Vizualizacija marketinških podataka portugalske banke

Tina, Josip and Mateo

1/8/2022

OPIS PROJEKTA

U ovom projektu radimo analizu podataka dobivenih anketiranjem od strane jedne portugalske banke. Ispitanici su dali svoje osobne podatke poput godina, države, spola, posla, razine obrazovanja.., te podatke o kreditima, tj. imaju li stambeni i privatni kredit.

Cilj projekta je analizirati dani skup podataka, očistiti ga i pronaći regresore koje stvaraju dovoljno dobar model za naše kategorične podatke. Kategorični podatci koje imamo jesu: loan, housing i subscribed. Subscribed je varijabla od najvećeg značaja banci.

Subscribed je krajnji cilj jer loan i housing utječu na isti. Ukoliko pronađemo povezanost za navedene prve dvije varijable, moći ćemo točno znati kome usmjeriti naše marketing pozive i moći ćemo predvidjeti hoće li osoba na temelju danih podataka se ostaviti depozit u našoj banci ili neće.

U ovom dijelu radit ćemo vizualizaciju podataka, da vidimo čime raspolažemo i kako su ti podaci međusobno povezani. Prikazivat ćemo uglavnom grafovima zavisnost.

Opis skupa podataka

Za početak ćemo opisati ćime raspolažemo.

#ucitavanje paketa Učitajmo potrebne pakete

```
library(dplyr)
```

#Učitavanje podataka

Učitajmo podatke iz .csv file-a

```
BankData = read.csv("../arrangedData.csv")
dim(BankData)
```

```
## [1] 30372    15
```

Podaci se sastoje od 30 372 testiranih ljudi i 15 varijabli koje promatramo.

Popis varijabli koje promatramo:

```
names(BankData)
```

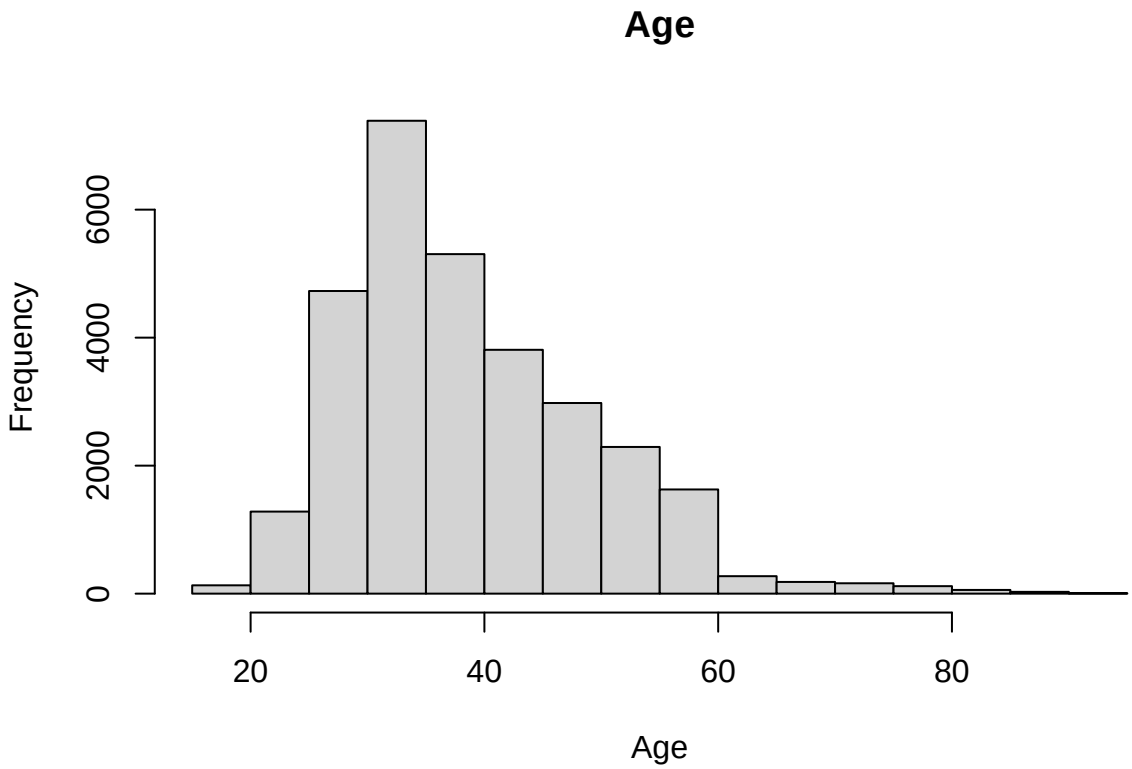
```
##  [1] "age"            "job"            "marital"        "education"
##  [5] "default"        "housing"        "loan"           "contact"
##  [9] "month"          "day_of_week"    "duration"       "campaign"
## [13] "cons.price.idx" "cons.conf.idx"  "subscribed"
```

Za testirane sudionike u tablici su navedeni njihovi podaci (godina, posao, stanje ženidbe, edukacija...) te podaci o kreditima... (stambeni i privatni)

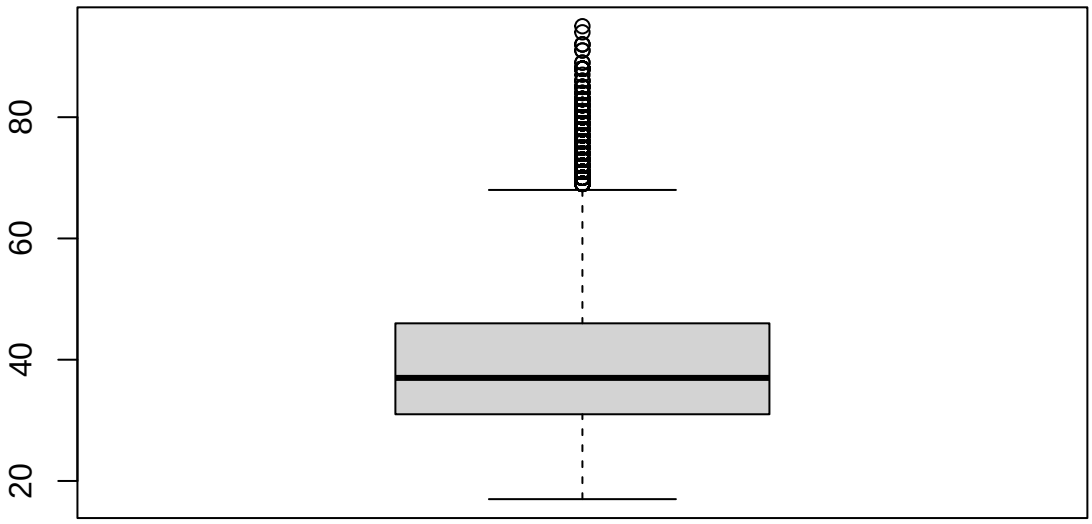
Promotrimo sada kako izgledaju varijable koje dobijemo kao informaciju od ispitanika (godine, posao, ženidbeno stanje) Gledat ćemo histograme da možemo vidjeti jesu li podaci barem približno normalni, te boxplot da vidimo prosjek i outliere.

AGE

```
hist(BankData$age,main='Age', xlab='Age', ylab='Frequency')
```



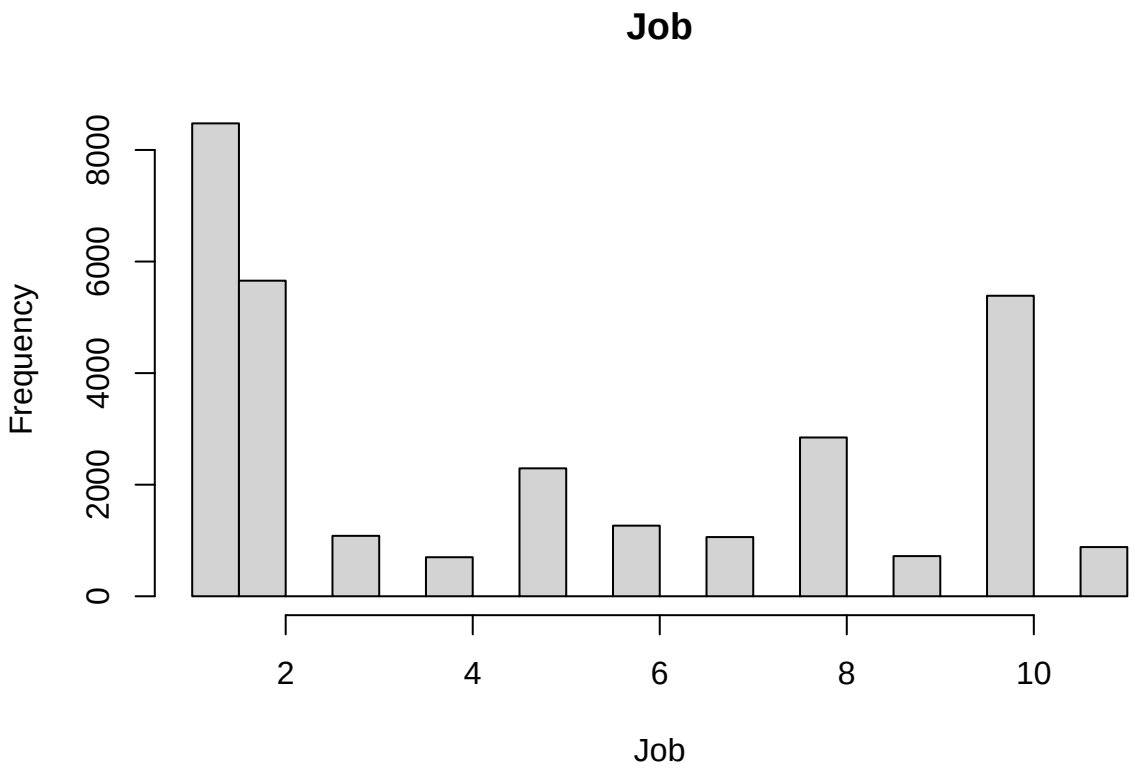

```
boxplot(BankData$age)
```



Vidimo da je prosjek malo ispod 40 godina te da ima nekoliko outliera, iako se većina kreće od 20 do 50 godina.

POSAO - admin:1 - bluecollar:2 - entrepreneur:3 - housemaid:4 - management:5 - retired:6 - self-employed:7 - services:8 - student:9 - technician:10 - unemployed:11

```
hist(BankData$job,main='Job', xlab='Job', ylab='Frequency')
```



Po ovom histogramu vidimo da je najviše zaposlenih u administraciji, a nakon toga u tehnologiji i fizičkim poslovima. Najmanje ispitanika su studenti i samozaposleni.

ŽENIDBENO STANJE - 1 = rastavljen / a - 2 = oženjen / udana - 3 = slobodan / slobodna

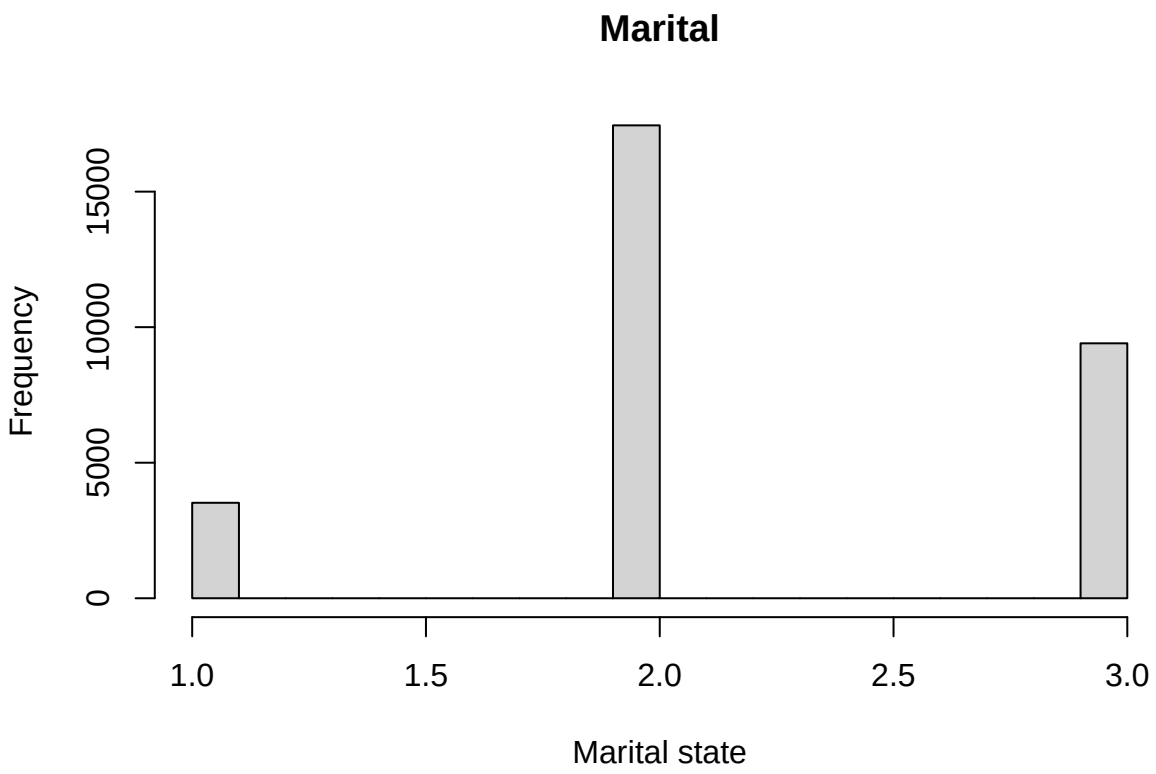
```
hist(BankData$marital,main='Marital', xlab='Marital state', ylab='Frequency', binwidth=12)
```

```
## Warning in plot.window(xlim, ylim, "", ...): "binwidth" is not a graphical
## parameter
```

```
## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## "binwidth" is not a graphical parameter
```

```
## Warning in axis(1, ...): "binwidth" is not a graphical parameter
```

```
## Warning in axis(2, ...): "binwidth" is not a graphical parameter
```

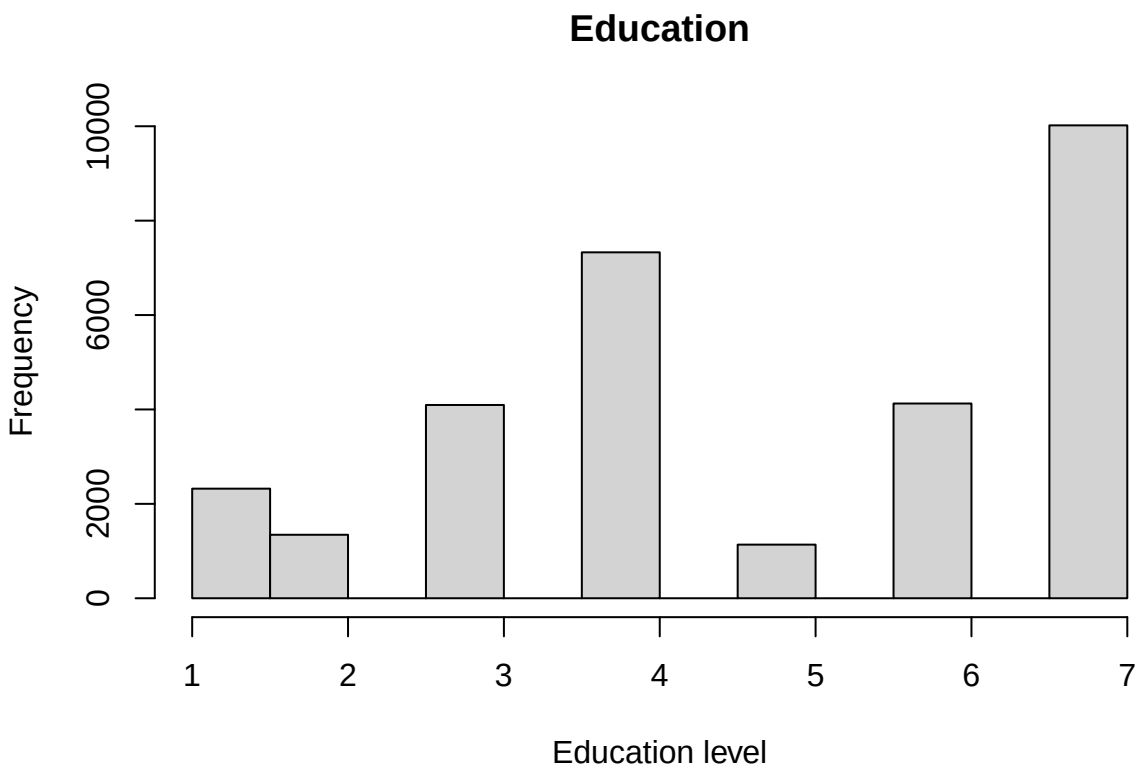


Vidimo da je najviše oženjenih, a najmanje rastavljenih među ispitinima.

EDUKACIJA

- basic.4y: 1
- basic.6y: 2
- basic.9y: 3
- high.school: 4
- illiterate: 5
- professional.course: 6
- university.degree: 7

```
hist(BankData$education,main='Education', xlab='Education level', ylab='Frequency')
```



Na danom histogramu vidimo da je najviše ispitanika završilo fakultet, a nakon toga barem srednju školu. Najmanje ih je nepismenih ili samo s osnovnom školom . Sada smo samo prikazali grafički neke osnovne podatke koje dobijemo za svakog ispitanika.

Sljedeće ćemo pokazati odnose svakih od tih svojstava s varijablama vezane za kredite (housing, loan, default) kao što smo rekli na početku.

DOB i KREDITI:

Podijelit ćemo dataset na mlađe, srednje mlade, te starije osobe i promatrati njihove kredite. Za mlađe smo uzeli osobe s 18 i manje godina, za srednje mlade one s 19 do 35, a za starije više od 35 godina starosti.

```
#View(BankData)

people_with_housing = BankData[which(BankData$housing == 1),]
people_without_housing = BankData[which(BankData$housing == 0),]
people_with_loan = BankData[which(BankData$loan == 1),]
people_without_loan = BankData[which(BankData$loan == 0),]
people_with_credit = BankData[which(BankData$default == 1),]
people_without_credit = BankData[which(BankData$default == 0),]
```

Također ćemo podatke podijeliti i po tome ima li osoba kredit ili ne.

```
young = BankData[which(BankData$age <= 25),]
middle = BankData[which(BankData$age > 25 & BankData$age <= 45),]
old = BankData[which(BankData$age > 45),]
```

Prosjeci po godinama za stambeni kredit:

```
cat('Prosječni postotak mladih ljudi koji imaju stambeni kredit ', mean(young$housing) * 100, '\n')
```

```
## Prosječni postotak mladih ljudi koji imaju stambeni kredit  55.02833
```

```
cat('Prosječni postotak srednje starih ljudi koji imaju stambeni kredit ', mean(middle$housing) * 100,
```

```
## Prosječni postotak srednje starih ljudi koji imaju stambeni kredit  53.73711
```

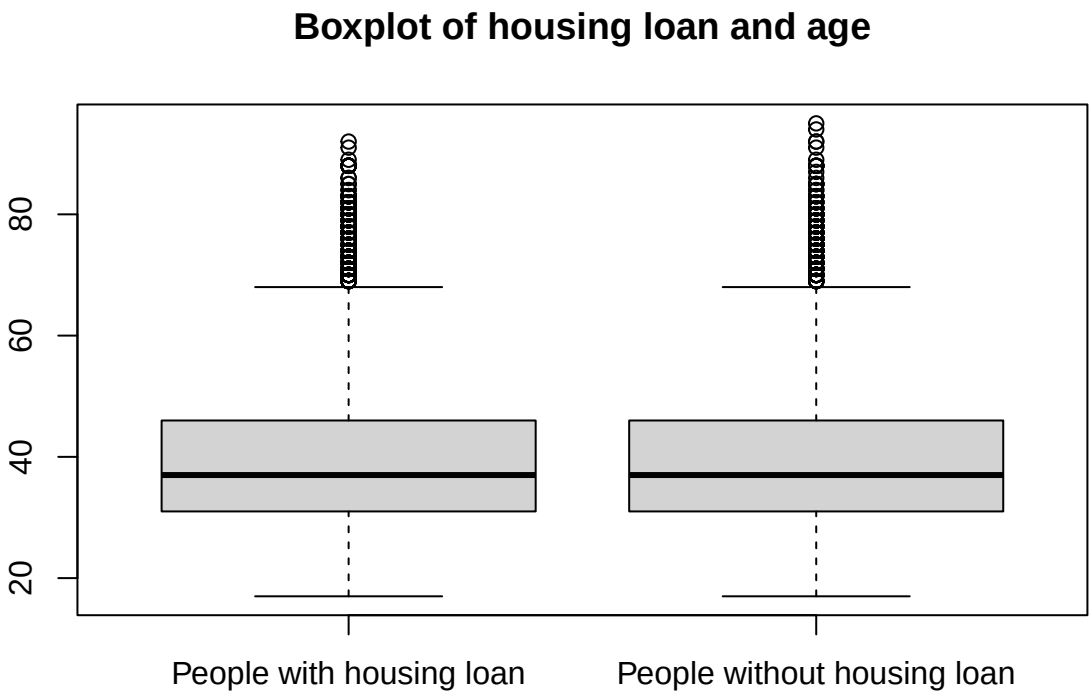
```
cat('Prosječni postotak starijih ljudi koji imaju stambeni kredit ', mean(old$housing) * 100, '\n')
```

```
## Prosječni postotak starijih ljudi koji imaju stambeni kredit  54.45839
```

Vidimo da najviše stambenog kredita imaju mladi ljudi, ali razlike nisu velike.

Boxplot starosti ljudi koji imaju stambeni kredit:

```
boxplot(people_with_housing$age, people_without_housing$age,
        names = c('People with housing loan','People without housing loan'),
        main = 'Boxplot of housing loan and age')
```



Prosjeci po godinama za privatni kredit:

```
cat('Prosječni postotak mladih ljudi koji imaju privatni kredit ', mean(young$loan) * 100, '\n')
```

```
## Prosječni postotak mladih ljudi koji imaju privatni kredit  15.72238
```

```
cat('Prosječni postotak srednje starih ljudi koji imaju privatni kredit ', mean(middle$loan) * 100, '\n
```

```
## Prosječni postotak srednje starih ljudi koji imaju privatni kredit  15.72081
```

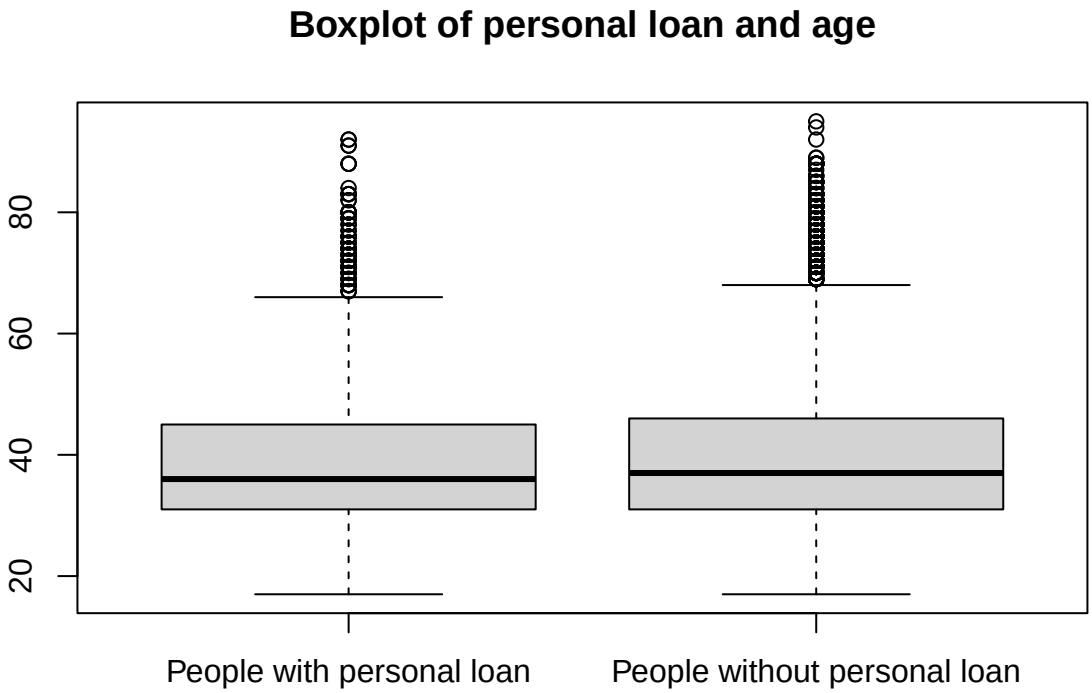
```
cat('Prosječni postotak starijih ljudi koji imaju privatni kredit ', mean(old$loan) * 100, '\n')
```

```
## Prosječni postotak starijih ljudi koji imaju privatni kredit  15.077
```

Za privatni kredit također možemo vidjeti da najviše imaju mlađi ljudi, ali nisu neke bitne razlike.

Boxplot starosti ljudi koji imaju privatni kredit

```
boxplot(people_with_loan$age, people_without_loan$age,
        names = c('People with personal loan','People without personal loan'),
        main = 'Boxplot of personal loan and age')
```



MARITAL STATUS

Podjela ispitanika ovisno o statusu ženidbe:

```
married_people = BankData[which(BankData$marital == 2),]
divorced_and_single_people = BankData[which(BankData$marital == 1 | BankData$marital == 3),]
#single_people = BankData[which(BankData$marital == 3),]
```

Prosjeci ljudi koji imaju privatni kredit:

```
cat('Prosjek oženjenih ljudi koji imaju privatni kredit ', mean(married_people$loan) * 100, '\n')
```

```
## Prosjek oženjenih ljudi koji imaju privatni kredit  15.66065
```

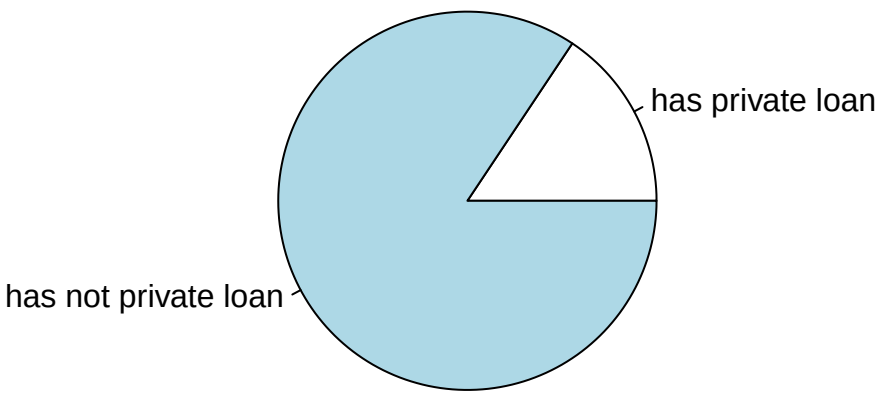
```
cat('Prosjek neoženjenih ljudi koji imaju privatni kredit ', mean(divorced_and_single_people$loan) * 10
```

```
## Prosjek neoženjenih ljudi koji imaju privatni kredit  15.41734
```

Pie chart za ženidbeni status i privatni kredit:

```
x <- c(15.66, 84.44)
labels <- c("has private loan", "has not private loan")

pie(x, labels, main = "Pie chart of married people")
```



Prosjeci ljudi koji imaju stambeni kredit:

```
cat('Prosjek oženjenih ljudi koji imaju stambeni kredit ', mean(married_people$housing) * 100, '\n')
```

```
## Prosjek oženjenih ljudi koji imaju stambeni kredit  53.63715
```

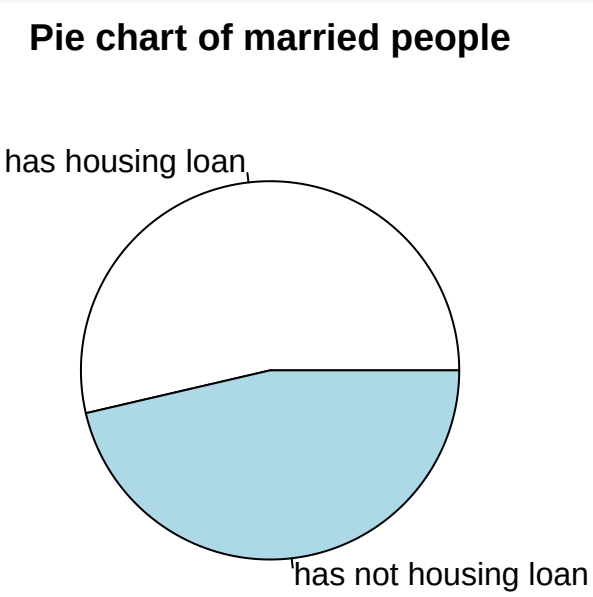
```
cat('Prosjek neoženjenih ljudi koji imaju stambeni kredit ', mean(divorced_and_single_people$housing) *
```

```
## Prosjek neoženjenih ljudi koji imaju stambeni kredit  54.44419
```

Pie chart za ženidbeni status i stambeni kredit:

```
x <- c(53.64, 46.36)
labels <- c("has housing loan", "has not housing loan")

pie(x, labels, main = "Pie chart of married people")
```



EDUKACIJA

Podjela ispitanika u ovisnosti o završenom statusu edukacije:

```
primary_school = BankData[which(BankData$education == 1 | BankData$education == 2 | BankData$education =
high_school = BankData[which(BankData$education == 4),]
university = BankData[which(BankData$education == 7),]
illiterate = BankData[which(BankData$education == 5),]
```

Pirvatni kredit u ovisnosti o edukaciji:

```
cat('Prosjek ljudi sa završenom samo osnovnom školom koji imaju privatni kredit ', mean(primary_school$l
```

```
## Prosjek ljudi sa završenom samo osnovnom školom koji imaju privatni kredit  15.15073
```

```
cat('Prosjek ljudi sa završenom srednjom školom koji imaju privatni kredit ', mean(high_school$loan) * 1
```

```
## Prosjek ljudi sa završenom srednjom školom koji imaju privatni kredit  15.51371
```

```
cat('Prosjek ljudi sa završenim fakulteteom koji imaju privatni kredit ', mean(university$loan) * 100,
```

```
## Prosjek ljudi sa završenim fakulteteom koji imaju privatni kredit  16.13934
```

```
cat('Prosjek nepismenih ljudi koji imaju privatni kredit ', mean(illiterate$loan) * 100, '\n')
```

```
## Prosjek nepismenih ljudi koji imaju privatni kredit  14.7007
```

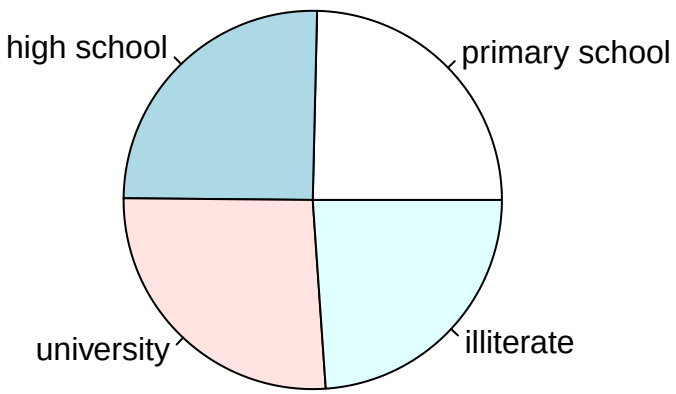
Vidimo da najveći privatni kredit imaju najobrazovaniji, tj sa završenim fakultetom, a najmanje nepismeni (uzmimo u obzir da je najmanji broj ispitanika nepismeno).

Pie chart prikaz:

```
mean = mean(15.15073, 15.51371 , 16.13934, 14.7007 )

x <- c(15.15073/mean, 15.51371/mean, 16.13934/mean, 14.7007/mean)
labels <- c("primary school", "high school", "university", "illiterate")

pie(x, labels, main = "Pie chart of of people having private loan")
```



Stambeni kredit u ovisnosti o edukaciji:

```
cat('Prosjek ljudi sa završenom samo osnovnom školom koji imaju stambeni kredit ', mean(primary_school$l
```

```
## Prosjek ljudi sa završenom samo osnovnom školom koji imaju stambeni kredit  53.24659
```

```
cat('Prosjek ljudi sa završenom srednjom školom koji imaju stambeni kredit ', mean(high_school$housing)
```

```
## Prosjek ljudi sa završenom srednjom školom koji imaju stambeni kredit  52.84486
```

```
cat('Prosjek ljudi sa završenim fakulteteom koji imaju stambeni kredit ', mean(university$housing)* 100
```

```
## Prosjek ljudi sa završenim fakulteteom koji imaju stambeni kredit  54.83581
```

```
cat('Prosjek nepismenih ljudi koji imaju stambeni kredit ', mean(illiterate$housing)* 100, '\n')
```

```
## Prosjek nepismenih ljudi koji imaju stambeni kredit  52.02465
```

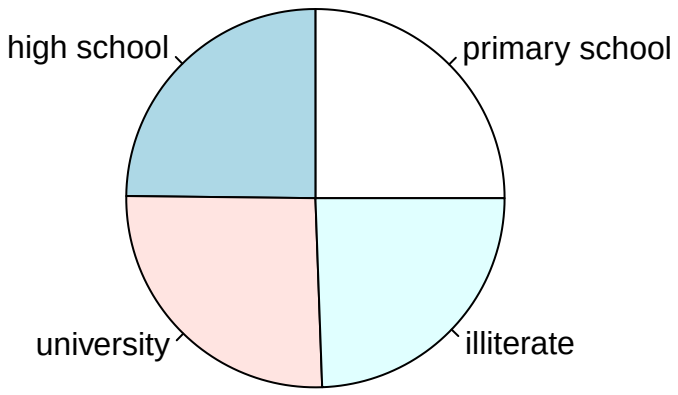
Iz ispisivanja srednjih vrijednosti možemo vidjeti da je stambeni kredit najmanje uzimalo nepismenih ljudi (kojih je inače jako malo u našem skupu podataka), a najviše je u stambenom kreditu ljudi s završenim fakultetom. Kad razmislimo to ima smisla jer ljudi s fakuletetom vjerojatno imaju, prema prosjecima, najveće sigurne plaće i isplaniraju isplatu tog kredita jer je njihov posao siguran.

Pie chart:

```
mean = mean(53.24659, 52.84486 , 54.83581, 52.02465 )

x <- c(53.24659/mean, 52.84486/mean, 54.83581/mean, 52.02465/mean)
labels <- c("primary school", "high school", "university", "illiterate")

pie(x, labels, main = "Pie chart of of people having housing loan")
```



Na grafu vidimo da te razlike nisu velike.. vidljive su ali ne toliko..

Sada smo provjerili zavisnost većine varijabli (edukacija, dob, ženidbeni status) o kreditima (housing i loan). Nakon toga pogledat ćemo ovisnost tih varijabli o glavnoj koju promatramo, a to je output varijabla subscribed, tj varijabla koja nam govori je li se ispitanik na kraju pretplatio, tj. hoće li u buduće ostavljati novce u banci..

Gledat ćemo ovisnost do sada promatranih varijabli o tome je li se ispitanik *pretplatio*.

Za početak ćemo podijeliti dataset na one koji su se prijavili i one koji nisu:

```
subscribed = BankData[which(BankData$subscribed == 1),]
not_subscribed = BankData[which(BankData$subscribed == 0),]
```

GODINE i PRETPLAĆIVANJE

Gledamo prosjeke pretplaćenih ovisno o godinama:

```
cat('Prosjek pretplaćenih mladih ljudi ', mean(young$subscribed) * 100, '\n')
```

```
## Prosjek pretplaćenih mladih ljudi  23.22946
```

```
cat('Prosjek pretplaćenih srednje starih ljudi ', mean(middle$subscribed) * 100, '\n')
```

```
## Prosjek pretplaćenih srednje starih ljudi  11.93425
```

```
cat('Prosjek pretplaćenih starijih ljudi ', mean(old$subscribed) * 100, '\n')
```

```
## Prosjek pretplaćenih starijih ljudi  16.03468
```

Vidimo da pretplaćivanje ovisi o dobi.. tj. da najviše mladih ljudi "postane dio banke". To ima smisla jer vjerojatno velik broj starijih ljudi već ima račune u nekoj drugoj banci..

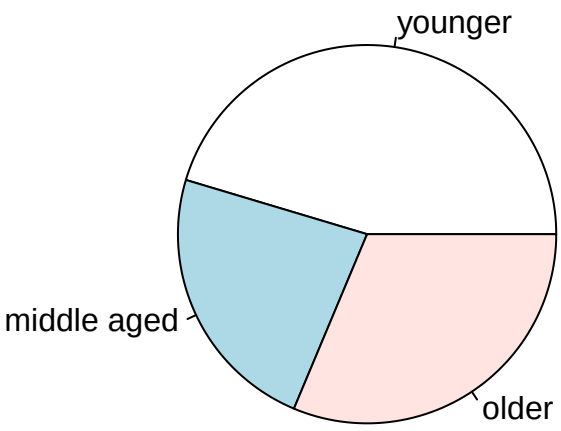
Pie chart za pretplaćene ovisno o godinama:

```
mean = mean(23.22946, 11.93425, 16.03468)

x <- c(23.22946/mean, 11.93425/mean, 16.03468/mean)

labels <- c("younger", "middle aged", "older")

pie(x, labels, main = "Pie chart of subscribed people")
```



Ovdje je također očito da je najviše pretplaćeno mladih ljudi.

EDUKACIJA i PRETPLAĆIVANJE

Gledamo prosjeke pretplaćenih ovisno o edukaciji:

```
cat('Prosječno pretplaćenih ljudi sa završenom osnovnom školom ', mean(primary_school$subscribed) * 100
```

```
## Prosječno pretplaćenih ljudi sa završenom osnovnom školom  10.95079
```

```
cat('Prosječno pretplaćenih ljudi sa završenom srednjom školom ', mean(high_school$subscribed) * 100, '
```

```
## Prosječno pretplaćenih ljudi sa završenom srednjom školom  12.77118
```

```
cat('Prosječno pretplaćenih ljudi sa završenim fakultetom ', mean(university$subscribed) * 100, '\n')
```

```
## Prosječno pretplaćenih ljudi sa završenim fakultetom  15.55045
```

```
cat('Prosječno pretplaćenih nepismenih ljudi ', mean(illiterate$subscribed) * 100, '\n')
```

```
## Prosječno pretplaćenih nepismenih ljudi  19.19014
```

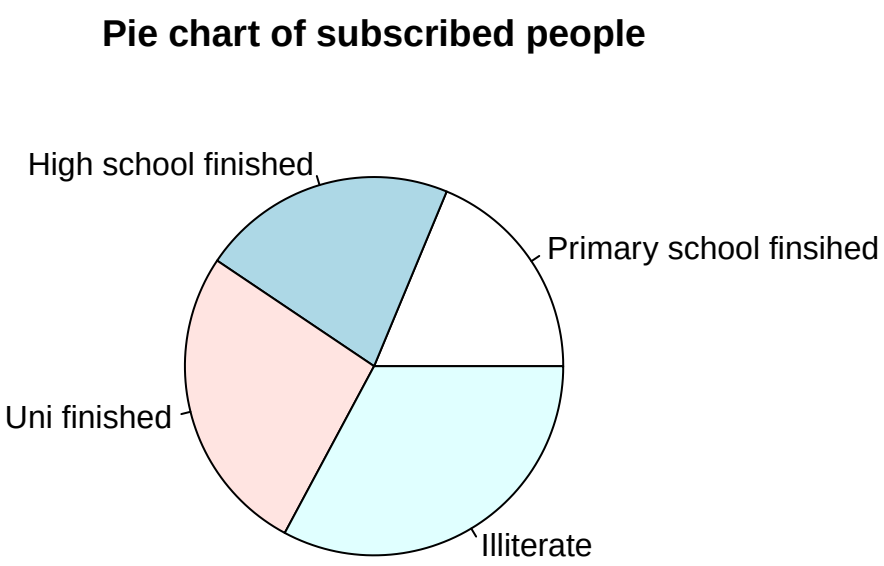
Vidimo da je pretplaćeno najviše nepismenih ljudi (uzmimo u obzir da je njih za razliku od ostalih jako malo testirano), a najmanje onih koji su završili samo osnovnu školu.

Pie chart za edukaciju i pretplaćenost:

```
mean = mean(10.95079, 12.77118, 15.55045, 19.19014)
x <- c(10.95079/mean, 12.77118/mean, 15.55045/mean, 19.19014/mean)

labels <- c("Primary school finsihed", "High school finished", "Uni finished", "Illiterate")

pie(x, labels, main = "Pie chart of subscribed people")
```



Pie chart samo opisuje ovo napisno gore.

ŽENIDBENI STATUS i PRETPLAĆIVANJE

Sada gledamo prosjek pretplaćenih ovisno o bračnom statusu:

```
cat('Prosjek preplaćenih ljudi koji su oženjeni ', mean(married_people$subscribed) * 100, '\n')
```

```
## Prosjek preplaćenih ljudi koji su oženjeni  12.43336
```

```
cat('Prosjek preplaćenih ljudi koji su neoženjeni ili rastavljeni ', mean(divorced_and_single_people$su
```

```
## Prosjek preplaćenih ljudi koji su neoženjeni ili rastavljeni  14.94546
```

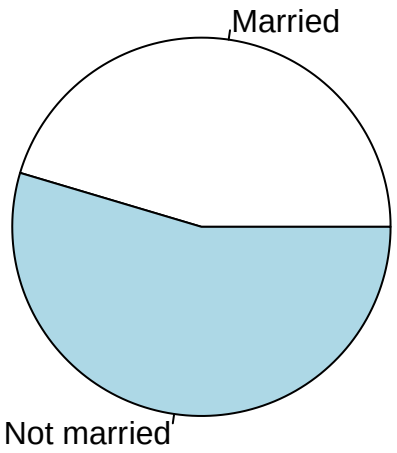
Vidimo da je više pretplaćeno ljudi koji su neoženjeni ili rastavljeni.

Pie chart ženidbenog statusa i pretplaćivanja:

```
x <- c(12.43, 14.95)
labels <- c("Married", "Not married")

pie(x, labels, main = "Pie chart of subscribed people")
```

Pie chart of subscribed people



Sada smo napravili i analizu pretplaćenih i nepretplaćenih ljudi ovisno o godinama, edukaciji i ženidbenom statusu i vidjeli neke potencijalne zaključke, ali to ćemo detaljnije vidjeti u kodu. Vizualizacija služi samo za bolje razumijevanje podataka i njihovog prikaza.