x = housing, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per housing") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_housing_y
```



La distinzione tra mutuatari o meno se,bra abbastanza discriminante sulla sottoscrizione; un gruppo sottoscrive al 16,5%, l'altro al 7,5%. Ha una logica il fatto che chi ha dovuto accolarsi un mutuo non abbia risorse finanziarie da investire per rendimento nel lungo periodo, data l'esigenza pressante di liquidità. Alla luce di ciò l'information value dovrebbe presentarci una variabile con potere predittivo medio.

```
housing_woe <- bank0 %>%
        select(housing, y) %>%
        group_by(housing) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
housing_woe$woe <- log(housing_woe$perc_no / housing_woe$perc_y)
housing_IV <- sum((housing_woe$perc_no - housing_woe$perc_y) * housing_woe$woe)
housing_IV
```

```
## [1] 0.1886815
```

Esatto.

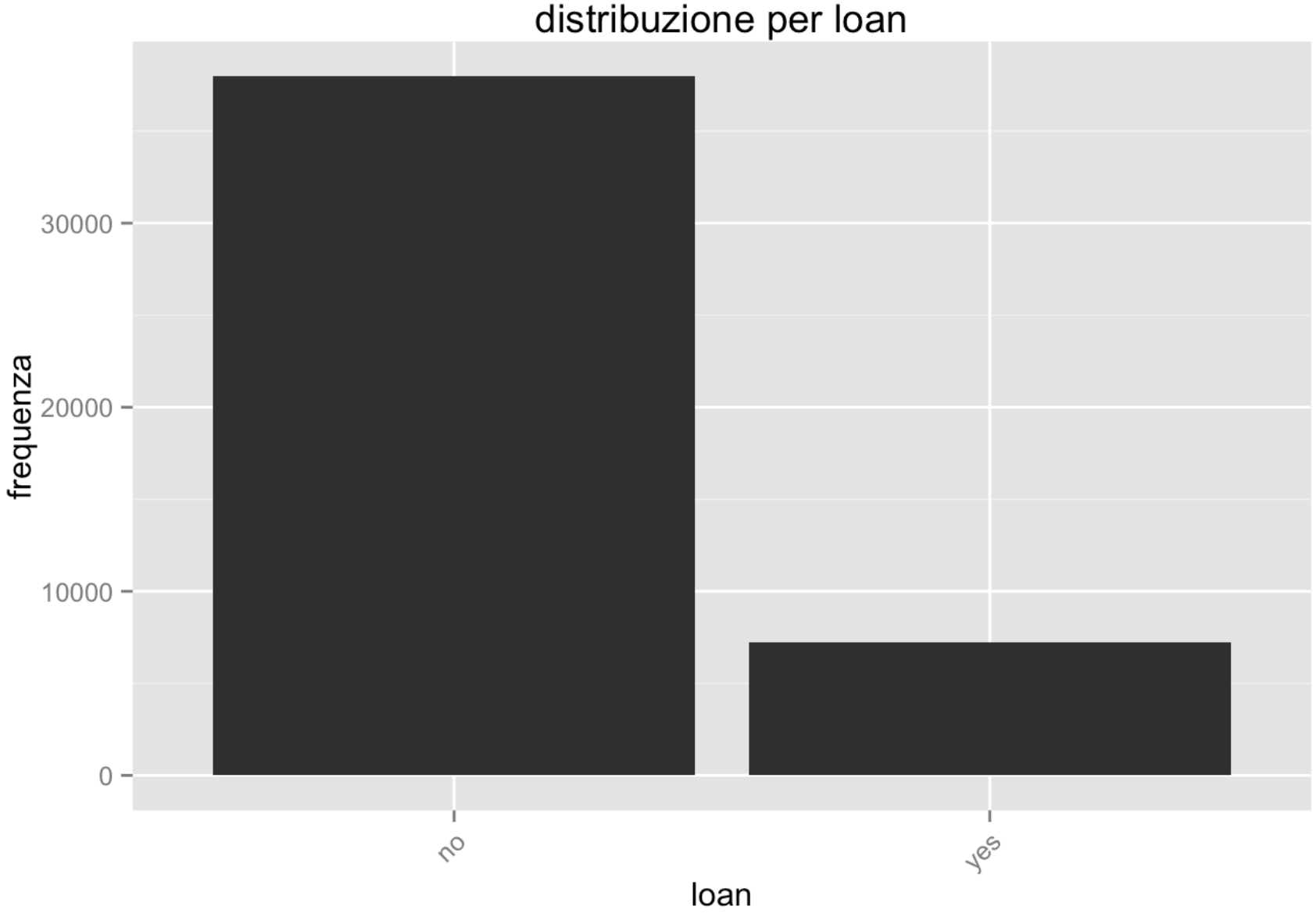
Loan

Variabile dicotomica che indica che il cliente ha richiesto e sta pagando un prestito.

```
t_loan  <- bank0 %>%
        group_by(loan) %>%
        summarise(frequenza = n()) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        arrange(desc(frequenza))
kable(t_loan, digits = 4, format = "markdown")
```

loan	frequenza	frequenza_relativa
no	37967	0.8398
yes	7244	0.1602

```
g_loan  <- ggplot(bank0, aes(x = loan)) +
        geom_bar() +
        ggtitle("distribuzione per loan") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_loan
```

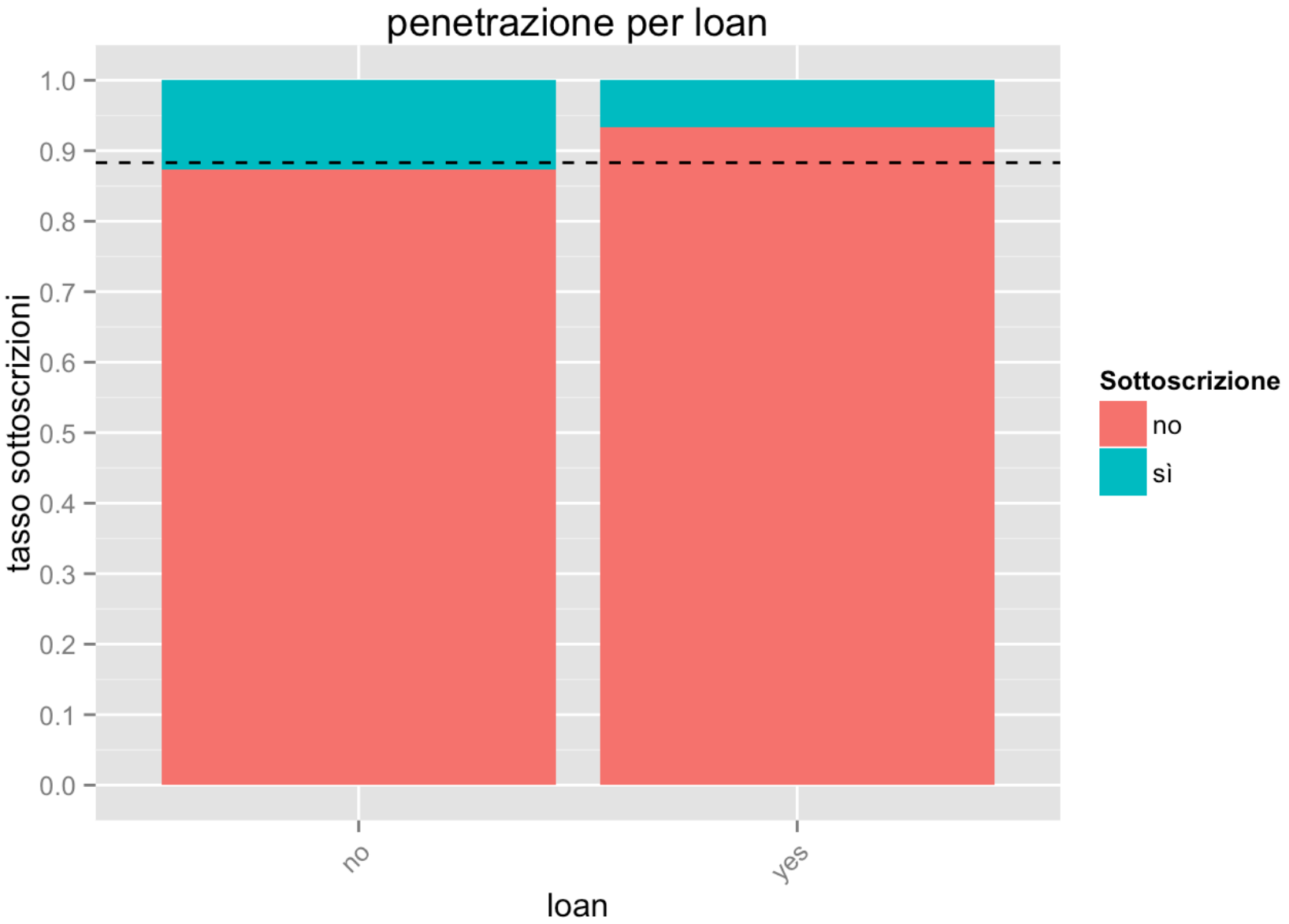


La quota di persone con prestiti è intorno al 16%. Vediamo come si distribuiscono i sottoscrittori

```
t_loan_y <- bank0 %>%
        group_by (loan) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(loan, frequenza, frequenza_relativa, tasso_sottoscrizioni) %>%
        arrange(desc(tasso_sottoscrizioni))
kable(t_loan_y, digits = 4, format = "markdown")
```

loan	frequenza	frequenza_relativa	tasso_sottoscrizioni
no	37967	0.8398	0.1266
yes	7244	0.1602	0.0668

```
g_loan_y  <- ggplot(bank0, aes(x = loan, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per loan") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_loan_y
```



Il 16% dei richiedenti prestito ha un tasso di sottoscrizione pari alla metà del restante 84%, che sottoscrive poco sopra la media di portafoglio. Ci si aspettava in effetti un comportamento simile a quello di `housing`, e sarà interessante vedere l'effetto di interazione in chi sta pagando sia mutuo che prestito.

```
loan_woe <- bank0 %>%
        select(loan, y) %>%
        group_by(loan) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
loan_woe$woe <- log(loan_woe$perc_no / loan_woe$perc_y)
loan_IV <- sum((loan_woe$perc_no - loan_woe$perc_y) * loan_woe$woe)
loan_IV
```

```
## [1] 0.05485853
```

In realtà l'information value restituisce una predittività debole.

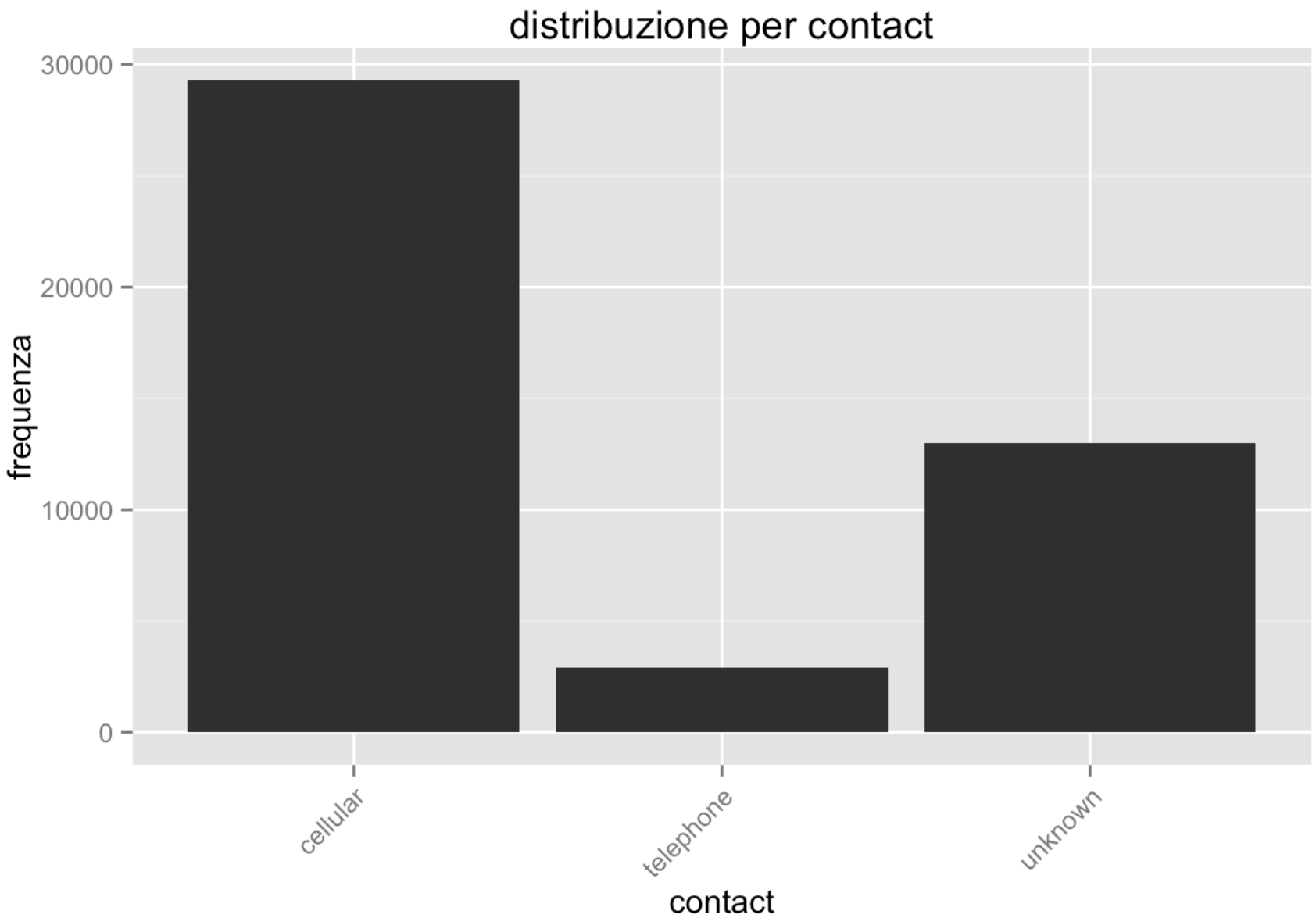
Contact

Variabile categoricha che indica la modalità di comunicazione col cliente.

```
t_contact  <- bank0 %>%
        group_by(contact) %>%
        summarise(frequenza = n()) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        arrange(desc(frequenza))
kable(t_contact, digits = 4, format = "markdown")
```

contact	frequenza	frequenza_relativa
cellular	29285	0.6477
unknown	13020	0.2880
telephone	2906	0.0643

```
g_contact  <- ggplot(bank0, aes(x = contact)) +
        geom_bar() +
        ggtitle("distribuzione per contact") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_contact
```

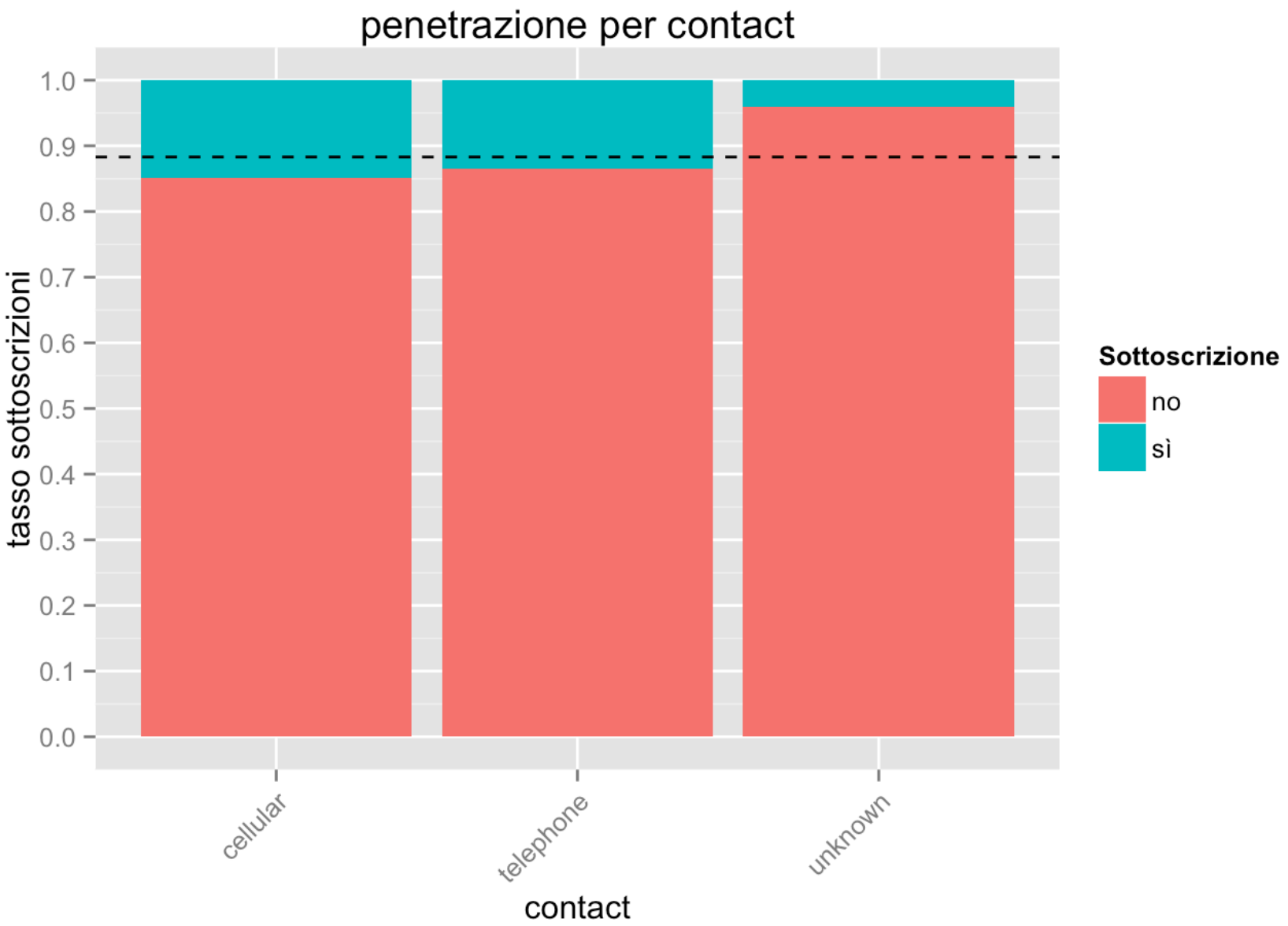


In sostanza abbiamo un 6% tramite fisso e il restante tramite cellullare. Purtroppo Un 29% dei clienti è stato contattato tramite modalità non rilevate.

```
t_contact_y <- bank0 %>%
        group_by (contact) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(contact, frequenza, frequenza_relativa, tasso_sottoscrizioni) %>%
        arrange(desc(tasso_sottoscrizioni))
kable(t_contact_y, digits = 4, format = "markdown")
```

contact	frequenza	frequenza_relativa	tasso_sottoscrizioni
cellular	29285	0.6477	0.1492
telephone	2906	0.0643	0.1342
unknown	13020	0.2880	0.0407

```
g_contact_y  <- ggplot(bank0, aes(x = contact, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per contact") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_contact_y
```



Interessantissimo notare che c'è una variabilità sistematica: se la modalità di contatto è sconosciuta la percentuale di sottoscrittori è bassissima. Questo vorrà dire qualcosa, ma è complicato immaginare il motivo. Il restante 70% di portafoglio sottoscrive al 14% circa, contro l'11,7% complessivo. La variabile sembra discriminante, vediamo se l'information value lo conferma.

```
contact_woe <- bank0 %>%
        select(contact, y) %>%
        group_by(contact) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
contact_woe$woe <- log(contact_woe$perc_no / contact_woe$perc_y)
contact_IV <- sum((contact_woe$perc_no - contact_woe$perc_y) * contact_woe$woe)
contact_IV
```

```
## [1] 0.3003961
```

Confermato, la predittività è di media / forte entità.

Day

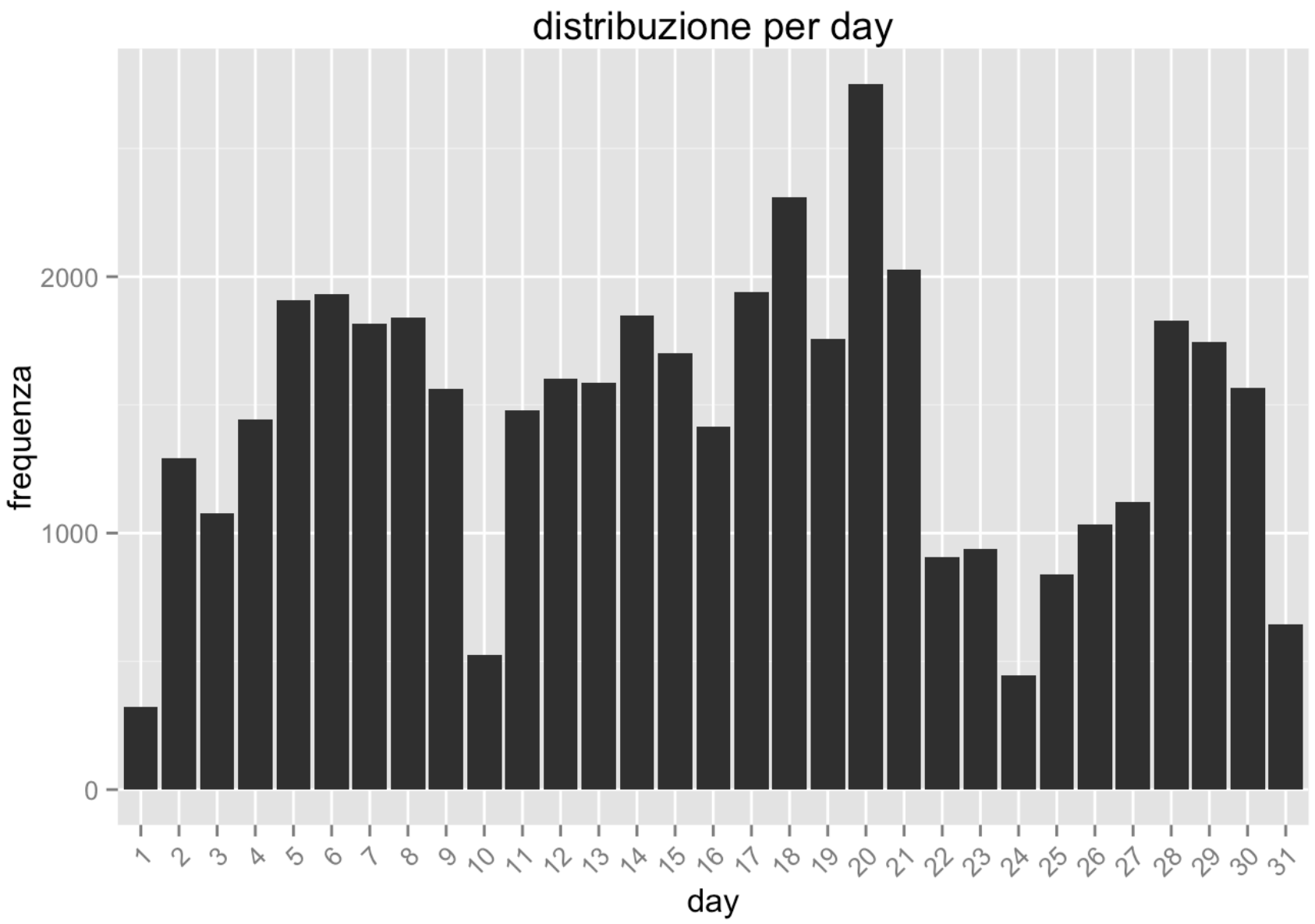
La variabile rileva il giorno del mese dell'ultimo contatto, che laddove `y = "yes"` significa la chiamata di vendita.

```
t_day  <- bank0 %>%
        group_by(day) %>%
        summarise(frequenza = n()) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        arrange(day)
kable(t_day, digits = 4, format = "markdown")
```

day	frequenza	frequenza_relativa
1	322	0.0071
2	1293	0.0286
3	1079	0.0239
4	1445	0.0320
5	1910	0.0422
6	1932	0.0427
7	1817	0.0402
8	1842	0.0407
9	1561	0.0345
10	524	0.0116
11	1479	0.0327
12	1603	0.0355
13	1585	0.0351
14	1848	0.0409
15	1703	0.0377
16	1415	0.0313
17	1939	0.0429
18	2308	0.0510
19	1757	0.0389
20	2752	0.0609
21	2026	0.0448
22	905	0.0200
23	939	0.0208
24	447	0.0099
25	840	0.0186
26	1035	0.0229
27	1121	0.0248

28	1830	0.0405
29	1745	0.0386
30	1566	0.0346
31	643	0.0142

```
g_day  <- ggplot(bank0, aes(x = factor(day))) +
        geom_bar() +
        ggtitle("distribuzione per day") +
        xlab("day") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_day
```



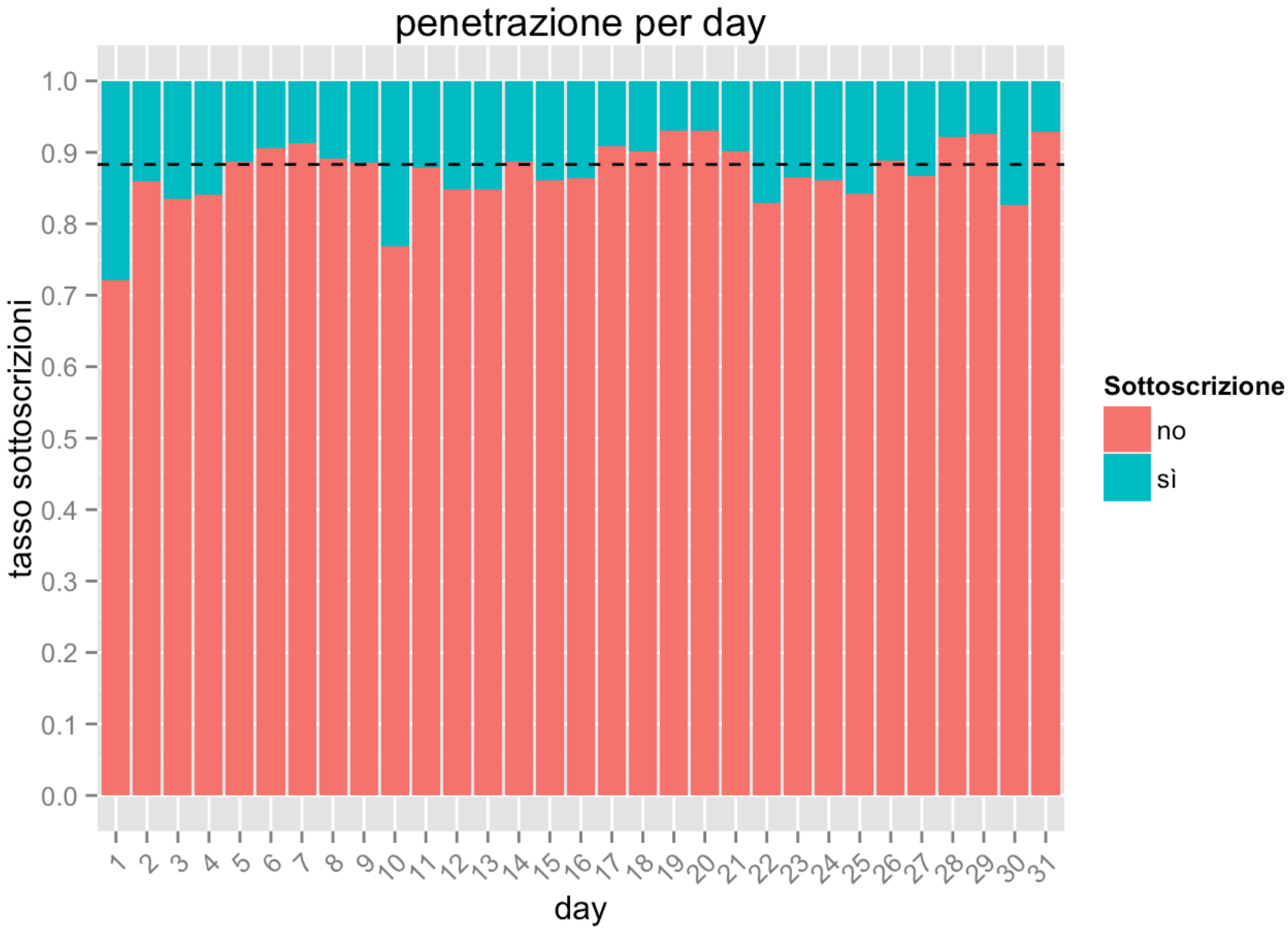
Il 1, 10, 24 e 31 del mese accade poche volte che ci sia l'ultima chiamata della campagna. Chissà perché.

```
t_day_y <- bank0 %>%
        group_by (day) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(day, frequenza, frequenza_relativa, tasso_sottoscrizioni) %>%
        arrange(day)
kable(t_day_y, digits = 4, format = "markdown")
```

day	frequenza	frequenza_relativa	tasso_sottoscrizioni
1	322	0.0071	0.2795
2	1293	0.0286	0.1408
3	1079	0.0239	0.1650
4	1445	0.0320	0.1592
5	1910	0.0422	0.1126
6	1932	0.0427	0.0937
7	1817	0.0402	0.0864
8	1842	0.0407	0.1091
9	1561	0.0345	0.1147
10	524	0.0116	0.2309
11	1479	0.0327	0.1224
12	1603	0.0355	0.1522
13	1585	0.0351	0.1521
14	1848	0.0409	0.1136
15	1703	0.0377	0.1398
16	1415	0.0313	0.1357
17	1939	0.0429	0.0908
18	2308	0.0510	0.0988
19	1757	0.0389	0.0694
20	2752	0.0609	0.0698
21	2026	0.0448	0.0992
22	905	0.0200	0.1702
23	939	0.0208	0.1342
24	447	0.0099	0.1387
25	840	0.0186	0.1583
26	1035	0.0229	0.1121

27	1121	0.0248	0.1338
28	1830	0.0405	0.0781
29	1745	0.0386	0.0739
30	1566	0.0346	0.1731
31	643	0.0142	0.0715

```
g_day_y  <- ggplot(bank0, aes(x = factor(day), fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per day") +
        xlab("day") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_day_y
```



I due giorni di maggior sottoscrizione sono quelli con il minor numero di contatti. Chissà perché. Sarebbe inoltre utile capire se c'è una interazione tra il giorno dell'ultima chiamata e il numero di contatti già effettuati, così da vedere se ci sono giorni dove basta una chiamata sola per vendere il prodotto. In generale c'è parecchia variabilità all'interno del mese sul tasso di sottoscrizioni.

```
day_woe <- bank0 %>%
        select(day, y) %>%
        group_by(day) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
day_woe$woe <- log(day_woe$perc_no / day_woe$perc_y)
day_IV <- sum((day_woe$perc_no - day_woe$perc_y) * day_woe$woe)
day_IV
```

```
## [1] 0.1177583
```

Media predittività.

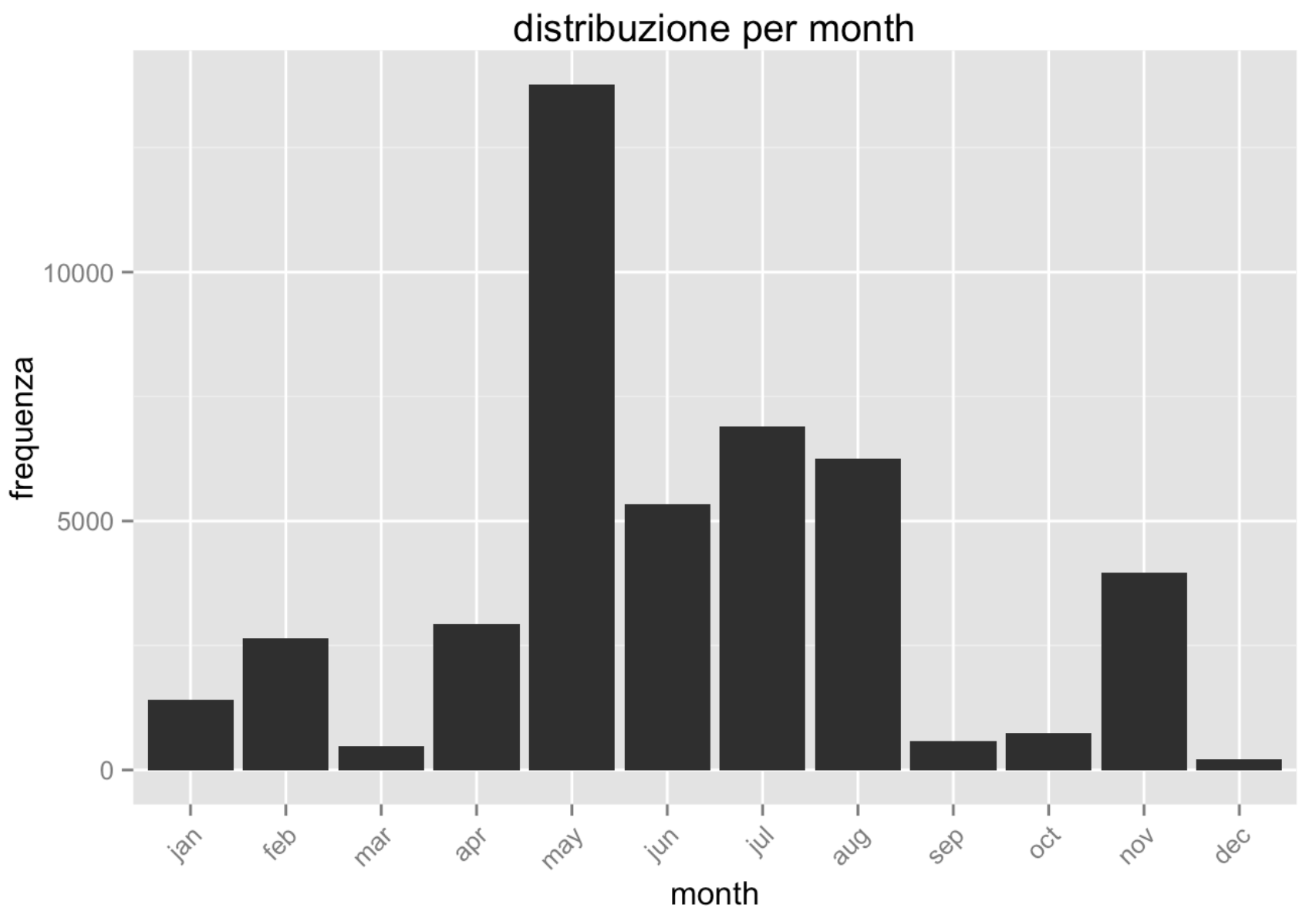
Month

La variabile rileva il mese del mese dell'ultimo contatto, che laddove `y = "yes` significa la chiamata di vendita.

```
bank0$month <- factor(bank0$month, levels = c("jan","feb","mar", "apr","may","jun", "jul
","aug", "sep", "oct", "nov","dec"))
t_month  <- bank0 %>%
        group_by(month) %>%
        summarise(frequenza = n()) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza))
kable(t_month, digits = 4, format = "markdown")
```

month	frequenza	frequenza_relativa
jan	1403	0.0310
feb	2649	0.0586
mar	477	0.0106
apr	2932	0.0649
may	13766	0.3045
jun	5341	0.1181
jul	6895	0.1525
aug	6247	0.1382
sep	579	0.0128
oct	738	0.0163
nov	3970	0.0878
dec	214	0.0047

```
g_month  <- ggplot(bank0, aes(x = month)) +
        geom_bar() +
        ggtitle("distribuzione per month") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_month
```



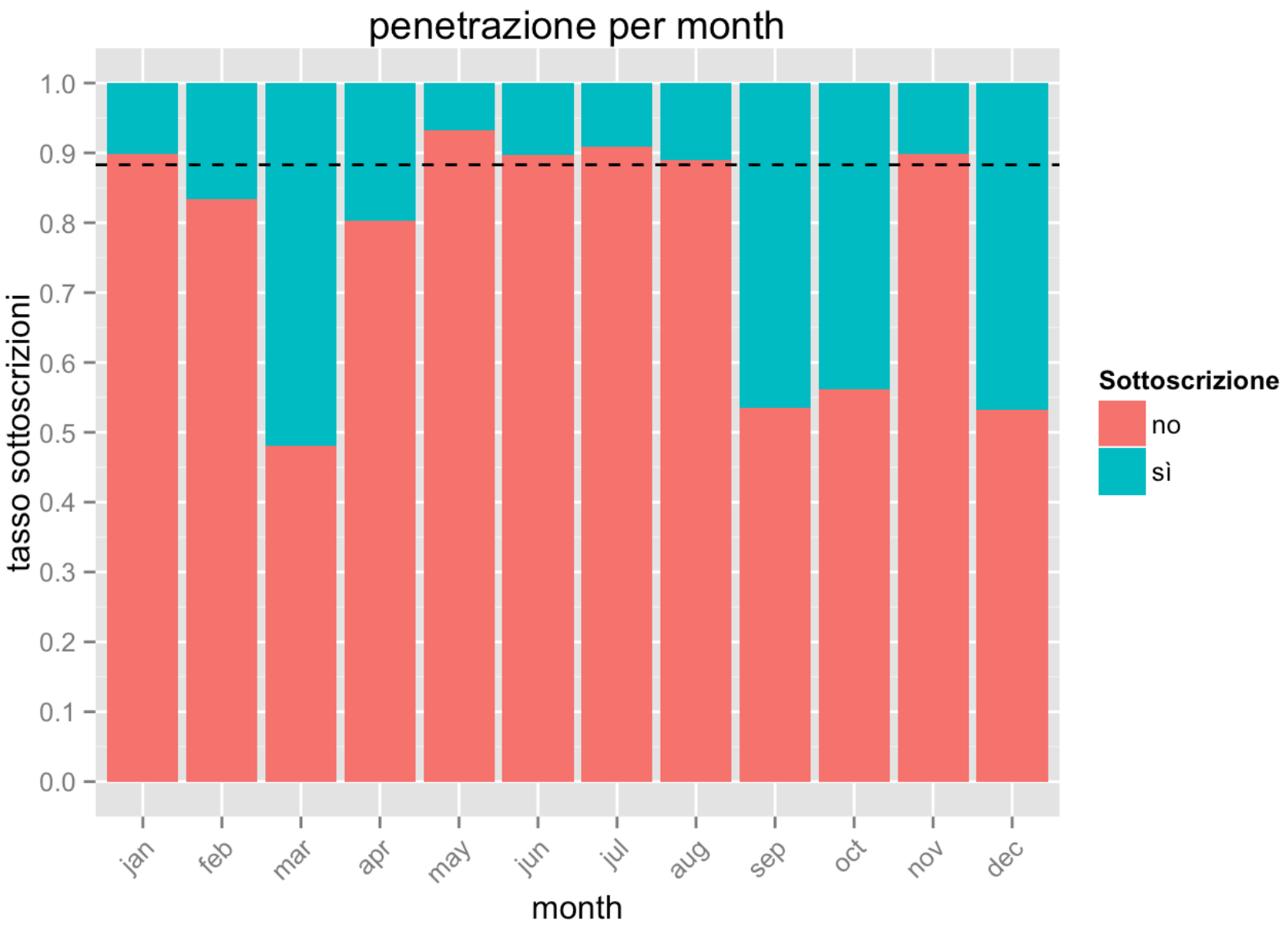
I mesi di ultima chiamata sono prevalentemente quelli estivi. La campagna di marketing è andata avant per due anni, quindi non è chiaro se ci sia stato o meno ogni anno un focus particolare in estate, quando magari l'operatività corrente è inferiore.

```
t_month_y <- bank0 %>%
        group_by (month) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(month, frequenza, frequenza_relativa, tasso_sottoscrizioni)
kable(t_month_y, digits = 4, format = "markdown")
```

month	frequenza	frequenza_relativa	tasso_sottoscrizioni
jan	1403	0.0310	0.1012
feb	2649	0.0586	0.1665
mar	477	0.0106	0.5199

apr	2932	0.0649	0.1968
may	13766	0.3045	0.0672
jun	5341	0.1181	0.1022
jul	6895	0.1525	0.0909
aug	6247	0.1382	0.1101
sep	579	0.0128	0.4646
oct	738	0.0163	0.4377
nov	3970	0.0878	0.1015
dec	214	0.0047	0.4673

```
g_month_y  <- ggplot(bank0, aes(x = month, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per month") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_month_y
```



I mesi di maggior sottoscrizione, in percentuale, sono quelli con il minor numero di contatti. Si sottoscrive pochissimo d'estate, infatti, e questo dovrebbe indurci a pensare che il management non ponga un focus particolare sul fare gli outbound d'estate. Anche qui l'interazione con il numero di contatti precedenti può essere importante, e idem per la durata dell'ultima chiamata: se ci sono ultime chiamate molto lunghe (magari perché si sta vendendo il prodotto) allora ci può stare che se ne facciano di meno in alcuni mesi.

```
month_woe <- bank0 %>%
        select(month, y) %>%
        group_by(month) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
month_woe$woe <- log(month_woe$perc_no / month_woe$perc_y)
month_IV <- sum((month_woe$perc_no - month_woe$perc_y) * month_woe$woe)
month_IV
```

```
## [1] 0.4361311
```

Predittività molto forte.

Duration

Durata in secondi dell'ultima chiamata effettuata al cliente. Se il valore è pari a zero significa che all'ultima chiamata il cliente non ha risposto. Analizziamo la distribuzione univariata:

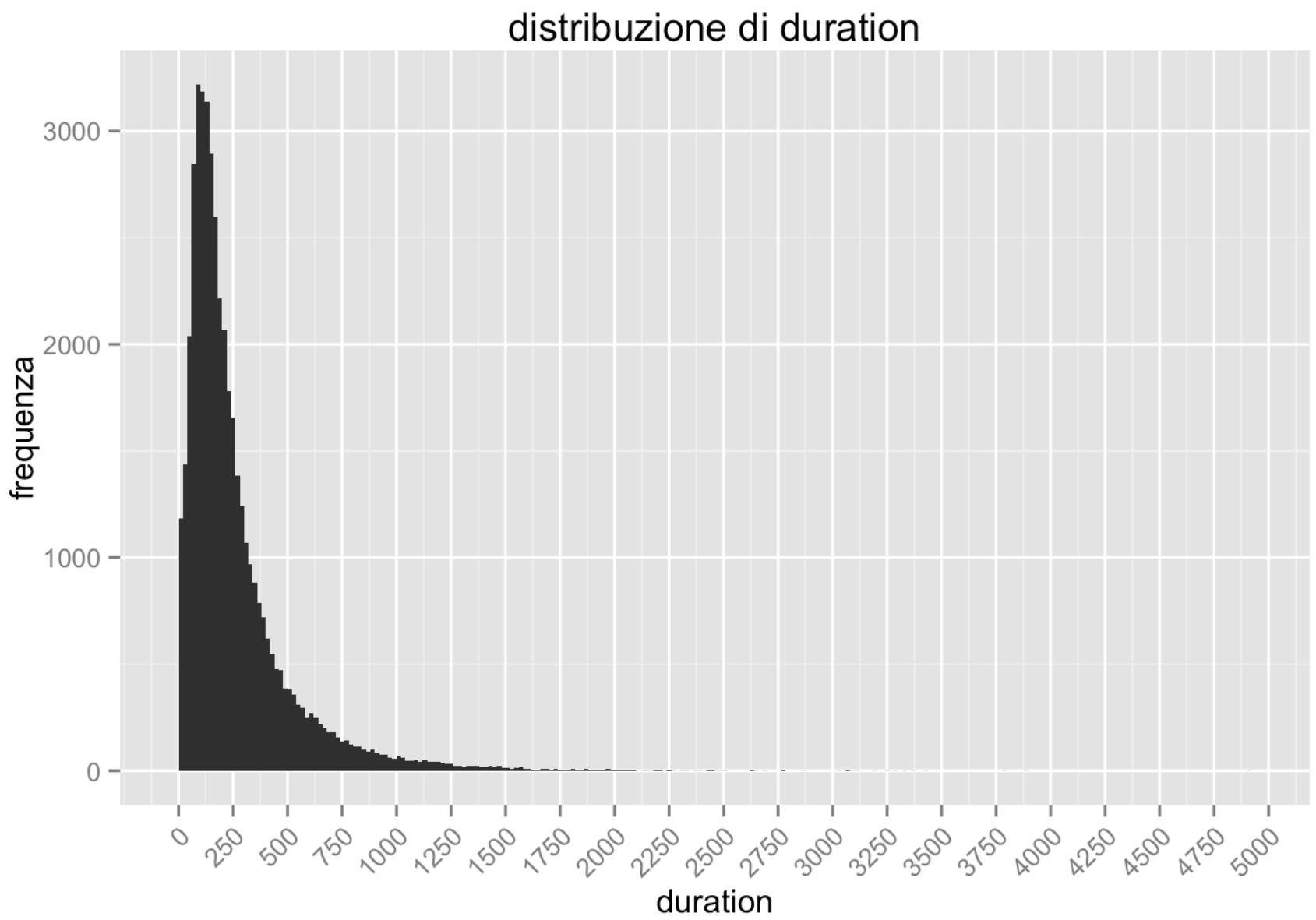
```
summary(bank0$duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   103.0   180.0   258.2   319.0  4918.0
```

```
duration_quantile <- quantile(bank0$duration, c(seq(0.1, 1, 0.05)))
duration_quantile
```

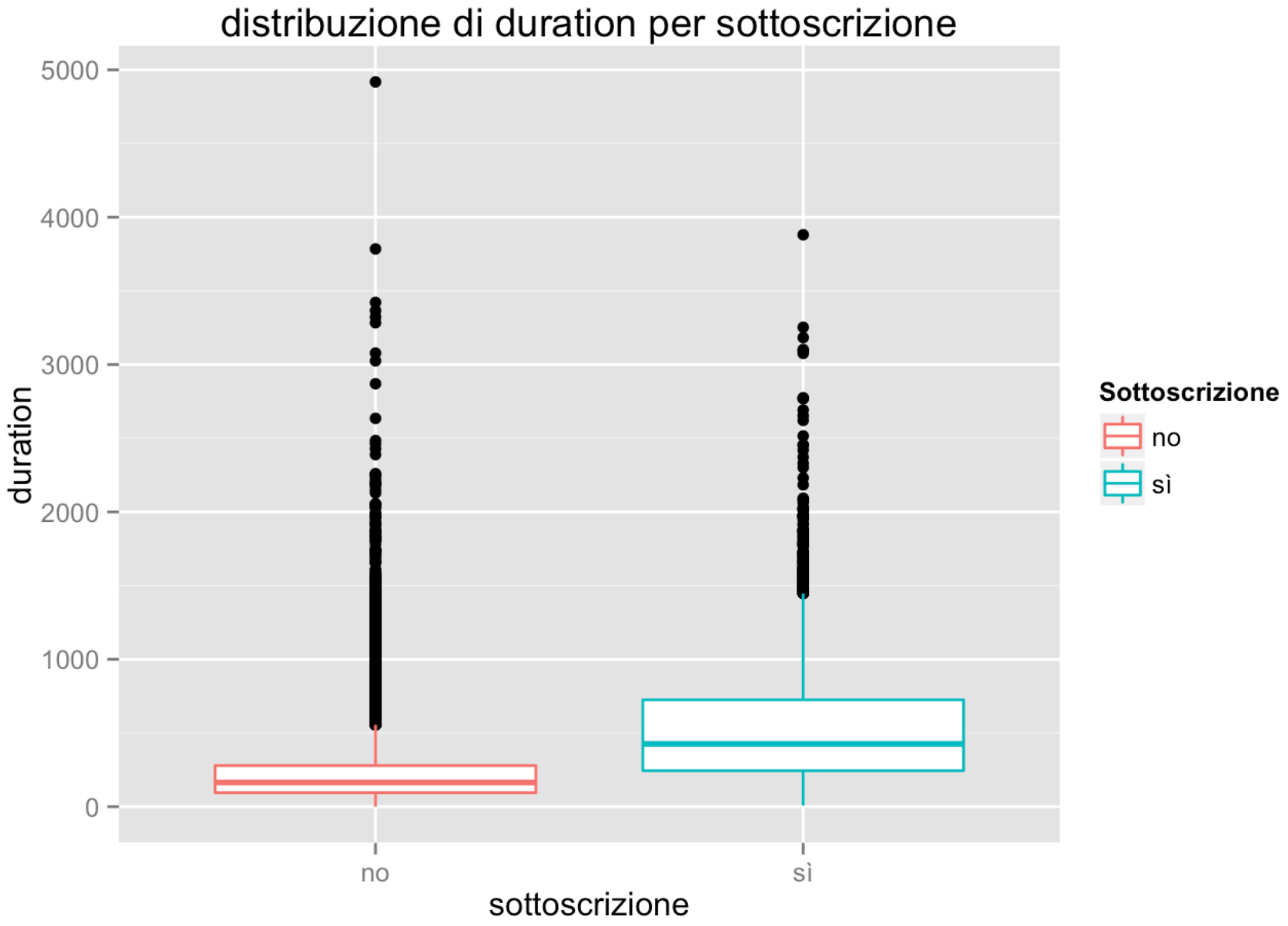
```
##  10%  15%  20%  25%  30%  35%  40%  45%  50%  55%  60%  65%  70%  75%  80%
##   58   75   89  103  117  132  147  163  180  200  223  249  280  319  368
##  85%  90%  95% 100%
##  437  548  751 4918
```

```
g_duration <- ggplot(bank0, aes(x = duration)) +
        geom_histogram(binwidth = 20) +
        ggtitle("distribuzione di duration") +
        xlab("duration") +
        ylab("frequenza") +
        scale_x_continuous(breaks = seq(0, 5000, 250))+
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_duration
```



Una distribuzione molto concentrata in un piccolo intervallo, il 70% delle chiamate non supera i 280 secondi. Vediamo se il comportamento della variabile cambia considerando chi sottoscrive e chi non sottoscrive.

```
g_duration_y <- ggplot(bank0, aes(x = y, y = duration)) +
        geom_boxplot(aes(col = y)) +
        ggtitle("distribuzione di duration per sottoscrizione") +
        xlab("sottoscrizione") +
        scale_x_discrete(labels = c("no", "sì")) +
        scale_color_discrete(name="Sottoscrizione", labels=c("no", "sì"))
g_duration_y
```



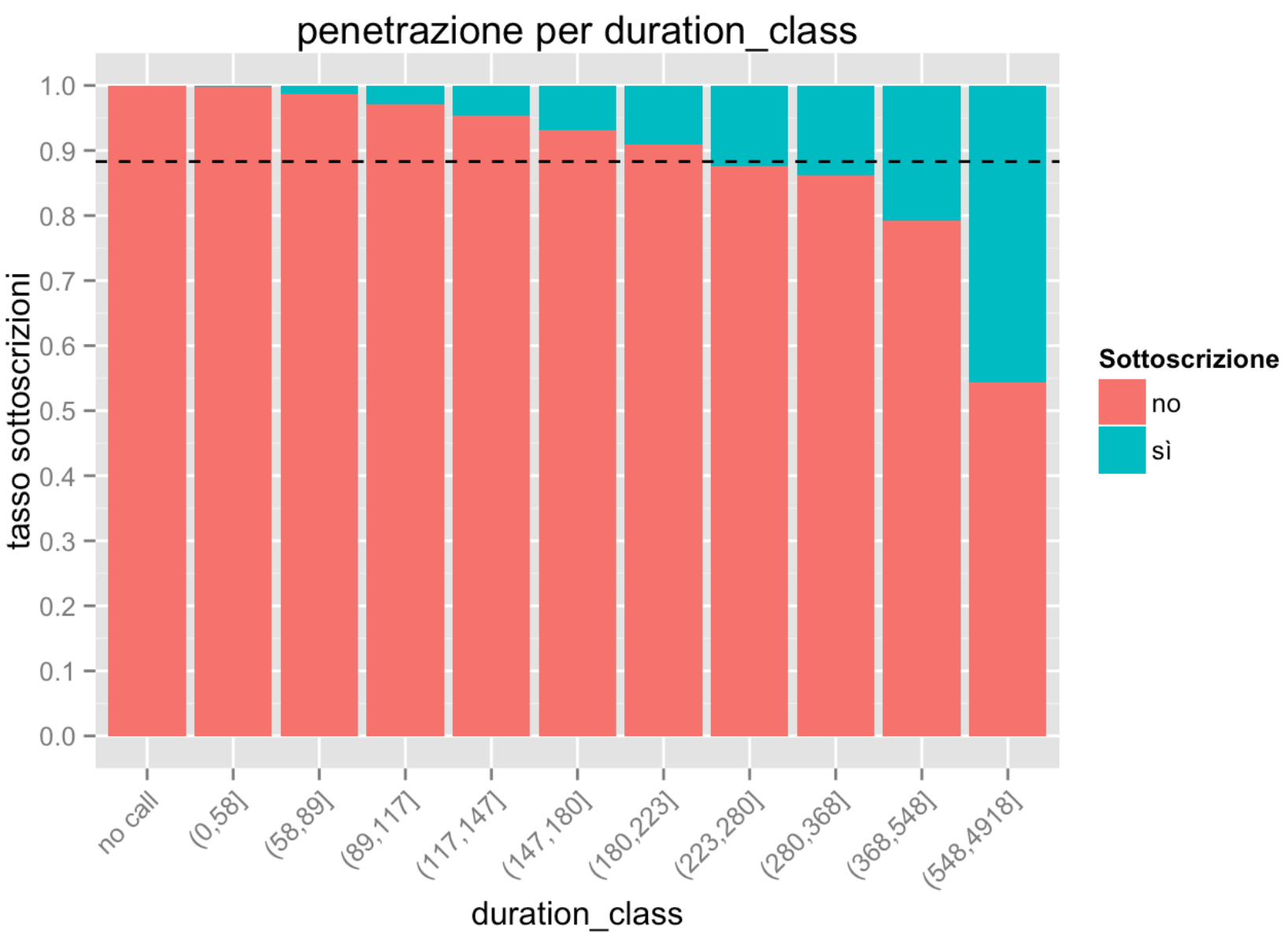

```
duration_decile <- quantile(bank0$duration[bank0$duration != 0], probs = seq(0.1,1,0.1))
bank0$duration_class <- cut(bank0$duration, breaks = c(-1, 0, duration_decile), right =
TRUE, labels = c("no call", "(0,58]",  "(58,89]", "(89,117]", "(117,147]"," (147,180]",
"(180,223]", "(223,280]", "(280,368]", "(368,548]", "(548,4918]"))

t_duration_class_y <- bank0 %>%
        group_by (duration_class) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(duration_class, frequenza, frequenza_relativa, tasso_sottoscrizioni)
kable(t_duration_class_y, digits = 4, format = "markdown")
```

duration_class	frequenza	frequenza_relativa	tasso_sottoscrizioni
no call	3	0.0001	0.0000
(0,58]	4526	0.1001	0.0020
(58,89]	4576	0.1012	0.0125
(89,117]	4495	0.0994	0.0285
(117,147]	4564	0.1009	0.0458
(147,180]	4496	0.0994	0.0681
(180,223]	4531	0.1002	0.0905

(223,280]	4517	0.0999	0.1233
(280,368]	4468	0.0988	0.1379
(368,548]	4527	0.1001	0.2085
(548,4918]	4508	0.0997	0.4554

```
g_duration_class_y  <- ggplot(bank0, aes(x = duration_class, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per duration_class") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_duration_class_y
```



Muovendosi all'interno dei 10 gruppi di dimensioni simili il tasso di sottoscrizione aumenta all'aumentare della durata della chiamata. La cosa ha senso, dato che una chiamata in cui si vende richiede tempo lunghi perché il cliente è molto interessato al prodotto e inoltre dovrà assolvere a dei questionari burocratici. Tuttavia la durata della chiamata non è nota in anticipo, quindi non ha senso includere la variabile in un modello predittivo.

```
#devo togliere i tre valori nulli di duration altrimenti la formula va in errore
duration_class_woe <- bank0 %>%
        filter(duration_class != "no call") %>%
        select(duration_class, y) %>%
        group_by(duration_class) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
duration_class_woe$woe <- log(duration_class_woe$perc_no / duration_class_woe$perc_y)
duration_class_IV <- sum((duration_class_woe$perc_no - duration_class_woe$perc_y) * dura
tion_class_woe$woe)
duration_class_IV
```

```
## [1] 1.610237
```

è infatti un valore di 1.6 è davvero sospetto, troppo alto per essere vero.

Campaign

Numero di contatti complessivi avuto con il cliente durante la campagna promozionale. Include l'ultimo contatto.

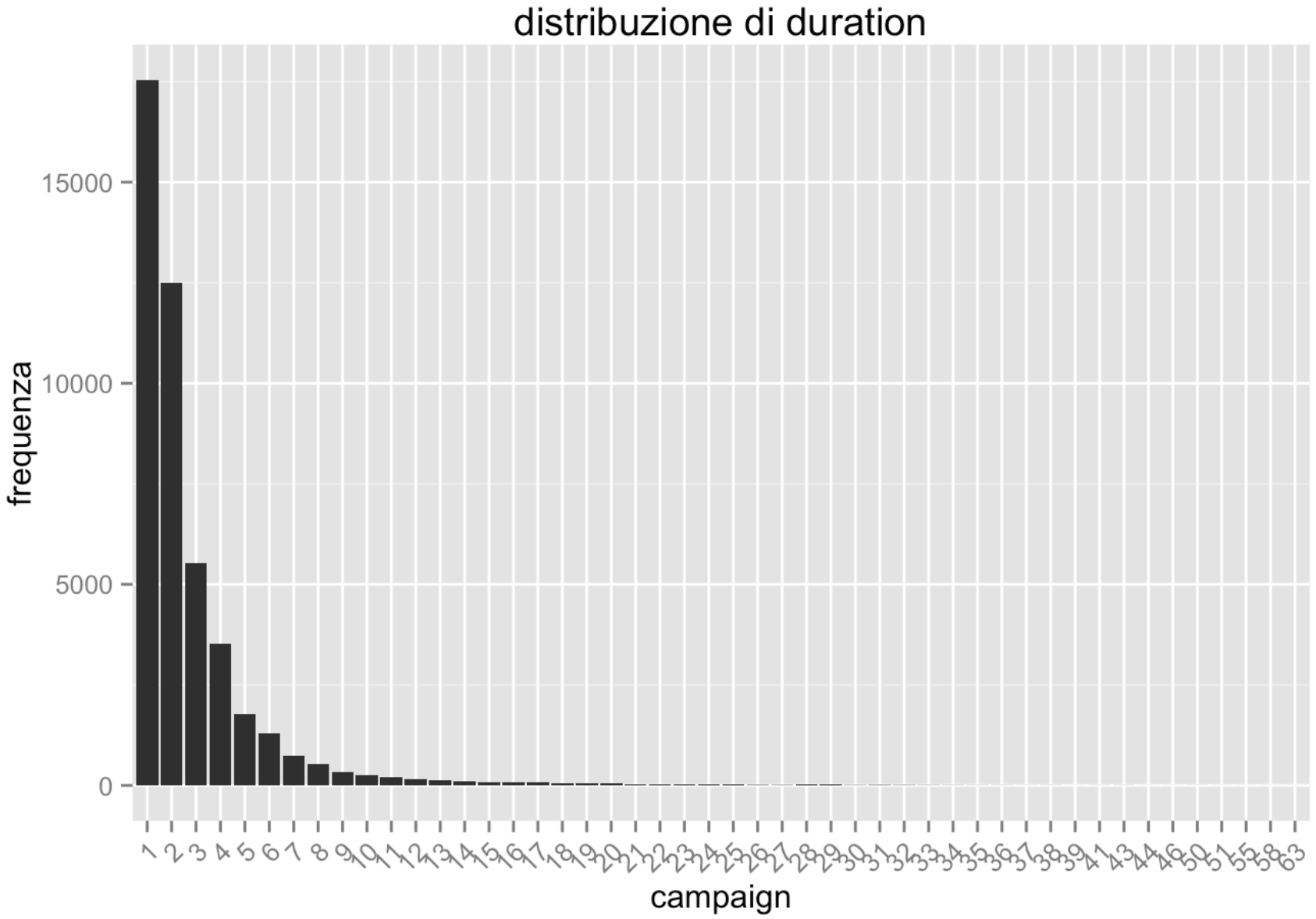
```
summary(bank0$campaign)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   2.764   3.000  63.000
```

```
campaign_quantile <- quantile(bank0$campaign, c(seq(0.1, 1, 0.05)))
campaign_quantile
```

```
##  10%  15%  20%  25%  30%  35%  40%  45%  50%  55%  60%  65%  70%  75%  80%
##    1    1    1    1    1    1    2    2    2    2    2    2    3    3    4
##  85%  90%  95% 100%
##    4    5    8   63
```

```
g_campaign <- ggplot(bank0, aes(x = factor(campaign))) +
        geom_histogram(binwidth = 1) +
        ggtitle("distribuzione di duration") +
        xlab("campaign") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_campaign
```



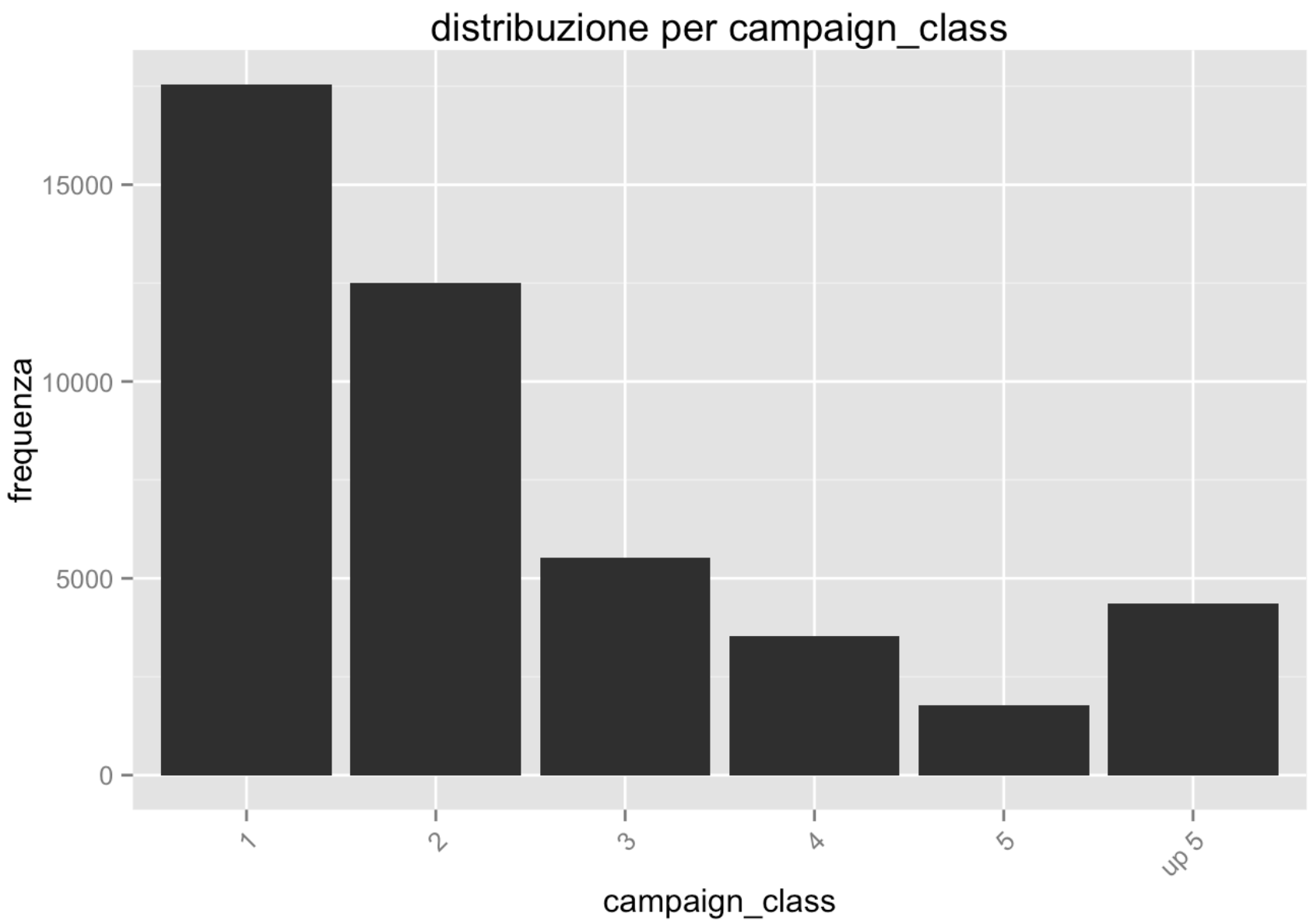
Il 95% dei valori è tra 1 e 8 contatti. Ha senso raggruppare la variabile in classi

```
bank0$campaign_class <- cut(bank0$campaign, breaks = c(0, 1, 2, 3, 4, 5, max(bank0$campa
ign)+1), right = TRUE, labels = c("1", "2", "3", "4", "5", "up 5"))

t_campaign_class  <- bank0 %>%
        group_by(campaign_class) %>%
        summarise(frequenza = n()) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza))
kable(t_campaign_class, digits = 4, format = "markdown")
```

campaign_class	frequenza	frequenza_relativa
1	17544	0.3880
2	12505	0.2766
3	5521	0.1221
4	3522	0.0779
5	1764	0.0390
up 5	4355	0.0963

```
g_campaign_class  <- ggplot(bank0, aes(x = campaign_class)) +
        geom_bar() +
        ggtitle("distribuzione per campaign_class") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_campaign_class
```



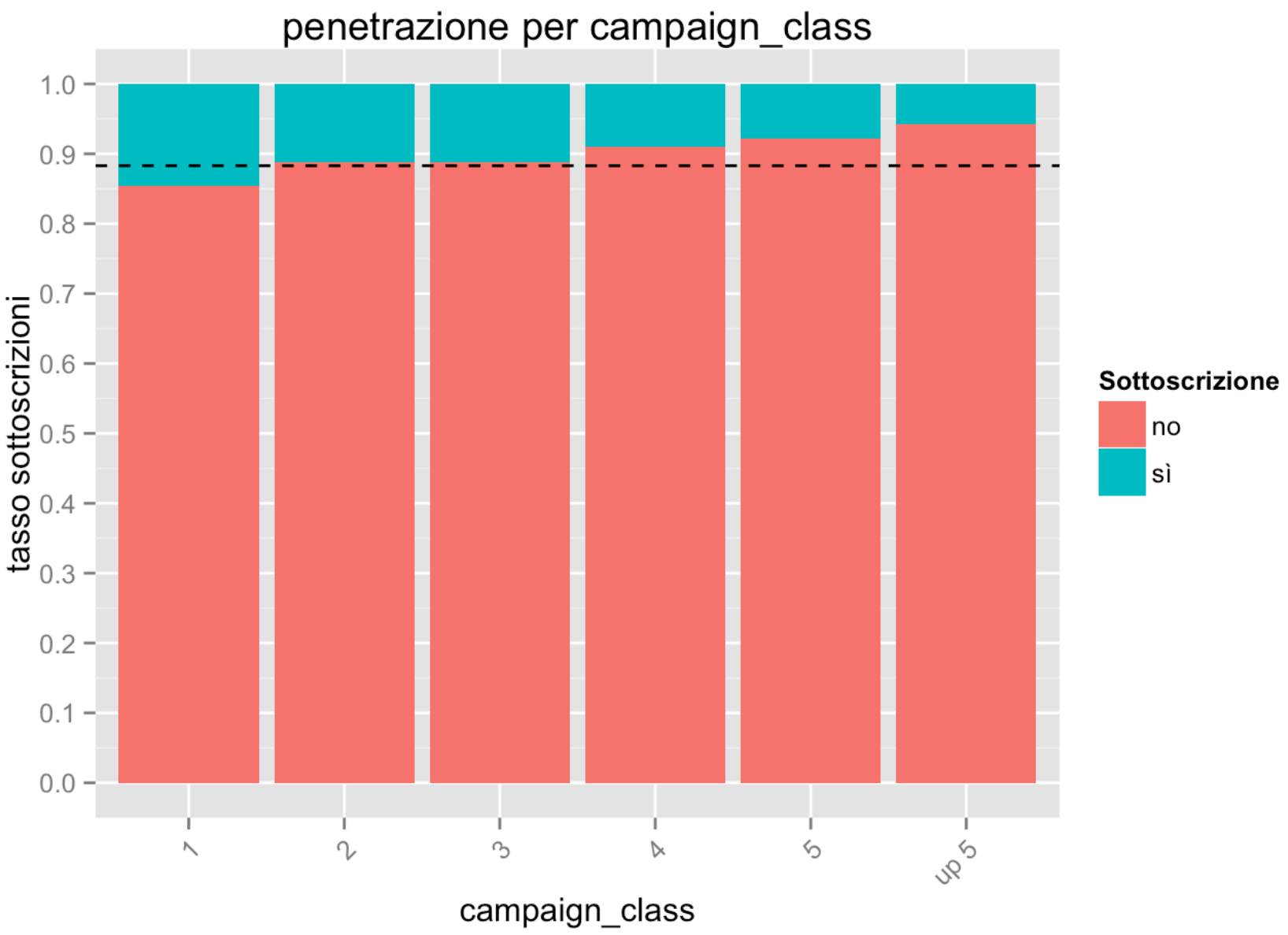
Vediamo come si comportano i sottoscrittori attraverso i vari livelli di `campaign`.

```
t_campaign_class_y <- bank0 %>%
        group_by (campaign_class) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(campaign_class, frequenza, frequenza_relativa, tasso_sottoscrizioni)
kable(t_campaign_class_y, digits = 4, format = "markdown")
```

campaign_class	frequenza	frequenza_relativa	tasso_sottoscrizioni
1	17544	0.3880	0.1460
2	12505	0.2766	0.1120
3	5521	0.1221	0.1119
4	3522	0.0779	0.0900

5	1764	0.0390	0.0788
up 5	4355	0.0963	0.0581

```
g_campaign_class_y  <- ggplot(bank0, aes(x = campaign_class, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per campaign_class") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_campaign_class_y
```



Meno chiamate vengono fatte ai clienti, più si vende (in proporzione). Significa che al cliente davvero interessato basta una chiamata. Solo la fascia “1 contatto” ha un tasso di sottoscrizione superiore a quello complessivo, questa è una informazione rilevante. Vediamo l’information value:

```
campaign_class_woe <- bank0 %>%
        select(campaign_class, y) %>%
        group_by(campaign_class) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
campaign_class_woe$woe <- log(campaign_class_woe$perc_no / campaign_class_woe$perc_y)
campaign_class_IV <- sum((campaign_class_woe$perc_no - campaign_class_woe$perc_y) * camp
aign_class_woe$woe)
campaign_class_IV
```

```
## [1] 0.08293285
```

0.08, predittore debole.

pdays

Variabile che rileva il numero di giorni trascorsi da quando il cliente è stato contattato per una precedente campagna. Se `pdays = -1` allora il cliente non è stato mai contattato per la precedente campagna.

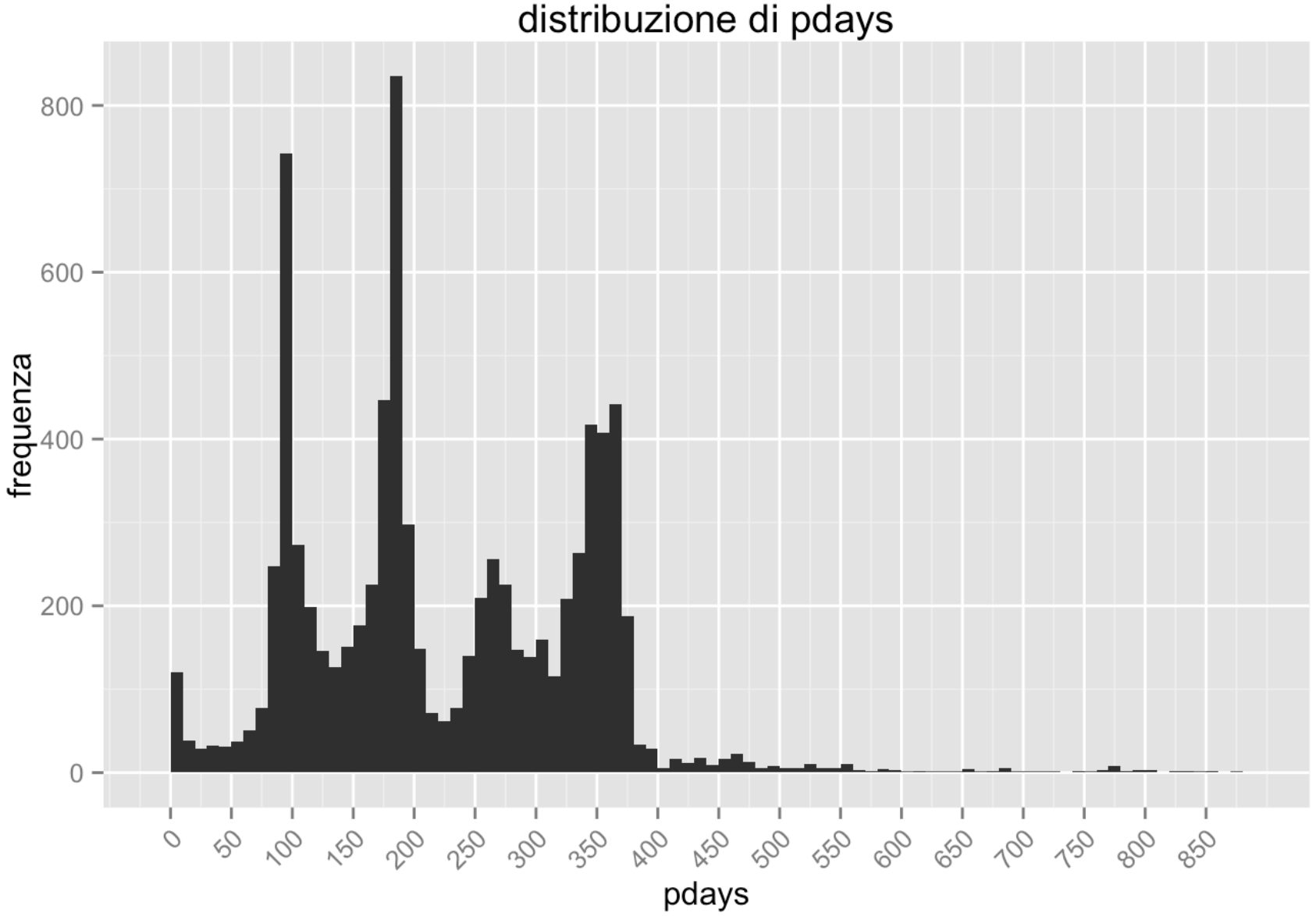
Vediamo come si distribuisce la variabile se il cliente è stato contattato.

```
summary(bank0$pdays[bank0$pdays != -1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   133.0   194.0   224.6   327.0   871.0
```

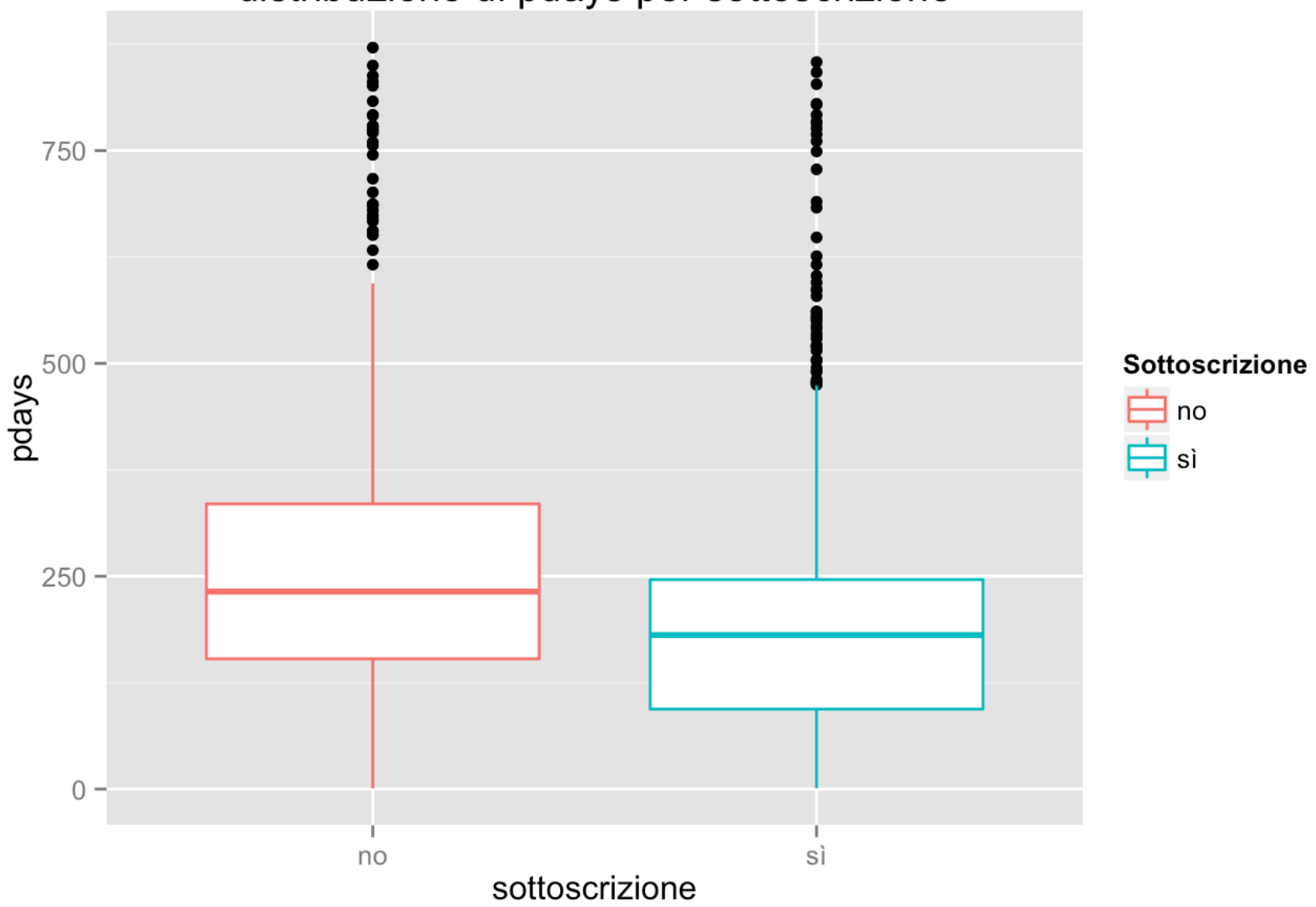
```
pdays_decile <- quantile(bank0$pdays[bank0$pdays != -1], c(seq(0.1, 1, 0.1)))

g_pdays_quant <- bank0  %>%
        filter(pdays != -1) %>%
        ggplot(aes(x = pdays)) +
        geom_histogram(binwidth = 10) +
        ggtitle("distribuzione di pdays") +
        xlab("pdays") +
        ylab("frequenza") +
        scale_x_continuous(breaks = seq(0, 871, 50))+
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_pdays_quant
```



Una distribuzione quasi trimodale; immagino che nell'attuale campagna ci siano stati dei periodi con maggiore concentrazione di chiamate, che hanno generato quelle tre mode. Anche la variabile `month` conferma che ci sono stati mesi con molta più concentrazione di chiamate. Vediamo come si dsitribuisce `pdays` condizionando per la sottoscrizione:

```
g_pdays_quant_y <- bank0 %>%
        filter(pdays != -1) %>%
        ggplot(aes(x = y, y = pdays)) +
        geom_boxplot(aes(col = y)) +
        ggtitle("distribuzione di pdays per sottoscrizione") +
        xlab("sottoscrizione") +
        scale_x_discrete(labels = c("no", "sì")) +
        scale_color_discrete(name="Sottoscrizione", labels=c("no", "sì"))
g_pdays_quant_y
```



Chi non sottoscrive è un cliente per cui sono trascorsi più giorni dall'ultimo contatto. Questo ha senso. Dobbiamo tuttavia indagare il comportamento dei sottoscrittori anche tra chi non è mai stato contatatto, e per questo raggruppiamo la variabile in classi, una per i non contattati e poi dieci in base ai decili.

```
bank0$pdays_class <- cut(bank0$pdays, breaks = c(-2, 0, pdays_decile), right = TRUE, lab
els = c("no campaign", "(0,91]",  "(91,108]", "(108,159]", "(159,181]"," (181,194]", "(1
94,258]", "(258,300]", "(300,343]", "(343,362]", "(362,871]"))
table(bank0$pdays_class)
```

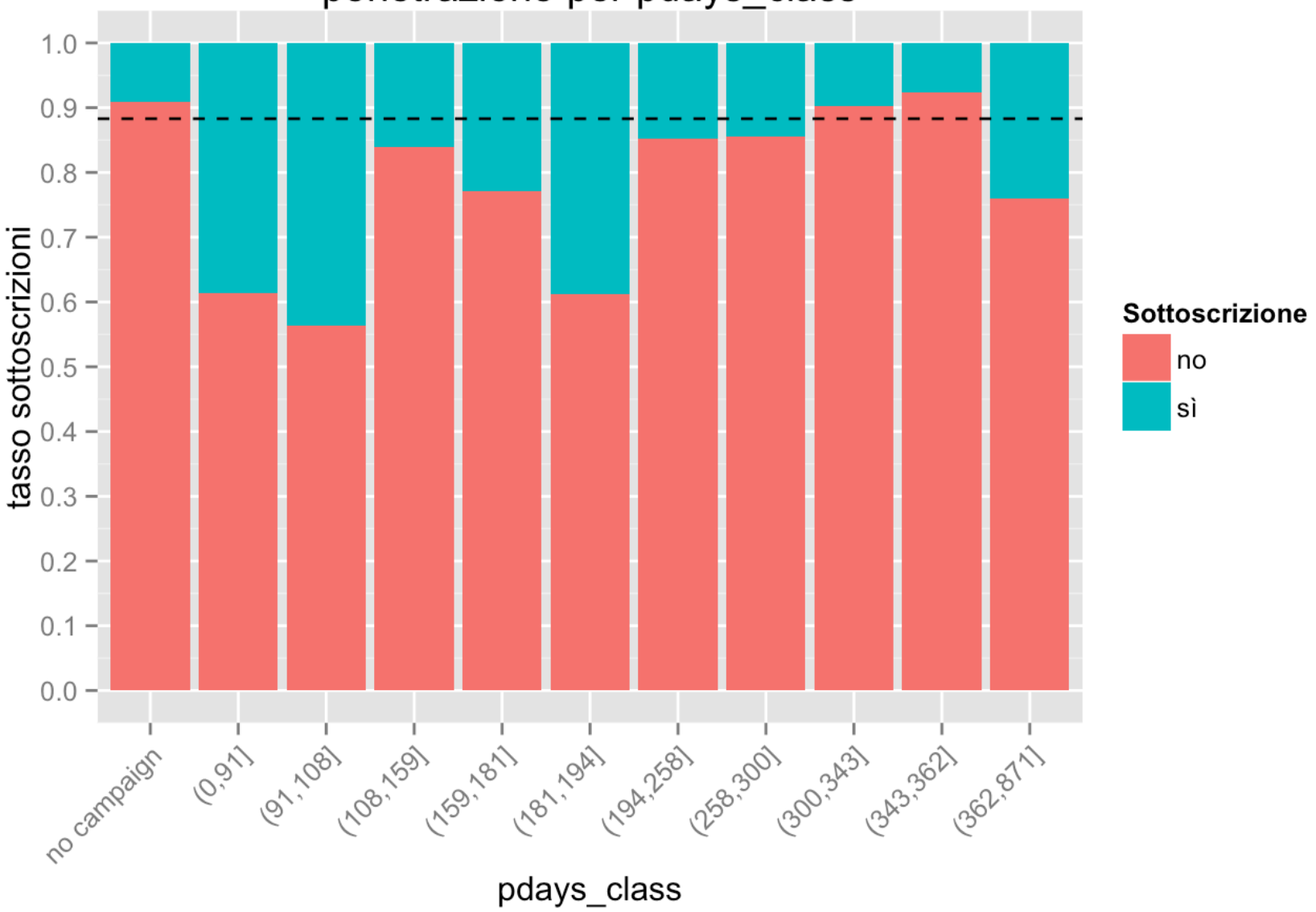
```
##
## no campaign      (0,91]    (91,108]   (108,159]   (159,181]   (181,194]
##       36954         844         817         819         835         814
##   (194,258]   (258,300]   (300,343]   (343,362]   (362,871]
##         833         825         879         765         826
```

Circa l'80% dei clienti contattati per questa campagna non è mai stato contattato precedentemente. Ora vediamo come il tasso di sottoscrizione si distribuisce all'interno delle 11 classi:

```
t_pdays_class_y <- bank0 %>%
        group_by (pdays_class) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(pdays_class, frequenza, frequenza_relativa, tasso_sottoscrizioni)
kable(t_pdays_class_y, digits = 4, format = "markdown")
```

pdays_class	frequenza	frequenza_relativa	tasso_sottoscrizioni
no campaign	36954	0.8174	0.0916
(0,91]	844	0.0187	0.3863
(91,108]	817	0.0181	0.4357
(108,159]	819	0.0181	0.1612
(159,181]	835	0.0185	0.2287
(181,194]	814	0.0180	0.3882
(194,258]	833	0.0184	0.1477
(258,300]	825	0.0182	0.1442
(300,343]	879	0.0194	0.0967
(343,362]	765	0.0169	0.0758
(362,871]	826	0.0183	0.2409

```
g_pdays_class_y  <- ggplot(bank0, aes(x = pdays_class, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per pdays_class") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_pdays_class_y
```



In generale se in passato si è stati contattati si sottoscrive di più che se non lo si è mai stati; il tasso non decresce in maniera monotona ma la distribuzione sembra comunque abbastanza informativa.

```
pdays_class_woe <- bank0 %>%
        select(pdays_class, y) %>%
        group_by(pdays_class) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
pdays_class_woe$woe <- log(pdays_class_woe$perc_no / pdays_class_woe$perc_y)
pdays_class_IV <- sum((pdays_class_woe$perc_no - pdays_class_woe$perc_y) * pdays_class_w
oe$woe)
pdays_class_IV
```

```
## [1] 0.3479284
```

Infatti 0.34 come information value significa predittore potente.

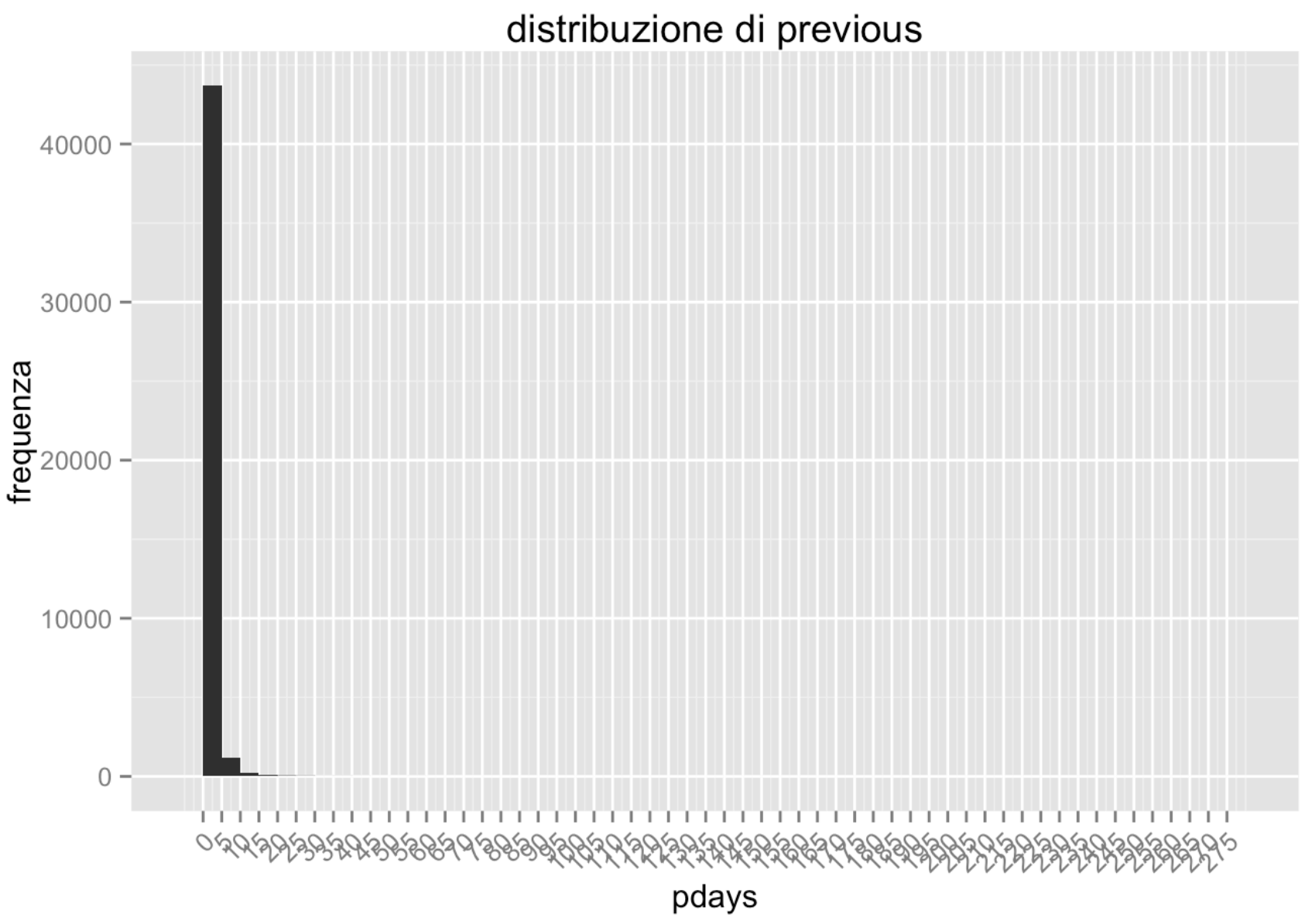
Previous

Numero di volte in cui il cliente è stato contattato prima di questa campagna.

```
previous_decile <- quantile(bank0$previous, c(seq(0.1, 1, 0.1)))
previous_decile
```

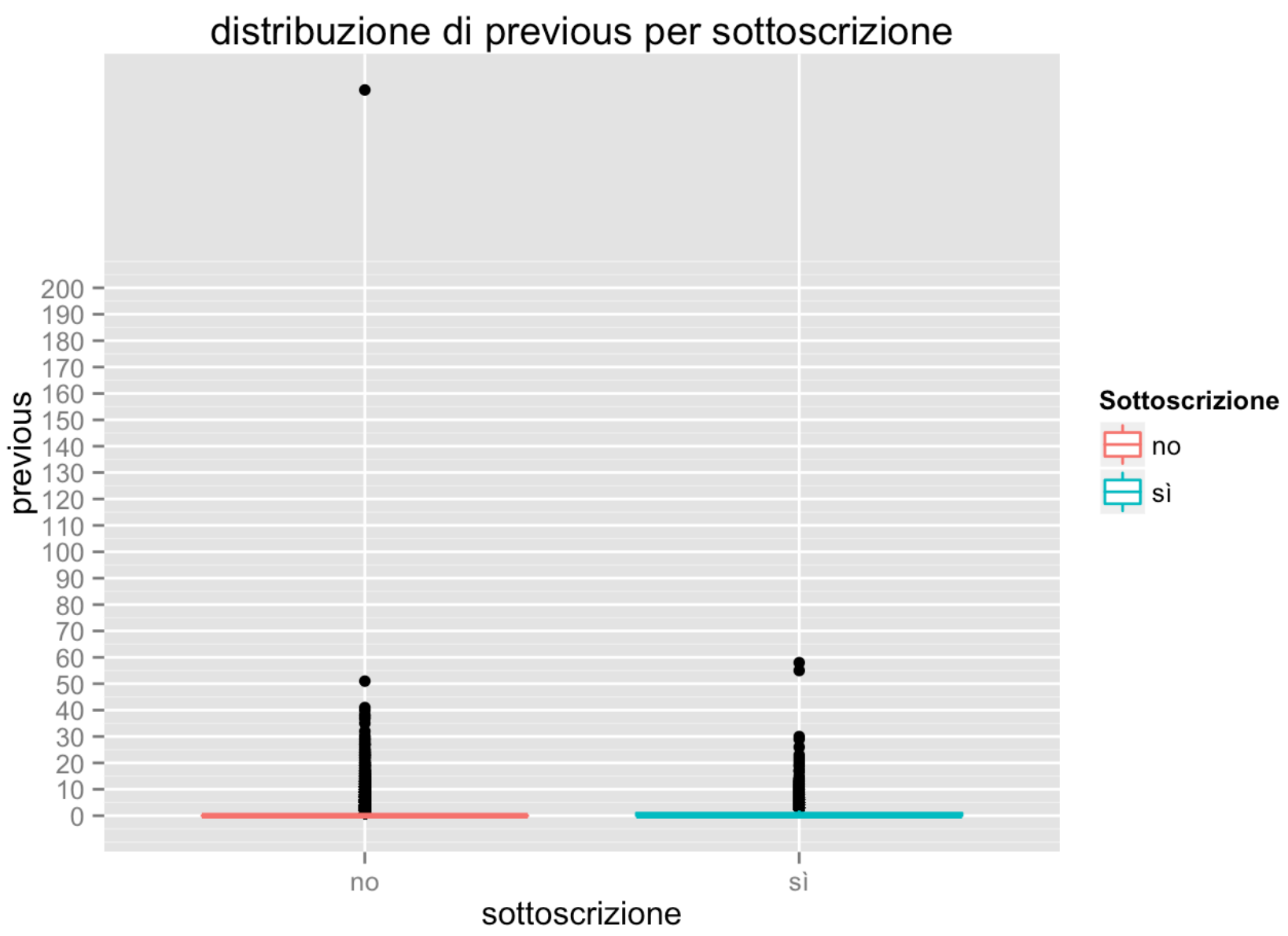
```
##  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    0    0    0    0    0    0    0    0    2  275
```

```
g_previous_quant <- ggplot(bank0, aes(x = previous)) +
        geom_histogram(binwidth = 5) +
        ggtitle("distribuzione di previous") +
        xlab("pdays") +
        ylab("frequenza") +
        scale_x_continuous(breaks = seq(0, 275, 5))+
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_previous_quant
```



L’80% dei clienti non è mai stato contattato, è una distribuzione molto strana. Vediamo distinguendo tra sottoscrittori e non in un boxplot:

```
g_previous_quant_y <- ggplot(bank0, aes(x = y, y = previous)) +
        geom_boxplot(aes(col = y)) +
        ggtitle("distribuzione di previous per sottoscrizione") +
        xlab("sottoscrizione") +
        scale_x_discrete(labels = c("no", "sì")) +
        scale_y_continuous(breaks = seq(0, 200, 10)) +
        scale_color_discrete(name="Sottoscrizione", labels=c("no", "sì"))
g_previous_quant_y
```



In questo modo individuare una associazione tra `previous` e sottoscrizione è impossibile.
Raggruppiamo in classi, anche se mi aspetto una forte correlazione con `pdays`.

```
previous_nz_quantile <- quantile(bank0$previous[bank0$previous > 0], probs = seq(0.1,1,0
.1))
#Considerando i decili, preferisco fare 7 classi di non pari frequenza: da 0 a 6 e maggi
ore di 6
bank0$previous_class <- cut(bank0$previous, breaks = c(0, 1, 2, 3, 4, 5, 7, max(bank0$pr
evious)+1), right = FALSE, labels = c("0 contact", "1 contact",  "2 contact", "3 contact
", "4 contact","5 or 6 contact", "+ 6 contact"))
table(bank0$previous_class)
```

```
##
##      0 contact      1 contact      2 contact      3 contact      4 contact
##          36954           2772           2106           1142            714
## 5 or 6 contact    + 6 contact
##            736            787
```

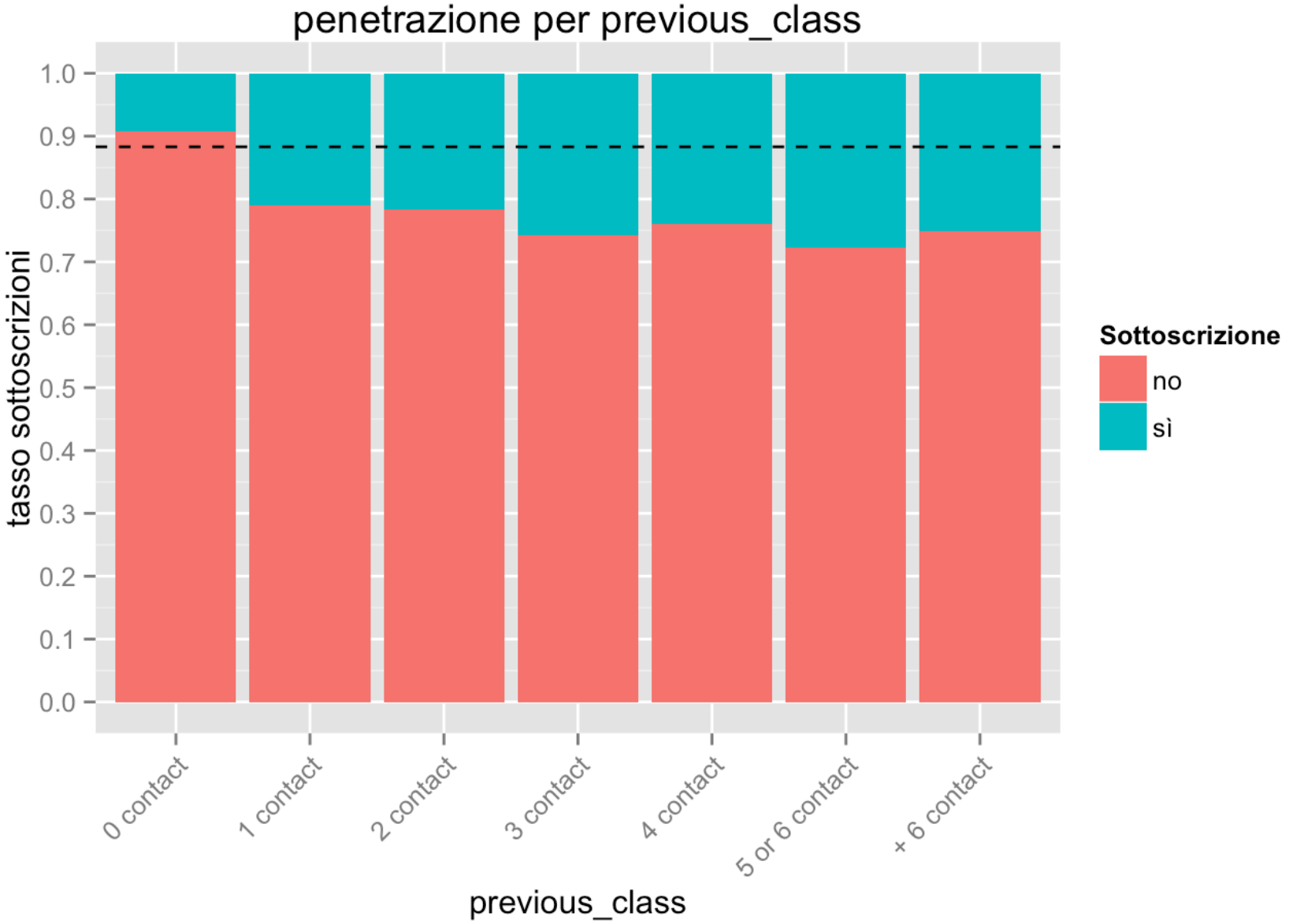
```
summary(bank0$previous_class)
```

```
##      0 contact      1 contact      2 contact      3 contact      4 contact
##          36954           2772           2106           1142            714
## 5 or 6 contact    + 6 contact
##            736            787
```

```
t_previous_class_y <- bank0 %>%
        group_by (previous_class) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(previous_class, frequenza, frequenza_relativa, tasso_sottoscrizioni)
kable(t_previous_class_y, digits = 4, format = "markdown")
```

previous_class	frequenza	frequenza_relativa	tasso_sottoscrizioni
0 contact	36954	0.8174	0.0916
1 contact	2772	0.0613	0.2103
2 contact	2106	0.0466	0.2165
3 contact	1142	0.0253	0.2574
4 contact	714	0.0158	0.2395
5 or 6 contact	736	0.0163	0.2772
+ 6 contact	787	0.0174	0.2503

```
g_previous_class_y  <- ggplot(bank0, aes(x = previous_class, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per previous_class") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_previous_class_y
```



Anche qui c'è un comportamento chiaro: se non si è mai stati contattati si sottoscrive meno (in proporzione) della media, altrimenti sensibilmente di più. Mi aspetto forte predittività:

```
previous_class_woe <- bank0 %>%
        select(previous_class, y) %>%
        group_by(previous_class) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
previous_class_woe$woe <- log(previous_class_woe$perc_no / previous_class_woe$perc_y)
previous_class_IV <- sum((previous_class_woe$perc_no - previous_class_woe$perc_y) * prev
ious_class_woe$woe)
previous_class_IV
```

```
## [1] 0.2234899
```

0.22 è vicina ad essere forte predittività.

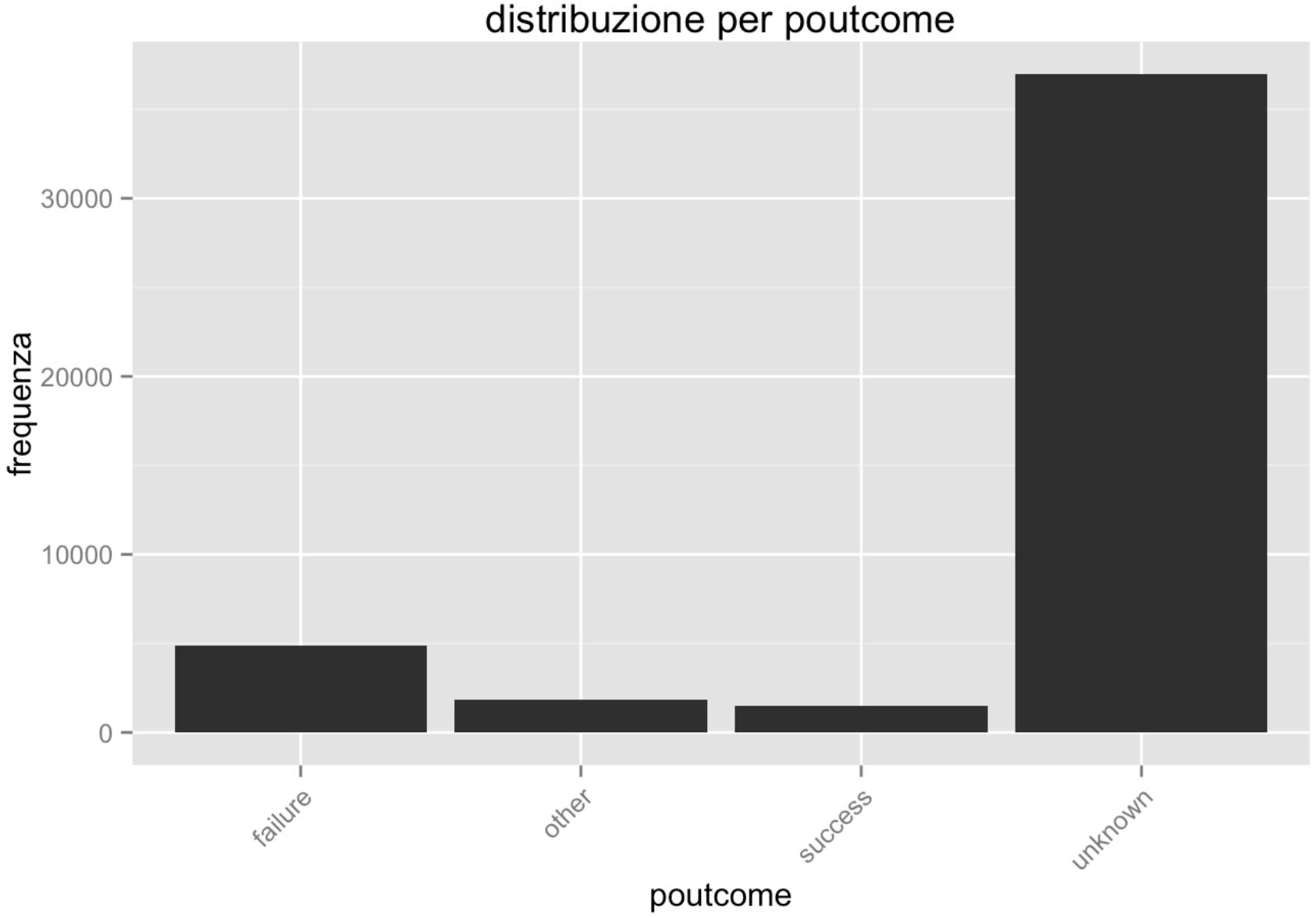
Poutcome

Esito della precedente campagna promozionale per il medesimo cliente. Analizziamo subito la distribuzione univariata:

```
t_poutcome <- bank0 %>%
        group_by (poutcome) %>%
        summarise (frequenza = n()) %>%
        mutate (frequenza_relativa = frequenza / sum(frequenza)) %>%
        arrange(desc(frequenza_relativa))
kable(t_poutcome, digits = 4, format = "markdown")
```

poutcome	frequenza	frequenza_relativa
unknown	36959	0.8175
failure	4901	0.1084
other	1840	0.0407
success	1511	0.0334

```
g_poutcome  <- ggplot(bank0, aes(x = poutcome)) +
        geom_bar() +
        ggtitle("distribuzione per poutcome") +
        ylab("frequenza") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_poutcome
```

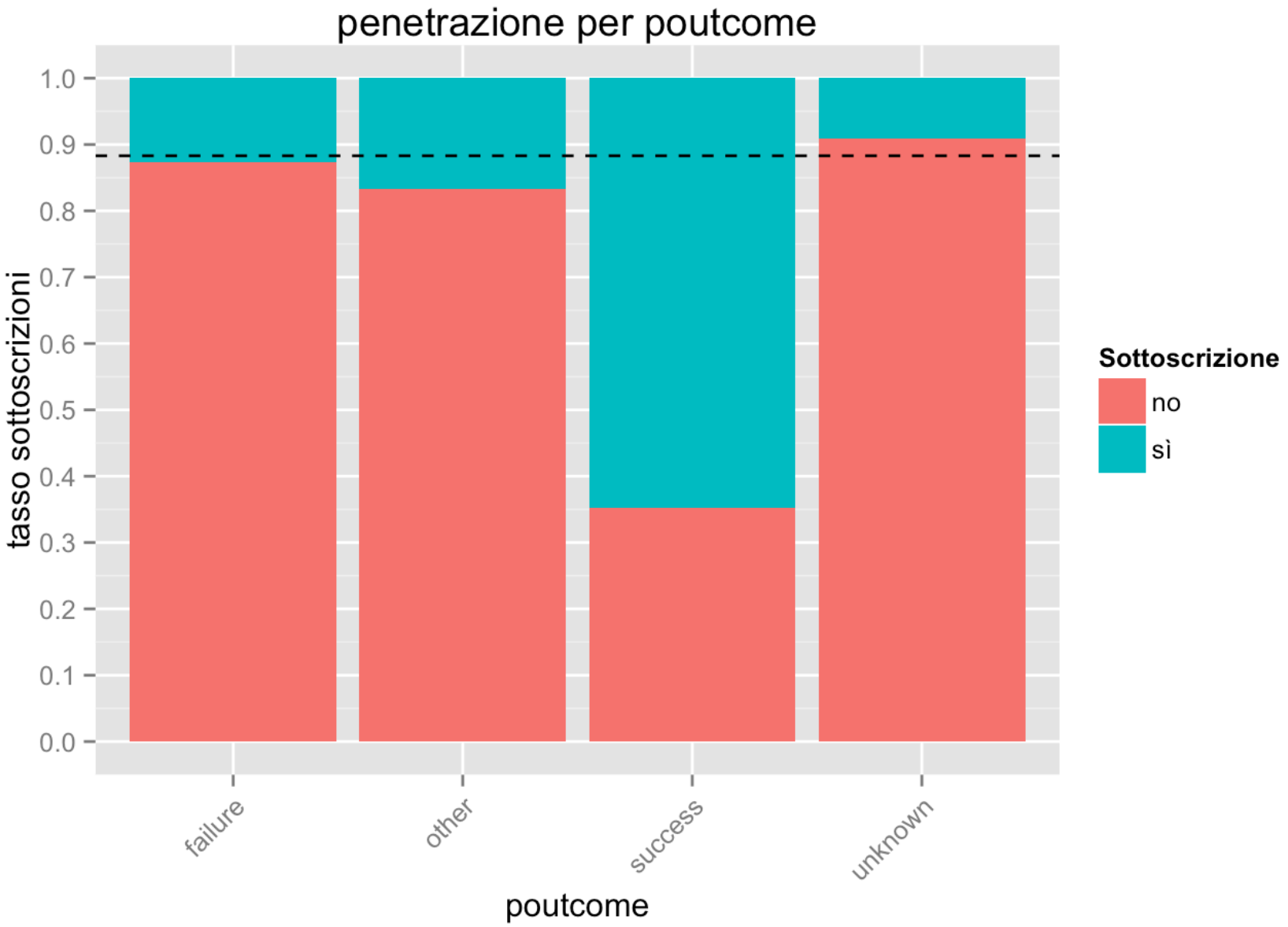


Gli “unknown” saranno senz’altro quell’80% di clienti mai contattati. Vediamo nelle quattro fascie come si sottoscrive:

```
t_poutcome_y <- bank0 %>%
        group_by (poutcome) %>%
        summarise (frequenza = n(), tasso_sottoscrizioni = mean(y=="yes")) %>%
        mutate(frequenza_relativa = frequenza / sum(frequenza)) %>%
        select(poutcome, frequenza, frequenza_relativa, tasso_sottoscrizioni) %>%
        arrange(desc(tasso_sottoscrizioni))
kable(t_poutcome_y, digits = 4, format = "markdown")
```

poutcome	frequenza	frequenza_relativa	tasso_sottoscrizioni
success	1511	0.0334	0.6473
other	1840	0.0407	0.1668
failure	4901	0.1084	0.1261
unknown	36959	0.8175	0.0916

```
g_poutcome_y  <- ggplot(bank0, aes(x = poutcome, fill = y)) +
        geom_bar(position = "fill") +
        geom_hline(yintercept = mean(bank0$y!="yes"), width = 2, col = "black", linetype
= 2) +
        ggtitle("penetrazione per poutcome") +
        ylab("tasso sottoscrizioni") +
        scale_y_continuous(breaks = seq(0, 1, 0.1)) +
        scale_fill_discrete(name="Sottoscrizione", labels=c("no", "sì")) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
g_poutcome_y
```



Se la precedente campagna ha avuto successo si sottoscrive in percentuali altissime. Persino se è stata fallimentare però si sottoscrive di più che se non si è mai stati contatatti; questo è coerente con l'analisi di `previous`, per cui il gruppo dei contattati sottoscriveva comunque di più dei mai contatatti.

```
poutcome_woe <- bank0 %>%
        select(poutcome, y) %>%
        group_by(poutcome) %>%
        summarise(n_no = sum(y == "no"), n_y = sum(y == "yes")) %>%
        mutate (perc_no = n_no / sum(n_no), perc_y = n_y / sum(n_y)) %>%
        select (starts_with("perc"))
poutcome_woe$woe <- log(poutcome_woe$perc_no / poutcome_woe$perc_y)
poutcome_IV <- sum((poutcome_woe$perc_no - poutcome_woe$perc_y) * poutcome_woe$woe)
poutcome_IV
```

```
## [1] 0.5146091
```

0.51, molto predittiva, ai limiti del sospetto.

Information values

Creiamo il vettore degli information values e lo rappresentiamo tramite grafico:

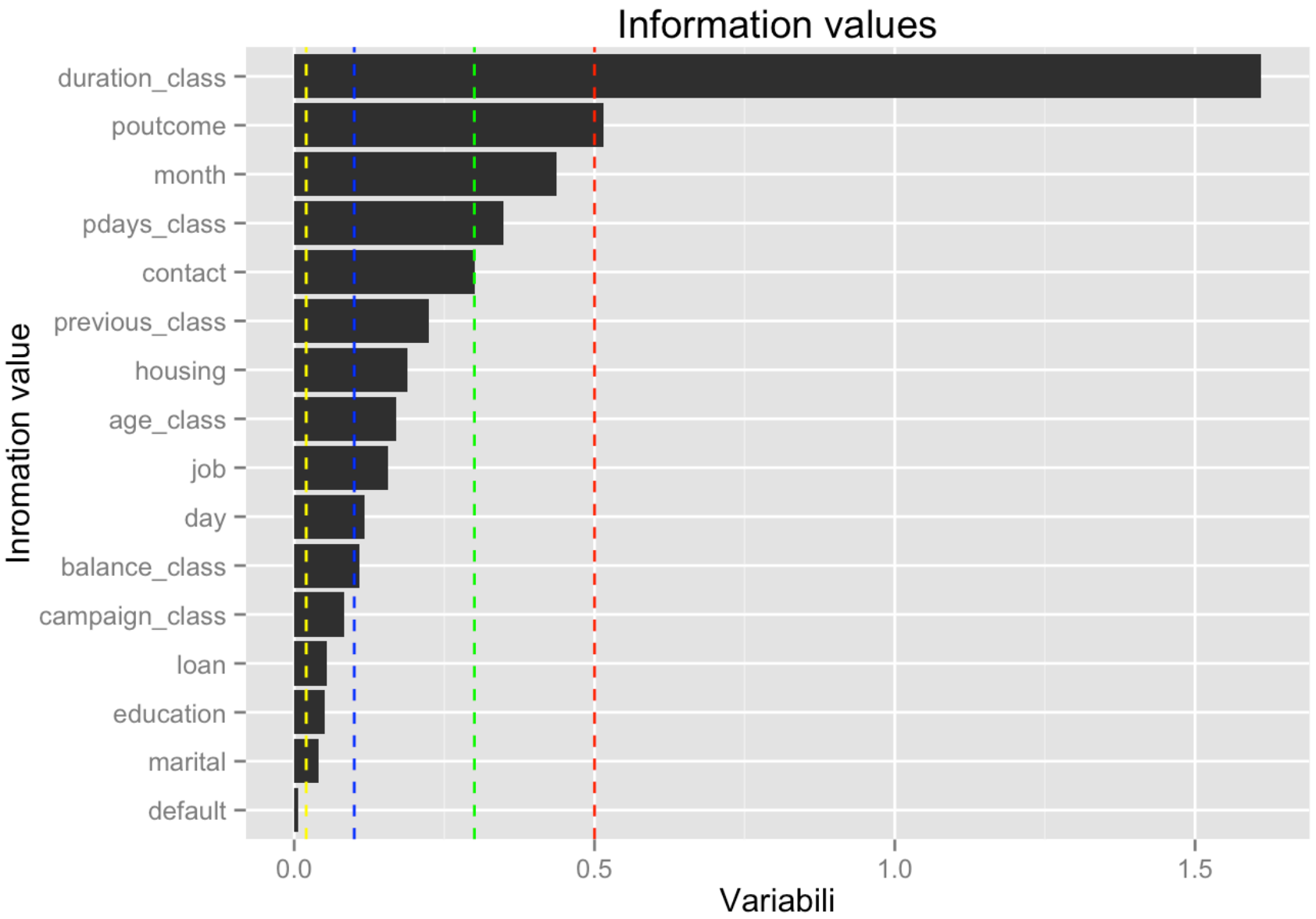
```
names(bank0)
```

```
##  [1] "age"           "job"           "marital"       "education"
##  [5] "default"       "balance"       "housing"       "loan"
##  [9] "contact"       "day"           "month"         "duration"
## [13] "campaign"      "pdays"         "previous"      "poutcome"
## [17] "y"             "age_class"     "balance_class" "duration_class"
## [21] "campaign_class" "pdays_class"   "previous_class"
```

```
IV <- c(age_class_IV, job_IV, marital_IV, education_IV, default_IV, balance_class_IV, ho
using_IV, loan_IV, contact_IV, day_IV, month_IV, duration_class_IV, campaign_class_IV, p
days_class_IV, previous_class_IV, poutcome_IV)
Variables <- c("age_class", "job", "marital", "education", "default", "balance_class", "
housing", "loan", "contact", "day", "month", "duration_class", "campaign_class", "pdays_
class", "previous_class", "poutcome")
t_IV_all <- data.frame(Variables, IV)
t_IV_all <- t_IV_all %>%
       arrange(desc(IV))
t_IV_all
```

```
##         Variables          IV
## 1  duration_class 1.610237410
## 2        poutcome 0.514609117
## 3           month 0.436131128
## 4     pdays_class 0.347928416
## 5         contact 0.300396103
## 6  previous_class 0.223489940
## 7         housing 0.188681483
## 8       age_class 0.170360068
## 9             job 0.155697294
## 10            day 0.117758343
## 11  balance_class 0.107971538
## 12 campaign_class 0.082932850
## 13           loan 0.054858527
## 14      education 0.050111946
## 15        marital 0.040126590
## 16        default 0.006256319
```

```
g_IV_all <- ggplot(t_IV_all, aes(x= reorder(Variables, IV), y= IV)) +
        geom_bar(stat='identity') +
        coord_flip() +
        ggtitle ("Information values") +
        ylab("Variabili") +
        xlab("Inromation value") +
        geom_hline(yintercept = c(0.02, 0.1, 0.3, 0.5), linewidth = 2, linetype = 2, col
= c("yellow", "blue", "green", "red"))
g_IV_all
```



Le 4 rette trattegiate indicano le soglie per: predittività assente (tra asse e retta gialla), debole (tra retta gialla e blue), media (tra retta blu e verde), forte (tra retta verde e rossa), sospetta (oltre retta rossa).