Projekt

Ivona Mustapić-Jogun

Podatci opisuju statistiku putnika u međunarodnoj zračnoj luci San Francisco. Svaki redak odnosi se na putnika zračne luke (ukupno 15007).

Varijable koje se koriste:
-Activity.Period : int
(godina i mjesec aktivnosti putnika)
-Operating.Airline : Factor 77 levels “Aer Lingus”,“Aeromexico”,...
(naziv zrakoplovne tvrtke za operatora zrakoplova za aktivnost putnika)
-Operating.Airline.IATA.Code: Factor 74 levels "“,”4T“,”5Y“,”9W“,...
(dvočlana oznaka za Operating.Airline)
-Published.Airline : Factor 68 levels”Aer Lingus“,”Aeromexico“,...
(naziv zrakoplovne tvrtke koja izdaje kartu i bilježi prihod za aktivnost putnika)
-Published.Airline.IATA.Code: Factor 65 levels”“,”4T“,”5Y“,”9W“,...
(dvočlana oznaka za Published.Airline)
-GEO.Summary : Factor 2 levels”Domestic“,”International"
(kategorizacija s obzirom na to je li se aktivnost putnika odvijala unutar United States ili ne)
-GEO.Region : Factor 9 levels “Asia”,“Australia / Oceania”,“Canada”,“Central America”,“Europe”,“Mexico”,“Middle East”,“South America”,“US”
(malo detaljnije od GEO.Summary, s obzirom na regije)
-Activity.Type.Code : Factor 3 levels “Deplaned”,“Enplaned”,“Thru / Transit”
(fizička aktivnost putnika, to jest ukrcaj na let, iskrcaj s leta ili prelazak kroz zračnu luku kako bi došli na drugu lokaciju, ali na isti zrakoplov i isti broj leta)
-Price.Category.Code : Factor 2 levels “Low Fare”,“Other”
(kategorizacija s obzirom na to je li zrakoplovna tvrtka niskobudžetna ili ne)
-Terminal : Factor 5 levels “International”,“Other”,“Terminal 1”,“Terminal 2”,“Terminal 3”
(oznaka terminala gdje se dogodila aktivnost putnika)
-Boarding.Area : Factor 8 levels “A”,“B”,“C”,“D”,“E”,“F”,“G”,“Other”
(oznaka područja gdje se dogodila aktivnost putnika, ovo je podskup terminala; npr. Terminal 1 sadrži područja B i C)
-Passenger.Count : int
(broj putnika s atributima od gore u mjesecu)
-Adjusted.Activity.Type.Code: Factor 3 levels “Deplaned”,“Enplaned”,“Thru / Transit * 2”
(pilagođena fizička aktivnost putnika kako bi se ukupan broj putnika izračunao po formuli Enplaned+Deplaned+Thru/Transit*2)*
-Adjusted.Passenger.Count : int
*(broj putnika po formuli Enplaned+Deplaned+Thru/Transit*2)
-Year : int
(godina aktivnosti putnika)
-Month : Factor 12 levels “January”,“February”,“March”,...
(mjesec aktivnosti putnika)

Cilj ovog projekta je vidjeti povezanost varijable GEO.Summary s ostalim čimbenicima ove zračne luke.

Za početak pogledajmo neke osnovne značajke.

```
setwd("C:/Users/Jogun/Desktop/rudarenje podataka/projekt")
podatci <- read.csv("Air_Traffic_Passenger_Statistics.csv")
head(podatci)
```

```
##   Activity.Period Operating.Airline Operating.Airline.IATA.Code
## 1          200507      ATA Airlines                          TZ
## 2          200507      ATA Airlines                          TZ
## 3          200507      ATA Airlines                          TZ
## 4          200507       Air Canada                           AC
## 5          200507       Air Canada                           AC
## 6          200507         Air China                          CA
##   Published.Airline Published.Airline.IATA.Code   GEO.Summary GEO.Region
## 1      ATA Airlines                          TZ      Domestic         US
## 2      ATA Airlines                          TZ      Domestic         US
## 3      ATA Airlines                          TZ      Domestic         US
## 4       Air Canada                           AC International     Canada
## 5       Air Canada                           AC International     Canada
## 6         Air China                          CA International       Asia
##   Activity.Type.Code Price.Category.Code      Terminal Boarding.Area
## 1           Deplaned            Low Fare    Terminal 1             B
## 2           Enplaned            Low Fare    Terminal 1             B
## 3    Thru / Transit            Low Fare    Terminal 1             B
## 4           Deplaned               Other    Terminal 1             B
## 5           Enplaned               Other    Terminal 1             B
## 6           Deplaned               Other International             G
##   Passenger.Count Adjusted.Activity.Type.Code Adjusted.Passenger.Count Year
## 1           27271                   Deplaned                    27271 2005
## 2           29131                   Enplaned                    29131 2005
## 3            5415          Thru / Transit * 2                    10830 2005
## 4           35156                   Deplaned                    35156 2005
## 5           34090                   Enplaned                    34090 2005
## 6            6263                   Deplaned                     6263 2005
##   Month
## 1  July
## 2  July
## 3  July
## 4  July
## 5  July
## 6  July
```

```
attach(podatci)
str(podatci)
```

```
## 'data.frame':    15007 obs. of  16 variables:
##  $ Activity.Period            : int  200507 200507 200507 200507 200507 200507 200507 200507 200507 1
##  $ Operating.Airline          : Factor w/ 77 levels "Aer Lingus","Aeromexico",..: 18 18 18 4 4 6 6 7
##  $ Operating.Airline.IATA.Code: Factor w/ 74 levels "","4T","5Y","9W",..: 62 62 62 8 8 17 17 9 9 48
##  $ Published.Airline          : Factor w/ 68 levels "Aer Lingus","Aeromexico",..: 16 16 16 4 4 5 5 6
##  $ Published.Airline.IATA.Code: Factor w/ 65 levels "","4T","5Y","9W",..: 56 56 56 8 8 17 17 9 9 45
##  $ GEO.Summary                : Factor w/ 2 levels "Domestic","International": 1 1 1 2 2 2 2 2 2 .
##  $ GEO.Region                 : Factor w/ 9 levels "Asia","Australia / Oceania",..: 9 9 9 3 3 1 1 5 5
##  $ Activity.Type.Code         : Factor w/ 3 levels "Deplaned","Enplaned",..: 1 2 3 1 2 1 2 1 2 1 ...
```

```
##  $ Price.Category.Code        : Factor w/ 2 levels "Low Fare","Other": 1 1 1 2 2 2 2 2 2 2 ...
##  $ Terminal                   : Factor w/ 5 levels "International",..: 3 3 3 3 3 1 1 1 1 1 ...
##  $ Boarding.Area              : Factor w/ 8 levels "A","B","C","D",..: 2 2 2 2 2 7 7 1 1 7 ...
##  $ Passenger.Count            : int  27271 29131 5415 35156 34090 6263 5500 12050 11638 4998 ...
##  $ Adjusted.Activity.Type.Code: Factor w/ 3 levels "Deplaned","Enplaned",..: 1 2 3 1 2 1 2 1 2 1 ...
##  $ Adjusted.Passenger.Count   : int  27271 29131 10830 35156 34090 6263 5500 12050 11638 4998 ...
##  $ Year                       : int  2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...
##  $ Month                      : Factor w/ 12 levels "April","August",..: 6 6 6 6 6 6 6 6 6 6 ...
```

```
dim(podatci)
```

```
## [1] 15007    16
```

Paketi koji se koriste u ovom projektu:
-RWeka
-party
-rpart
-cluster
-nnet

```
library(RWeka)
```

```
## Warning: package 'RWeka' was built under R version 3.6.3
```

```
library(cluster)
library(nnet)
```

```
## Warning: package 'nnet' was built under R version 3.6.3
```

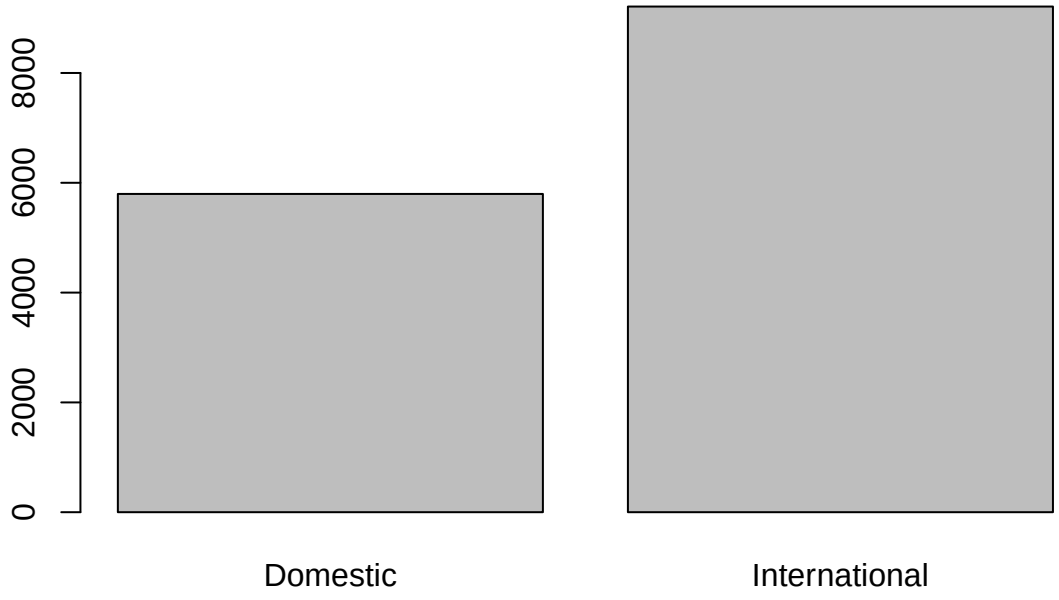
Neke varijable sam zanemarila jer su nepotrebne. To su Operating.Airline, Operating.Airline.IATA.Code, Published.Airline.IATA.Code, Adjusted.Activity.Type.Code i Adjusted.Passenger.Count.

```
podatci1<-podatci[c(1,4,6:12,15,16)]
head(podatci1)
```

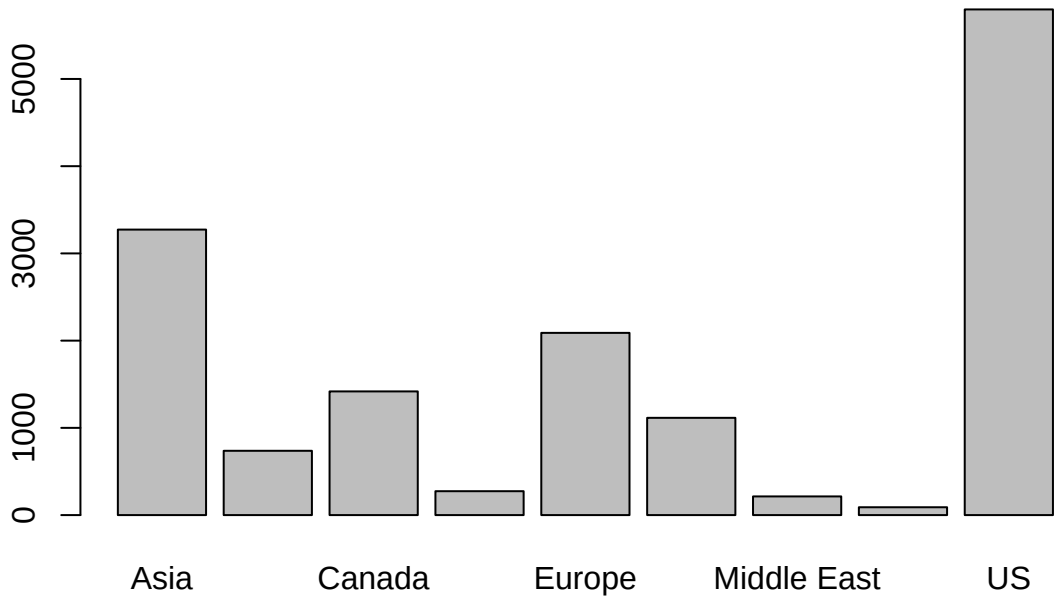
```
##   Activity.Period Published.Airline   GEO.Summary GEO.Region Activity.Type.Code
## 1          200507      ATA Airlines      Domestic         US           Deplaned
## 2          200507      ATA Airlines      Domestic         US           Enplaned
## 3          200507      ATA Airlines      Domestic         US     Thru / Transit
## 4          200507       Air Canada  International     Canada           Deplaned
## 5          200507       Air Canada  International     Canada           Enplaned
## 6          200507         Air China International       Asia           Deplaned
##   Price.Category.Code      Terminal Boarding.Area Passenger.Count Year Month
## 1            Low Fare    Terminal 1             B           27271 2005  July
## 2            Low Fare    Terminal 1             B           29131 2005  July
## 3            Low Fare    Terminal 1             B            5415 2005  July
## 4               Other    Terminal 1             B           35156 2005  July
## 5               Other    Terminal 1             B           34090 2005  July
## 6               Other International             G            6263 2005  July
```

Odlučila sam varijable prikazati pomoću barplotova kako bih vidjela neka njihova svojstva. Uočila sam da je:
-više putnika s internacionalnih letova.
-najviše putnika putovalo unutar US-a.
-približno jednako putnika koji se iskrcavaju i ukrcavaju dok je jako mali broj putnika koji prolaze kroz zračnu luku i ostaju u istom zrakoplovu i istom broju leta što je sasvim logično.
-više putnika koji leti zrakoplovnim tvrtkama koje nisu niskobudžetne.
-najviše putnika na terminalu za internacionalne letove.
-najviše putnika na području A. To je područje internacionalnog terminala. Nakon njega najviše putnika ima na području G koje zajedno s područjem A čini internacionalni terminal. Ostala područja imaju znatno manji broj putnika pa iz ovoga možemo zaključiti kako bi zračna luka trebala drukčije rasporediti upotrebu područja, odnosno terminala. Npr. koristeći dva terminala za internacionalne letove.
-za 2005. i 2016. godinu manje putnika no to je zato što imamo podatke od 07.2005. do 03.2016., to jest za 2005. i 2016. godinu nemamo sve mjesece.
-zastupljenost mjeseci prilbližno jednaka. Najmanji broj putnika imaju 4., 5. i 6. mjesec koji nisu zastupljeni onoliko puta koliko i svi ostali.

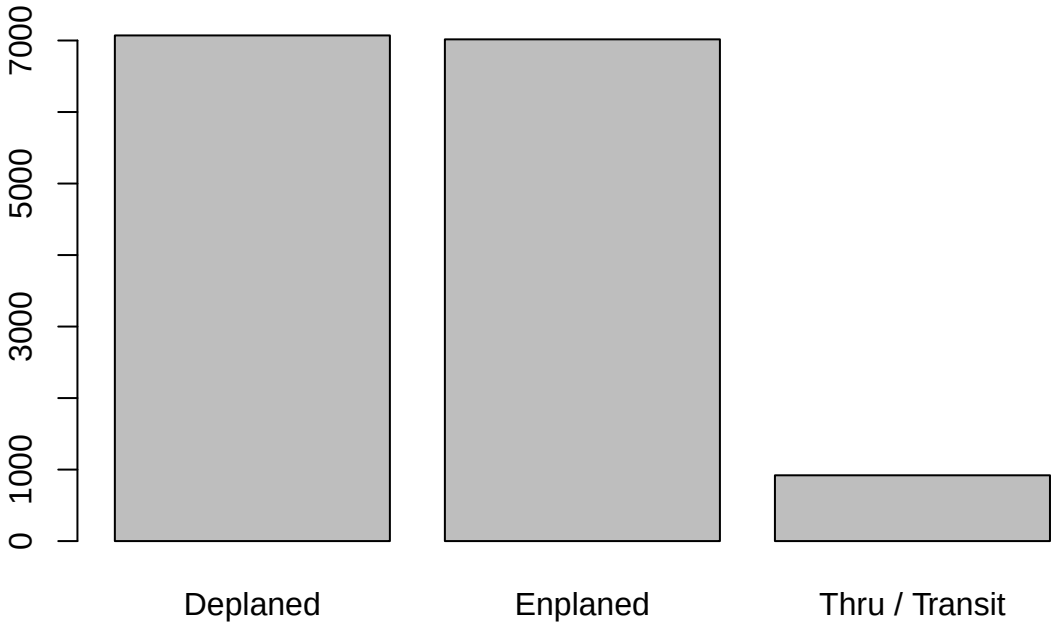
```
barplot(table(podatci$GEO.Summary))
```



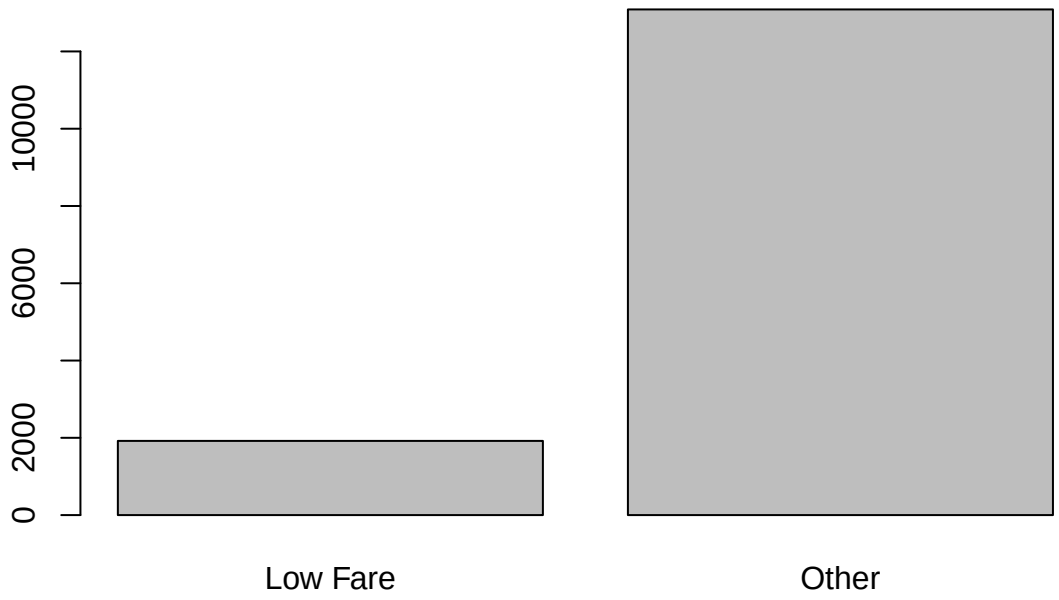

```
barplot(table(podatci$GEO.Region))
```



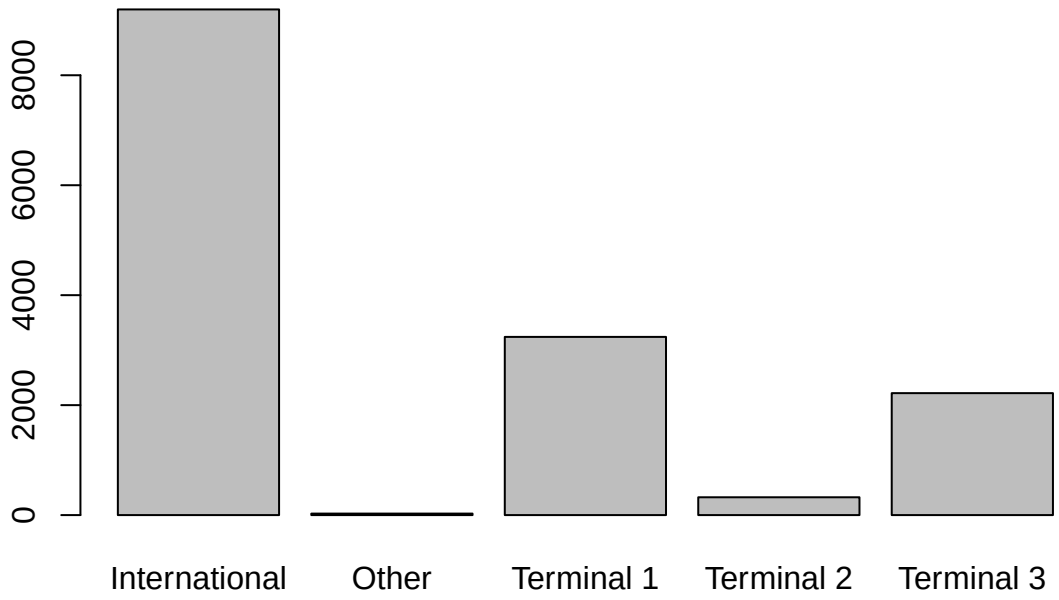

```
barplot(table(podatci$Activity.Type.Code))
```



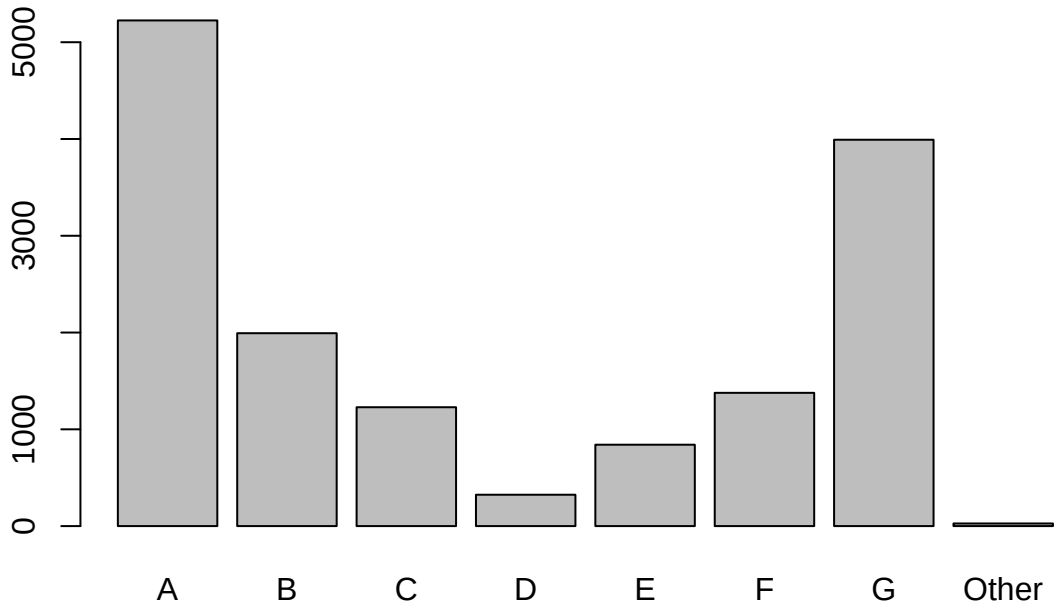

```
barplot(table(podatci$Price.Category.Code))
```



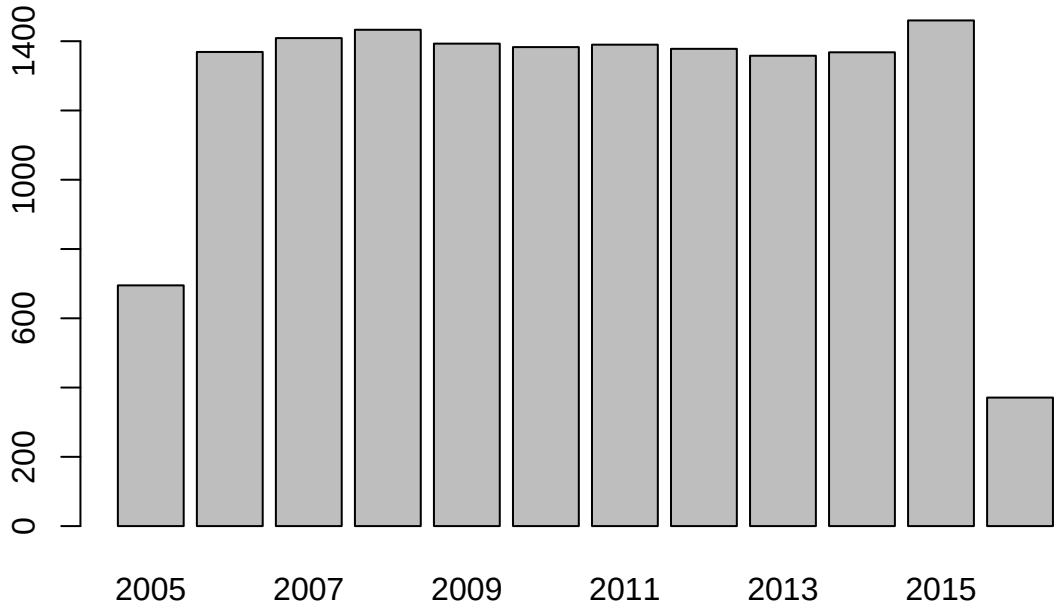

```
barplot(table(podatci$Terminal))
```



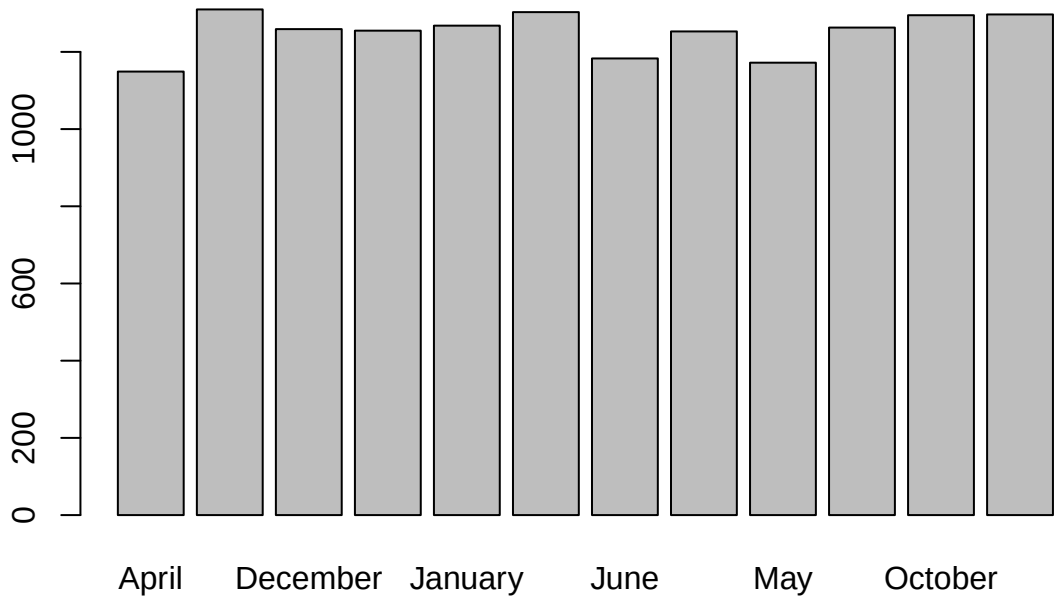

```
barplot(table(podatci$Boarding.Area))
```




```
barplot(table(podatci$Year))
```




```
barplot(table(podatci$Month))
```



Sada sam odlučila zanemariti i varijable Year i Month jer mi neće više biti potrebne, a njihovu vrijednost imam u zajedničkoj varijabli Activity.Period.

```
podatci1<-podatci1[c(-10,-11)]
head(podatci1)
```

```
##   Activity.Period Published.Airline   GEO.Summary GEO.Region Activity.Type.Code
## 1          200507      ATA Airlines      Domestic         US           Deplaned
## 2          200507      ATA Airlines      Domestic         US           Enplaned
## 3          200507      ATA Airlines      Domestic         US     Thru / Transit
## 4          200507       Air Canada  International     Canada           Deplaned
## 5          200507       Air Canada  International     Canada           Enplaned
## 6          200507         Air China International       Asia           Deplaned
##   Price.Category.Code      Terminal Boarding.Area Passenger.Count
## 1            Low Fare    Terminal 1             B           27271
## 2            Low Fare    Terminal 1             B           29131
## 3            Low Fare    Terminal 1             B            5415
## 4               Other    Terminal 1             B           35156
## 5               Other    Terminal 1             B           34090
## 6               Other International             G            6263
```

Activity.Period i Passenger.Count nisu ovisne, to jest broj putnika ne ovisi o vremenu u godini.

```
cor(podatci[c(1,12)])
```

```
##                  Activity.Period Passenger.Count
## Activity.Period      1.00000000      0.06031063
## Passenger.Count      0.06031063      1.00000000
```

Sada dolazimo do glavnog dijela.

U sljedećem dijelu prikazat ću stabla s obzirom na GEO.Summary izbacivanjem varijabli koje su sigurno povezane s tom varijablom i vidjeti kakvo će razvrstavanje biti nakon svih tih izbacivanja. Koristit ću metodu J48. Upotrijebit ću ju prvo na cijelom skupu podataka koji sada imam, zatim na skupu podataka bez varijable GEO.Region, nakon toga bez varijabla GEO.Region i Terminal te naposljetku bez varijabla GEO.Region, Terminal i Boarding.Area.

Koristeći stablo na cijelom skupu podataka koji sada imam dobila sam točno razvrstan skup podataka. To je zbog toga što su GEO.Region i GEO.Summary u potpunosti povezane varijable.

```
set.seed(123)
skup_no1 <- sample(2, nrow(podatci1), replace=TRUE, prob=c(0.7, 0.3))
skup_za_treniranje1 <- podatci1[skup_no1==1,]
skup_za_testiranje1 <- podatci1[skup_no1==2,]

podatci_j48_stablo1 <- J48(GEO.Summary ~ ., data = skup_za_treniranje1)
treniranje_j481 <- summary(podatci_j48_stablo1)$confusionMatrix
treniranje_j481
```

```
##                predicted
##                 Domestic International
##   Domestic          4086             0
##   International        0          6460
```

```
eval_j481 <- evaluate_Weka_classifier(podatci_j48_stablo1, newdata = skup_za_testiranje1)$confusionMatr
eval_j481
```

```
##                predicted
##                 Domestic International
##   Domestic          1711             0
##   International        0          2750
```

Sada sam iz skupa podataka izbacila varijablu GEO.Region kako bih vidjela kako će metoda J48 razvrstati ove podatke. Ovdje već postoji nekoliko grešaka no razvrstavanje nije toliko loše.

```
podatci2 <- podatci1[-4]

set.seed(123)
skup_no2 <- sample(2, nrow(podatci2), replace=TRUE, prob=c(0.7, 0.3))
skup_za_treniranje2 <- podatci2[skup_no2==1,]
skup_za_testiranje2 <- podatci2[skup_no2==2,]

podatci_j48_stablo2 <- J48(GEO.Summary ~ ., data = skup_za_treniranje2)
treniranje_j482 <- summary(podatci_j48_stablo2)$confusionMatrix
treniranje_j482
```

```
##                predicted
##                 Domestic International
##   Domestic          3990            96
##   International       66          6394
```

```
eval_j482 <- evaluate_Weka_classifier(podatci_j48_stablo2, newdata = skup_za_testiranje2)$confusionMatri
eval_j482
```

```
##                predicted
##                 Domestic International
##   Domestic          1500           211
##   International      129          2621
```

Izbacivši i varijablu Terminal, dobila sam malo bolje razvrstavanje na skupu za treniranje, ali malo gore na skupu za testiranje.

```
podatci3 <- podatci2[-6]
set.seed(123)
skup_no3 <- sample(2, nrow(podatci3), replace=TRUE, prob=c(0.7, 0.3))
skup_za_treniranje3 <- podatci3[skup_no3==1,]
skup_za_testiranje3 <- podatci3[skup_no3==2,]

podatci_j48_stablo3 <- J48(GEO.Summary ~ ., data = skup_za_treniranje3)
treniranje_j483 <- summary(podatci_j48_stablo3)$confusionMatrix
treniranje_j483
```

```
##                predicted
##                 Domestic International
##   Domestic          4023            63
##   International       84          6376
```

```
eval_j483 <- evaluate_Weka_classifier(podatci_j48_stablo3, newdata = skup_za_testiranje3)$confusionMatri
eval_j483
```

```
##                predicted
##                 Domestic International
##   Domestic          1579           132
##   International      304          2446
```

Na kraju sam izbacila i varijablu Boarding.Area. Sada više u skupu podataka nemamo nijednu varijablu naizgled povezanu s GEO.Summary. Ovdje već postoji puno više grešaka, posebno na skupu za testiranje.

```
podatci4 <- podatci3[-6]
set.seed(123)
skup_no4 <- sample(2, nrow(podatci4), replace=TRUE, prob=c(0.7, 0.3))
skup_za_treniranje4 <- podatci4[skup_no4==1,]
skup_za_testiranje4 <- podatci4[skup_no4==2,]

podatci_j48_stablo4 <- J48(GEO.Summary ~ ., data = skup_za_treniranje4)
treniranje_j484 <- summary(podatci_j48_stablo4)$confusionMatrix
treniranje_j484
```

```
##                predicted
##                 Domestic International
##   Domestic          3903           183
##   International      136          6324
```

```
eval_j484 <- evaluate_Weka_classifier(podatci_j48_stablo4, newdata = skup_za_testiranje4)$confusionMatri
eval_j484
```

```
##                predicted
##                 Domestic International
##   Domestic          1237           474
##   International      800          1950
```

Zaključujem kako je varijabla GEO.Summary pOvezana i s Terminal i Boarding.Area, ali ne toliko koliko s GEO.Region. Izbacivanjem svih tih varijabli ne dobivam baš dobre rezultate razvrstavanja.

Sada sam ostavila varijable koje naizgled nisu povezane s GEO.Summary.

```
novi_podatci<-podatci[c(1,4,6,8,9,12)]
head(novi_podatci)
```

```
##   Activity.Period Published.Airline   GEO.Summary Activity.Type.Code
## 1          200507      ATA Airlines      Domestic           Deplaned
## 2          200507      ATA Airlines      Domestic           Enplaned
## 3          200507      ATA Airlines      Domestic     Thru / Transit
## 4          200507       Air Canada  International           Deplaned
## 5          200507       Air Canada  International           Enplaned
## 6          200507         Air China International           Deplaned
##   Price.Category.Code Passenger.Count
## 1            Low Fare           27271
## 2            Low Fare           29131
## 3            Low Fare            5415
## 4               Other           35156
## 5               Other           34090
## 6               Other            6263
```

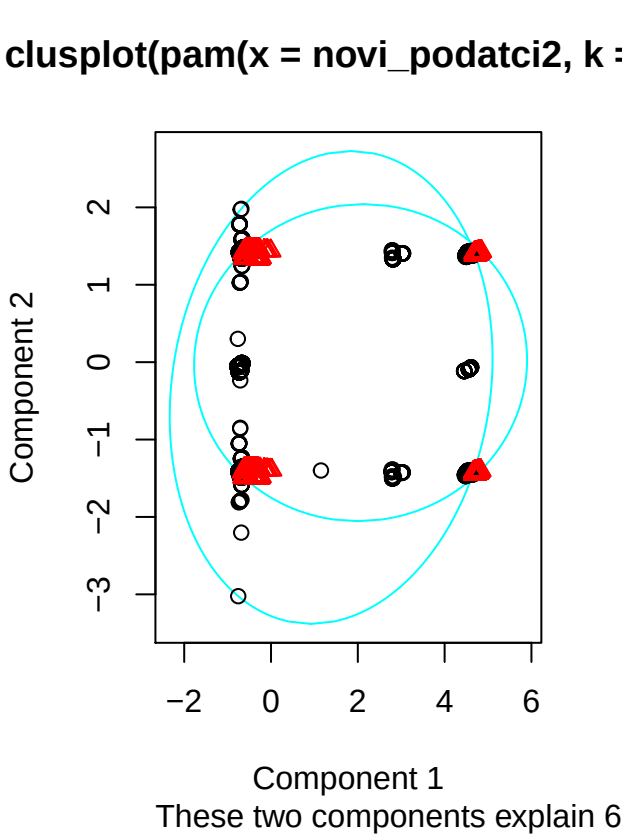
Želim vidjeti hoće li ovi podatci biti dobro grupirani korištenjem k-medoida (k=2). Kao rezultat dobivam da podatci nisu dobro grupirani.

```
novi_podatci1<-novi_podatci
novi_podatci1 <- cbind(novi_podatci1, model.matrix( ~ 0 + Activity.Type.Code, novi_podatci1))
novi_podatci1 <- cbind(novi_podatci1, model.matrix( ~ 0 + Price.Category.Code, novi_podatci1))
novi_podatci1 <- cbind(novi_podatci1, model.matrix( ~ 0 + Published.Airline, novi_podatci1))
novi_podatci1<-novi_podatci1[c(-2,-4,-5)]
novi_podatci2<-novi_podatci1
novi_podatci2$GEO.Summary<-NULL

km1 <- pam(novi_podatci2, 2)
table(novi_podatci1$GEO.Summary, km1$clustering)
```

```
##
##                    1    2
##   Domestic      4229 1568
##   International 9094  116
```

```
layout(matrix(c(1,2),1,2))
clusplot(km1, col.p=km1$clustering)
```



Ovdje sam željela vidjeti kako će podatci biti grupirani kada ostavim jedine dvije numeričke varijable s varijablom GEO.Summary. Rezultati grupiranja su jednaki kao gore. Zaključujem kako varijable koje imam gore, a ne ovdje ne utječu na grupiranje (te varijable su faktori).

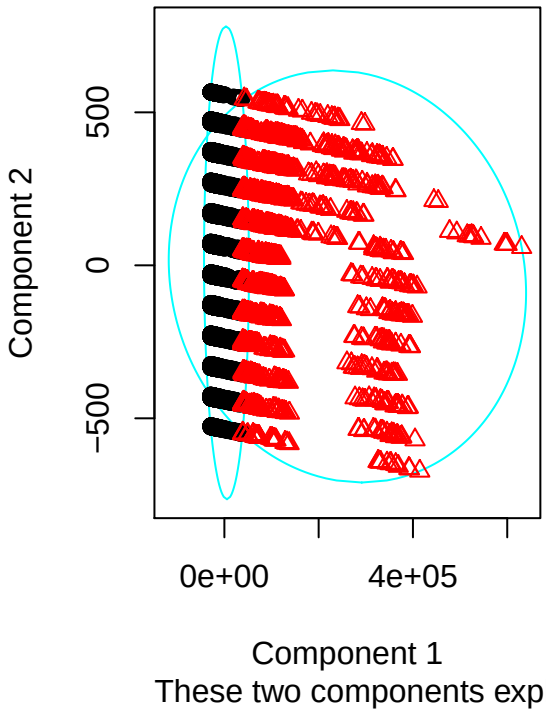
```
novi_podatci3<-podatci[c(1,6,12)]
novi_podatci4<-novi_podatci3
novi_podatci4$GEO.Summary<-NULL

km2 <- pam(novi_podatci4, 2)
table(novi_podatci3$GEO.Summary, km2$clustering)
```

```
##
##                     1    2
##   Domestic      4229 1568
##   International 9094  116
```

```
layout(matrix(c(1,2),1,2))
clusplot(km2, col.p=km2$clustering)
```

clusplot(pam(x = novi_podatci4, k :



Sada sam koristila neuronske mreže. Nisam uspjela dobiti točne rezultate.

```
set.seed(123)
skup_no5 <- sample(2, nrow(novi_podatci), replace=TRUE, prob=c(0.7, 0.3))
skup_za_treniranje5 <- novi_podatci[skup_no5==1,]
skup_za_testiranje5 <- novi_podatci[skup_no5==2,]

nn1_model <- nnet(GEO.Summary ~ ., data = skup_za_treniranje5, size = 13, rang = 0.1, decay = 0.00001, r
```

```
## # weights:  963
## initial  value 7622.920381
## iter  10 value 6490.811433
## iter  20 value 5585.569569
## iter  30 value 5472.081575
## iter  40 value 4477.104172
## iter  50 value 3514.177965
## iter  60 value 3107.783432
## iter  70 value 2858.629074
## iter  80 value 2736.162418
## iter  90 value 2733.811414
## iter 100 value 2729.602365
## iter 110 value 2633.488293
## iter 120 value 2615.191676
## iter 130 value 2614.505283
## iter 140 value 2614.149049
## iter 150 value 2613.879061
```

```
## iter 160 value 2612.231567
## iter 170 value 2611.765009
## iter 180 value 2603.403493
## iter 190 value 2536.624738
## iter 200 value 2457.078020
## iter 210 value 2456.002482
## final  value 2456.001630
## converged
```

```
table(skup_za_treniranje5$GEO.Summary,
      predict(nn1_model, skup_za_treniranje5, type="class"))
```

```
##
##                 Domestic International
##   Domestic          3544           542
##   International      330          6130
```

```
table(skup_za_testiranje5$GEO.Summary, predict(nn1_model, skup_za_testiranje5, type="class"))
```

```
##
##                 Domestic International
##   Domestic          1494           217
##   International      154          2596
```

Zaključujem da varijabla GEO.Summary nije povezana s varijablama koje nisu naizgled s njom povezane. Općenito mislim da se iz ovog skupa podataka ne može puno zaključiti osim nekih osnovnih svojstava koje sam navela u početku.