INSTITUT ZA MATEMATIKU I INFORMATIKU

PRIRODNO-MATEMATIČKOG FAKULTETA

UNIVERZITET U KRAGUJEVCU

Seminarski rad

Predstavljanje i tumačenje skupa podataka „University recommendation“

Studenti:

Uroš Petronijević 73/2019

Miodrag Ranđelović 76/2019

Profesor:

dr Branko Arsić

Sadržaj

Za budućeg diplomiranog studenta, izbor univerziteta na koje će se prijaviti je zagonetka. Često se studenti pitaju da li je njihov profil dovoljno dobar za određeni univerzitet. Ovo pitanje smo rešili tako što smo izgradili sistem preporuka zasnovan na različitim algoritmima klasifikacije.. Podaci nisu bili lako dostupni, ali zahvaljujući devojci pod imenom Aditya Sureshkumar, sakupljeni su podaci sa *Edulik.com* i napravljen je skup podataka koji sadrži profile studenata koji su primljeni/odbijeni na 45 različitih univerziteta u SAD. Na osnovu ovog skupa podataka, obučeni su različiti modeli i predloženi su univerziteti koji maksimiziraju šanse da student dobije prijem sa tog univerziteta. U ovom radu analiziran je skup podataka „university recommendation" koji predstavlja profile studenata koji su primljeni/odbijeni na 45 različitih univerziteta u SAD-a. Ciljevi istraživanja su bili sledeći:

1. Da se izvrši adekvatan opis obeležja, i detaljna analiza uticaja/veza/zavisnosti između obeležja i u skupu podataka
2. Da se na principijalan način izvrši formiranje, odabir i tumačenje najadekvatnijeg modela mašinskog učenja za predviđanje prijema studenata na fakultetima u SAD-a.
3. Da se sirovi skup podataka dovede do nivoa kvaliteta koji omogućava dovoljno pouzdano statističko zaključivanje o vezama između obeležja. kao i formiranje adekvatnih modela mašinskog učenja za predviđanje prijema studenata na fakultetima u SAD-a.

Učitavanje biblioteka

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(dplyr)
library(magrittr)

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract

library(Amelia)

## Loading required package: Rcpp

## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.0, built: 2021-05-26)
## ## Copyright (C) 2005-2022 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##

library(ggpubr)
library(Cairo)
library(broom)
library(randomForest)

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

library(rpart)
library(rpart.plot)
```

Uvod

```
fajl = read.csv("original_data.csv")
#View(fajl)
```

Skup podataka koji će biti obrađen dat je u CSV formatu, u fajlu *original_data.csv*. Nakon učitavanja podataka, funkcija **dim** daje informacije o dimenzijama okvira podataka. Vidi se da skup podataka sadrži 53644 redova i 26 kolona/obeležja. Funkcija **summary** daje detaljnu statistiku o svakoj koloni/obeležju to jest: maksimum, minimum, medijanu, broj nedostajućih vrednosti, prvi kvartil, treći kvartil.

```
dim(fajl)
```

```
## [1] 53644    26
```

```
summary(fajl)
```

```
##    userName            major            researchExp       industryExp
##  Length:53644       Length:53644       Min.   : 0.0000   Min.   :  0.000
##  Class :character   Class :character   1st Qu.: 0.0000   1st Qu.:  0.000
##  Mode  :character   Mode  :character   Median : 0.0000   Median :  0.000
##                                        Mean   : 0.3395   Mean   :  4.057
##                                        3rd Qu.: 0.0000   3rd Qu.:  0.000
##                                        Max.   :53.0000   Max.   :138.000
##
##  specialization       toeflScore       program           department
##  Length:53644       Min.   :   0.0   Length:53644       Length:53644
##  Class :character   1st Qu.: 101.0   Class :character   Class :character
##  Mode  :character   Median : 107.0   Mode  :character   Mode  :character
##                     Mean   : 109.6
##                     3rd Qu.: 111.0
##                     Max.   :1350.0
##                     NA's   :4414
##   toeflEssay          internExp           greV             greQ
##  Length:53644       Min.   : 0.0000   Min.   :   0.0   Min.   :   0.0
##  Class :character   1st Qu.: 0.0000   1st Qu.: 152.0   1st Qu.: 162.0
##  Mode  :character   Median : 0.0000   Median : 159.0   Median : 168.0
##                     Mean   : 0.4543   Mean   : 324.5   Mean   : 422.5
##                     3rd Qu.: 0.0000   3rd Qu.: 550.0   3rd Qu.: 780.0
##                     Max.   :96.0000   Max.   :5560.0   Max.   :7990.0
##                     NA's   :14        NA's   :1256     NA's   :1220
##  userProfileLink    journalPubs            greA            topperCgpa
##  Length:53644       Length:53644       Min.   :   0.000   Min.   :  0.00
##  Class :character   Class :character   1st Qu.:   3.000   1st Qu.:  8.10
##  Mode  :character   Mode  :character   Median :   3.500   Median :  9.60
##                                        Mean   :   5.065   Mean   : 35.75
##                                        3rd Qu.:   4.000   3rd Qu.: 80.00
##                                        Max.   :1470.000   Max.   :100.00
##                                        NA's   :2858       NA's   :3
##  termAndYear          confPubs          ugCollege            gmatA
##  Length:53644       Length:53644       Length:53644       Min.   :  3.00
##  Class :character   Class :character   Class :character   1st Qu.:  4.00
##  Mode  :character   Mode  :character   Mode  :character   Median :  5.00
##                                                           Mean   :  6.12
##                                                           3rd Qu.:  5.00
##                                                           Max.   :102.00
##                                                           NA's   :53525
##       cgpa             gmatQ          cgpaScale          gmatV
##  Min.   :  0.00   Min.   :  8.00   Min.   :  0.00   Min.   : 19.00
##  1st Qu.:  8.17   1st Qu.: 46.00   1st Qu.: 10.00   1st Qu.: 27.00
##  Median : 10.00   Median : 48.00   Median :100.00   Median : 31.00
##  Mean   : 39.34   Mean   : 49.33   Mean   : 55.26   Mean   : 34.89
##  3rd Qu.: 71.73   3rd Qu.: 50.00   3rd Qu.:100.00   3rd Qu.: 34.00
##  Max.   :833.00   Max.   :168.00   Max.   :100.00   Max.   :152.00
```

```
##                   NA's   :53521                      NA's   :53530
##    univName             admit
##  Length:53644       Min.   :0.0000
##  Class :character   1st Qu.:0.0000
##  Mode  :character   Median :1.0000
##                     Mean   :0.5211
##                     3rd Qu.:1.0000
##                     Max.   :1.0000
##
```

Funkcijom **str** proveravamo kakva je struktura datih kolona/obeležja. Možemo videti da postoji 12 obeležja znakovnog tipa(chr) i 14 obeležja numeričkog tipa, a od toga su 11 obeležja tipa (int) i 3 obeležja tipa (num).

```
str(fajl)
```

```
## 'data.frame':    53644 obs. of  26 variables:
##  $ userName       : chr  "143saf" "7790ashish" "AB25" "abhijitg" ...
##  $ major          : chr  "Systems and Control" "Manufacturing Engineering"
"(MIS / MSIM / MSIS / MSIT)" "" ...
##  $ researchExp    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ industryExp    : int  18 0 66 0 0 0 0 0 0 0 ...
##  $ specialization : chr  "Robotics" "" "" "" ...
##  $ toeflScore     : int  112 NA 94 NA 81 273 104 95 101 107 ...
##  $ program        : chr  "MS" "MS" "MS" "" ...
##  $ department     : chr  "Instrumentation & Control" "0" "Computer
Engineering" "0" ...
##  $ toeflEssay     : chr  "26" "" "21" "" ...
##  $ internExp      : int  5 0 0 0 0 0 0 0 0 0 ...
##  $ greV           : int  160 NA 146 NA 420 410 150 147 490 550 ...
##  $ greQ           : int  167 NA 157 NA 770 1010 161 156 740 780 ...
##  $ userProfileLink: chr
"http://www.edulix.com/unisearch/user.php?uid=252766"
"http://www.edulix.com/unisearch/user.php?uid=196141"
"http://www.edulix.com/unisearch/user.php?uid=226830"
"http://www.edulix.com/unisearch/user.php?uid=10967" ...
##  $ journalPubs    : chr  "0" "0" "0" "" ...
##  $ greA           : num  4.5 NA 3 NA 2.5 600 4.5 3 3 4.5 ...
##  $ topperCgpa     : num  8.9 0 81 0 70 ...
##  $ termAndYear    : chr  "Fall - 2015" "Fall - 2013" "Fall - 2015" "" ...
##  $ confPubs       : chr  "0" "0" "0" "" ...
##  $ ugCollege      : chr  "Dharamsinh Desai University" "" "IET DAVV" "" ...
##  $ gmatA          : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ cgpa           : num  8.5 0 78.3 0 57 ...
##  $ gmatQ          : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ cgpaScale      : int  10 0 100 0 100 100 100 100 100 100 ...
##  $ gmatV          : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ univName       : chr  "Worcester Polytechnic Institute" "Worcester
Polytechnic Institute" "Worcester Polytechnic Institute" "Worcester
Polytechnic Institute" ...
##  $ admit          : int  1 1 1 1 1 1 1 1 1 1 ...
```

Obeležja i njihov opis koje sadrži okvir podataka *university recommendation*

- userName - Korisničko ime u edulik.com
- major - Smer koji je korisnik pohadjao/pohadja
- researchExp - Istraživačko iskustvo u mesecima
- industryExp - Industrijsko iskustvo u mesecima
- specialization - Namenjena specijalizacija za visoke studije
- toeflScore -TOEFL (jedan od dva glavna testa znanja engleskog jezika prihvaćena na univerzitetima širom sveta, predstavlja skracenicu od recenice "the Test Of English as a Foreign Language" odnosno "Test engleskog kao stranog jezika")
- program - Nameravani diplomski program
- department - Odeljenje u kojem je korisnik bio/je upisan
- toeflEssay - Ocena korisnickog eseja za test TOEFL.
- internExp - staž iskustvo u mesecima
- greV - predstavlja GRE-ov verbalni rezultat (GRE je standardizovani test koji postoji od 1936. godine i koji izračunava tri veoma važna parametra: verbalno i kvantitativno rezonovanje i analitičko pisanje. Ovo obeležje se odnosi na ocenu verbalnog rezonovanja)
- greQ - GRE kvantitativan rezultat (deo GRE testa koji se odnosi na znanje iz matematike)
- userProfileLink - Link do korisničkog profila na edulik.com
- journalPubs - broj publikacija časopisa
- greA - GRE AWA rezultat (deo GRE testa koji se odnosi na znanje iz analitičkog pisanja)
- termAndYear - Predviđeni termin pridruživanja. Npr.: jesen - 2022
- confPubs - broj publikacija na konferenciji
- ugCollege - koledži odnosno srednje škole iz koje đak dolazi.
- gmatA - rezulatat na GMAT AWA testu (test za koji se koristi za analiziranje obrazloženja datog argumenta i da napišete kritiku tog argumenta)
- cgpa - srednja prosečna ocena koja se koristi za procenu akademskog učinka
- gmatQ - rezulati na GMAT quant testu (test koji meri sposobnosti matematičkog zaključivanja, rešavanja kvantitaivnih problema i tumačenja grafičkih podatataka)
- cgpaScale - CGPA(Kumulativni prosek ocena) skala za studentov prosek ocena
- topperCgpa - vrednost CGPA u najvisem delu rang liste
- gmatV - rezulati na GMAT verbal testu (test koji meri sposobnosti za čitanje i shvatanje napisanog materijala)
- univName - Naziv univerziteta za koji je student aplicirao
- admit - Rezultat aplikacije na fakultet(0/1 - odbijen/prihvaćen)

Moramo proveriti koliko svaka kolona ima nedostajućih vrednosti i došli smo do zaključka da obeležja: **gmatA**, **gmatQ**, **gmatV** imaju više od 99% nedostajućih vrednosti, i nemoguće je popuniti te nedostajuće vrednosti već je najbolje ukloniti ih. Razloh zašto većina studenata nemaju rezultate jeste jer *gmat* test nije značajan pri upisu na željene univerzitete, tako da većina studenata i nije radilo ove testove.

```
(colMeans(is.na(fajl)))*100
```

```
##        userName           major     researchExp     industryExp
specialization
##     0.000000000     0.000000000     0.000000000     0.000000000
0.005592424
##      toeflScore         program      department      toeflEssay
internExp
##     8.228320036     0.000000000     0.000000000     0.000000000
0.026097979
##            greV            greQ userProfileLink     journalPubs
greA
##     2.341361569     2.274252479     0.000000000     0.000000000
5.327716054
##      topperCgpa     termAndYear        confPubs       ugCollege
gmatA
##     0.005592424     0.000000000     0.000000000     0.000000000
99.778167176
##            cgpa           gmatQ       cgpaScale           gmatV
univName
##     0.000000000    99.770710611     0.000000000    99.787487883
0.000000000
##           admit
##     0.000000000
```

```
novi_univerziteti = subset(fajl, select = -c(gmatA, gmatQ, gmatV))
```

Analizirajući okvir podataka, primećeno je da obeležje **termAndYear** sadrži informaciju koje godine i kog semestra je predviđeni termin pridruživanja studenta željenom fakultetu u formatu *jesen - 2022*.

```
head(novi_univerziteti$termAndYear,n=50)
```

```
##  [1] "Fall - 2015"   "Fall - 2013"   "Fall - 2015"   ""
##  [5] "Fall - 2011"   "Fall - 2006"   "Fall - 2015"   "Fall - 2012"
##  [9] "Fall - 2011"   "Fall - 2011"   "Fall - 2012"   "Fall - 2011"
## [13] "Fall - 2011"   "Fall - 2011"   "Fall - 2016"   "Fall - 2015"
## [17] "Spring - 2011" "Fall - 2015"   "Fall - 2014"   "Fall - 2014"
## [21] "Fall - 2012"   "Fall - 2013"   "Fall - 2012"   "Fall - 2012"
## [25] "Fall - 2011"   "Fall - 2013"   "Fall - 2015"   "Fall - 2014"
## [29] "Fall - 2012"   "Fall - 2013"   "Fall - 2012"   "Fall - 2013"
## [33] "Fall - 2014"   "Fall - 2013"   "Spring - 2012" "Spring - 2014"
## [37] "Spring - 2016" "Fall - 2011"   "Fall - 2014"   "Fall - 2013"
## [41] "Fall - 2009"   "Fall - 2012"   "Fall - 2013"   "Fall - 2011"
## [45] "Fall - 2011"   "Fall - 2011"   "Fall - 2012"   "Fall - 2011"
## [49] "Fall - 2014"   "Fall - 2013"
```

Da bismo dobili tačnije informacije, kreiraćemo 2 nova obeležja, tako što ćemo trenutni razdvojiti po karakteru "-". Kreiraćemo obeležje **year** koje će sadržati godinu pridruživanja, i **term** koji će sadržati semestar pridruživanja.

```
novi_univerziteti$term = sapply(novi_univerziteti$termAndYear,
                                FUN = function(x) unlist(strsplit(x, split =
" - "))[1])

novi_univerziteti$year = sapply(novi_univerziteti$termAndYear,
                                FUN = function(x) unlist(strsplit(x, split =
" - "))[2])

novi_univerziteti=subset(novi_univerziteti, select = -c(termAndYear))
```

Jedinstvene vrednosti novonastale kategorije *year*

```
unique(novi_univerziteti$year)

##  [1] "2015"    "2013"    NA        "2011"    "2006"    "2012"    "2016"
##  [8] "2014"    "2009"    "2008"    "2007"    "20133"   "1992"    "1989"
## [15] "2010"    "1990"    "6"       "2005"    "14"      "2"       "215"
## [22] "11"      "15"      "13"      "2103"    "12"      "3024"    "2112"
## [29] "201"     "20113"   "2105"    "2017"    "20131"   "2012101"
```

Jedinstvene vrednosti novonastale kategorije *term*

```
unique(novi_univerziteti$term)

## [1] "Fall"   NA       "Spring" "Summer" "0"      "8.89"   "81"
```

Potrebno je odrediti koliko sva obeležja karakternog tipa (*chr* tip) imaju jedninstvenih vrednosti kako bismo ih pretvorili u faktor obeležja (*fct* tip). Deo tih obeležja moguće je direktno preobraziti u faktor promenljive ukoliko nemaju preveliki broj jedinstvenih vrednosti, dok obeležja sa velikim brojem jedinstvenih vrednosti potrebno je dodatno analizirati.

```
tempDF <- as.data.frame(lengths(lapply(fajl %>%
select(where(is.character)),unique)))
colnames(tempDF) = c("broj_jedinstvenih_vrednosti")

tempDF

##                 broj_jedinstvenih_vrednosti
## userName                              14798
## major                                   245
## specialization                         3622
## program                                   5
## department                             1487
## toeflEssay                               37
## userProfileLink                       14798
## journalPubs                              14
## termAndYear                              57
## confPubs                                 13
## ugCollege                              1823
## univName                                 54
```

Međutim pregledajući podatke uz pomoć funkcija *View* i *str* primetili smo da obeležje confPubs da uglavnom sadrži numeričke vrednosti, zato ćemo to sada detaljnije proveriti.

```
xtabs(~novi_univerziteti$confPubs)
```

```
## novi_univerziteti$confPubs
##                       0           1          15           2           3
##         322       51656        1046           1         353         135
##           4           5           6           8 Fall - 2012 Fall - 2014
##          71          28           8          10           4           3
## Fall - 2015
##           7
```

Kao što možemo primeti, većina podataka su numerička, dok karakterni primeri deluju kao greske odnosno izuzeci. Iako ćemo se kasnije detaljnije baviti izuzecima, radi lakseg daljeg rada, odmah ćemo ukloniti te uzorke.

```
novi_univerziteti=novi_univerziteti[!grepl("Fall",
novi_univerziteti$confPubs),]
```

Nakon pregleda jedinstvenih vrednosti za karakterna obeležja, zaključeno je da *program*, *toeflEssay*, *journalPubs*, *univName*, *term* i *year* ispunjavaju potreban kriterijum za transormisanje u kategorijsko obeležje.

```
novi_univerziteti$program = as.factor(novi_univerziteti$program)
novi_univerziteti$toeflEssay = as.factor(novi_univerziteti$toeflEssay)
novi_univerziteti$journalPubs = as.factor(novi_univerziteti$journalPubs)
novi_univerziteti$univName = as.factor(novi_univerziteti$univName)
novi_univerziteti$term = as.factor(novi_univerziteti$term)
novi_univerziteti$year = as.factor(novi_univerziteti$year)

str(novi_univerziteti)
```

```
## 'data.frame':    53630 obs. of  24 variables:
##  $ userName       : chr  "143saf" "7790ashish" "AB25" "abhijitg" ...
##  $ major          : chr  "Systems and Control" "Manufacturing Engineering"
"(MIS / MSIM / MSIS / MSIT)" "" ...
##  $ researchExp    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ industryExp    : int  18 0 66 0 0 0 0 0 0 0 ...
##  $ specialization : chr  "Robotics" "" "" "" ...
##  $ toeflScore     : int  112 NA 94 NA 81 273 104 95 101 107 ...
##  $ program        : Factor w/ 5 levels "","Both MS and PhD",..: 3 3 3 1 3 3
3 3 3 3 ...
##  $ department     : chr  "Instrumentation & Control" "0" "Computer
Engineering" "0" ...
##  $ toeflEssay     : Factor w/ 35 levels "","0","1.5","10",..: 15 1 10 1 1
30 16 11 13 1 ...
##  $ internExp      : int  5 0 0 0 0 0 0 0 0 0 ...
##  $ greV           : int  160 NA 146 NA 420 410 150 147 490 550 ...
##  $ greQ           : int  167 NA 157 NA 770 1010 161 156 740 780 ...
##  $ userProfileLink: chr
```

```
"http://www.edulix.com/unisearch/user.php?uid=252766"
"http://www.edulix.com/unisearch/user.php?uid=196141"
"http://www.edulix.com/unisearch/user.php?uid=226830"
"http://www.edulix.com/unisearch/user.php?uid=10967" ...
##  $ journalPubs    : Factor w/ 11 levels "","0","1","10",..: 2 2 2 1 2 2 2 2
2 2 ...
##  $ greA           : num  4.5 NA 3 NA 2.5 600 4.5 3 3 4.5 ...
##  $ topperCgpa     : num  8.9 0 81 0 70 ...
##  $ confPubs       : chr  "0" "0" "0" "" ...
##  $ ugCollege      : chr  "Dharamsinh Desai University" "" "IET DAVV" "" ...
##  $ cgpa           : num  8.5 0 78.3 0 57 ...
##  $ cgpaScale      : int  10 0 100 0 100 100 100 100 100 100 ...
##  $ univName       : Factor w/ 54 levels "Arizona State University",..: 54
54 54 54 54 54 54 54 54 54 ...
##  $ admit          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ term           : Factor w/ 3 levels "Fall","Spring",..: 1 1 1 NA 1 1 1 1
1 1 ...
##  $ year           : Factor w/ 33 levels "11","12","13",..: 25 21 25 NA 17
11 25 19 17 17 ...
```

Takođe potrebno je odrediti i za sva obeležja numeričkog tipa (*int* i *num* tipovi) imaju jedninstvenih vrednosti kako bismo ih pretvorili u faktor obeležja (*fct* tip).

```
tempDF <- as.data.frame(lengths(lapply(fajl %>%
select(where(is.numeric)),unique)))
colnames(tempDF) = c("broj_jedinstvenih_vrednosti")

tempDF
```

```
##             broj_jedinstvenih_vrednosti
## researchExp                          33
## industryExp                          97
## toeflScore                          110
## internExp                            27
## greV                                145
## greQ                                169
## greA                                 66
## topperCgpa                          534
## gmatA                                 6
## cgpa                               1911
## gmatQ                                13
## cgpaScale                             5
## gmatV                                18
## admit                                 2
```

Nakon pregleda jedinstvenih vrednosti za numerička obeležja, zaključeno je da *cgpaScale* i *admit* ispunjavaju potreban kriterijum za transormisanje u kategorijsko obeležje.

```
novi_univerziteti$cgpaScale = as.factor(novi_univerziteti$cgpaScale)
novi_univerziteti$admit = as.factor(novi_univerziteti$admit)
```

```
str(novi_univerziteti)
```

```
## 'data.frame':    53630 obs. of  24 variables:
##  $ userName       : chr  "143saf" "7790ashish" "AB25" "abhijitg" ...
##  $ major          : chr  "Systems and Control" "Manufacturing Engineering"
"(MIS / MSIM / MSIS / MSIT)" "" ...
##  $ researchExp    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ industryExp    : int  18 0 66 0 0 0 0 0 0 0 ...
##  $ specialization : chr  "Robotics" "" "" "" ...
##  $ toeflScore     : int  112 NA 94 NA 81 273 104 95 101 107 ...
##  $ program        : Factor w/ 5 levels "","Both MS and PhD",..: 3 3 3 1 3 3
3 3 3 3 ...
##  $ department     : chr  "Instrumentation & Control" "0" "Computer
Engineering" "0" ...
##  $ toeflEssay     : Factor w/ 35 levels "","0","1.5","10",..: 15 1 10 1 1
30 16 11 13 1 ...
##  $ internExp      : int  5 0 0 0 0 0 0 0 0 0 ...
##  $ greV           : int  160 NA 146 NA 420 410 150 147 490 550 ...
##  $ greQ           : int  167 NA 157 NA 770 1010 161 156 740 780 ...
##  $ userProfileLink: chr
"http://www.edulix.com/unisearch/user.php?uid=252766"
"http://www.edulix.com/unisearch/user.php?uid=196141"
"http://www.edulix.com/unisearch/user.php?uid=226830"
"http://www.edulix.com/unisearch/user.php?uid=10967" ...
##  $ journalPubs    : Factor w/ 11 levels "","0","1","10",..: 2 2 2 1 2 2 2 2
2 2 ...
##  $ greA           : num  4.5 NA 3 NA 2.5 600 4.5 3 3 4.5 ...
##  $ topperCgpa     : num  8.9 0 81 0 70 ...
##  $ confPubs       : chr  "0" "0" "0" "" ...
##  $ ugCollege      : chr  "Dharamsinh Desai University" "" "IET DAVV" "" ...
##  $ cgpa           : num  8.5 0 78.3 0 57 ...
##  $ cgpaScale      : Factor w/ 5 levels "0","4","5","10",..: 4 1 5 1 5 5 5 5
5 5 ...
##  $ univName       : Factor w/ 54 levels "Arizona State University",..: 54
54 54 54 54 54 54 54 54 54 ...
##  $ admit          : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ term           : Factor w/ 3 levels "Fall","Spring",..: 1 1 1 NA 1 1 1 1
1 1 ...
##  $ year           : Factor w/ 33 levels "11","12","13",..: 25 21 25 NA 17
11 25 19 17 17 ...
```

Takođe ono što smo zaključili zahvaljući funkciji **str** jeste da nam obeležja *userName* i *userProfileLink* neće biti korisna u daljem radu zato što samo predstavljaju nalog studenta na *edulik.com*, kao i sam link ka nalogu, zbog čega će navedena obeležja biti uklonjena i neće se koristiti u daljoj analizi.

```
novi_univerziteti = subset(novi_univerziteti, select = -c(userName,
userProfileLink))
dim(novi_univerziteti)
```

```
## [1] 53630    22
```

Redukovanje kategorija

Znakovna obeležja *major*, *specialization*, *ugCollege* i *department*, odnosno obeležja sa velikim brojem jedinstvenih vrednosti, potrebno je redukovati putem funkcije **fct_lump** koja zadržava zadati broj kategorija, a ostale kategorije spaja u novu kategoriju. Nakon toga, iskoristićemo funkciju **fct_infreq** koja će sortirati kategorije obeležja prema broju pojavljivanja.

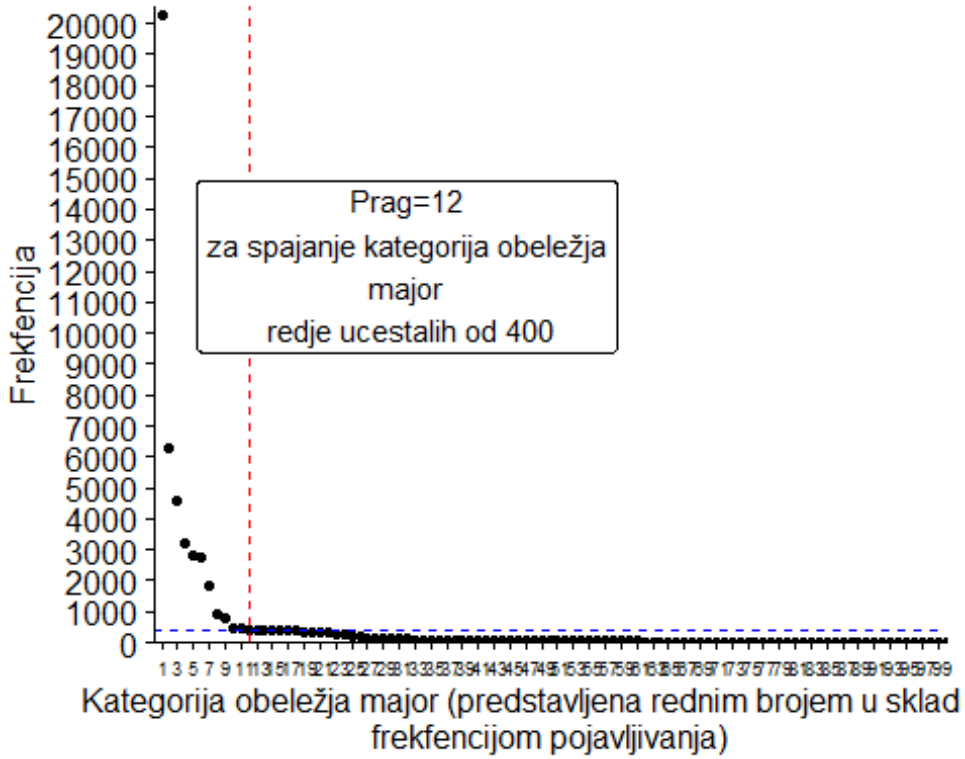
Da bi se odredio broj kategorija koji bi trebalo zadržati, korišćena je grafička metoda *Cleveland tačkasti dijagram*, koja predstavlja alternativu za stubičaste dijagrame, ali njome dobijamo manju prenatrpanost na dijagramu. Posmatranjem dijagrama i korišćenjem heuristike vezanoj za određivanje broja klastera, možemo proceniti minimalnu dozvoljenu učestalost za kategorije posmatranog obeležja. One kategorije koje imaju učestalnost ispod dozvoljene potrebno je spojiti u jednu novu kategoriju. S obzirom da funkcija **fct_lump** zahteva kao ulazni parametar broj najčešćih kategorija koje treba zadržati, taj parametar je izračunat na osnovu minimalne učestalosti datog obeležja, i nazvan **prag**.

Nakon što se prag izračuna, potrebno je proveriti da li na grafiku desno od izračunatog praga ne postoje prevelike varijacije u učestalnosti preostalih kategorija. Ukoliko ih nema, postojeći odnosno izračunati prag se zadržava. U suprotnom, potrebno je zadati novu minimalnu učestalost, i ponavlja heuristika za novi prag. Prag je na dijagramu predstavljen vertikalnom crvenom linijom.

```
# Pomoćna metoda za zaokruživanje na gore za svrhu određivanja limita y-ose
round_up <- function(df, kol, base=1000){
vektor_kolona <- df %>% pull({{kol}})
maks = max(vektor_kolona)
return(base*ceiling(maks/base))
}

tempDF <- as_tibble(as.data.frame(table(novi_univerziteti$major))) %>%
arrange(desc(Freq)) %>%
mutate(rbr=seq_along(Var1)) %>% rename_with(~c("major","frekvencija","rbr"))
tempDF
```

```
## # A tibble: 245 x 3
##    major                                frekvencija   rbr
##    <fct>                                      <int> <int>
##  1 Computer Science                           20269     1
##  2 Electrical Engineering                      6287     2
##  3 MIS                                         4557     3
##  4 Electronics and Communication               3224     4
##  5 Mechanical Engineering                      2806     5
##  6 Computer Engineering                        2734     6
##  7 Industrial Engineering                      1807     7
##  8 Electrical and Computer Engineering          894     8
##  9 Civil Engineering                            811     9
```

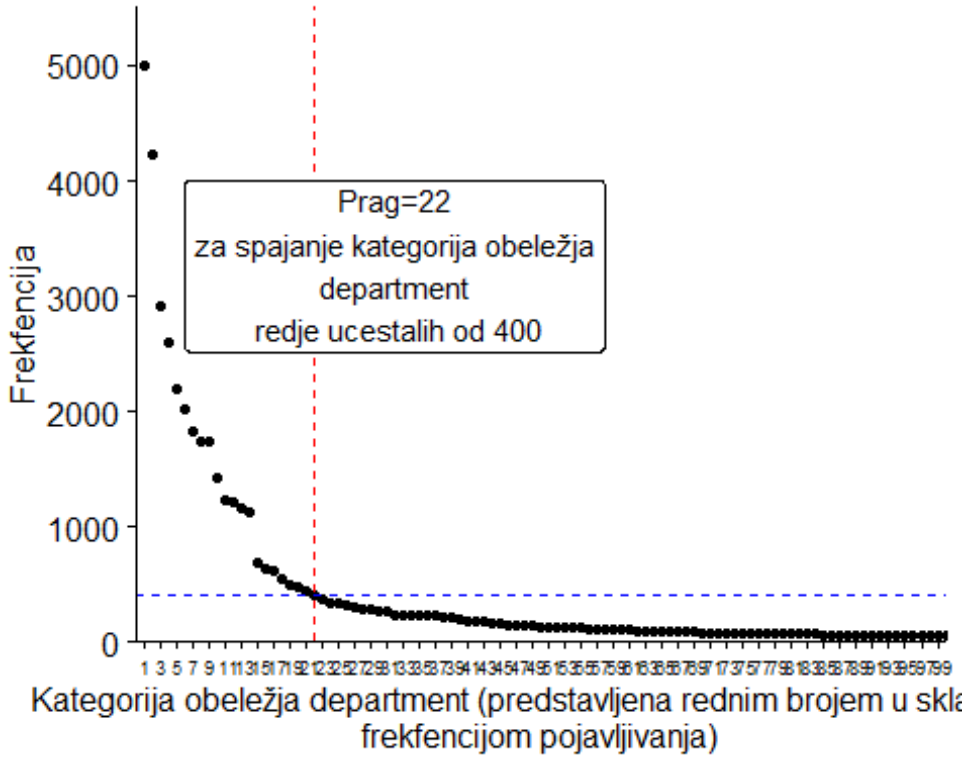
```
## 10 Telecommunication                          434    10
## # ... with 235 more rows

yLim <- round_up(tempDF,frekvencija,500)
# Prag predstavlja broj kategorija koje će biti zadržane
# Sve kategorije ispod praga bi'e spojene u jednu novu kategoriju
prag <- tempDF %>% filter(frekvencija>400) %>% pull(rbr) %>% .[length(.)]
# Cleveland tačkasti dijagram za obeležje major
mojplot <- ggdotchart(tempDF, x = "rbr", y = "frekvencija", sorting =
"descending")+
geom_vline(xintercept = prag, linetype = 2, color = "red")+
geom_hline(yintercept = 400, linetype = 2, color = "blue")+
theme(axis.text.x = element_text(angle = 0, vjust = 0.5, hjust=1,size=8))+
scale_y_continuous(expand=c(0,0),breaks = seq(0, yLim, by=1000))+
scale_x_discrete(expand=c(0,0),breaks = seq(1, 100,
by=2))+annotate(x=32,y=+Inf,
label=paste0("Prag=",prag,"\nza spajanje kategorija obeležja
major\n redje učestalih od 400"),vjust=2,geom="label")+
coord_cartesian(ylim = c(0,yLim),xlim = c(0,100))+
labs(x="Kategorija obeležja major (predstavljena rednim brojem u skladu sa
frekfencijom pojavljivanja)",
y="Frekfencija")+
theme(axis.text.x=element_text(size=7, angle=0,hjust=0.3,vjust=0.8,
color="black"))
mojplot
```



```r
# Združivanje retkih kategorija u novu kategoriju
novi_univerziteti <- novi_univerziteti %>% mutate(major = fct_lump(major,
prag), major = fct_infreq(major))

tempDF <- as_tibble(as.data.frame(table(novi_univerziteti$department))) %>%
arrange(desc(Freq)) %>%
mutate(rbr=seq_along(Var1)) %>%
rename_with(~c("department","frekvencija","rbr"))
tempDF
```

```
## # A tibble: 1,486 x 3
##    department                    frekvencija   rbr
##    <fct>                               <int> <int>
##  1 Computer Science                     5003     1
##  2 ECE                                  4231     2
##  3 Information Technology               2915     3
##  4 CSE                                  2598     4
##  5 Computer Engineering                 2190     5
##  6 0                                    2021     6
##  7 Mechanical Engineering               1822     7
##  8 IT                                   1741     8
##  9 Electronics and Communication        1732     9
## 10 Mechanical                           1425    10
## # ... with 1,476 more rows
```

```r
yLim <- round_up(tempDF,frekvencija,500)
# Prag predstavlja broj kategorija koje će biti zadržane
# Sve kategorije ispod praga biće spojene u jednu novu kategoriju
prag <- tempDF %>% filter(frekvencija>400) %>% pull(rbr) %>% .[length(.)]
# Cleveland tačkasti dijagram za obeležje department
mojplot <- ggdotchart(tempDF, x = "rbr", y = "frekvencija", sorting =
"descending")+
geom_vline(xintercept = prag, linetype = 2, color = "red")+
geom_hline(yintercept = 400, linetype = 2, color = "blue")+
theme(axis.text.x = element_text(angle = 0, vjust = 0.5, hjust=1,size=8))+
scale_y_continuous(expand=c(0,0),breaks = seq(0, yLim, by=1000))+
scale_x_discrete(expand=c(0,0),breaks = seq(1, 100,
by=2))+annotate(x=32,y=+Inf,
label=paste0("Prag=",prag,"\nza spajanje kategorija obeležja
department\n redje učestalih od 400"),vjust=2,geom="label")+
coord_cartesian(ylim = c(0,yLim),xlim = c(0,100))+
labs(x="Kategorija obeležja department (predstavljena rednim brojem u skladu
sa
frekfencijom pojavljivanja)",
y="Frekfencija")+
theme(axis.text.x=element_text(size=7, angle=0,hjust=0.3,vjust=0.8,
color="black"))
mojplot
```

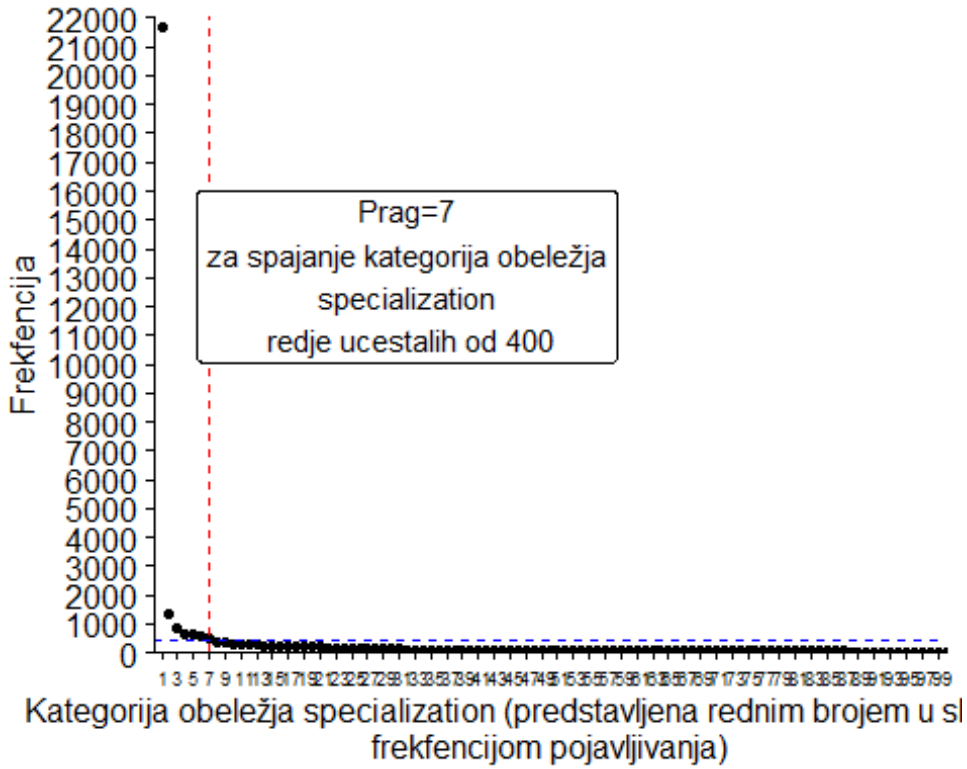



```
# Združivanje retkih kategorija u novu kategoriju
novi_univerziteti <- novi_univerziteti %>% mutate(department =
fct_lump(department, prag), department = fct_infreq(department))

tempDF <- as_tibble(as.data.frame(table(novi_univerziteti$specialization)))
%>%
arrange(desc(Freq)) %>%
mutate(rbr=seq_along(Var1)) %>%
rename_with(~c("specialization","frekvencija","rbr"))
tempDF
```

```
## # A tibble: 3,620 x 3
##    specialization            frekvencija   rbr
##    <fct>                           <int> <int>
##  1 ""                              21677     1
##  2 "VLSI"                           1321     2
##  3 "Embedded Systems"                827     3
##  4 "Computer Networks"               638     4
##  5 "Networks"                        614     5
##  6 "Networking"                      596     6
##  7 "Artificial Intelligence"         493     7
##  8 "Software Engineering"            368     8
##  9 "Systems"                         354     9
## 10 "General"                         301    10
## # ... with 3,610 more rows
```

```
yLim <- round_up(tempDF,frekvencija,500)
# Prag predstavlja broj kategorija koje će biti zadržane
```

```
# Sve kategorije ispod praga biće spojene u jednu novu kategoriju
prag <- tempDF %>% filter(frekvencija>400) %>% pull(rbr) %>% .[length(.)]
# Cleveland tačkasti dijagram za obeležje specialization
mojplot <- ggdotchart(tempDF, x = "rbr", y = "frekvencija", sorting =
"descending")+
geom_vline(xintercept = prag, linetype = 2, color = "red")+
geom_hline(yintercept = 400, linetype = 2, color = "blue")+
theme(axis.text.x = element_text(angle = 0, vjust = 0.5, hjust=1,size=8))+
scale_y_continuous(expand=c(0,0),breaks = seq(0, yLim, by=1000))+
scale_x_discrete(expand=c(0,0),breaks = seq(1, 100,
by=2))+annotate(x=32,y=+Inf,
label=paste0("Prag=",prag,"\nza spajanje kategorija obeležja
specialization\n redje učestalih od 400"),vjust=2,geom="label")+
coord_cartesian(ylim = c(0,yLim),xlim = c(0,100))+
labs(x="Kategorija obeležja specialization (predstavljena rednim brojem u
skladu sa
frekfencijom pojavljivanja)",
y="Frekfencija")+
theme(axis.text.x=element_text(size=7, angle=0,hjust=0.3,vjust=0.8,
color="black"))
mojplot
```

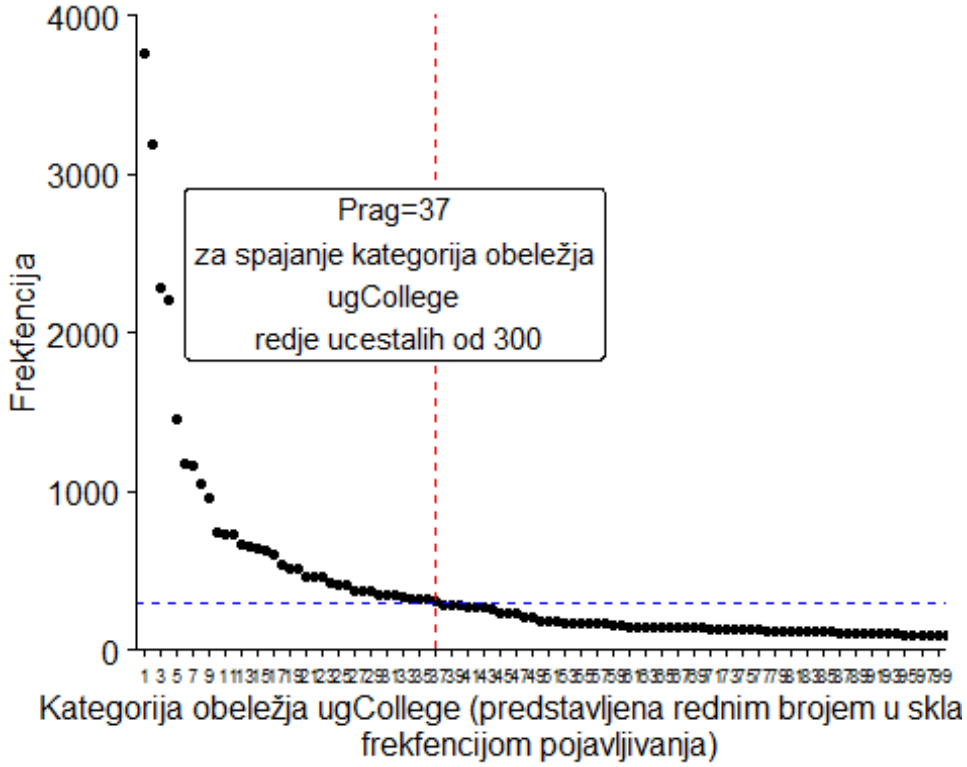



```
# Združivanje retkih kategorija u novu kategoriju
novi_univerziteti <- novi_univerziteti %>% mutate(specialization =
fct_lump(specialization, prag), specialization = fct_infreq(specialization))

tempDF <- as_tibble(as.data.frame(table(novi_univerziteti$ugCollege))) %>%
arrange(desc(Freq)) %>%
```

```
mutate(rbr=seq_along(Var1)) %>%
rename_with(~c("ugCollege","frekvencija","rbr"))
tempDF
```

```
## # A tibble: 1,823 x 3
##    ugCollege                    frekvencija   rbr
##    <fct>                              <int> <int>
##  1 "MU"                                3763     1
##  2 "VTU"                               3189     2
##  3 ""                                  2278     3
##  4 "Anna University"                   2206     4
##  5 "Pune University"                   1458     5
##  6 "JNTU"                              1176     6
##  7 "BITS Pilani"                       1166     7
##  8 "University of Mumbai"              1048     8
##  9 "University of Pune"                 953     9
## 10 "SSN College of Engineering"         746    10
## # ... with 1,813 more rows
```

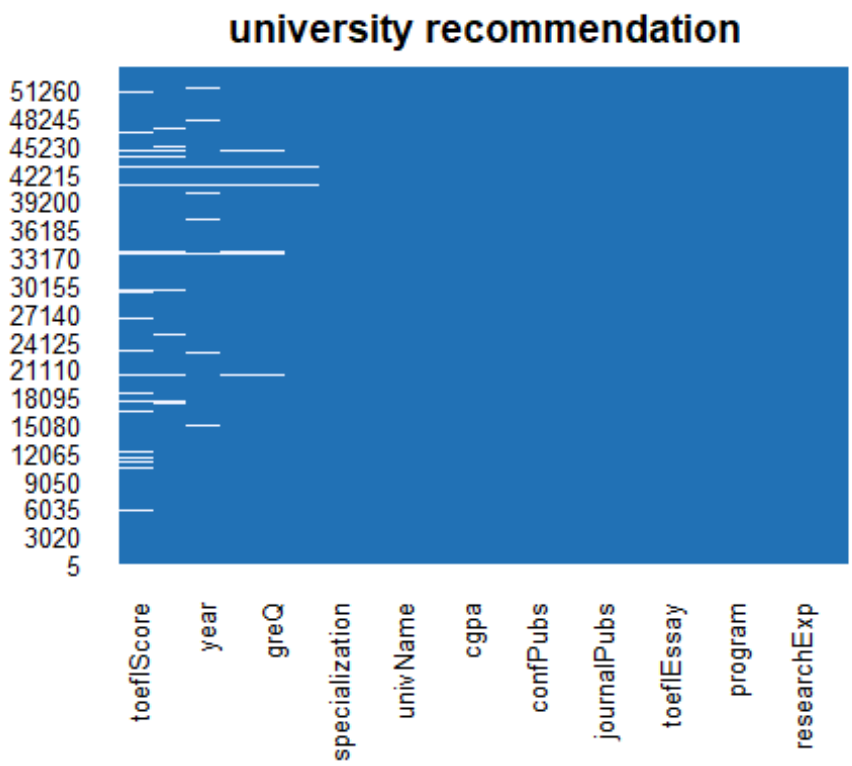
```
yLim <- round_up(tempDF,frekvencija,500)
# Prag predstavlja broj kategorija koje će biti zadržane
# Sve kategorije ispod praga biće spojene u jednu novu kategoriju
prag <- tempDF %>% filter(frekvencija>300) %>% pull(rbr) %>% .[length(.)]
# Cleveland tačkasti dijagram za obeležje ugCollege
mojplot <- ggdotchart(tempDF, x = "rbr", y = "frekvencija", sorting =
"descending")+
geom_vline(xintercept = prag, linetype = 2, color = "red")+
geom_hline(yintercept = 300, linetype = 2, color = "blue")+
theme(axis.text.x = element_text(angle = 0, vjust = 0.5, hjust=1,size=8))+
scale_y_continuous(expand=c(0,0),breaks = seq(0, yLim, by=1000))+
scale_x_discrete(expand=c(0,0),breaks = seq(1, 100,
by=2))+annotate(x=32,y=+Inf,
label=paste0("Prag=",prag,"\nza spajanje kategorija obeležja
ugCollege\n redje učestalih od 300"),vjust=2,geom="label")+
coord_cartesian(ylim = c(0,yLim),xlim = c(0,100))+
labs(x="Kategorija obeležja ugCollege (predstavljena rednim brojem u skladu sa
frekfencijom pojavljivanja)",
y="Frekfencija")+
theme(axis.text.x=element_text(size=7, angle=0,hjust=0.3,vjust=0.8,
color="black"))
mojplot
```




```
# Združivanje retkih kategorija u novu kategoriju
novi_univerziteti <- novi_univerziteti %>% mutate(ugCollege =
fct_lump(ugCollege, prag), ugCollege = fct_infreq(ugCollege))
```

Nedostajuće vrednosti

```
missmap(obj = novi_univerziteti, main = "university recommendation", legend =
FALSE)
```




```
(colMeans(is.na(novi_univerziteti)))*100
```

```
##          major    researchExp    industryExp specialization     toeflScore
##    0.000000000    0.000000000    0.000000000    0.005593884    8.230468022
##        program     department     toeflEssay      internExp           greV
##    0.000000000    0.000000000    0.000000000    0.000000000    2.341972776
##           greQ    journalPubs           greA     topperCgpa       confPubs
##    2.274846168    0.000000000    5.329106843    0.000000000    0.000000000
##      ugCollege           cgpa      cgpaScale       univName          admit
##    0.000000000    0.000000000    0.000000000    0.000000000    0.000000000
##           term           year
##    0.600410218    2.690658214
```

Specialization

```
xtabs(~novi_univerziteti$specialization)
```

```
## novi_univerziteti$specialization
##                   Other                                            VLSI
##                   27461                   21677                    1321
##        Embedded Systems       Computer Networks                Networks
##                     827                     638                     614
##              Networking Artificial Intelligence
##                     596                     493
```

```
length(which(is.na(novi_univerziteti$specialization)))
```

```
## [1] 3
```

Obeležje specialization sadrži samo 3 reda sa nedostajućim vrednostima, što predstavlja samo 0.005592424 procenata našeg okvira podataka, tako da te uzorke možemo slobodno obrisati.

```
novi_univerziteti <- novi_univerziteti[-
which(is.na(novi_univerziteti$specialization)), ] #funkcija za brisanje
uzoraka
```

Posmatranjem nivoa datog obeležja, primetili smo da više od trećine podataka imaju vrednost praznog stringa (““). Međutim, to nećemo gledati kao NA vrednosti, zato što nedostatak tih vrednosti u stvari predstavlja da dati studenti nemaju namenjenu specijalizaciju za visoke studije, što je sasvim legitimna stvar.

Jedino radi boljeg razumevanja te uzorke ćemo preimenovati u **No specialization**, ali da bismo to uradili vratićemo na trenutak obeležje u karakterni tip.

```
novi_univerziteti$specialization=as.character(novi_univerziteti$specialization
)

novi_univerziteti$specialization[which(novi_univerziteti$specialization=="")]=
"No specialization"

novi_univerziteti$specialization=as.factor(novi_univerziteti$specialization)
```

ToeflEssay

```
length(which(novi_univerziteti$toeflEssay==""))/dim(novi_univerziteti)[1]*100

## [1] 77.87868
```

Obeležje **toeflEssay** nema nedostajuće vrednosti, međutim, analizirajući podatke zaključili smo da čak 77 posto uzoraka imaju vrednost praznog stringa, odnosno (““). Međutim prikupljajući domensko znanje, saznali smo da celokupni TOEFL test se sastoji od 4 oblasti. Čitanja engleskog jezika, slušanja, pričanja i pisanja. Oblast vezana za pisanje može se polagati na dva načina. Jedan predstavlja čitanje kratkog odlomka i slušanja kratkog predavanja, a zatim se treba napisati odgovor na ono što je pročitano i saslušano. Drugi način predstavlja pisanje eseja na osnovu ličnog iskustva ili mišljenja kao odgovor na temu pisanja, čiji rezultati upravo predstavljaju dato obeležje. To znači da je u redu da korisnici nemaju vrednost za ovo obeležje, jer to znači su oblast pisanja polagali na prvi način.

Radi lakšeg manipulisanja, prazne string vrednosti konvertovaćemo u 0.

```
novi_univerziteti$toeflEssay[novi_univerziteti$toeflEssay==""]=0

length(which(novi_univerziteti$toeflEssay==""))

## [1] 0
```

ToeflScore

Kod ovog obeležja, nedostajuće vrednosti su sasvim slučajne, i nema načina detaljnije pretpostaviti nedostajuće vrednosti za svaki uzorak. Zbog toga ćemo nedostajuće vrednosti popuniti medijanom ili prosečnom vrednošću obeležja, a normalnost podataka ćemo zaključiti Shapiro-Wilk" testom. S obzirom da je nemoguće ovu metodu primeniti nad velikim brojem uzoraka, primenićemo sledeći postupak. Hiljadu puta ćemo uzorkovati bez ponavljanja 3000 nasumičnih opservacija iz celog okvira podataka. Nad svakim od hiljadu uzoraka biće primenjen Shapiro-Wilk" test. Prosečna p-vrednost svih 1000 testova biće izabrana kao merodavna za statističko zaključivanje.

```
p_vrednosti = replicate(1000,
shapiro.test(sample(novi_univerziteti$toeflScore,3000))$p.val)
prosecna_pvrednost = mean(p_vrednosti)
prosecna_pvrednost
```

```
## [1] 2.721737e-70
```

Nakon uspešno izvršenog eksperimenta, s obzirom da je prosečna p vrednost znatno manja od 0.05, to nam govori da nam raspodela podataka nije normalna, i da treba koristiti medijanu. Tako da ćemo sve nedostajuće vrednosti popuniti medijanom obeležja.

```
novi_univerziteti$toeflScore[is.na(novi_univerziteti$toeflScore)]=median(novi_
univerziteti$toeflScore,na.rm = TRUE) #funkcija za popunjavanje medijanom
length(which(is.na(novi_univerziteti$toeflScore))) #provera
```

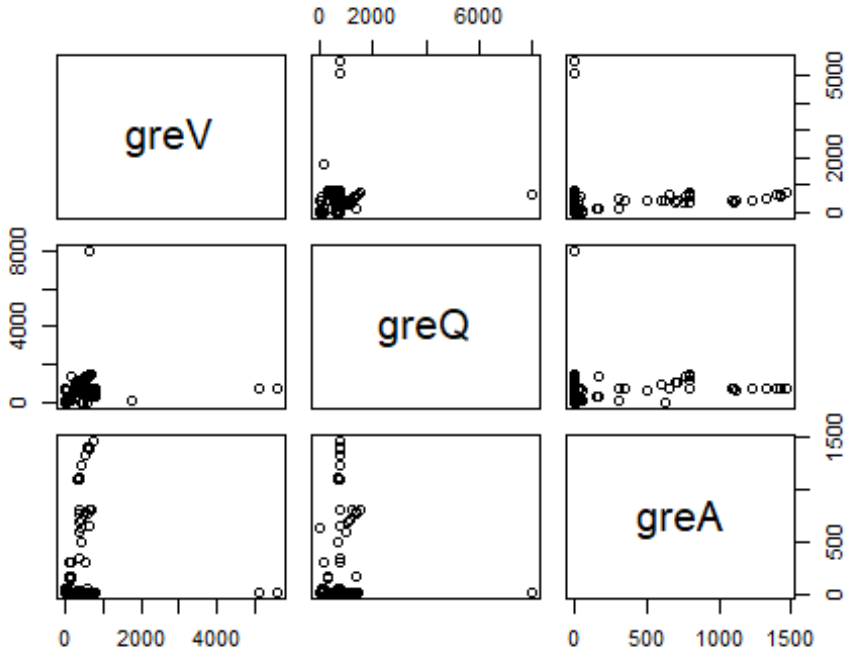
```
## [1] 0
```

GreV, greQ i greA

GRE je standardizovani test koji postoji od 1936. godine i koji izračunava tri veoma važna parametra: verbalno i kvantitativno rezonovanje i analitičko pisanje. Obeležje greV jeste ocena za verbalno rezonovanje, greQ za kvantitavino, dok greA obeležava analitičko pisanje.

S obzirom da imamo dosta uzoraka gde postoji nedostajuća vrednost samo za jedno od obeležja, proverićemo da li postoji korelacija između ova tri obeležja.

```
pairs(novi_univerziteti[c(10,11,13)])
```



Na osnovu plotova možemo jasno zaključiti da ne postoji nikakva korelacija između greA i druga dva obeležja, dok na osnovu slike nismo sto posto sigurni da li postoji korelacija između obeležja greV i greQ, tako da ćemo funkcijom koja izračunava korelaciju zaključiti da li korelacija postoji.

Za početak moramo pripremiti podatke, odnosno da izbacimo sve nedostajuće vrednosti da bismo mogli da izračunamo korelaciju. Pošto broj uzoraka mora biti identičan za oba obeležja, uporedićemo onoliko uzoraka koliko kraće obeležje ima uzoraka.

```
greV= na.omit(novi_univerziteti$greV) #izbacivanje NA vrednosti
greQ= na.omit(novi_univerziteti$greQ)

l1= length(greV) #greV ima manje uzoraka nakon izbacivanja NA vrednosti
l2= length(greQ)

cor(greV[1:l1],greQ[1:l1])

## [1] 0.02716113
```

Nakon određivanja korelacije, rezultat pokazuje da ne postoji korelacija između ova dva obeležja. To znači da nikako na osnovu jednog obeležja ne možemo tačnije pretpostaviti vrednost drugog obeležja.

```
length(which(is.na(novi_univerziteti$greV)))

## [1] 1256
```

```
length(which(is.na(novi_univerziteti$greQ)))
```

```
## [1] 1220
```

```
length(which(is.na(novi_univerziteti$greA)))
```

```
## [1] 2858
```

S obzirom da broj uzoraka sa nedostajućim vrednostima nije mali, nećemo ih obrisati, jer nam mogu biti značajni u daljoj analizi, već ćemo ih popuniti na osnovu medijane ili prosečne vrednosti svih uzoraka za to obeležje, a to ćemo zaključiti na osnovu Shapiro-Wilk" testa. Postupak će biti sličan kao za jedno od prethodnih obeležja, hiljadu puta ćemo uzorkovati bez ponavljanja 3000 nasumičnih opservacija iz celog okvira podataka. Nad svakim od hiljadu uzoraka biće primenjen Shapiro-Wilk" test. Prosečna p-vrednost svih 1000 testova biće izabrana kao merodavna za statističko zaključivanje.

```
p_vrednosti = replicate(1000,
shapiro.test(sample(novi_univerziteti$greV,3000))$p.val)
prosecna_pvrednost = mean(p_vrednosti)
prosecna_pvrednost
```

```
## [1] 2.885855e-53
```

```
p_vrednosti = replicate(1000,
shapiro.test(sample(novi_univerziteti$greQ,3000))$p.val)
prosecna_pvrednost = mean(p_vrednosti)
prosecna_pvrednost
```

```
## [1] 3.023392e-59
```

```
p_vrednosti = replicate(1000,
shapiro.test(sample(novi_univerziteti$greA,3000))$p.val)
prosecna_pvrednost = mean(p_vrednosti)
prosecna_pvrednost
```

```
## [1] 2.010341e-65
```

S obzirom da je kod sva tri obeležja prosečna p vrednost znatno manja od 0.05, to nam govori da nam raspodela podataka nije normalna, i da treba koristiti medijanu. Tako da ćemo sve nedostajuće vrednosti popuniti medijanom obeležja.

```
novi_univerziteti$greQ[is.na(novi_univerziteti$greQ)]=median(novi_univerziteti
$greQ,na.rm = TRUE) #funkcija za popunjavanje medijanom
length(which(is.na(novi_univerziteti$greQ))) #provera za greQ
```

```
## [1] 0
```

```
novi_univerziteti$greV[is.na(novi_univerziteti$greV)]=median(novi_univerziteti
$greV,na.rm = TRUE) #funkcija za popunjavanje medijanom
length(which(is.na(novi_univerziteti$greV))) #provera za greV
```

```
## [1] 0
```

```
novi_univerziteti$greA[is.na(novi_univerziteti$greA)]=median(novi_univerziteti
$greA,na.rm = TRUE) #funkcija za popunjavanje medijanom
length(which(is.na(novi_univerziteti$greA))) #provera za greA

## [1] 0
```

Term i year

S obzirom da nema previše nedostajućih vrednosti iz ovih obeležja, a ne postoji dobar način za njihovo popunjavanje, obrisaćemo sve uzorke koje sadrže NA vrednosti iz ovog obeležja.

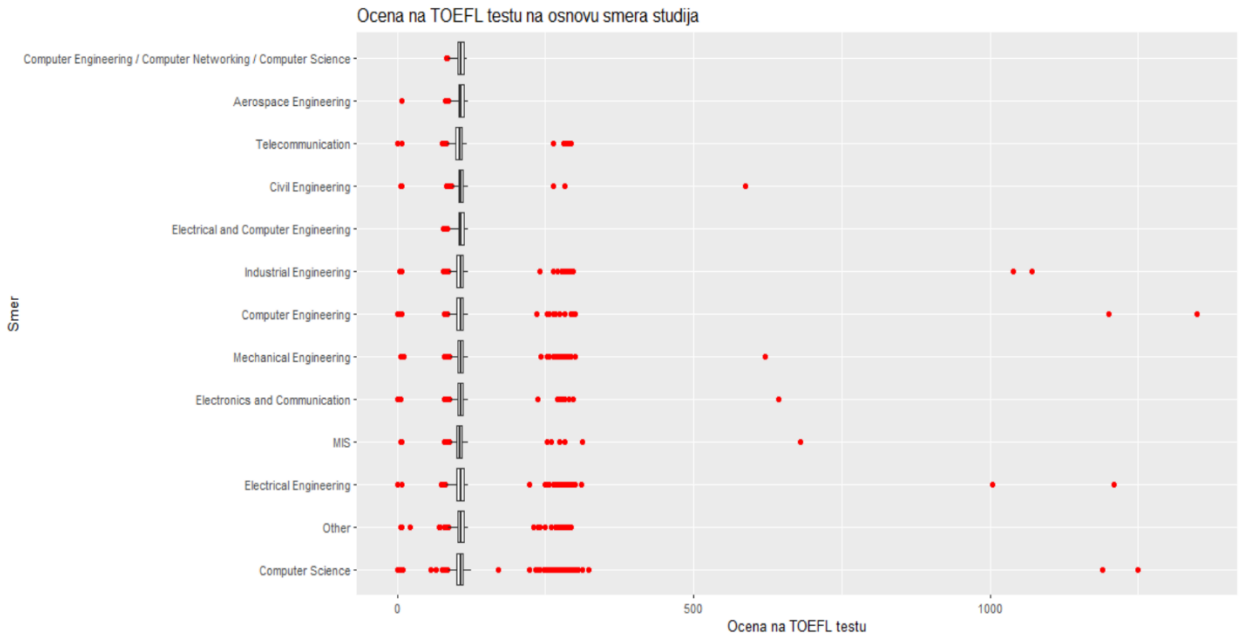
```
novi_univerziteti= na.omit(novi_univerziteti)
```

Izuzeci

Veoma je važno detektovati i otkloniti izuzetke kako bismo napravili što bolji model i kako bismo smanjili grešku.

toeflScore

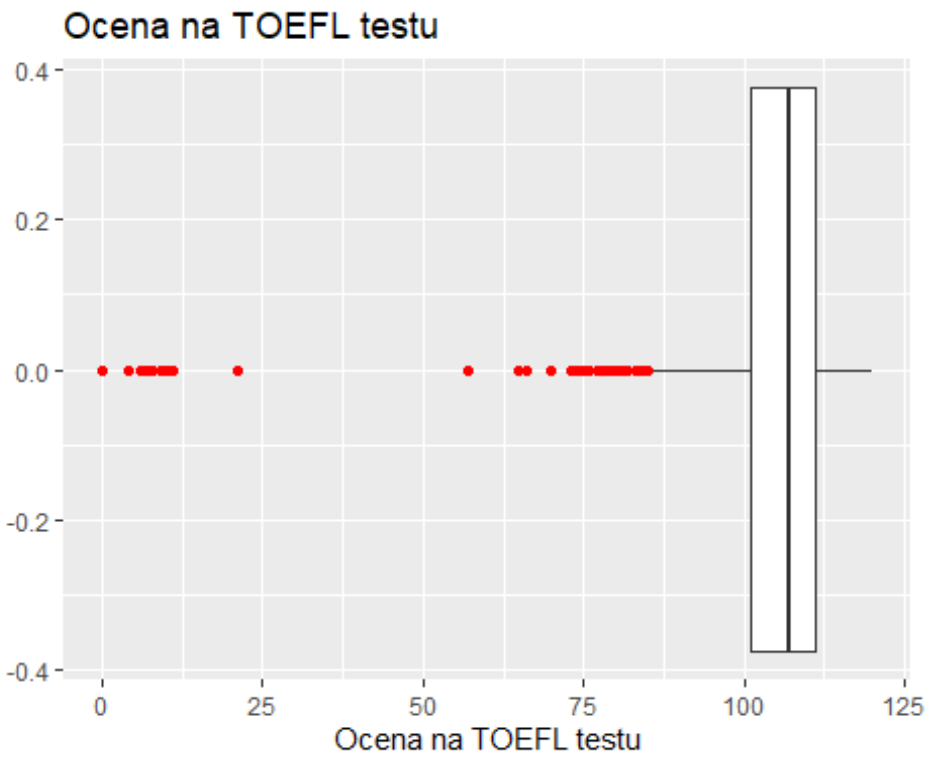
```
ggplot(novi_univerziteti, aes(x = toeflScore, y = major)) +
geom_boxplot(outlier.colour = "red") + labs(title = "Ocena na TOEFL testu na
osnovu smera studija",
     x = "Ocena na TOEFL testu",
     y = "Smer")
```



```
xtabs(~novi_univerziteti$toeflScore)

## novi_univerziteti$toeflScore
##    0    4    6    7    8    9   10   11   21   57   65   66   70   73   74
75
##   22    6   23  146   43    6    2    1    2    2    2    1    1    1    4
1
##   76   77   78   79   80   81   82   83   84   85   86   87   88   89   90
91
##    4    1    3   13   91   85  116   94  131  168  218  228  248  245  486
591
##   92   93   94   95   96   97   98   99  100  101  102  103  104  105  106
107
##  596  676  710  882  815 1075 1160 1147 1790 1715 1813 1770 2342 2383 2115
6398
##  108  109  110  111  112  113  114  115  116  117  118  119  120  124  170
223
## 2592 2206 2726 2212 2330 1920 1929 1523 1226  915  573  316   87    1    1
4
##  230  233  235  237  240  243  247  250  253  256  257  260  263  267  270
273
##    2   11    1    9    5    1    9    8   24    1   31   25   38   47   77
125
##  275  277  280  283  287  290  293  297  300  306  310  312  313  322  587
620
##    7   79  108  172  128  117   96   53   44    1    1    4    2    2    1
1
##  643  680 1004 1040 1070 1190 1200 1210 1250 1350
##    1    2    5    1    1    1    3    4    2    1
```

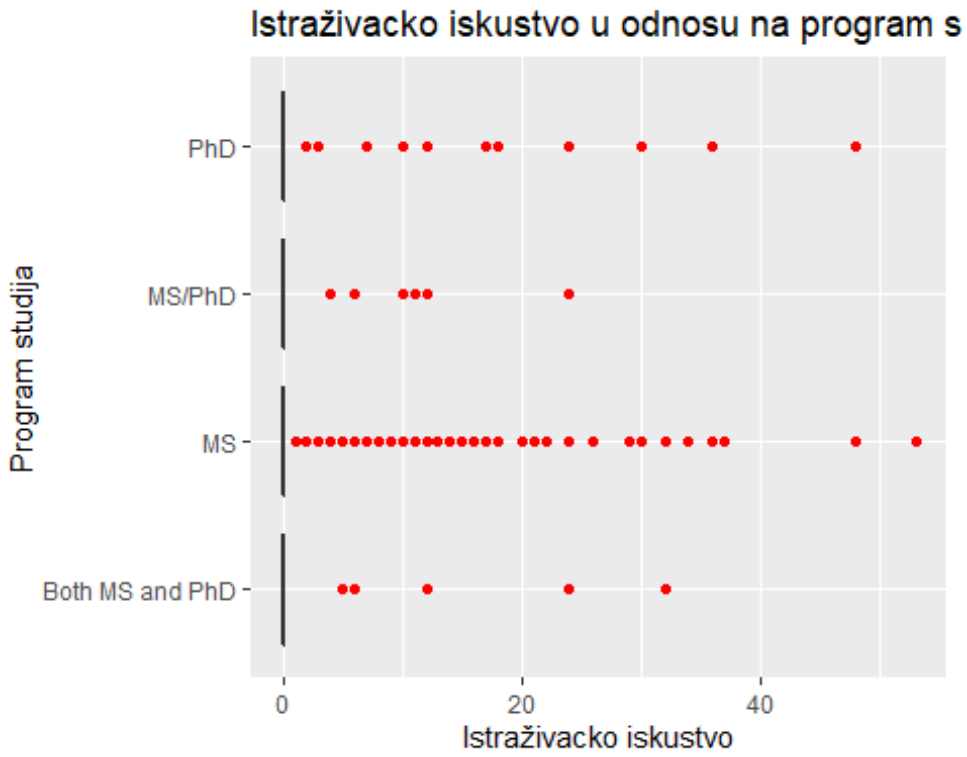
Sa grafika odnosa ocene na TOEFL testu i smera studija vidimo da ne postoji značajna razlika u oceni na TOEFL testu po smerovima studija. Kao što možemo videti, postoji veliki broj izuzetaka, međutim od toga nisu zaista sve izuzeci, ali nam je tako grafikom predstavljeno jer 80 posto podataka u opsegu između 91 i 117 poena. Po literaturi, broj poena na ovom testu može biti između 0 i 120 poena, tako da ćemo sve ostale uzorke odbaciti.

```
novi_univerziteti = novi_univerziteti[which(novi_univerziteti$toeflScore>=0 &
novi_univerziteti$toeflScore<=120),]

ggplot(novi_univerziteti, aes(x = toeflScore )) + geom_boxplot(outlier.colour
= "red") + labs(title = "Ocena na TOEFL testu",
     x = "Ocena na TOEFL testu",
     y = "")
```



researchExp

```
ggplot(novi_univerziteti, aes(x = researchExp, y = program)) +
geom_boxplot(outlier.colour = "red") + labs(title = "Istraživačko iskustvo u
odnosu na program studija",
     x = "Istraživačko iskustvo",
     y = "Program studija")
```



Sa grafika odnosa istraživačkog iskustva i programa studija vidimo da ne postoji značajna razlika u istraživačkom iskustvu po programu studija. Ono što možemo zaključiti na osnovu grafika jeste da veliki broj uzoraka nemaju istraživačko iskustvo, i zato sve one osobe koje ga poseduju smatraju se izuzecima. Međutim mi te vrednosti nećemo ukloniti, jer u tom slučaju u našem okviru podataka imali bi samo one uzorke bez istraživačkog iskustva i samim tim ni to obeležje ne bi predstavljalo nikakav validan parametar za kasnije kreiranje modela.

Prvo ćemo proveriti koje sve jedinstvene vrednosti posedujemo, i u kom broju.

```
xtabs(~novi_univerziteti$researchExp)
```

```
## novi_univerziteti$researchExp
##     0     1     2     3     4     5     6     7     8     9    10    11
12
## 49411    24   178    88    62    37   299    39    36    24    43    10
235
##    13    14    15    16    17    18    20    21    22    24    26    29
30
##     3    11     8    23    21    77     8     8     7   150     1    12
24
##    32    34    36    37    48    53
##     7     3    55     2    19     2
```

```
length(unique(novi_univerziteti$researchExp))
```

```
## [1] 32
```

```
length(which(novi_univerziteti$researchExp==0))/dim(novi_univerziteti)[1]*100
```

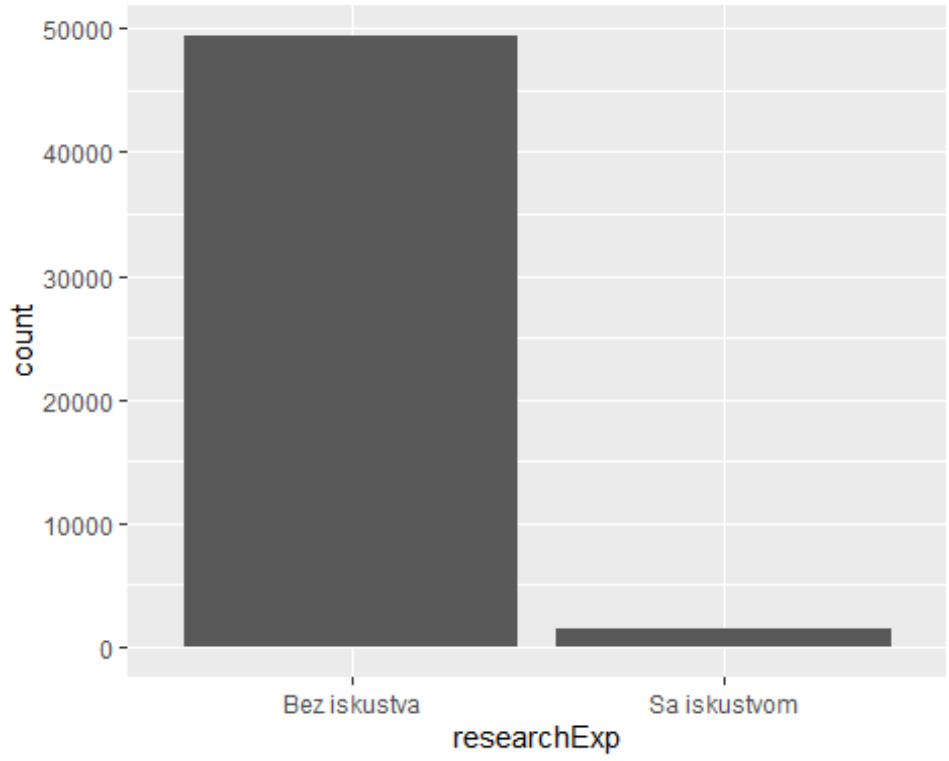
```
## [1] 97.02319
```

Primećujemo da postoje 32 jedinstvene vrednosti, međutim čak 97% uzoraka nemaju istaživačko iskustvo. Zbog toga, sve ostale osobe koje imaju istraživačko iskustvo grupisaćemo u jednu celinu, i dato obeležje ćemo podeliti u 2 kategorije: 1. Osobe bez istraživačkog iskustva. 2. Osobe sa istraživačkim iskustvom.

```
novi_univerziteti$researchExp =
as.numeric(as.character(novi_univerziteti$researchExp))
novi_univerziteti$researchExp[novi_univerziteti$researchExp > 0] = 1
unique(novi_univerziteti$researchExp)
```

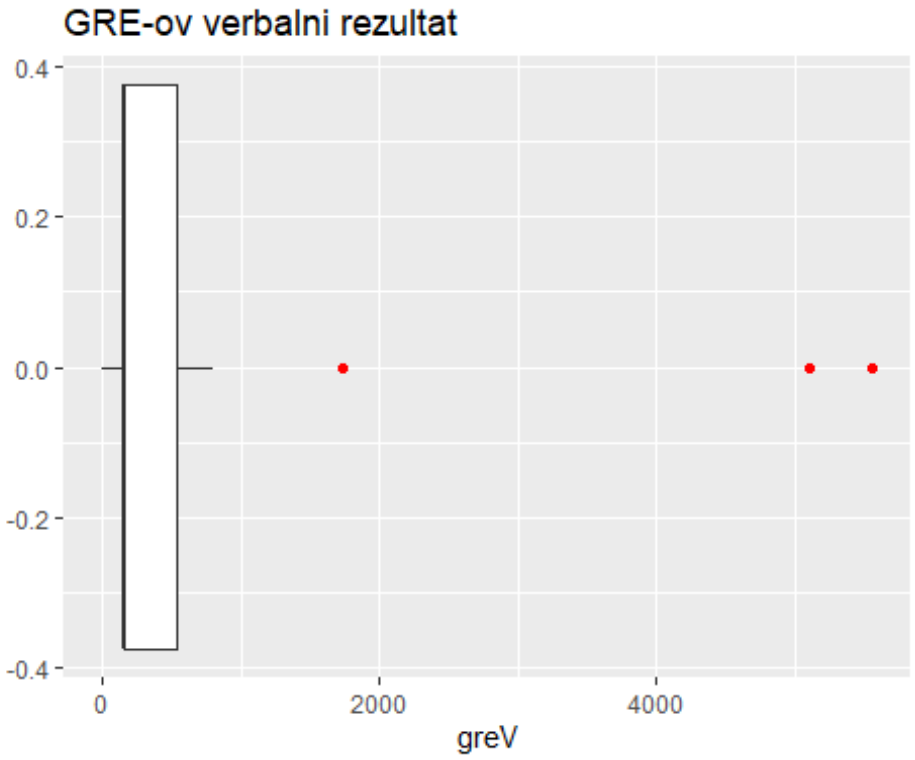
```
## [1] 0 1
```

```
novi_univerziteti$researchExp = factor(novi_univerziteti$researchExp,
                                       levels = c(0,1), labels = c("Bez
iskustva","Sa iskustvom"))
#novi_univerziteti$researchExp[is.na(novi_univerziteti$researchExp)] = 0 cemu
ovo

novi_univerziteti$researchExp=as.factor(novi_univerziteti$researchExp)

ggplot(novi_univerziteti) + geom_bar(aes(x = researchExp))
```

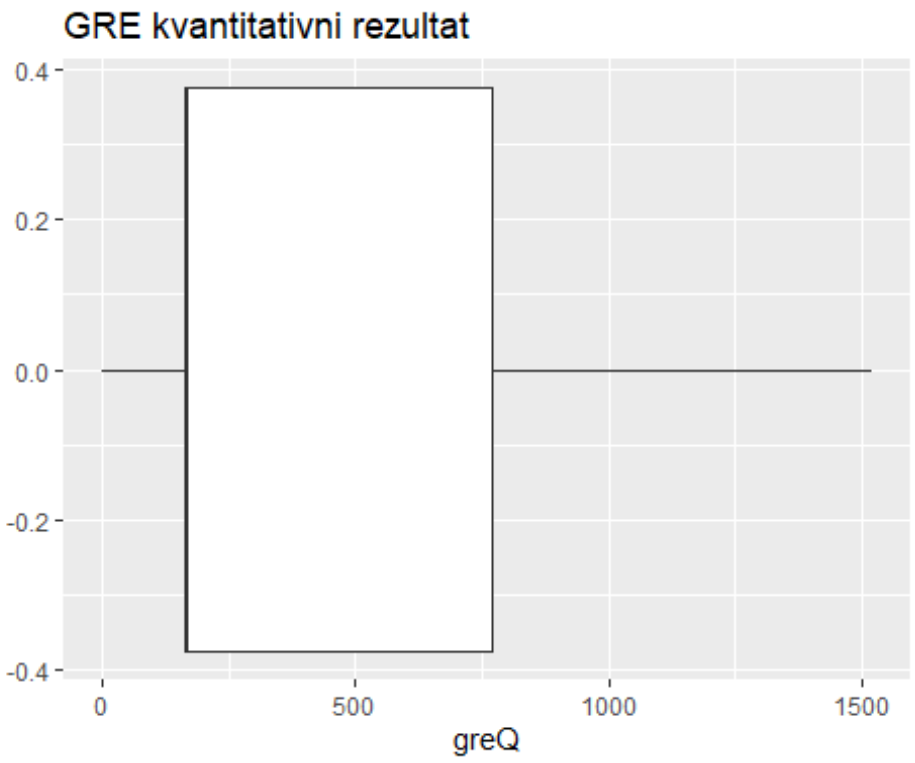


greV, greQ, greA

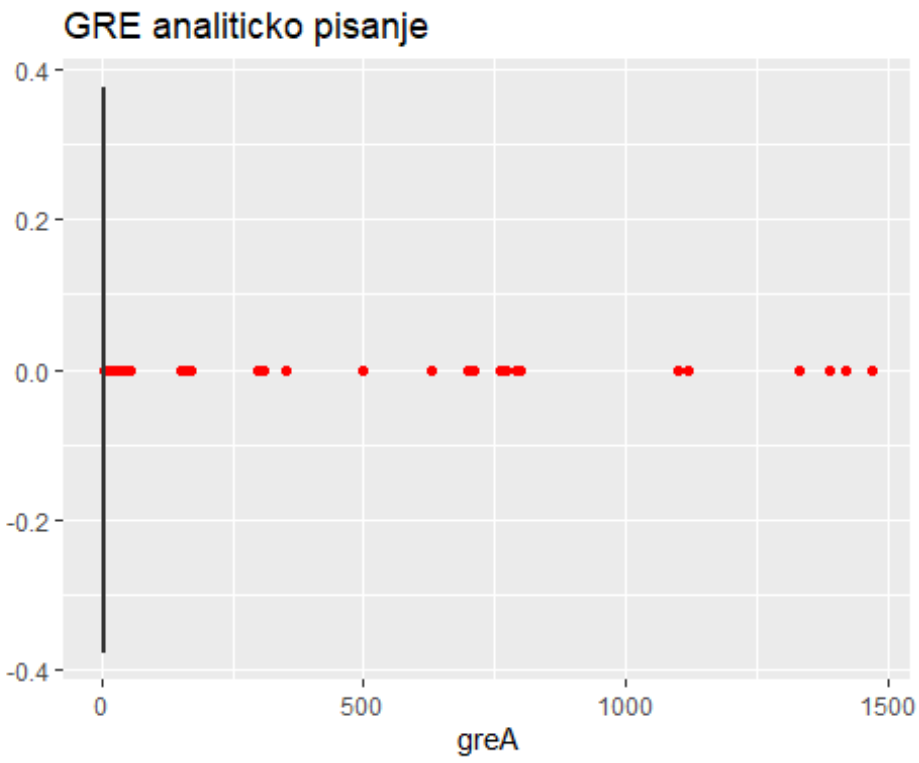
```
ggplot(novi_univerziteti, aes(x = greV)) + geom_boxplot(outlier.colour =
"red") +
  labs(title = "GRE-ov verbalni rezultat",
       x = "greV",
       y = "")
```




```
ggplot(novi_univerziteti, aes(x = greQ)) + geom_boxplot(outlier.colour =
"red") +
  labs(title = "GRE kvantitativni rezultat",
       x = "greQ",
       y = "")
```




```
ggplot(novi_univerziteti, aes(x = greA)) + geom_boxplot(outlier.colour =
"red") +
  labs(title = "GRE analitičko pisanje",
       x = "greA",
       y = "")
```

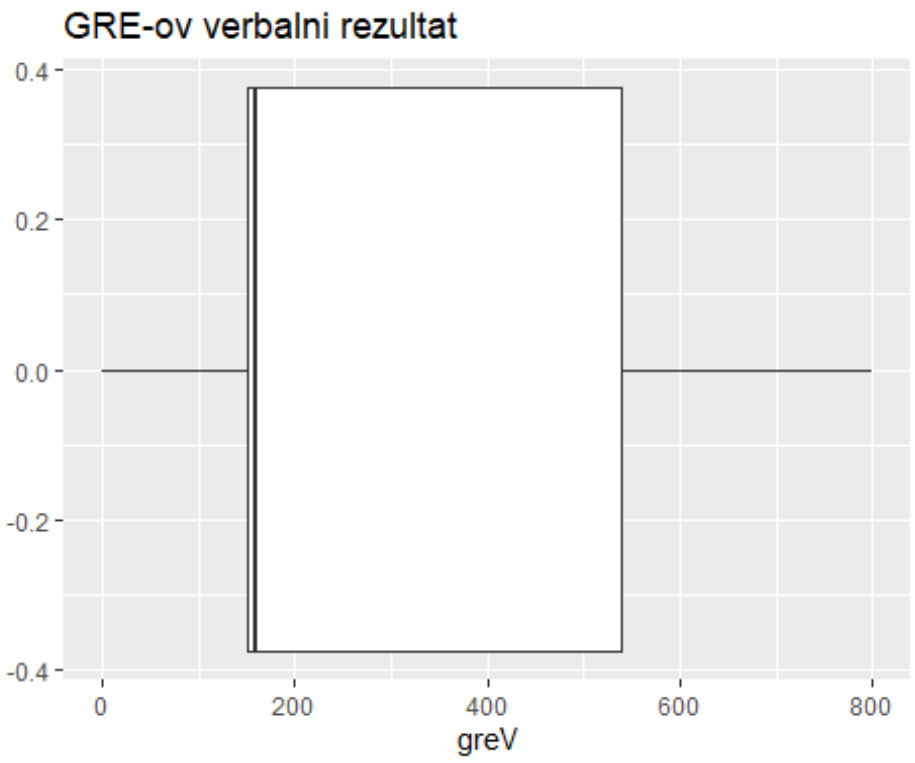


Što se tiče verbalnih i kvantitativnih rezultata GRE testa, današnja vrednost je u opsegu od 0 do 170 poena. Međutim, do 2011. godine, najveći mogući rezultat za ova dva testa iznosio je 800 poena, a uzorci iz našeg okvira podataka pripadaju tom periodu, tako da će izuzeci za ovo obeležje biti sve vrednosti van pomenutog opsega.

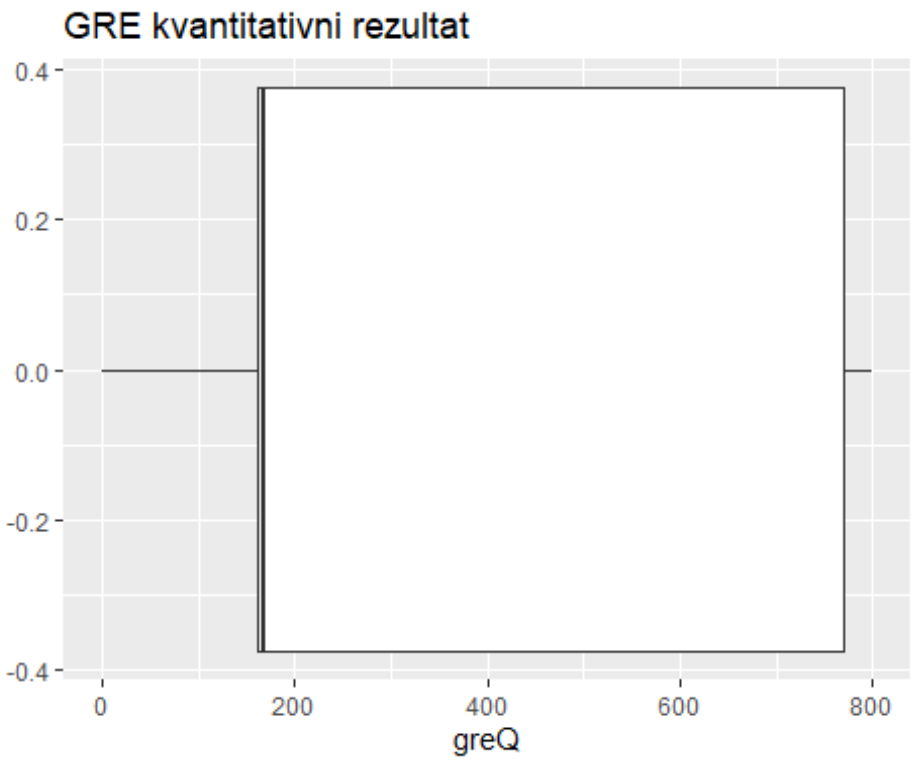
```
novi_univerziteti=novi_univerziteti[which(novi_univerziteti$greV>=0 &
novi_univerziteti$greV<=800),]

novi_univerziteti=novi_univerziteti[which(novi_univerziteti$greQ>=0 &
novi_univerziteti$greQ<=800),]
```

Sada data obeležja izgledaju ovako.

```
ggplot(novi_univerziteti, aes(x = greV)) + geom_boxplot(outlier.colour =
"red") +
  labs(title = "GRE-ov verbalni rezultat",
       x = "greV",
       y = "")
```




```
ggplot(novi_univerziteti, aes(x = greQ)) + geom_boxplot(outlier.colour =
"red") +
  labs(title = "GRE kvantitativni rezultat",
       x = "greQ",
       y = "")
```

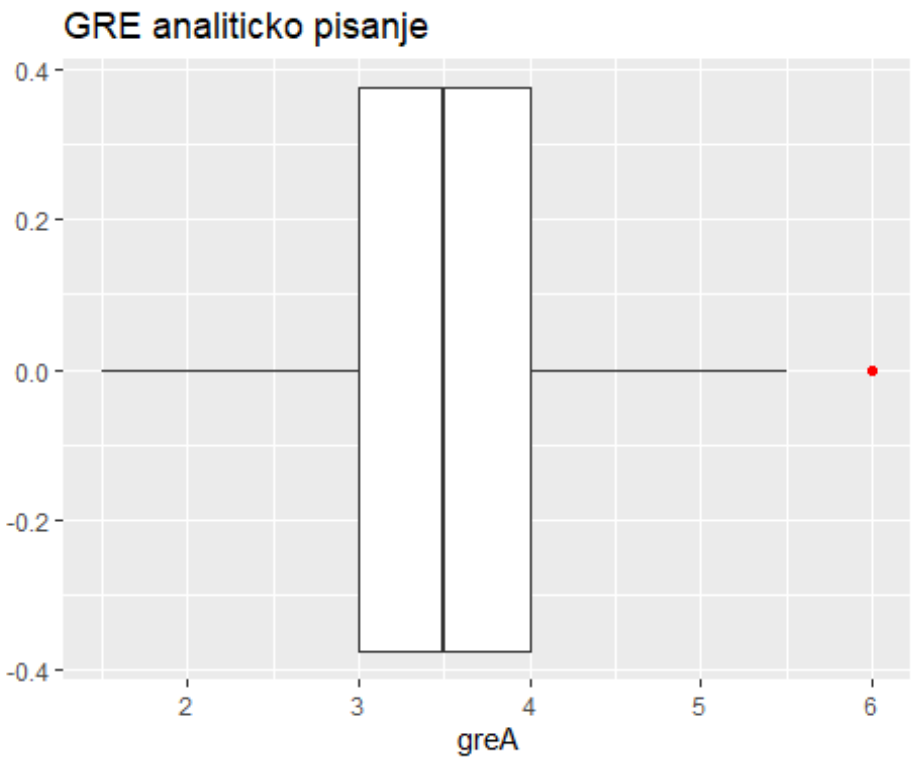


Što se tiče GRE testa za analitičko pisanje, opseg poena se nije menjao i maksimalni broj poena je i dalje 6. Tako da sve vrednosti veće od te predstavljaju izuzetke i biće uklonjeni.

```
novi_univerziteti=novi_univerziteti[which(novi_univerziteti$greA>=0 &
novi_univerziteti$greA<=6),]
```

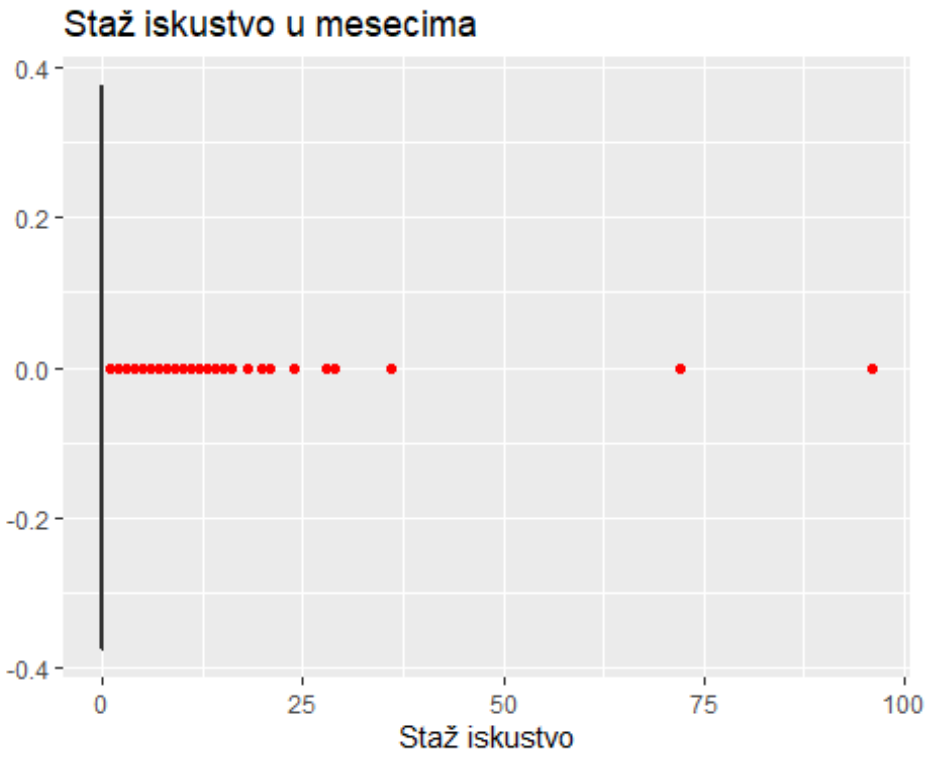
Sada dato obeležje ima drugačiji izgled.

```
ggplot(novi_univerziteti, aes(x = greA)) + geom_boxplot(outlier.colour =
"red") +
  labs(title = "GRE analitičko pisanje",
       x = "greA",
       y = "")
```



internExp

```
ggplot(novi_univerziteti) + geom_boxplot( aes(x = internExp),outlier.colour =
"red") + labs(title = "Staž iskustvo u mesecima",
     x = "Staž iskustvo",
     y = "")
```



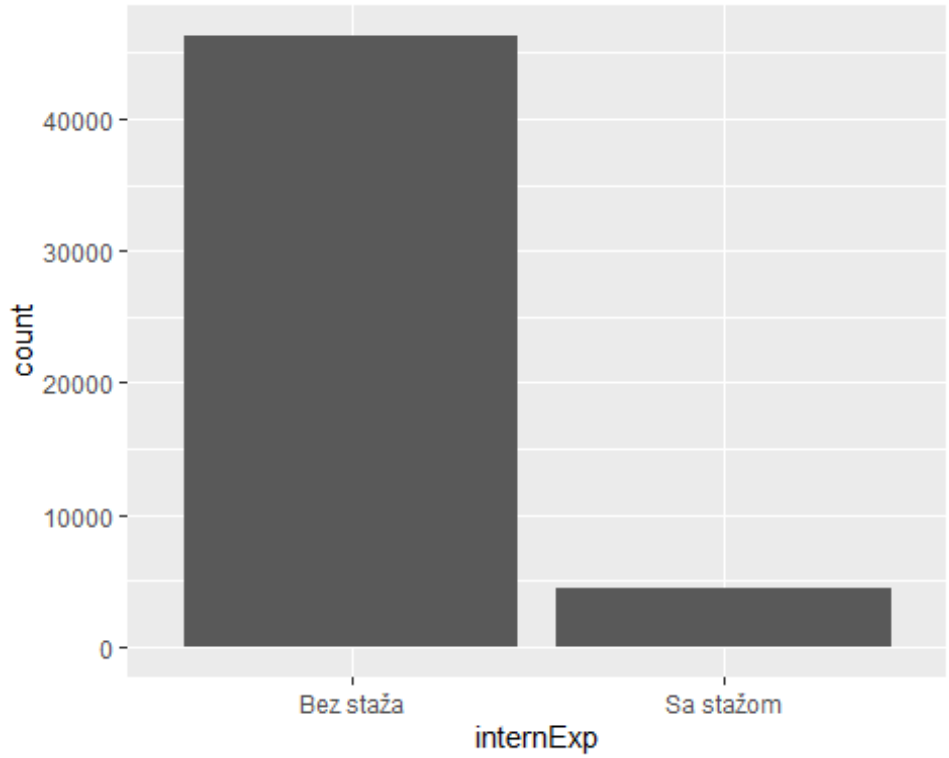
Na osnovu grafika kod iskustva u stažu, možemo primetiti istu stvar kao kod obeležja koje predstavlja istraživačko iskustvo, a to je da veliki broj uzoraka nemaju staža, i zato sve one osobe koje ga poseduju smatraju se izuzecima. Međutim mi te vrednosti nećemo ukloniti, jer u tom slučaju u našem okviru podataka imali bi samo one uzorke bez istraživačkog iskustva i samim tim ni to obeležje ne bi predstavljalo nikakav validan parametar za kasnije kreiranje modela.

Prvo ćemo proveriti koje sve jedinstvene vrednosti posedujemo, i u kom broju.

```
xtabs(~novi_univerziteti$internExp)

## novi_univerziteti$internExp
##     0     1     2     3     4     5     6     7     8     9    10    11
12
## 46297   409   967   640   398   214  1000   105   232    64    52    12
254
##    13    14    15    16    18    20    21    24    28    29    36    72
96
##     7    21     9     4     9     1     8    27     7     3     8     4
15

length(unique(novi_univerziteti$internExp))

## [1] 26

length(which(novi_univerziteti$internExp==0))/dim(novi_univerziteti)[1]*100

## [1] 91.19507
```

Primećujemo da postoje 26 jedinstvenih vrednosti, međutim čak 91% uzoraka nemaju staž iskustvo. Zbog toga, sve ostale osobe koje imaju iskustvo staža grupisaćemo u jednu celinu, i dato obeležje ćemo podeliti u 2 kategorije: 1. Osobe bez staža 2. Osobe sa stažom.

```
novi_univerziteti$internExp =
as.numeric(as.character(novi_univerziteti$internExp))
novi_univerziteti$internExp[novi_univerziteti$internExp > 0] = 1
unique(novi_univerziteti$internExp)

## [1] 1 0

novi_univerziteti$internExp = factor(novi_univerziteti$internExp,
                                     levels = c(0,1), labels = c("Bez
staža","Sa stažom"))
#novi_univerziteti$internExp[is.na(novi_univerziteti$internExp)] = 0

novi_univerziteti$internExp=as.factor(novi_univerziteti$internExp)

ggplot(novi_univerziteti) + geom_bar(aes(x = internExp))
```

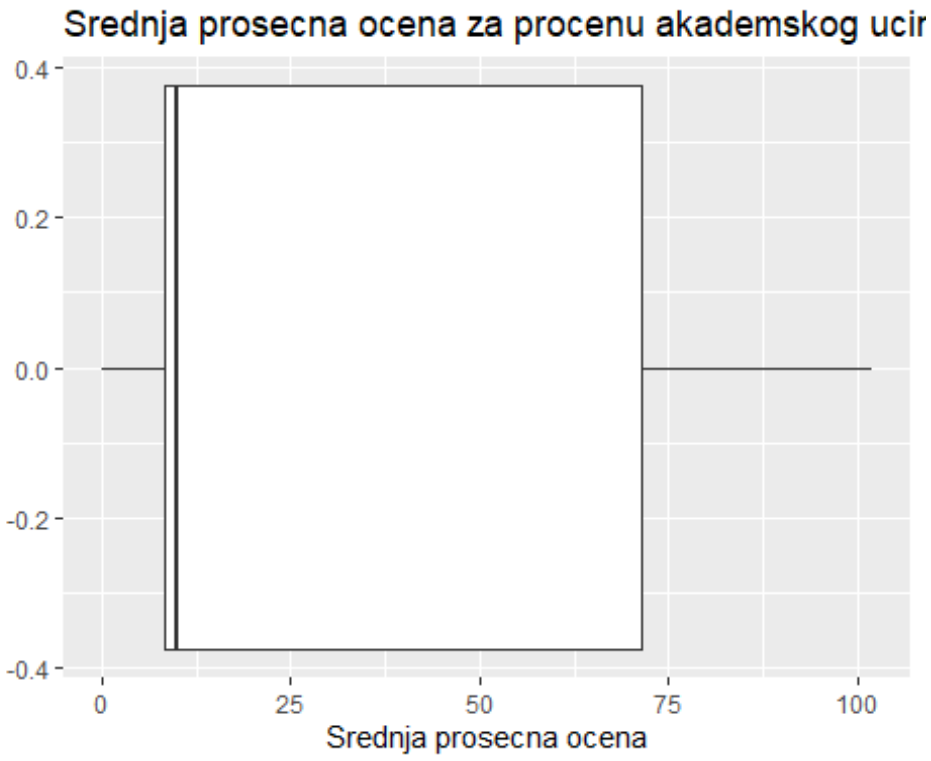


Cgpa

```
ggplot(novi_univerziteti) + geom_boxplot( aes(x = cgpa ),outlier.colour =
"red") + labs(title = "Srednja prosečna ocena za procenu akademskog učinka",
     x = "Srednja prosečna ocena",
     y = "")
```

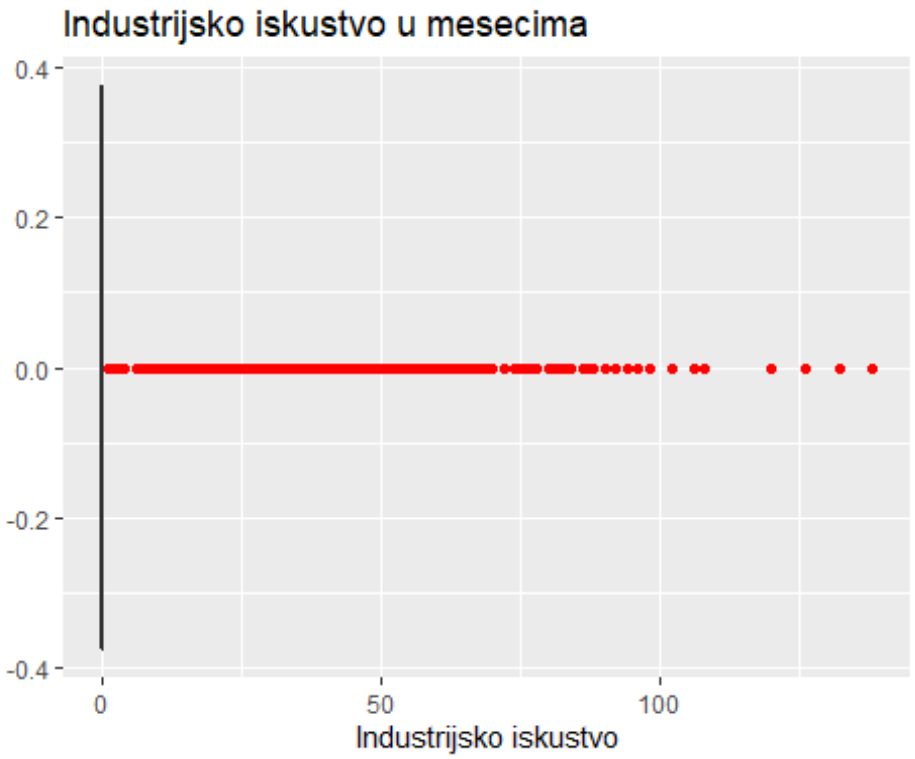
Primećujemo da postoji veliki izuzetak koji ćemo ukloniti.

```
outlajeri = boxplot(novi_univerziteti$cgpa, plot = FALSE)$out

novi_univerziteti = novi_univerziteti[-which(novi_univerziteti$cgpa %in%
outlajeri),]
ggplot(novi_univerziteti, aes(x = cgpa )) + geom_boxplot(outlier.colour =
"red") + labs(title = "Srednja prosečna ocena za procenu akademskog učinka",
     x = "Srednja prosečna ocena",
     y = "")
```



industryExp

```
ggplot(novi_univerziteti) + geom_boxplot( aes(x = industryExp ),outlier.colour
= "red") + labs(title = "Industrijsko iskustvo u mesecima",
     x = "Industrijsko iskustvo",
     y = "")
```



Ni ovo obeležje koje predstavlja iskustvo ne razlikuje se sa dosadašnjim. Međutim, ovde već na osnovu grafika možemo zaključiti da postoje veliki izuzeci koje bi trebalo ukloniti, a to su uzorci čiji je staž preko 110 nedelja.

```
novi_univerziteti=novi_univerziteti[-
which(novi_univerziteti$industryExp>110),]
```

Preostale podatke ćemo obraditi po istom principu.

```
xtabs(~novi_univerziteti$industryExp)

## novi_univerziteti$industryExp
##     0     1     2     3     4     6     7     8     9    10    11    12
13
## 43772    34    58    77    22   125    47    56    71    57    23   400
37
##    14    15    16    17    18    19    20    21    22    23    24    25
26
##    78   106   122   105   255   106   157    88   191    54   936    95
85
##    27    28    29    30    31    32    33    34    35    36    37    38
39
##    97    91   191   301    88   152   122   105    30   517    46    57
35
##    40    41    42    43    44    45    46    47    48    49    50    51
52
##    73    73   214    74    92    65    58    15   250    40    33    59
42
```

```
##    53    54    55    56    57    58    59    60    61    62    63    64
65
##    39   100    32    31    38    60    12    87    14     6    12     6
12
##    66    67    68    69    70    72    74    75    76    77    78    80
81
##    38     1     6    17    14    16     3     8     6    15    12     1
7
##    82    83    84    86    87    88    90    92    94    96    98   102
106
##     1     5    28     1    10     2    16     2     1     5     1     2
6
##   108
##     3
```

```
length(unique(novi_univerziteti$industryExp))
```

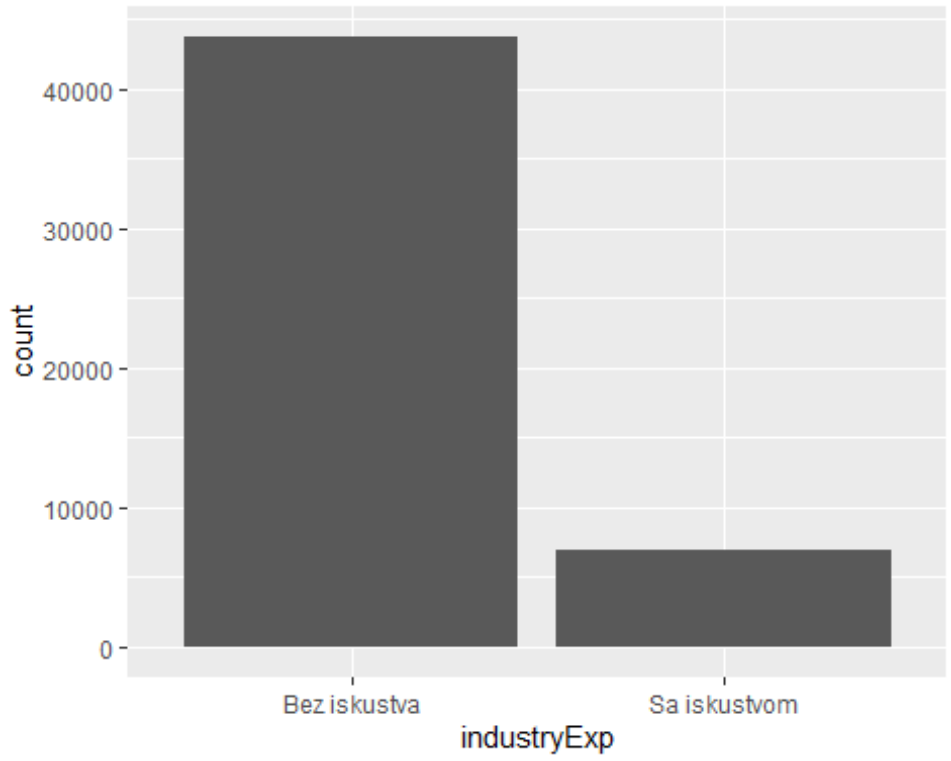
```
## [1] 92
```

```
length(which(novi_univerziteti$industryExp==0))/dim(novi_univerziteti)[1]*100
```

```
## [1] 86.24515
```

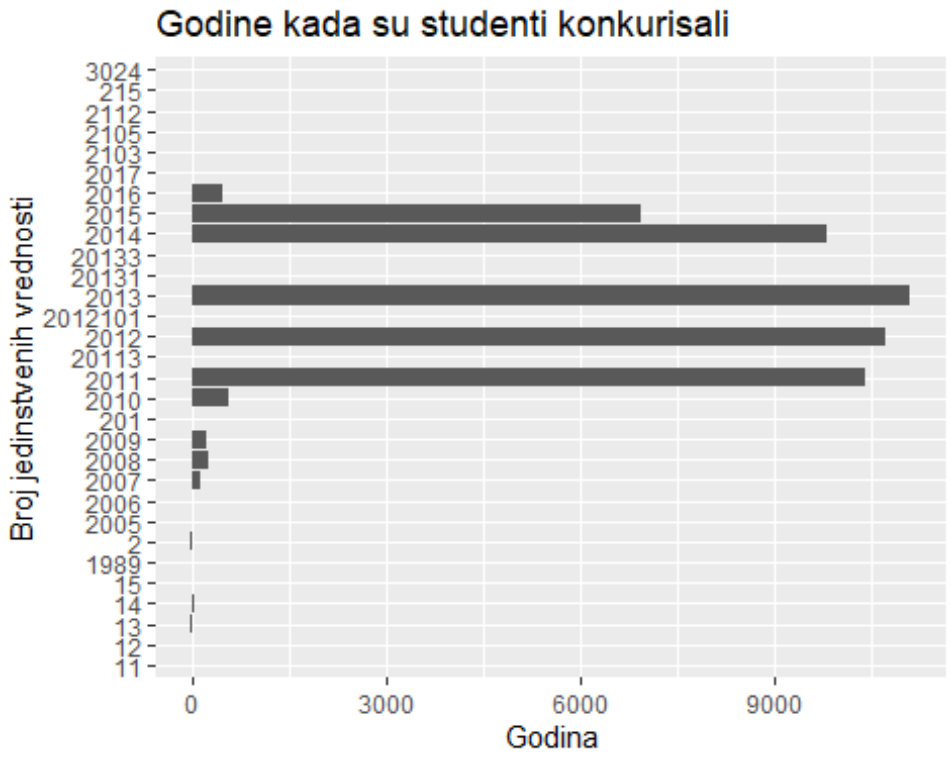
Primećujemo da ovde ipak postoji čak 92 jedinstvene vrednosti, međutim ponovo više od 86% uzoraka nemaju industrijsko iskustvo. Zbog toga, sve ostale osobe koje imaju iskustvo staža grupisaćemo u jednu celinu, i dato obeležje ćemo podeliti u 2 kategorije: 1. Sa industrijskim iskustvom. 2. Bez industrijskog iskustva.

```
novi_univerziteti$industryExp =
as.numeric(as.character(novi_univerziteti$industryExp))
novi_univerziteti$industryExp[novi_univerziteti$industryExp > 0] = 1
#unique(novi_univerziteti$industryExp)
novi_univerziteti$industryExp = factor(novi_univerziteti$industryExp,
                                       levels = c(0,1), labels = c("Bez
iskustva","Sa iskustvom"))
#novi_univerziteti$industryExp[is.na(novi_univerziteti$industryExp)] = 0

novi_univerziteti$industryExp=as.factor(novi_univerziteti$industryExp)

ggplot(novi_univerziteti) + geom_bar(aes(x = industryExp))
```



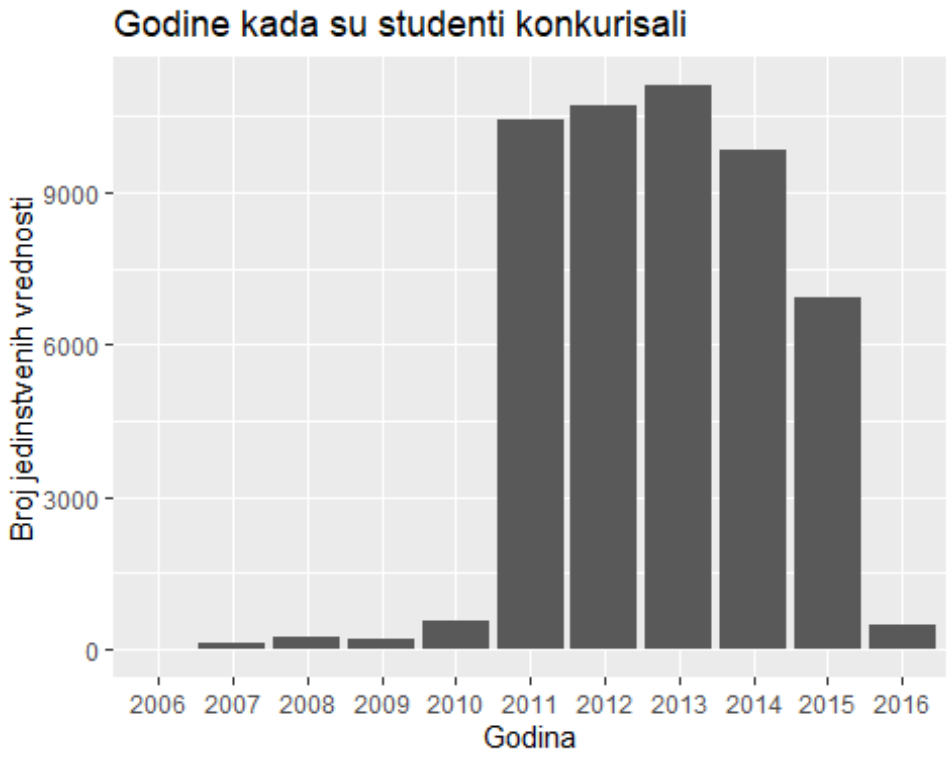
year

```
ggplot(novi_univerziteti, aes(y = year)) + geom_bar() + labs(title = "Godine kada su studenti konkurisali", x = "Godina", y = "Broj jedinstvenih vrednosti")
```



Na osnovu barplota iznad vidimo da obeležje *year*(godine kada su studenti konkurisali za upis na fakultet) sadrži neke vrednosti koje su nelogične kada su u pitanju godine. Na osnovu logike isključićemo sve vrednosti koje su manje od 2006 i sve vrednosti koje su veće od 2016.

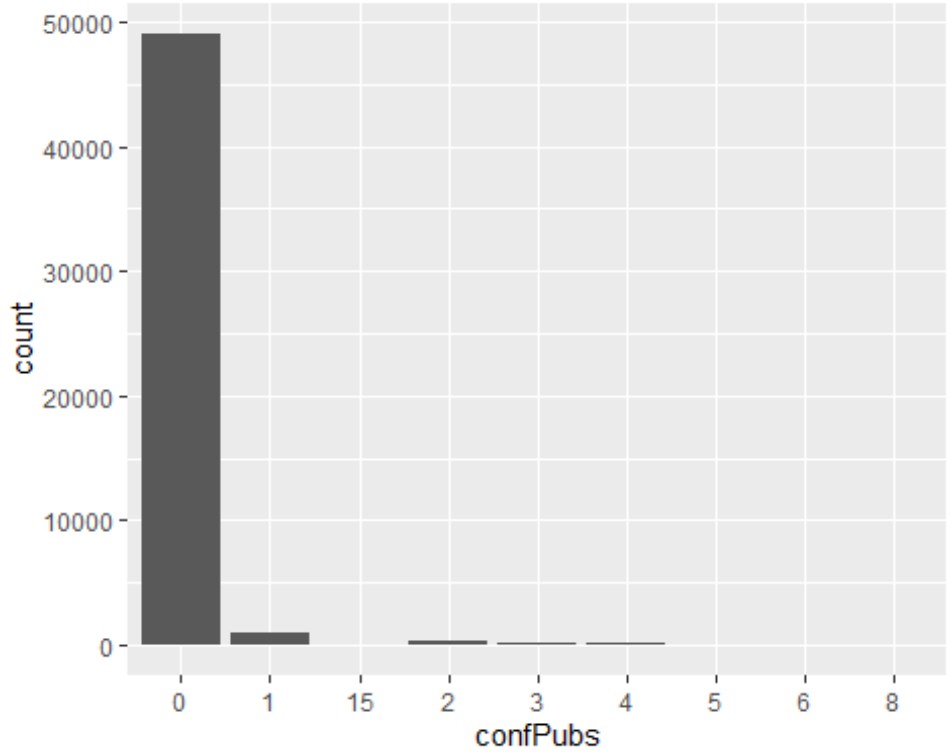
```
novi_univerziteti = novi_univerziteti[-which(novi_univerziteti$year != 2006 &
novi_univerziteti$year != 2007 & novi_univerziteti$year != 2008 &
novi_univerziteti$year != 2009
                                                    & novi_univerziteti$year
!= 2010 & novi_univerziteti$year != 2011 & novi_univerziteti$year != 2012 &
novi_univerziteti$year != 2013
                                                    & novi_univerziteti$year
!= 2014 & novi_univerziteti$year != 2015 & novi_univerziteti$year != 2016),]
ggplot(novi_univerziteti, aes(x = year)) + geom_bar() + labs(title = "Godine
kada su studenti konkurisali", x = "Godina", y = "Broj jedinstvenih
vrednosti")
```




```
novi_univerziteti$year = droplevels(novi_univerziteti$year)
```

confPubs

```
ggplot(novi_univerziteti) + geom_bar(aes(x = confPubs))
```

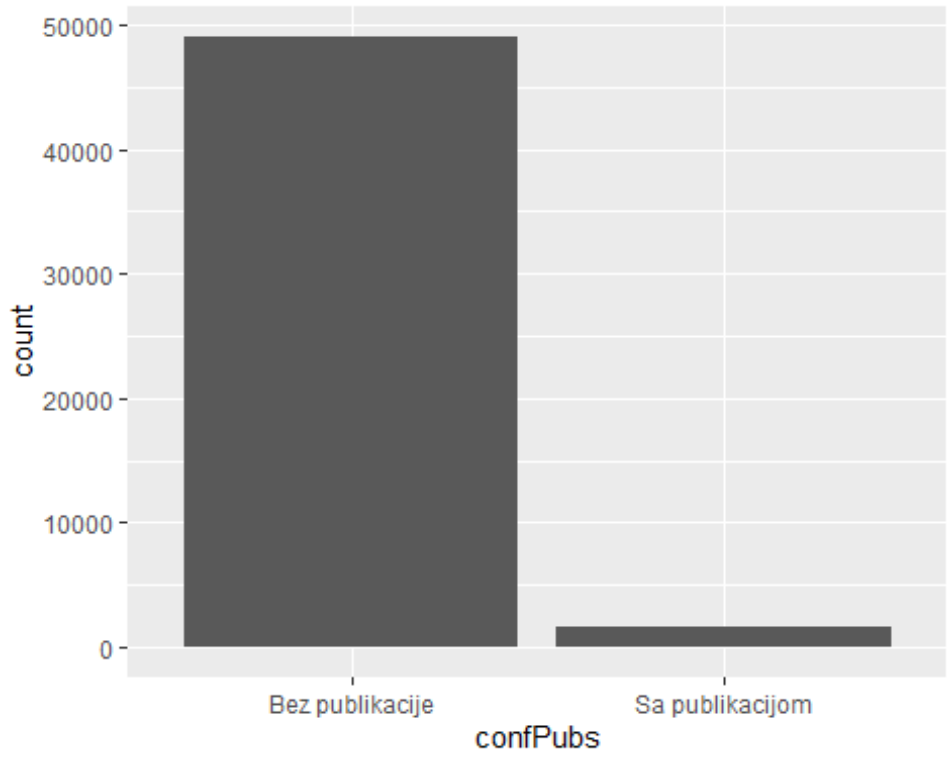


```
xtabs(~novi_univerziteti$confPubs)

## novi_univerziteti$confPubs
##     0     1    15     2     3     4     5     6     8
## 49039  1031     1   326   129    71    22     8    10

length(unique(novi_univerziteti$confPubs))

## [1] 9

length(which(novi_univerziteti$confPubs==0))/dim(novi_univerziteti)[1]*100

## [1] 96.8442
```

I ovde možemo zaključiti da većina nije imalo publikacije, tako da ćemo i ovo obeležje konvertovati u faktor promenljivu sa 2 kategorije:

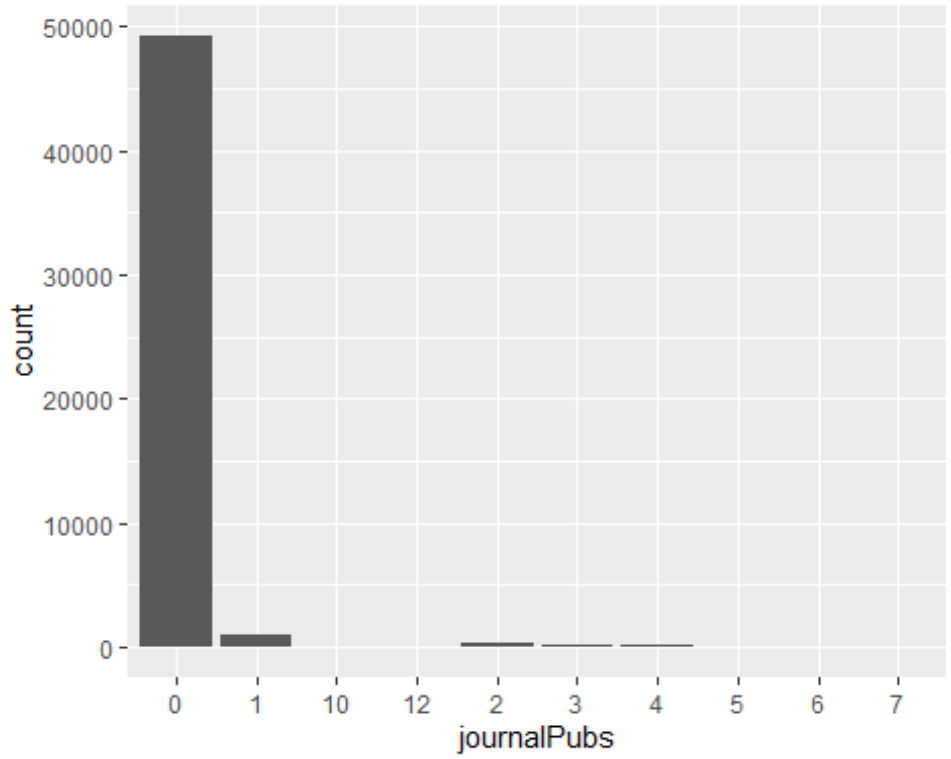
1. Bez publikacije
2. Sa publikacijom

```
novi_univerziteti$confPubs =
as.numeric(as.character(novi_univerziteti$confPubs))
novi_univerziteti$confPubs[novi_univerziteti$confPubs > 0] = 1
#unique(novi_univerziteti$confPubs)
novi_univerziteti$confPubs = factor(novi_univerziteti$confPubs,
                                    levels = c(0,1), labels = c("Bez
publikacije","Sa publikacijom"))
#novi_univerziteti$confPubs[is.na(novi_univerziteti$confPubs)] = 0

novi_univerziteti$confPubs=as.factor(novi_univerziteti$confPubs)

ggplot(novi_univerziteti) + geom_bar(aes(x = confPubs))
```



journalPubs

```
ggplot(novi_univerziteti) + geom_bar(aes(x = journalPubs))
```



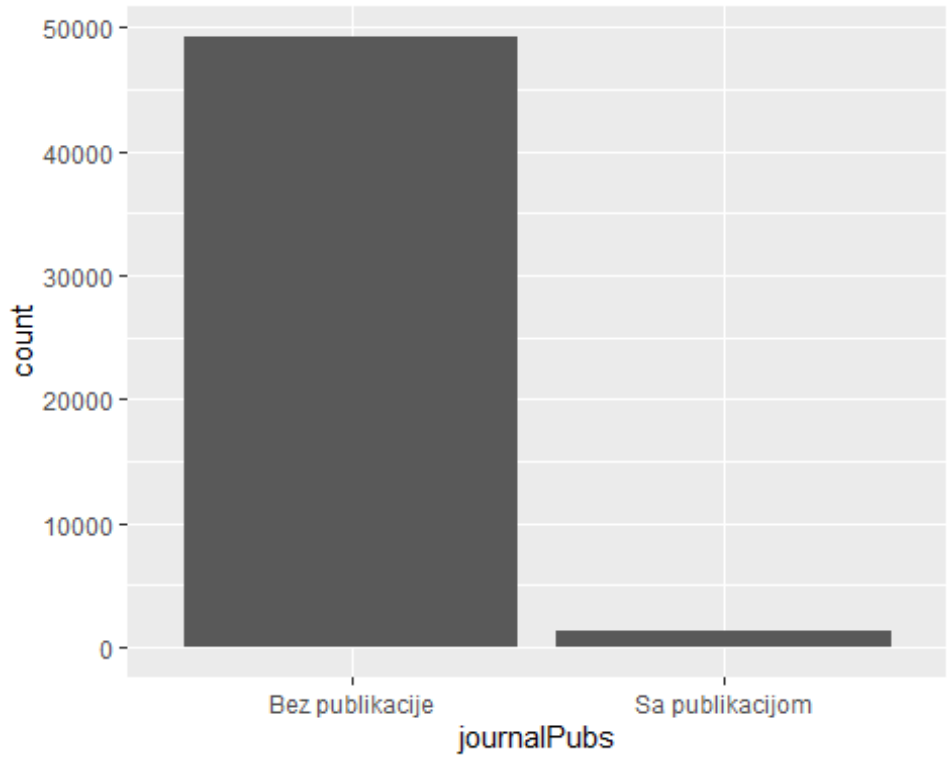
Potpuno istu stvar uočavamo i sa publikacijom časopisa.

```
xtabs(~novi_univerziteti$journalPubs)

## novi_univerziteti$journalPubs
##           0     1    10    12     2     3     4     5     6     7
##     0 49273   955     1     4   270    63    60     8     2     1

length(unique(novi_univerziteti$journalPubs))

## [1] 10

length(which(novi_univerziteti$journalPubs==0))/dim(novi_univerziteti)[1]*100

## [1] 97.30632
```

Kategorije:

1. Bez publikacije
2. Sa publikacijom

```
novi_univerziteti$journalPubs =
as.numeric(as.character(novi_univerziteti$journalPubs))
novi_univerziteti$journalPubs[novi_univerziteti$journalPubs > 0] = 1
#unique(novi_univerziteti$journalPubs)
novi_univerziteti$journalPubs = factor(novi_univerziteti$journalPubs,
                                       levels = c(0,1), labels = c("Bez
publikacije","Sa publikacijom"))
#novi_univerziteti$journalPubs[is.na(novi_univerziteti$journalPubs)] = 0

novi_univerziteti$journalPubs=as.factor(novi_univerziteti$journalPubs)

ggplot(novi_univerziteti) + geom_bar(aes(x = journalPubs))
```



EDA - Exploratory Data Analysis

Da bismo napravili dobar model moramo da vidimo u kakvoj su zavisnosti data obeležja sa obeležjem *admit*(student je prihvaćen/odbijen).

Analiza kategorijskih podataka naspram numeričkih

Kada bismo samo upoređivali obeležja admit u odnosu na druga numerička obeležja, ne bi primetili nikakvu povezanost između njih, već bi za primljene i odbijene đake grafik izgledao identično. Zbog toga ćemo pri upoređivanju obeležja admit naspram numeričkih obeležja, upoređivaćemo i sa obeležjem univName, zato što jedino na taj način možemo zaključiti koliko neka numerička promenljiva utiče na upis u zavisnosti od fakulteta.

```
length(unique(novi_univerziteti$univName))
```

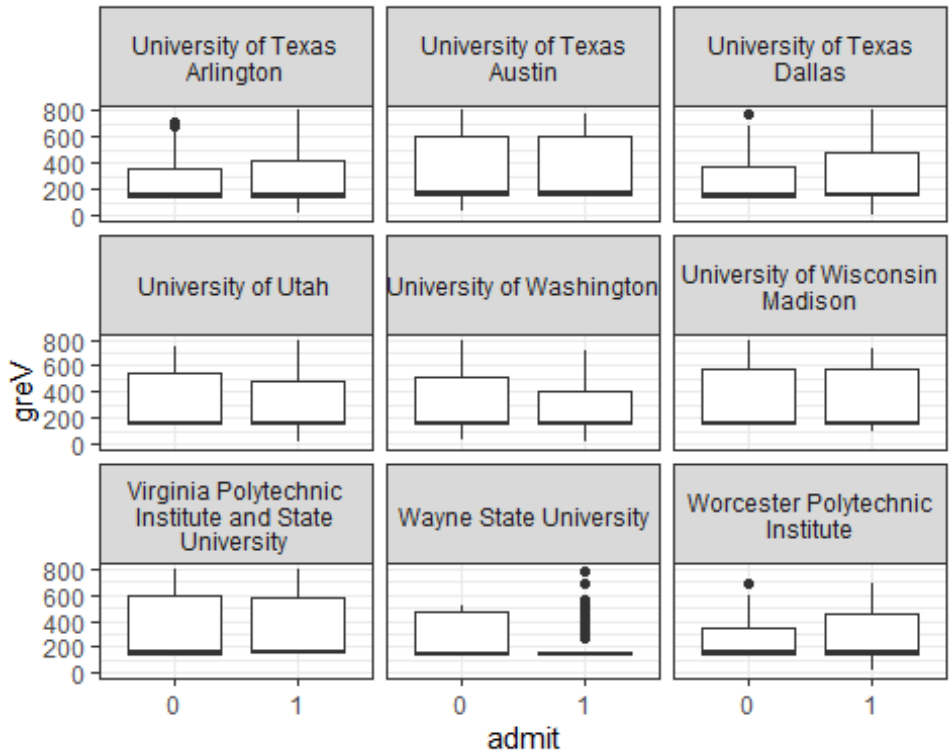
```
## [1] 54
```

S obzirom da je broj fakulteta 54, prikaz podataka u odnosu na svaki fakultet na jednom grafiku bilo bi previše nepregledno. Zbog toga ćemo morati naš okvir podataka podeliti u 6 manjih okvira, tako što će svaki okvir sadržati samo one uzorke koje za obeležje univName imaju neki od 9 odabranih univerziteta.

```
names=unique(novi_univerziteti$univName)


univ1=names[1:9]
univ2=names[10:18]
univ3=names[19:27]
univ4=names[28:36]
univ5=names[37:45]
univ6=names[46:54]


okvir1=novi_univerziteti[novi_univerziteti$univName %in%  univ1,]
okvir2=novi_univerziteti[novi_univerziteti$univName %in%  univ2,]
okvir3=novi_univerziteti[novi_univerziteti$univName %in%  univ3,]
okvir4=novi_univerziteti[novi_univerziteti$univName %in%  univ4,]
okvir5=novi_univerziteti[novi_univerziteti$univName %in%  univ5,]
okvir6=novi_univerziteti[novi_univerziteti$univName %in%  univ6,]
```

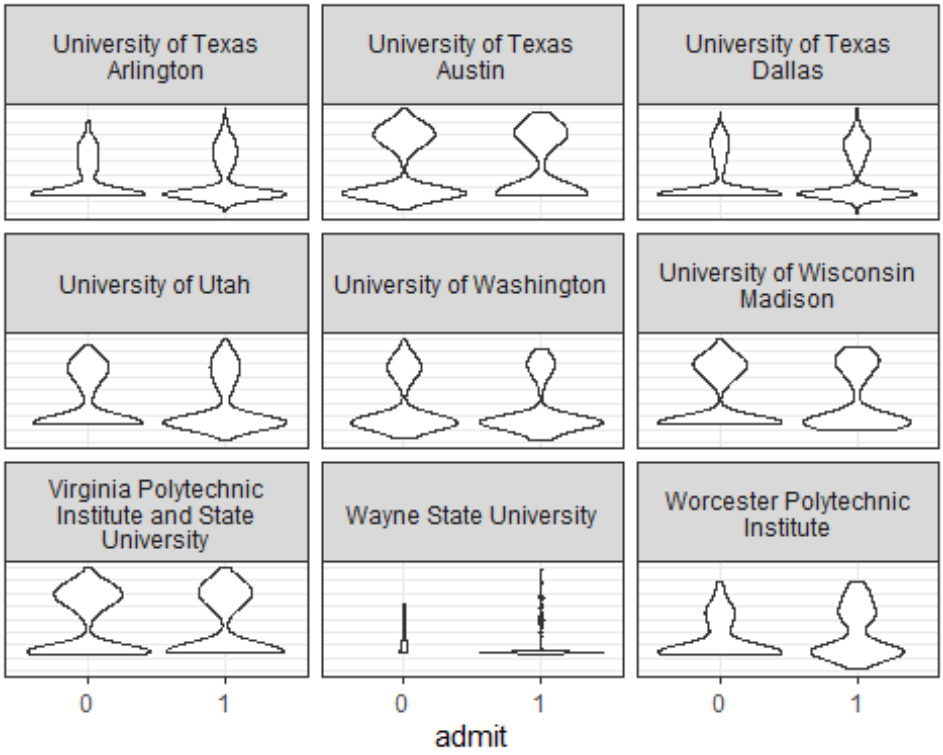
greV i admit

```
ggplot(okvir1, aes(x=admit,y=greV)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```

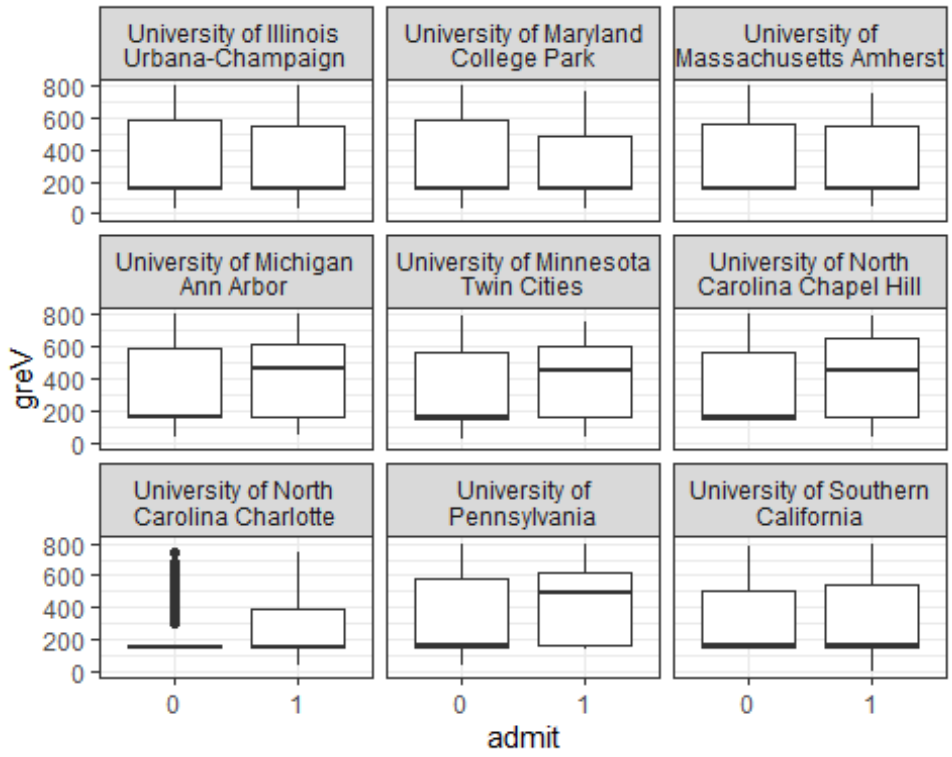



```
ggplot(okvir1, aes(x=admit,y=greV)) +
geom_violin(alpha=1) +
```

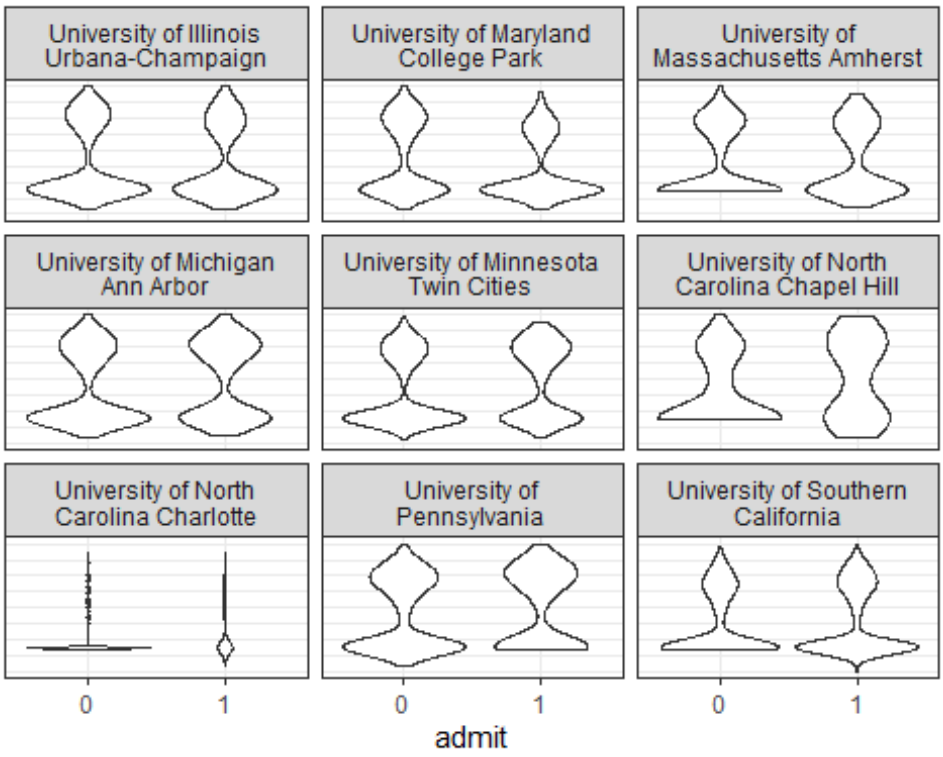
```
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



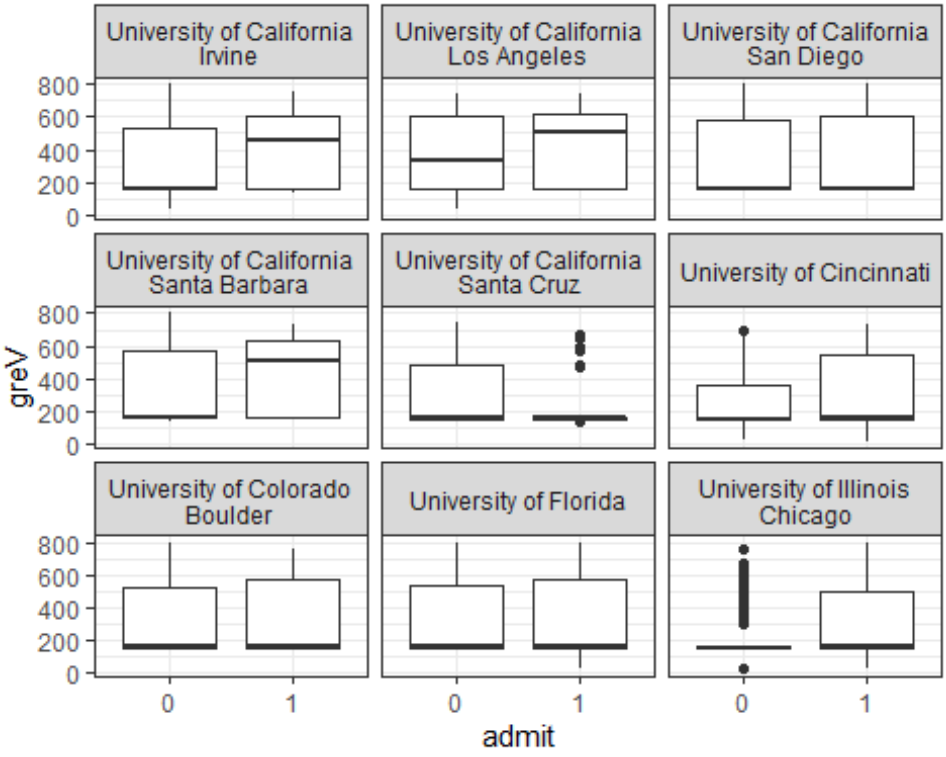

```
ggplot(okvir2, aes(x=admit,y=greV)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



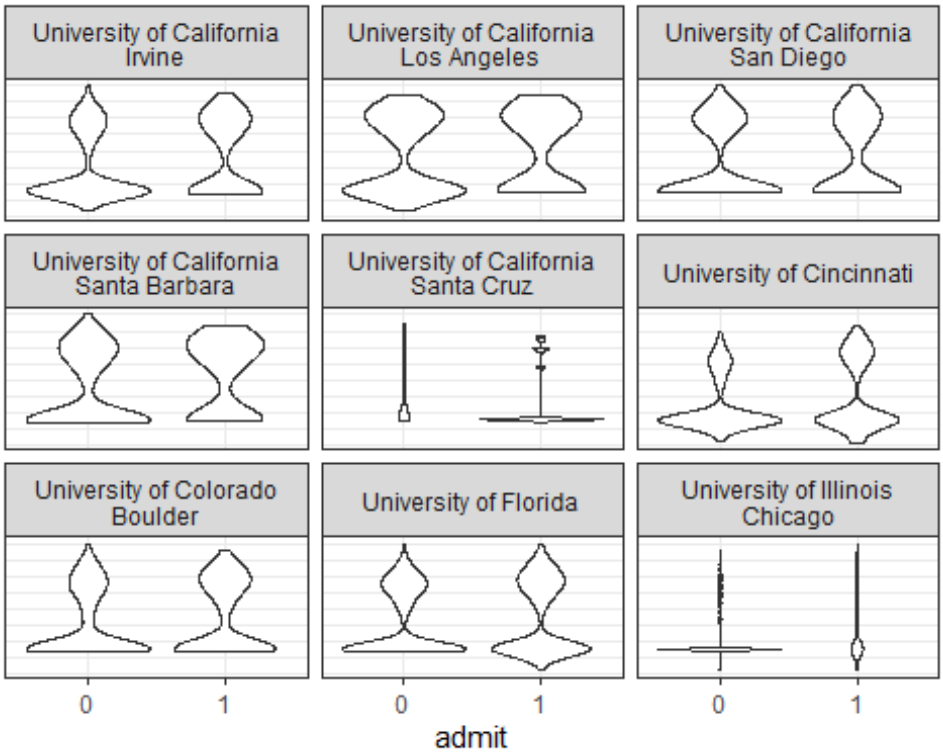

```
ggplot(okvir2, aes(x=admit,y=greV)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



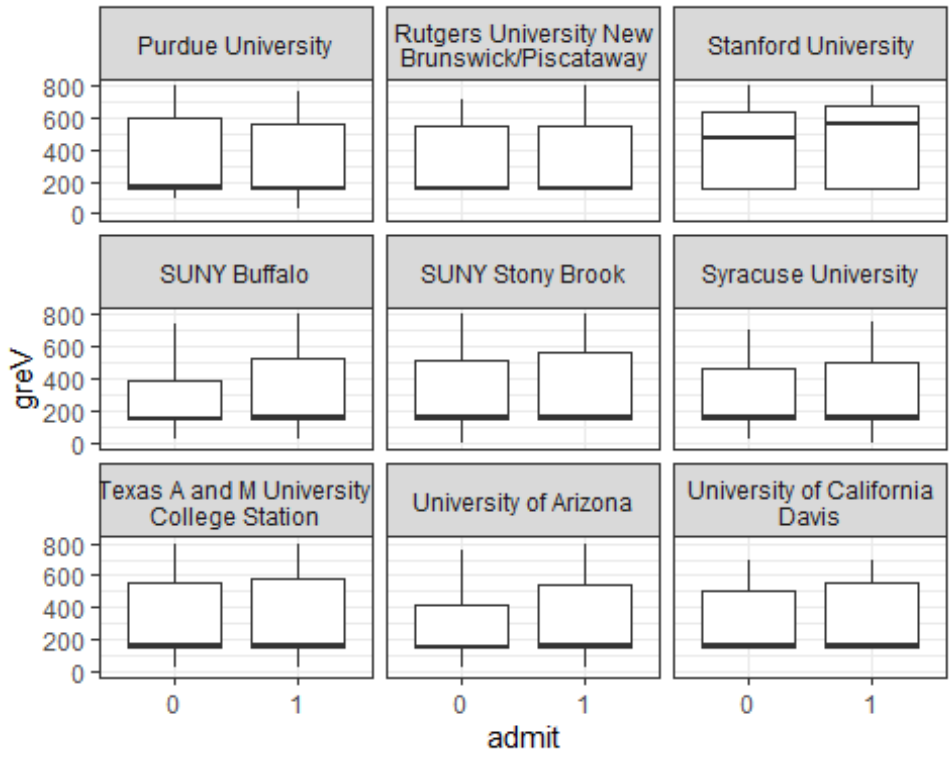

```
ggplot(okvir3, aes(x=admit,y=greV)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



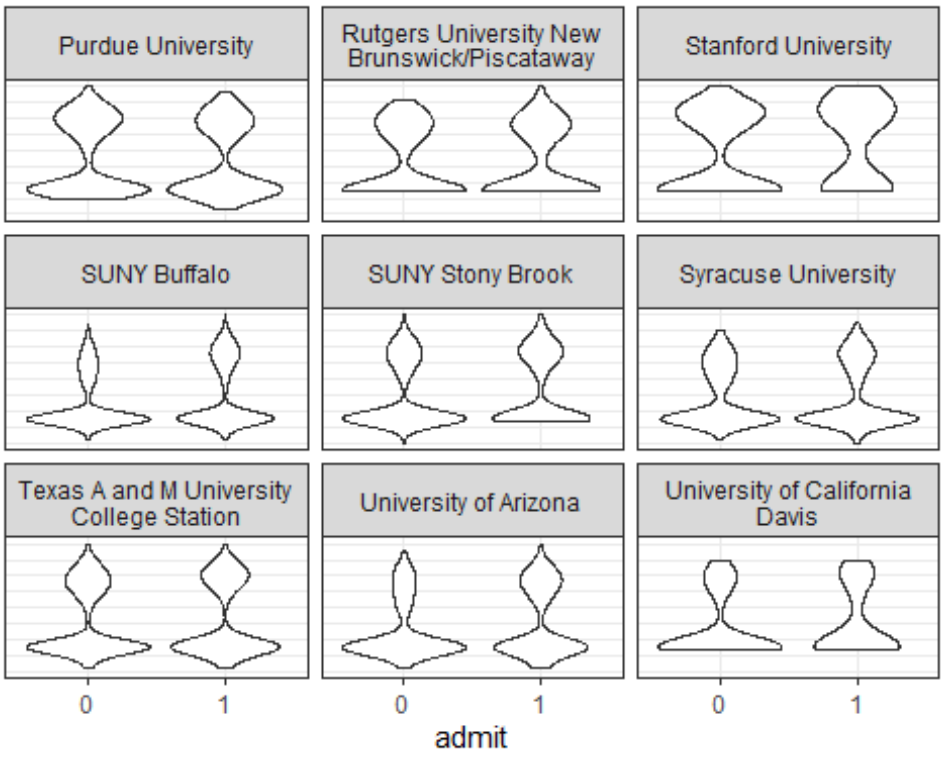
```
ggplot(okvir3, aes(x=admit,y=greV)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



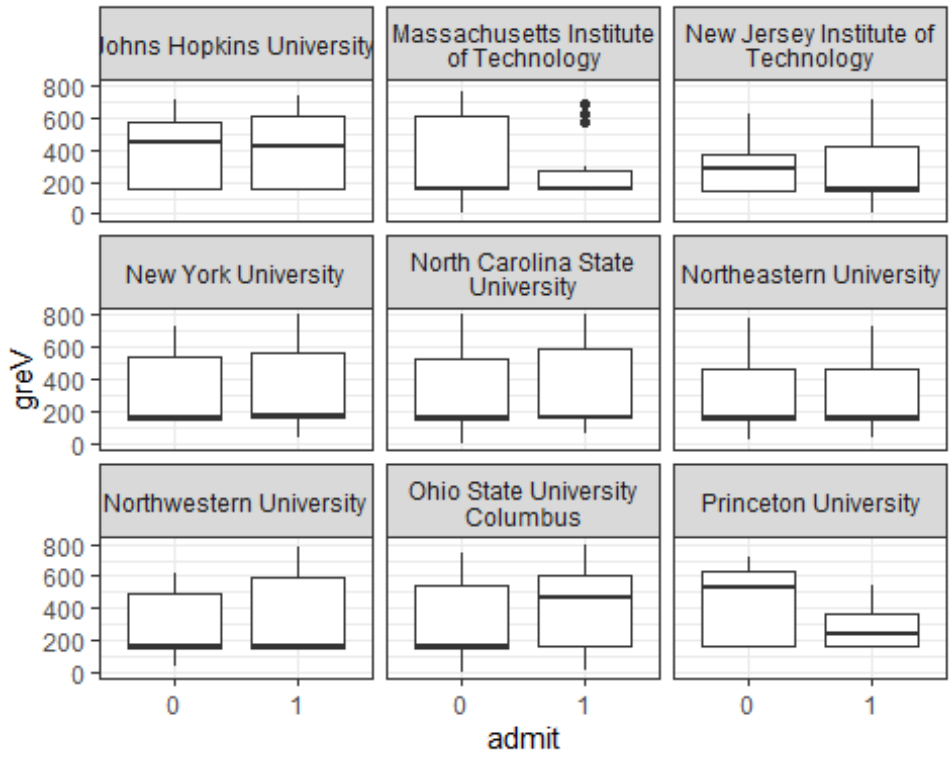

```
ggplot(okvir4, aes(x=admit,y=greV)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



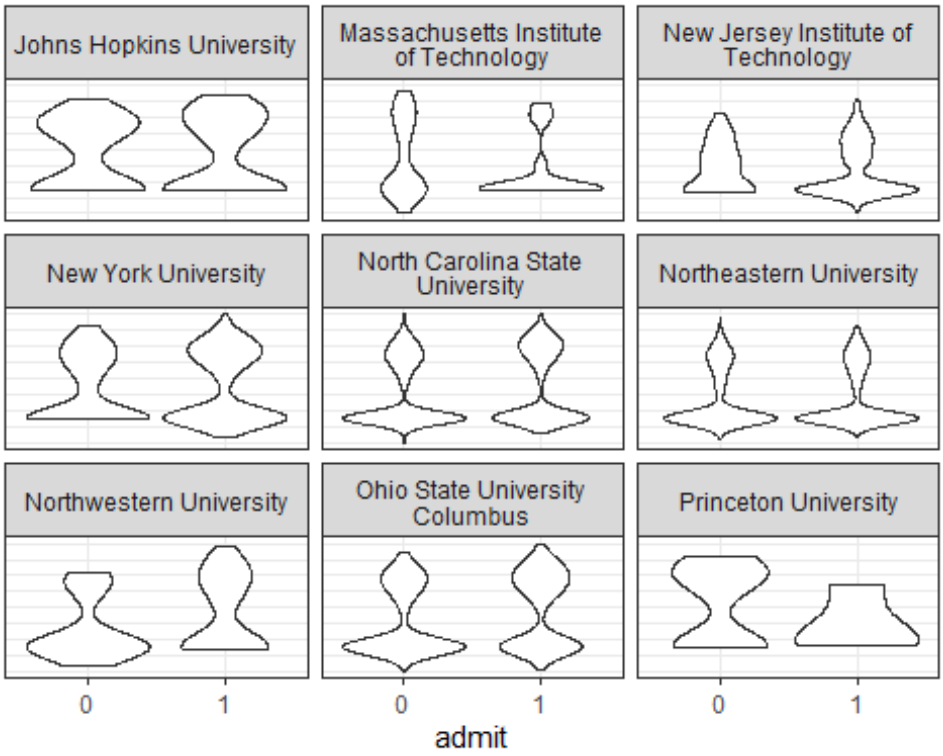

```
ggplot(okvir4, aes(x=admit,y=greV)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



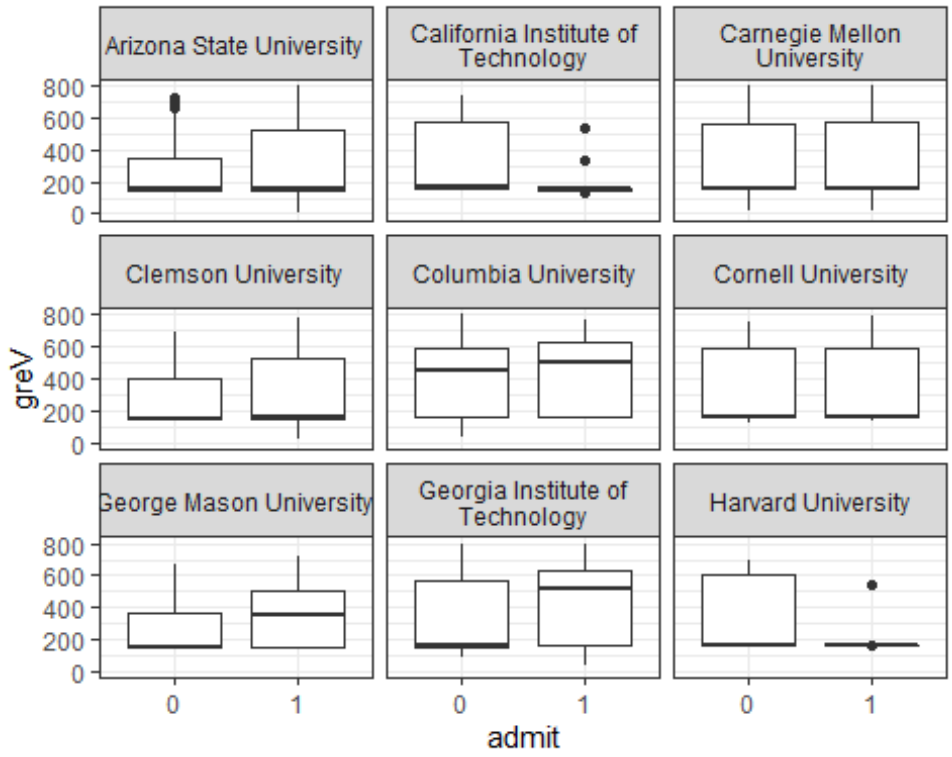

```
ggplot(okvir5, aes(x=admit,y=greV)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



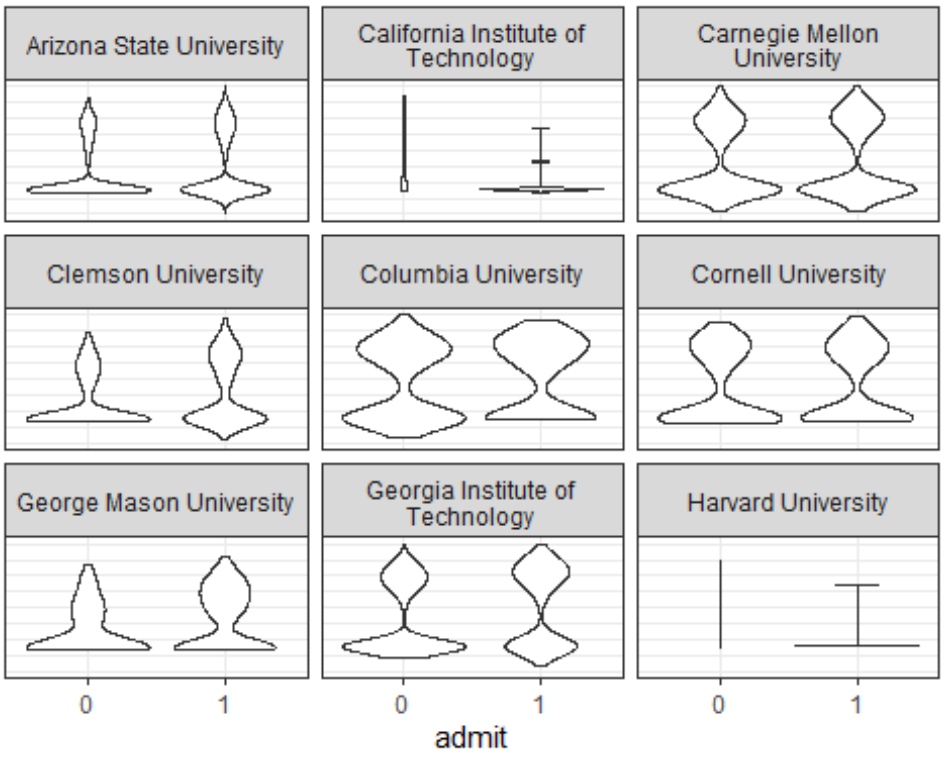
```
ggplot(okvir5, aes(x=admit,y=greV)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```




```
ggplot(okvir6, aes(x=admit,y=greV)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```




```
ggplot(okvir6, aes(x=admit,y=greV)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```

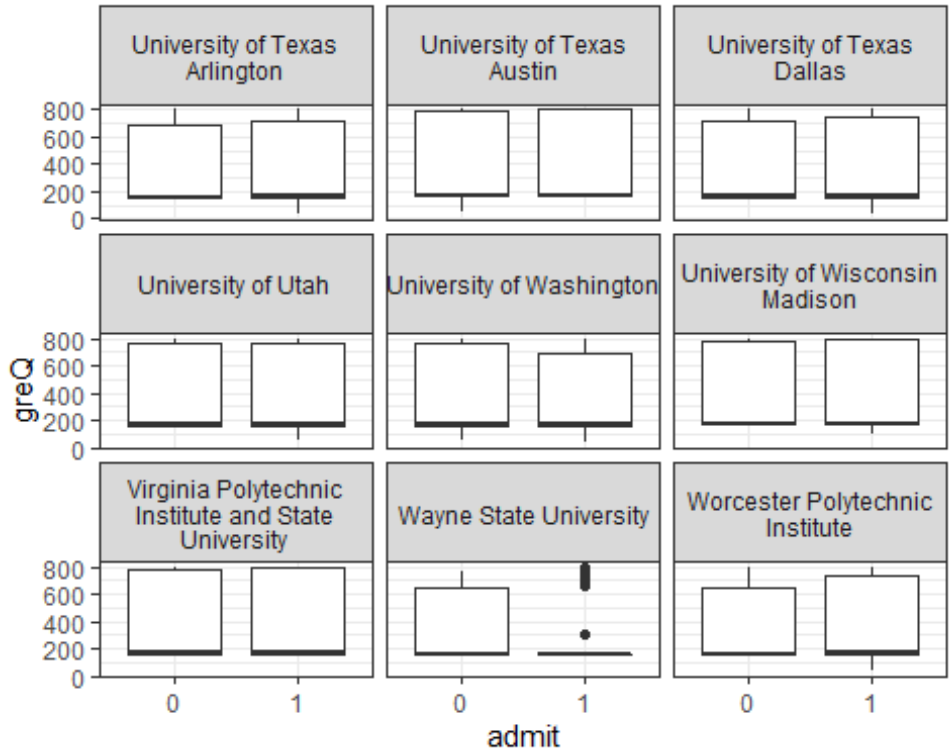



```
#K-S test normalnosti
novi_univerziteti %>% group_by(admit) %>%
summarise(izlaz = list(ks.test(greV, "pnorm", mean=mean(greV, na.rm = T),
sd=sd(greV, na.rm = T)) %>% tidy), .groups = 'drop') %>% unnest(c(izlaz))

## Warning in ks.test(greV, "pnorm", mean = mean(greV, na.rm = T), sd =
sd(greV, :
## ties should not be present for the Kolmogorov-Smirnov test

## Warning in ks.test(greV, "pnorm", mean = mean(greV, na.rm = T), sd =
sd(greV, :
## ties should not be present for the Kolmogorov-Smirnov test

## # A tibble: 2 x 5
##   admit statistic p.value method                                alternative
##   <fct>     <dbl>   <dbl> <chr>                                 <chr>
## 1 0         0.359       0 One-sample Kolmogorov-Smirnov test two-sided
## 2 1         0.345       0 One-sample Kolmogorov-Smirnov test two-sided
```

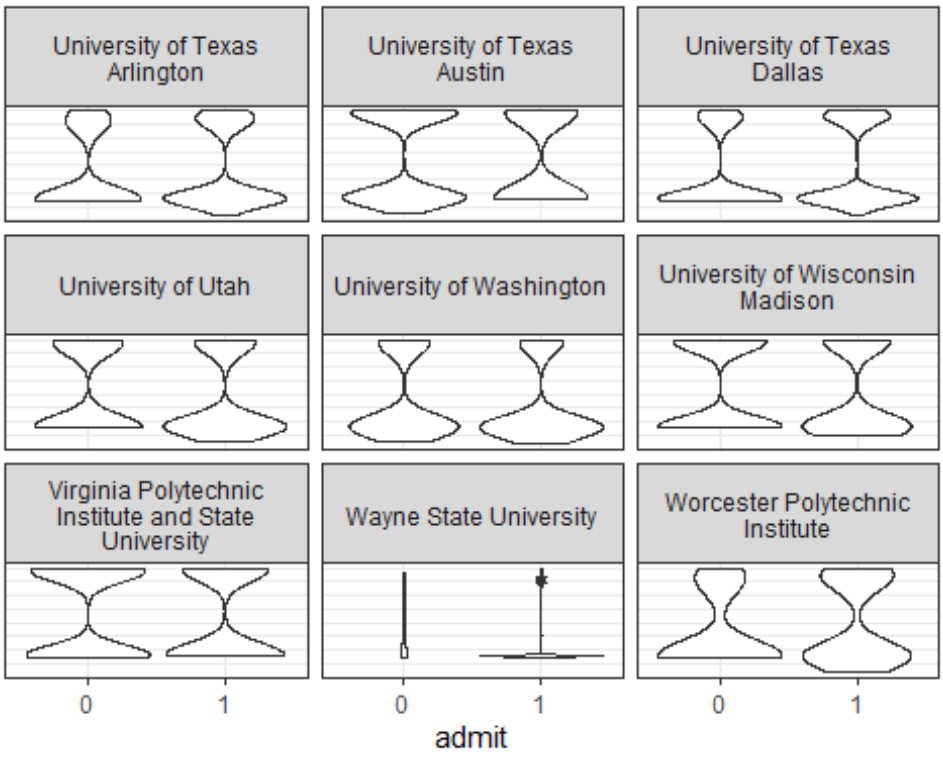
Na osnovu grafika iznad vidimo da rezultati greV testa utiče na to da li je osoba primljena na fakultet ili nije, ali zavisi od fakulteta. Testiranjem normalnosti Kolmogorov-Smirnov testom pokazano je da ne postoji normalnost unutar obe grupe obeležja (p = 0.0 < α=0.05, p = 0.0 < α=0.05).

greQ i admit

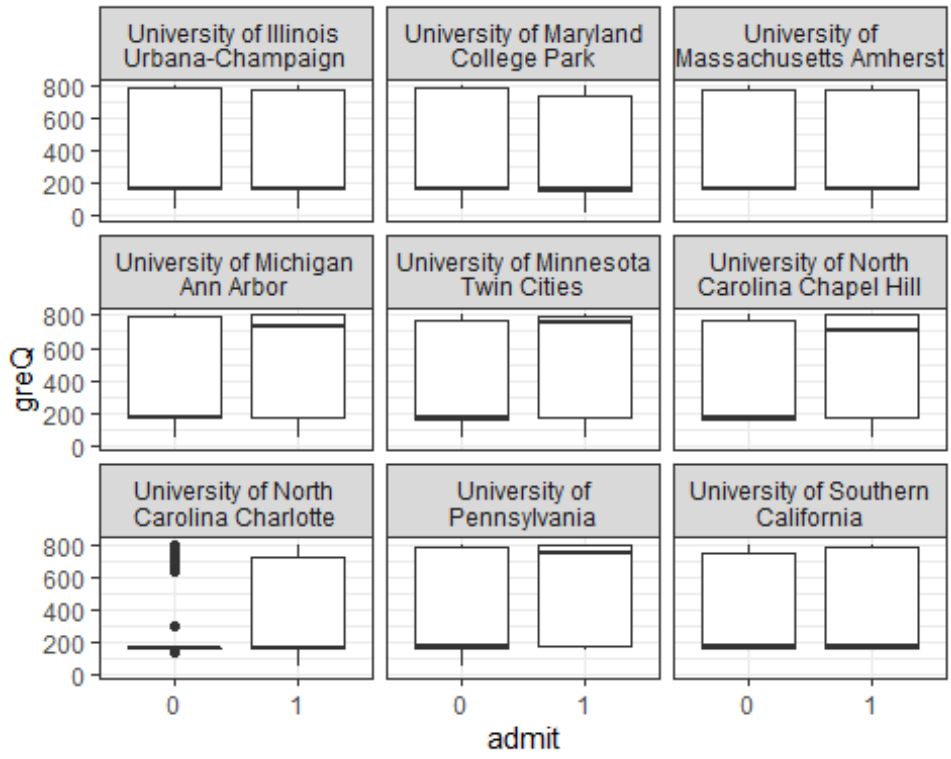
```
ggplot(okvir1, aes(x=admit,y=greQ)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



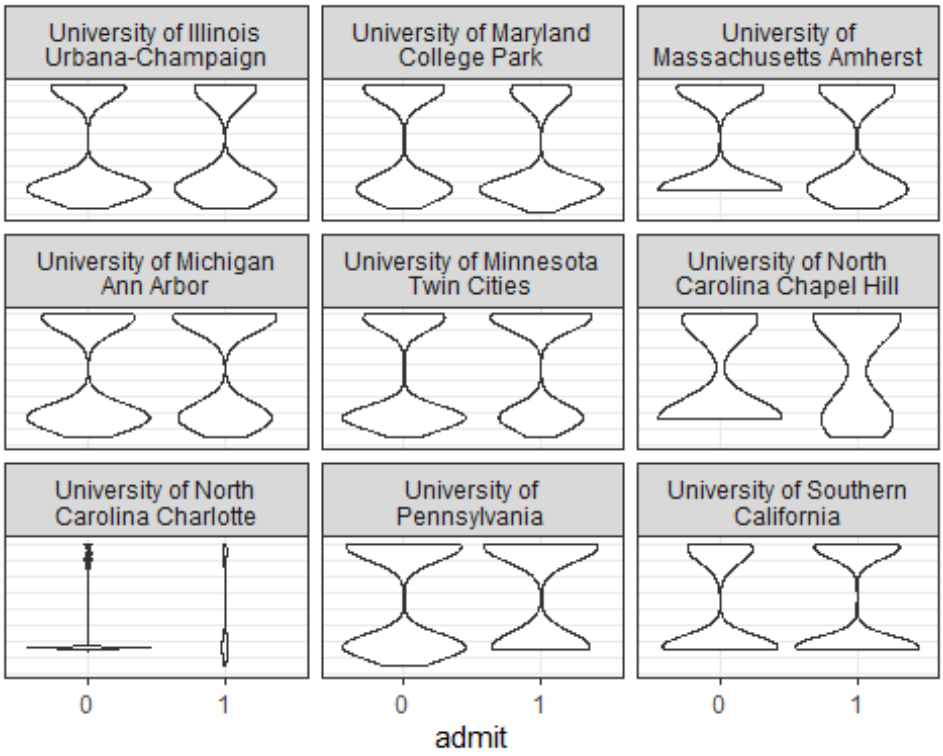

```
ggplot(okvir1, aes(x=admit,y=greQ)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



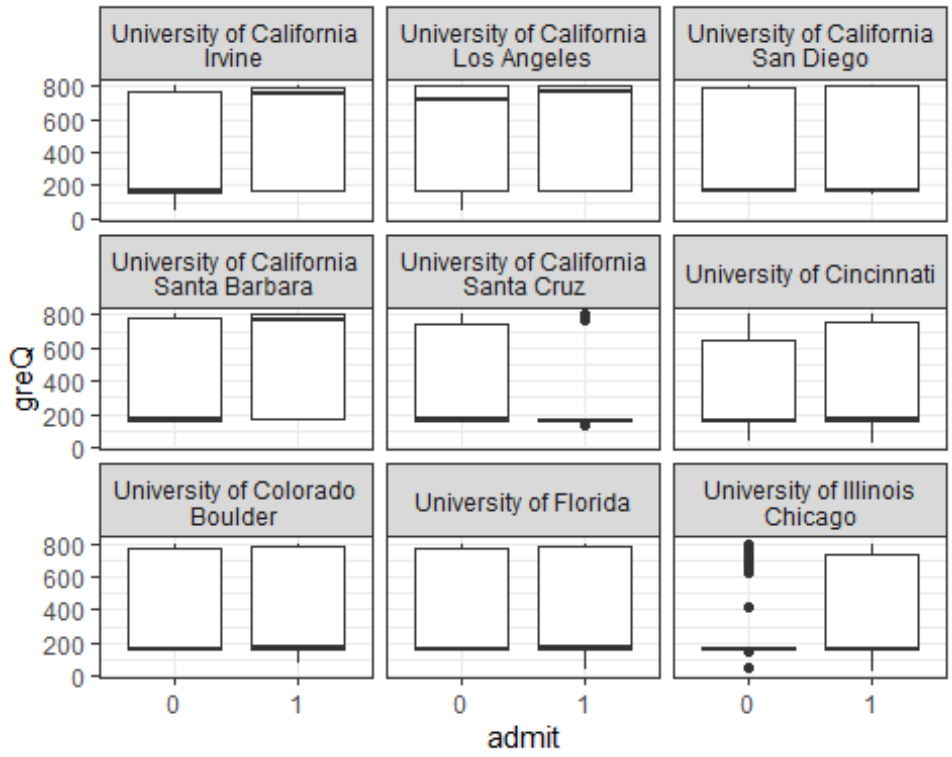

```
ggplot(okvir2, aes(x=admit,y=greQ)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



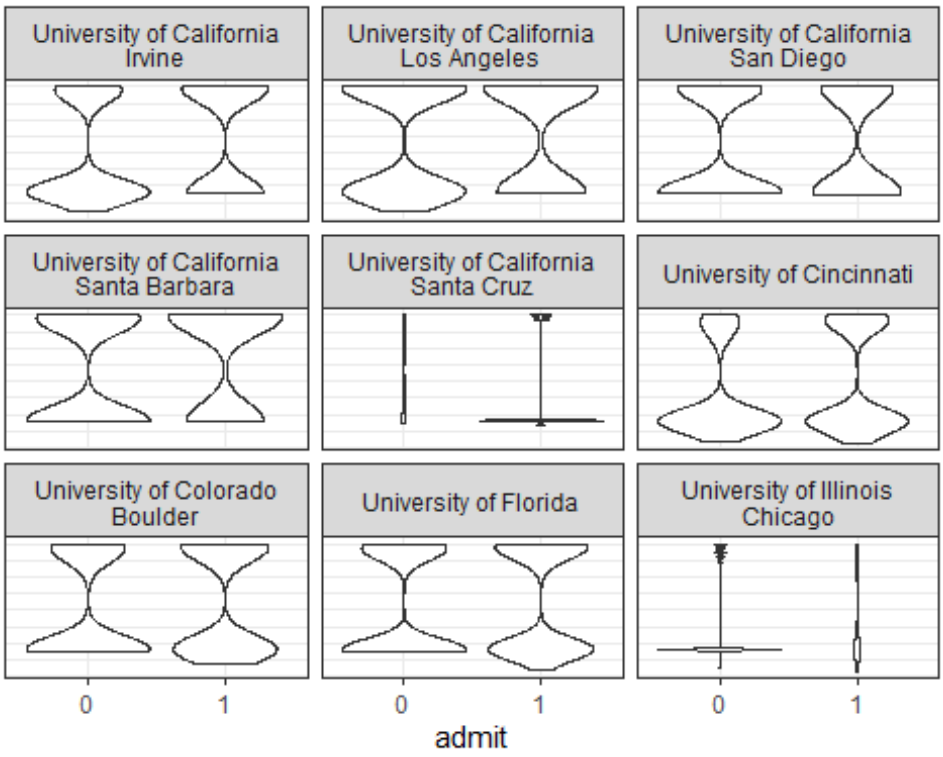
```
ggplot(okvir2, aes(x=admit,y=greQ)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



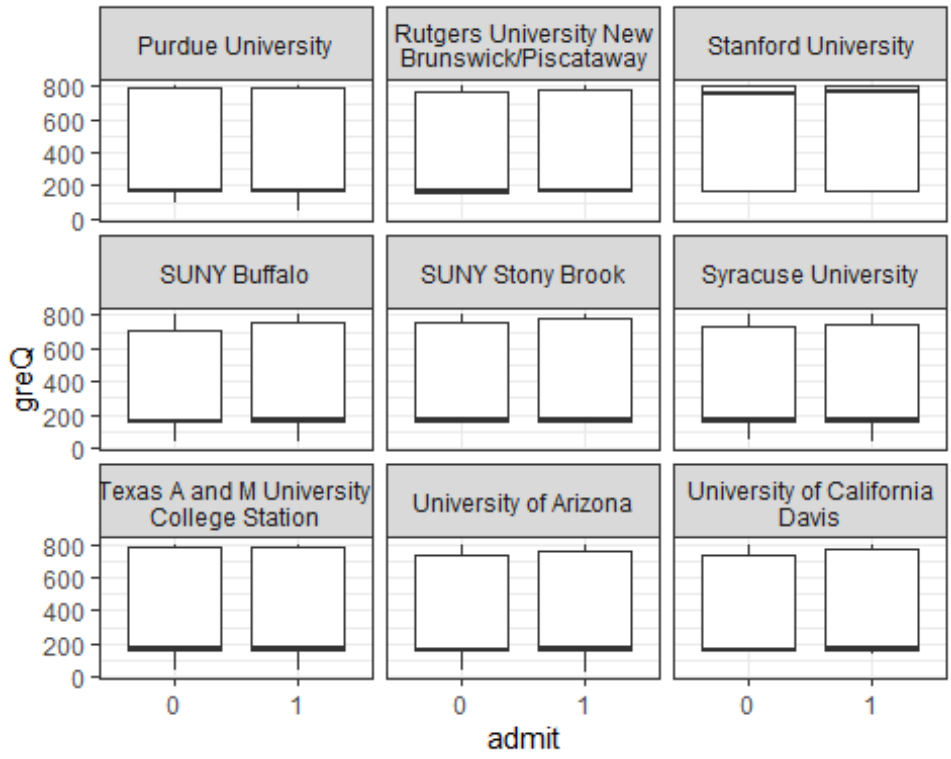

```
ggplot(okvir3, aes(x=admit,y=greQ)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



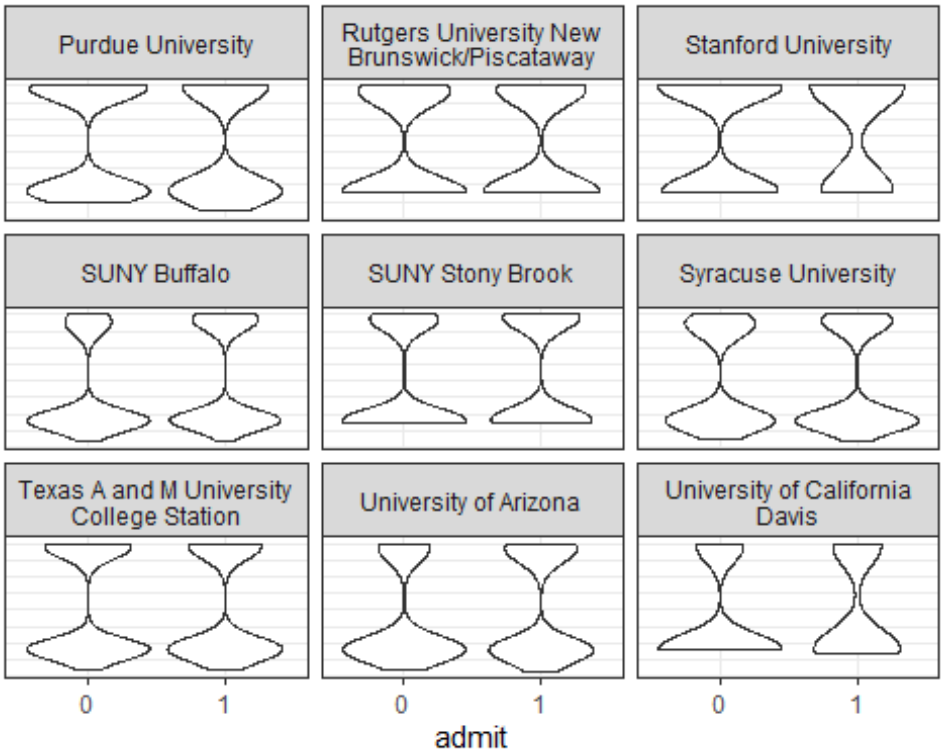

```
ggplot(okvir3, aes(x=admit,y=greQ)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



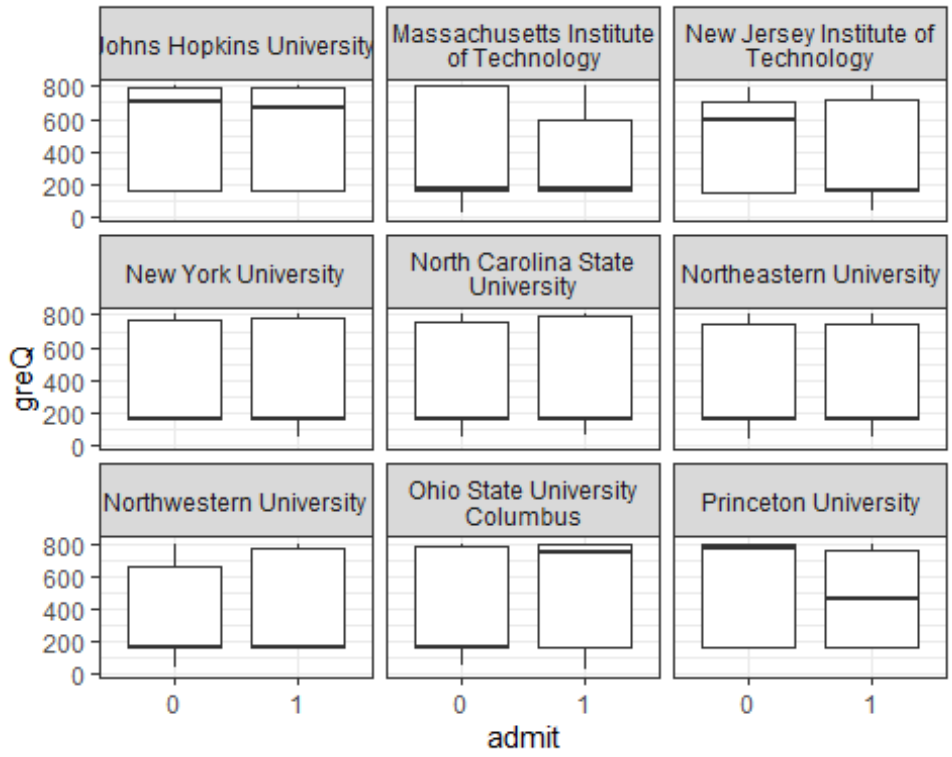

```
ggplot(okvir4, aes(x=admit,y=greQ)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



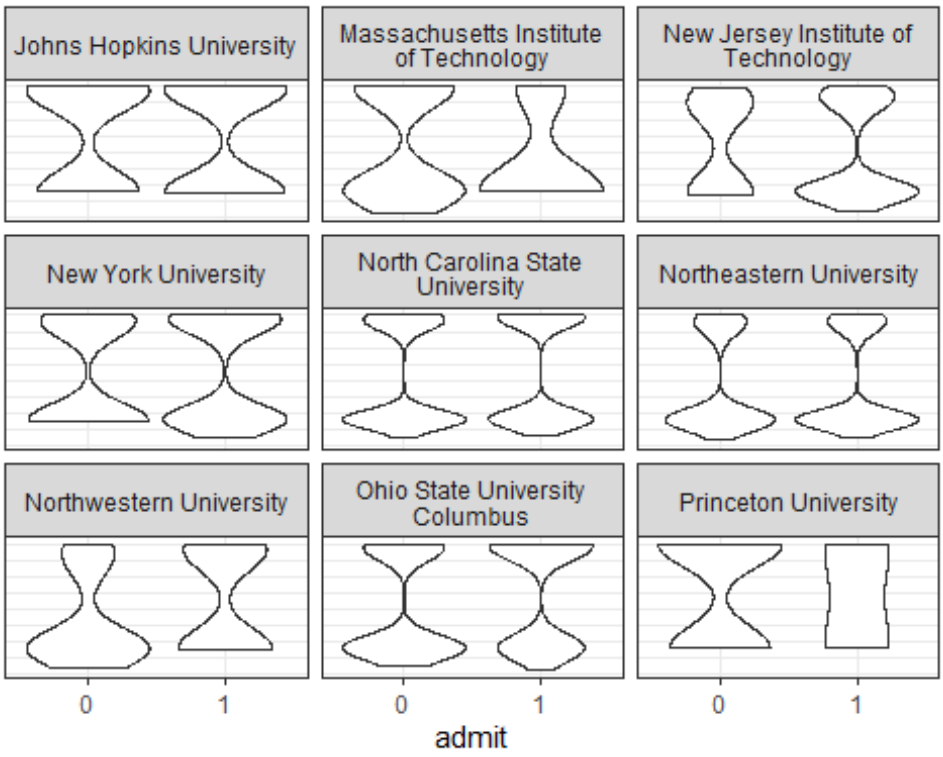
```
ggplot(okvir4, aes(x=admit,y=greQ)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



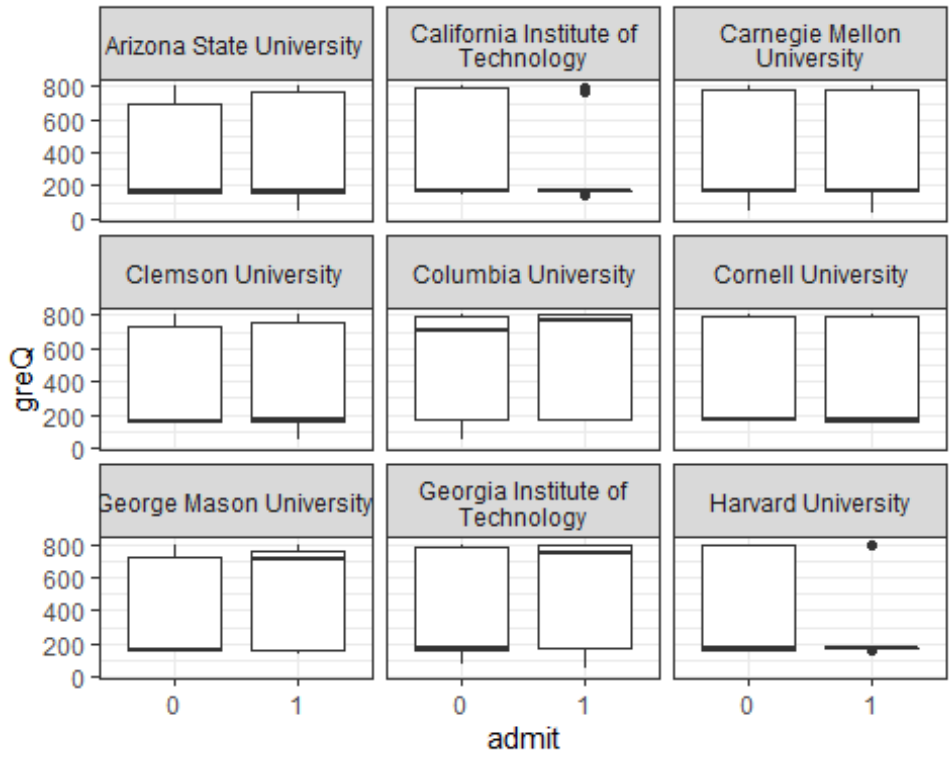

```
ggplot(okvir5, aes(x=admit,y=greQ)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



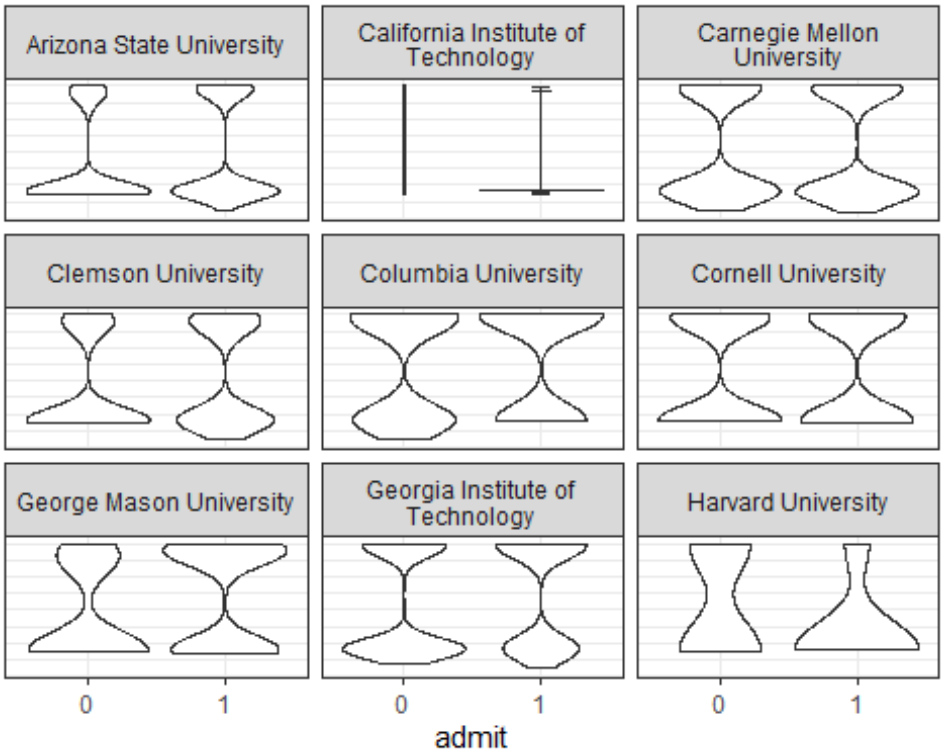

```
ggplot(okvir5, aes(x=admit,y=greQ)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```




```
ggplot(okvir6, aes(x=admit,y=greQ)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



```
ggplot(okvir6, aes(x=admit,y=greQ)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



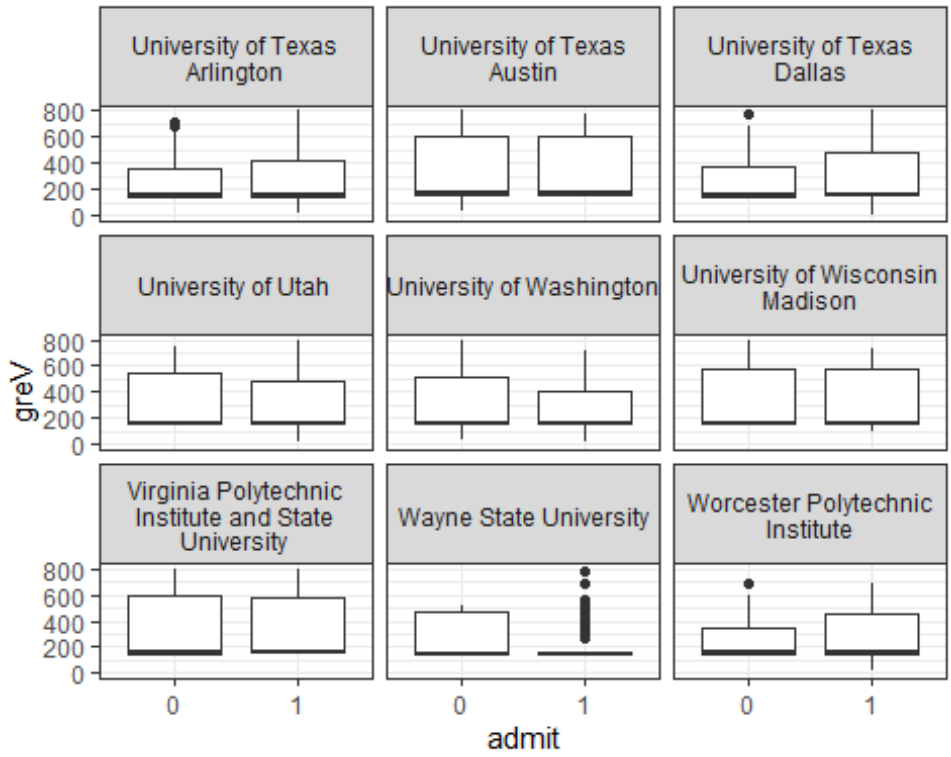

```
#K-S test normalnosti
novi_univerziteti %>% group_by(admit) %>%
summarise(izlaz = list(ks.test(greQ, "pnorm", mean=mean(greQ, na.rm = T),
sd=sd(greQ, na.rm = T)) %>% tidy), .groups = 'drop') %>% unnest(c(izlaz))

## Warning in ks.test(greQ, "pnorm", mean = mean(greQ, na.rm = T), sd =
sd(greQ, :
## ties should not be present for the Kolmogorov-Smirnov test

## Warning in ks.test(greQ, "pnorm", mean = mean(greQ, na.rm = T), sd =
sd(greQ, :
## ties should not be present for the Kolmogorov-Smirnov test

## # A tibble: 2 x 5
##   admit statistic p.value method                                alternative
##   <fct>     <dbl>   <dbl> <chr>                                 <chr>
## 1 0         0.384       0 One-sample Kolmogorov-Smirnov test two-sided
## 2 1         0.372       0 One-sample Kolmogorov-Smirnov test two-sided
```

Na osnovu grafika iznad vidimo da rezultati greV testa utiče na to da li je osoba primljena na fakultet ili nije, ali zavisi od fakulteta. Testiranjem normalnosti Kolmogorov-Smirnov testom

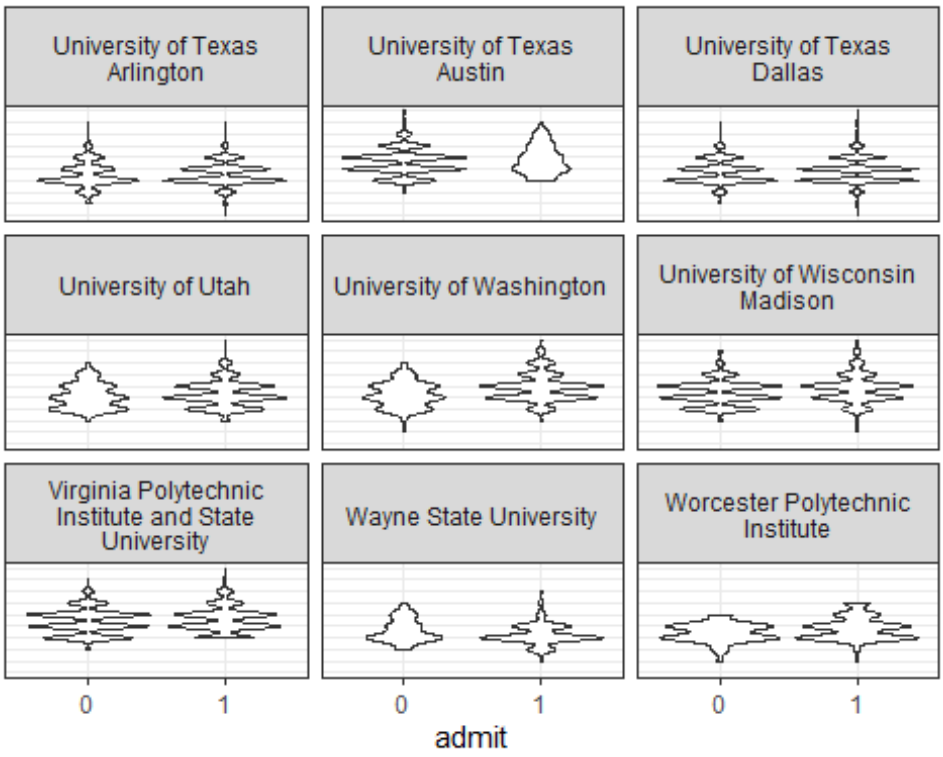
pokazano je da ne postoji normalnost unutar obe grupe obeležja (p = 0.0 < α=0.05, p = 0.0 < α=0.05).

greA i admit

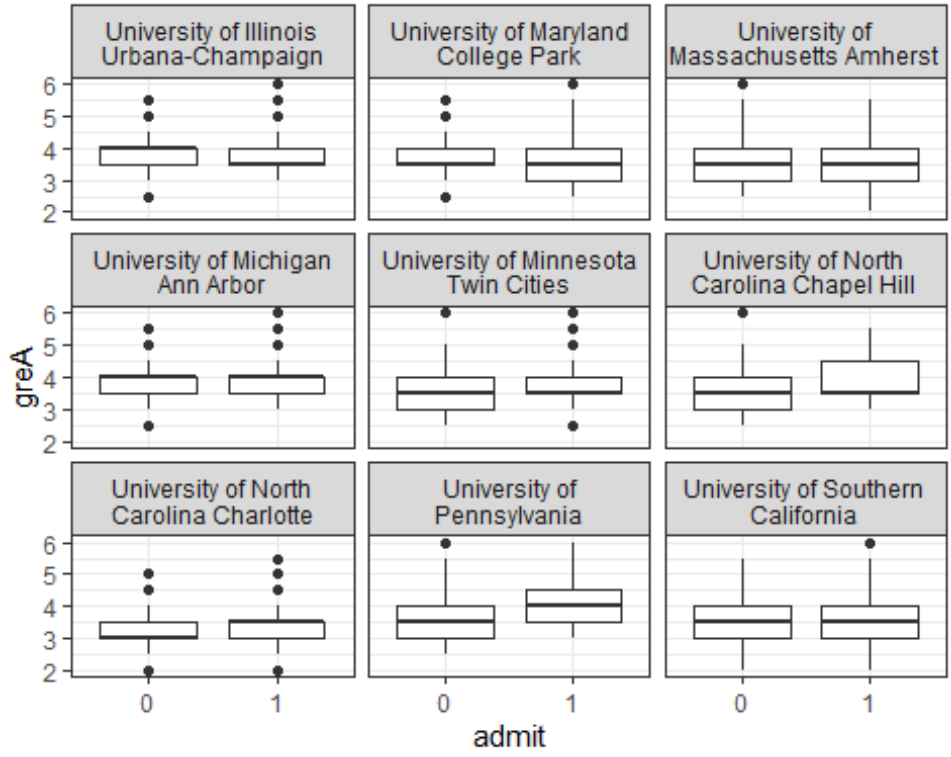
```
ggplot(okvir1, aes(x=admit,y=greV)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



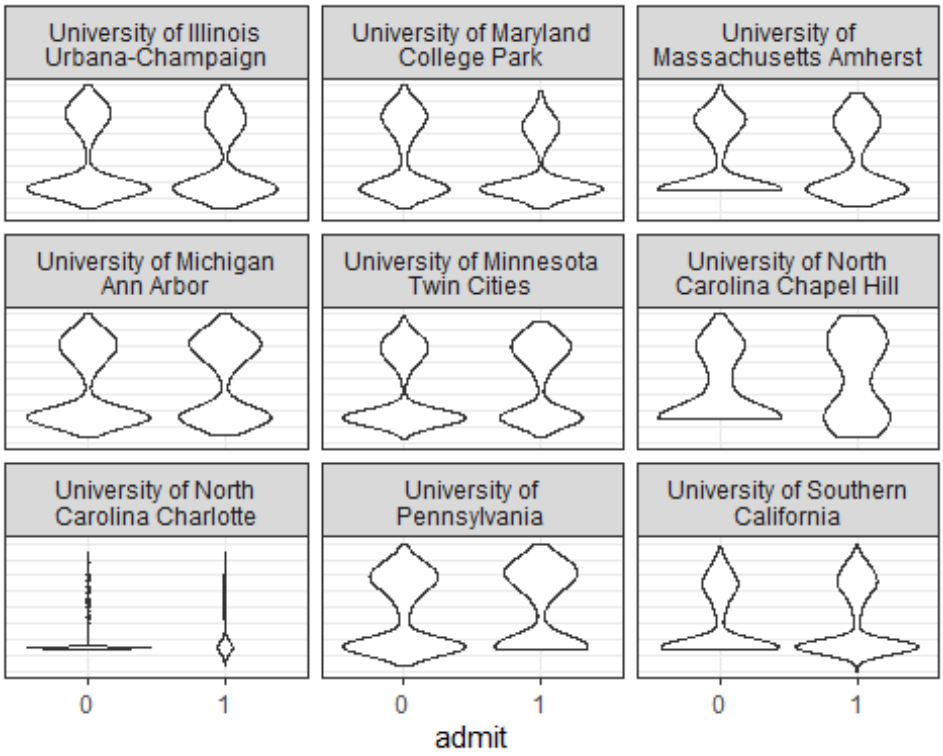

```
ggplot(okvir1, aes(x=admit,y=greA)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



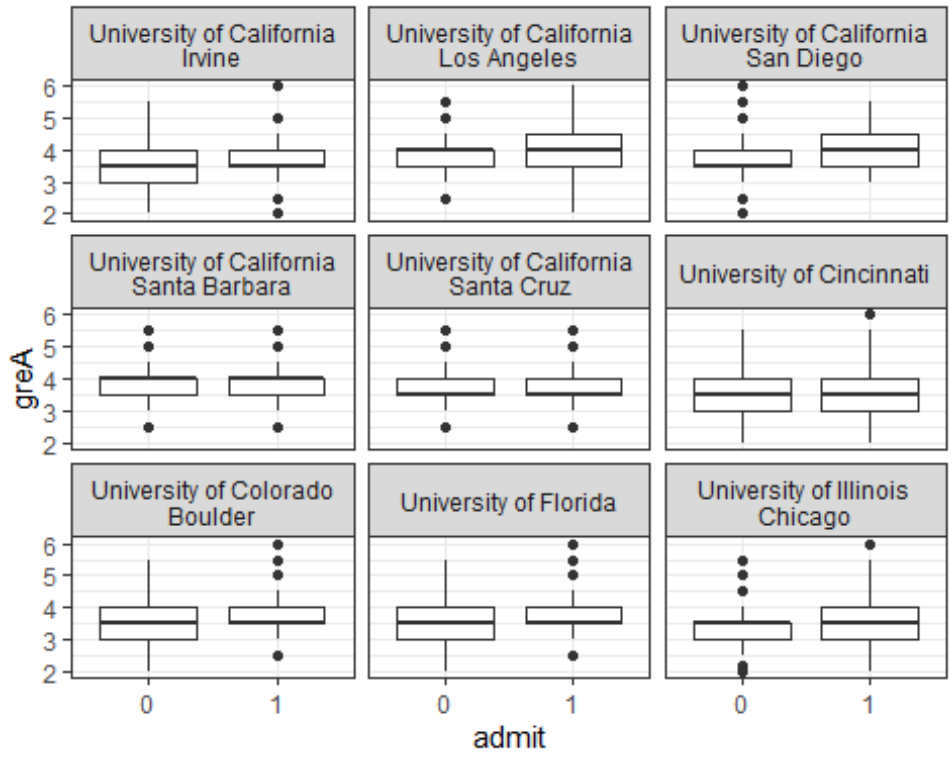

```
ggplot(okvir2, aes(x=admit,y=greA)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



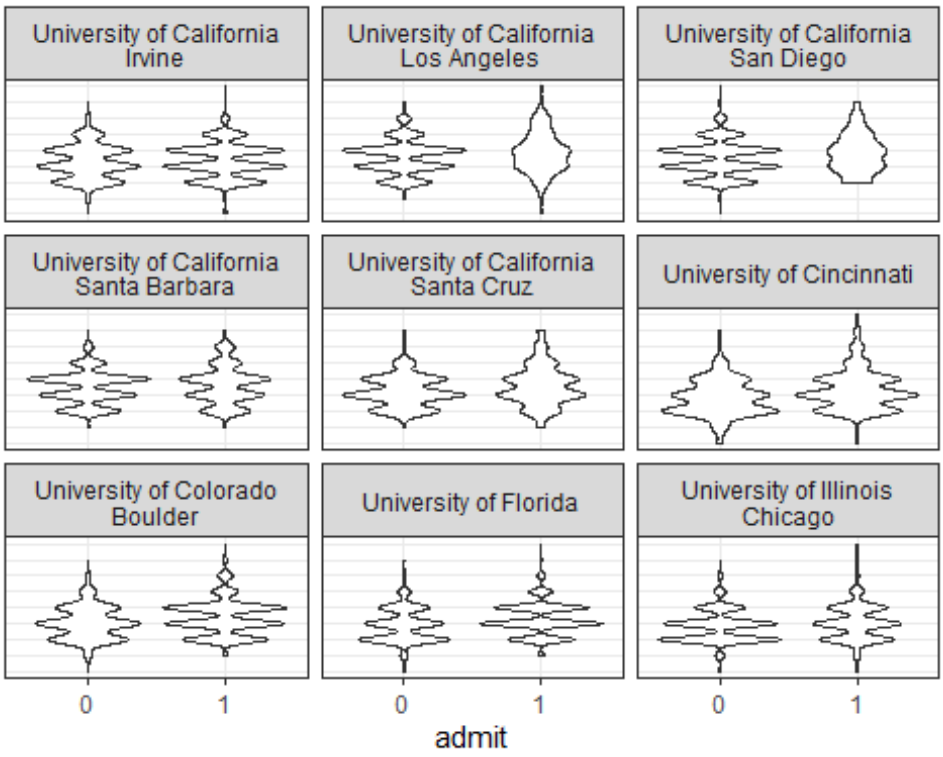
```
ggplot(okvir2, aes(x=admit,y=greV)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



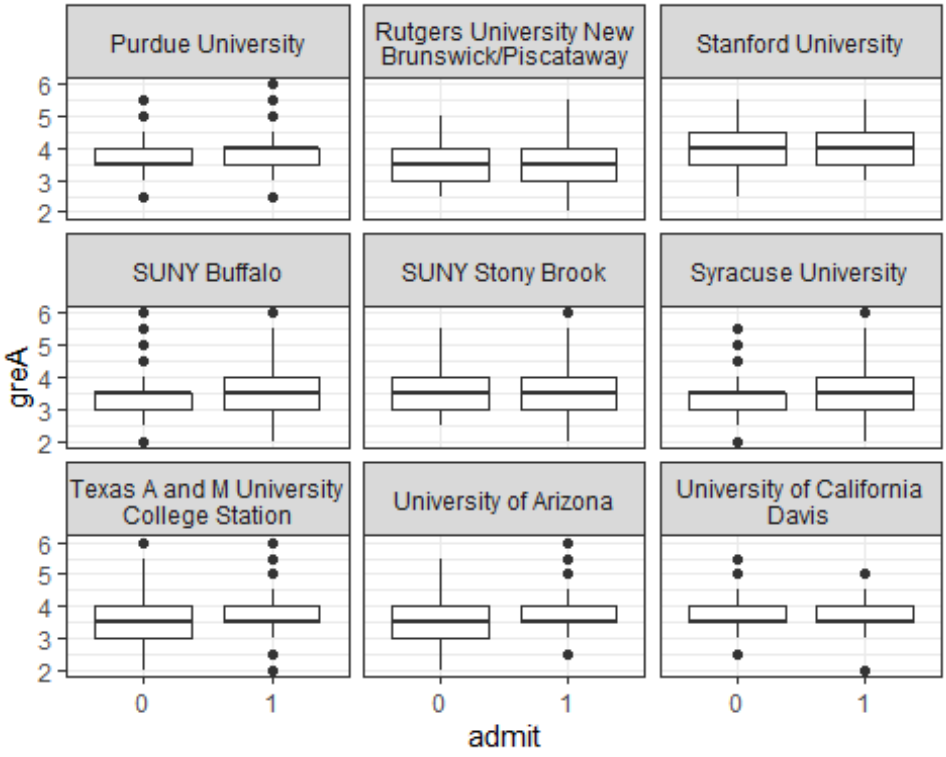

```
ggplot(okvir3, aes(x=admit,y=greA)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



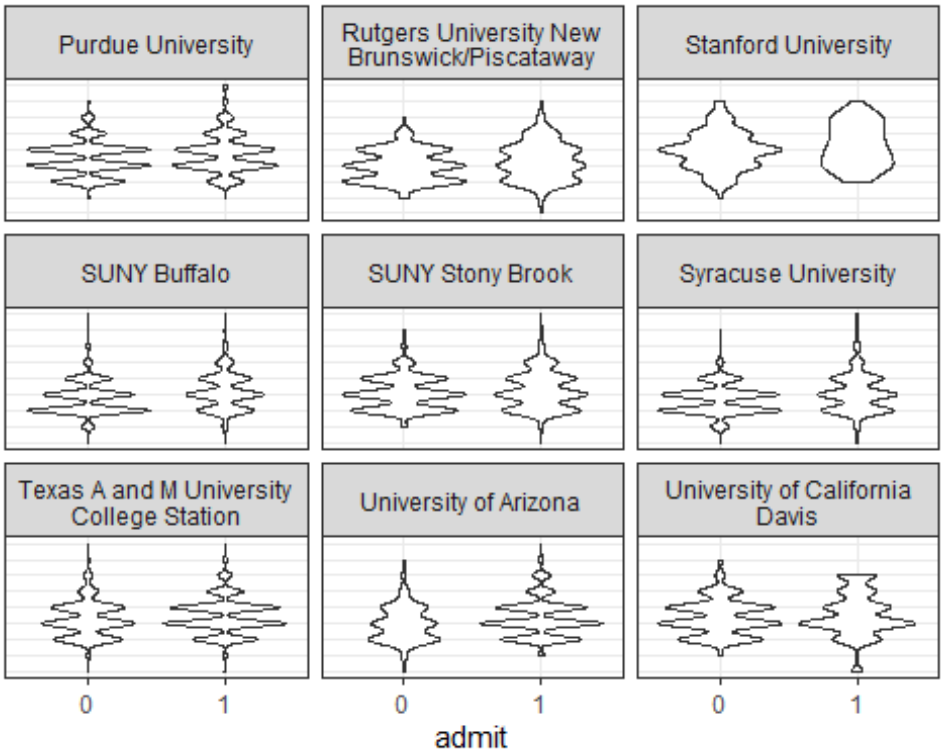

```
ggplot(okvir3, aes(x=admit,y=greA)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



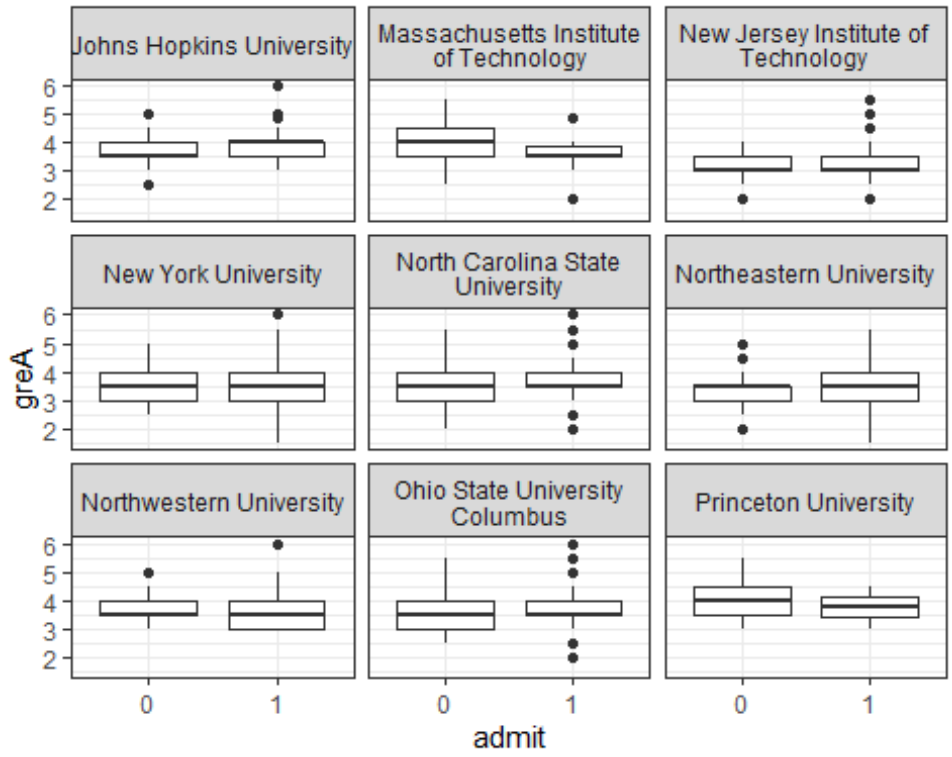

```
ggplot(okvir4, aes(x=admit,y=greA)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



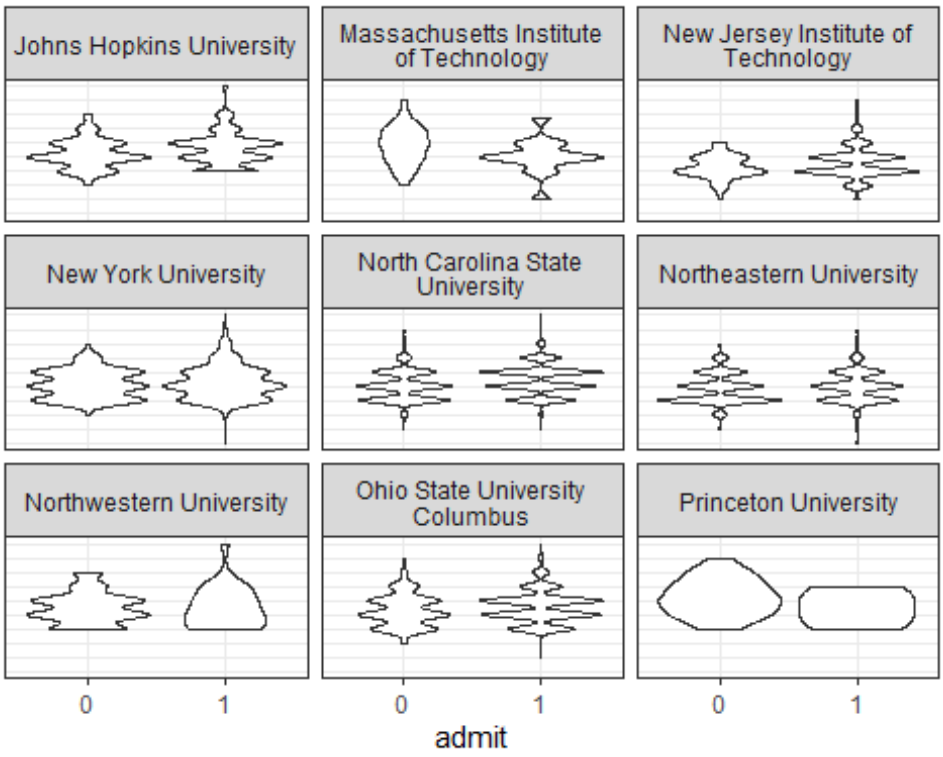
```
ggplot(okvir4, aes(x=admit,y=greA)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



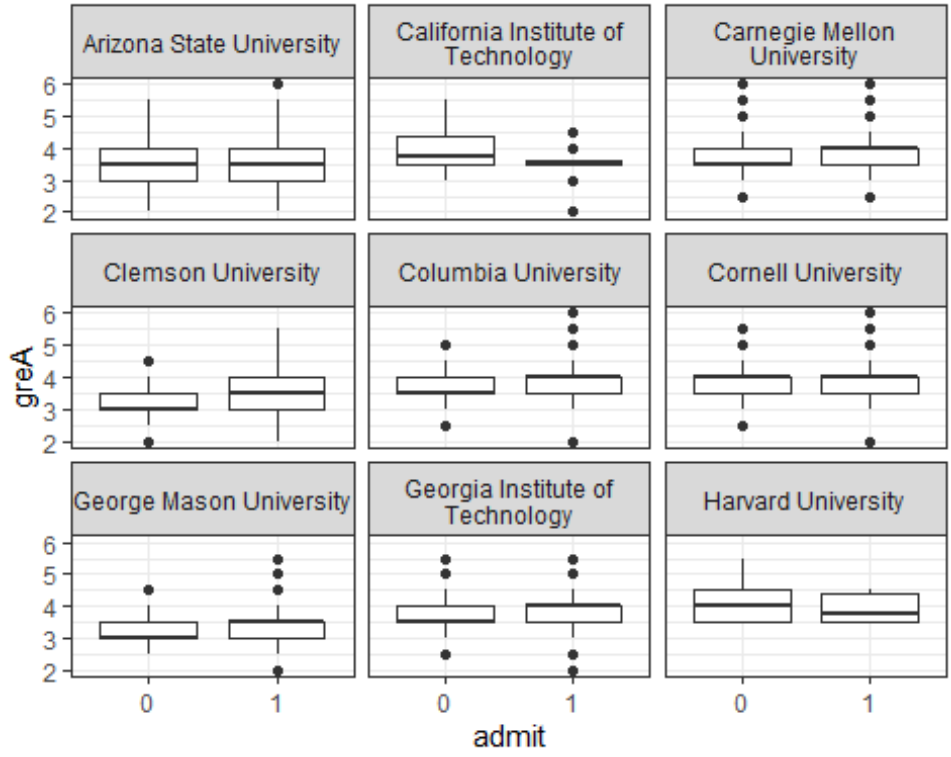

```
ggplot(okvir5, aes(x=admit,y=greA)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



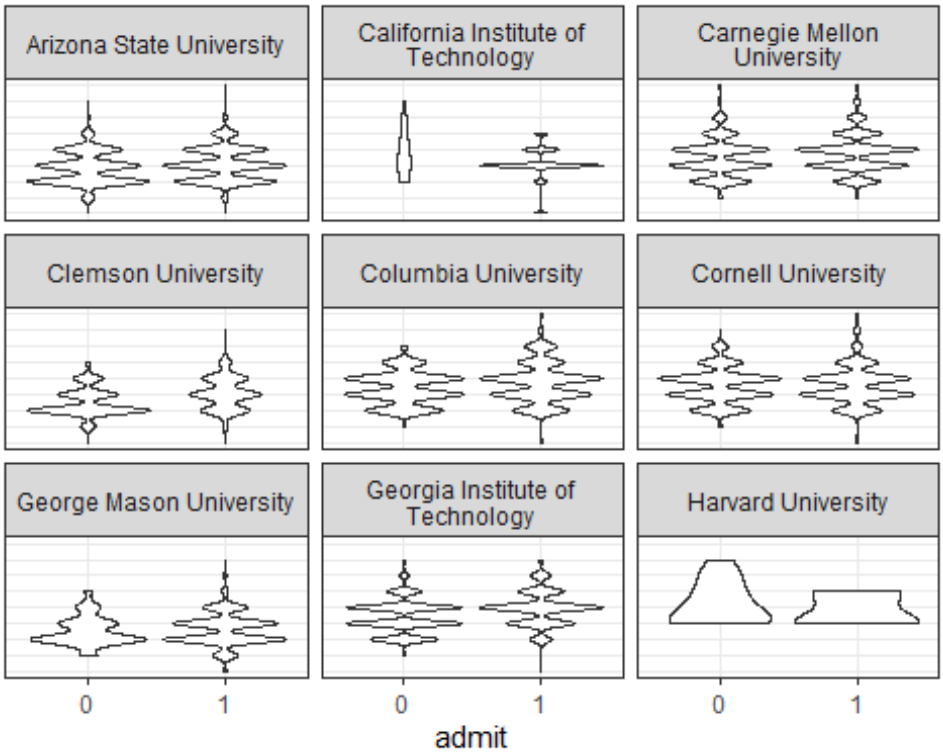

```
ggplot(okvir5, aes(x=admit,y=greA)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```




```
ggplot(okvir6, aes(x=admit,y=greA)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



```
ggplot(okvir6, aes(x=admit,y=greA)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



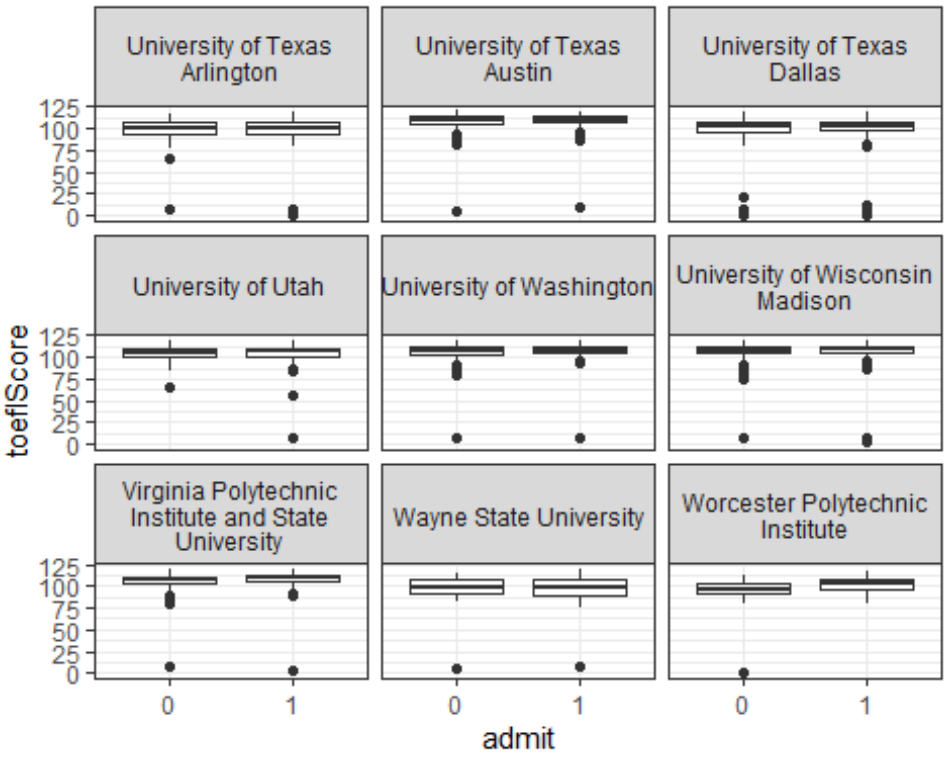

```
#K-S test normalnosti
novi_univerziteti %>% group_by(admit) %>%
summarise(izlaz = list(ks.test(greA, "pnorm", mean=mean(greA, na.rm = T),
sd=sd(greA, na.rm = T)) %>% tidy), .groups = 'drop') %>% unnest(c(izlaz))

## Warning in ks.test(greA, "pnorm", mean = mean(greA, na.rm = T), sd =
sd(greA, :
## ties should not be present for the Kolmogorov-Smirnov test

## Warning in ks.test(greA, "pnorm", mean = mean(greA, na.rm = T), sd =
sd(greA, :
## ties should not be present for the Kolmogorov-Smirnov test

## # A tibble: 2 x 5
##   admit statistic p.value method                                alternative
##   <fct>     <dbl>   <dbl> <chr>                                 <chr>
## 1 0         0.193       0 One-sample Kolmogorov-Smirnov test two-sided
## 2 1         0.202       0 One-sample Kolmogorov-Smirnov test two-sided
```

Na osnovu grafika iznad vidimo da rezultati greV testa utiče na to da li je osoba primljena na fakultet ili nije, ali zavisi od fakulteta. Testiranjem normalnosti Kolmogorov-Smirnov testom

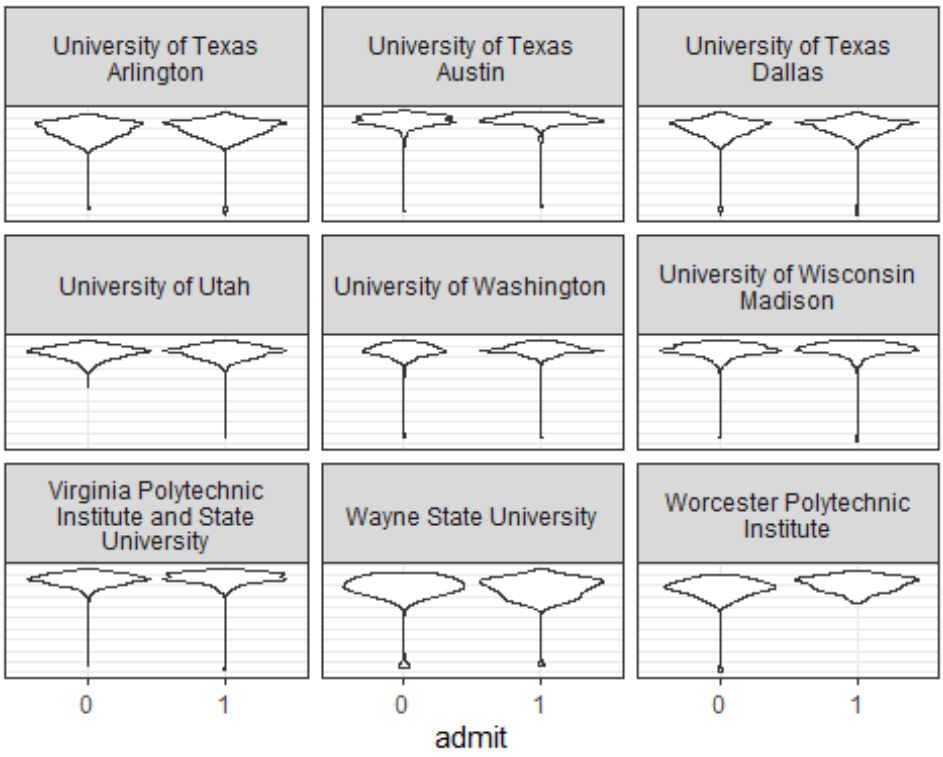
pokazano je da ne postoji normalnost unutar obe grupe obeležja (p = 0.0 < α=0.05, p = 0.0 < α=0.05).

toeflScore i admit

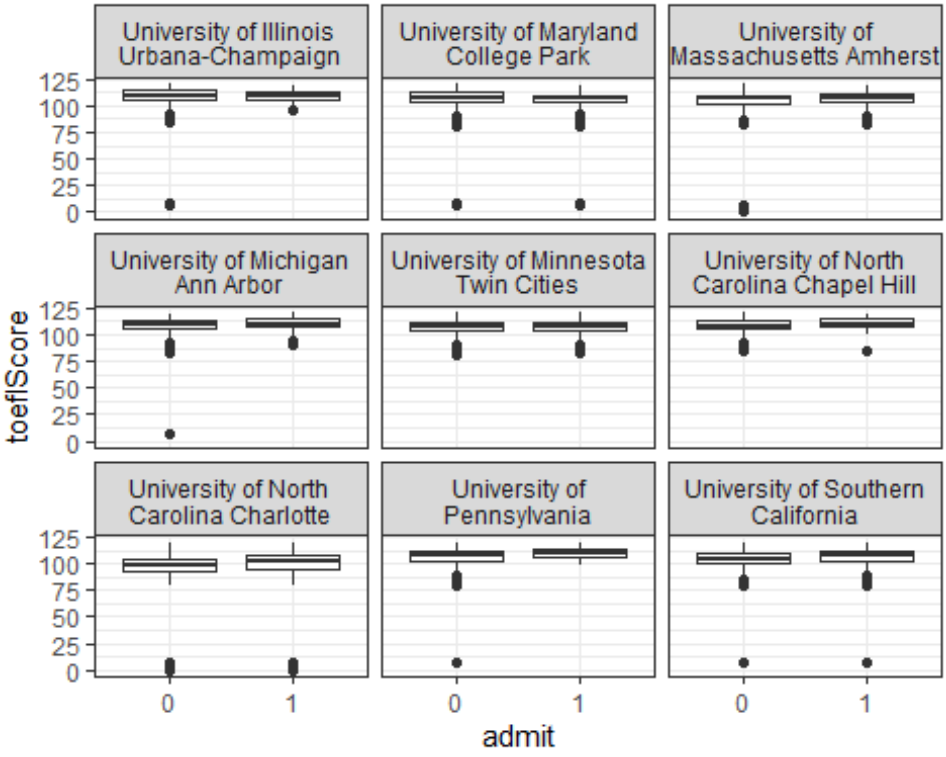
```
ggplot(okvir1, aes(x=admit,y=toeflScore)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



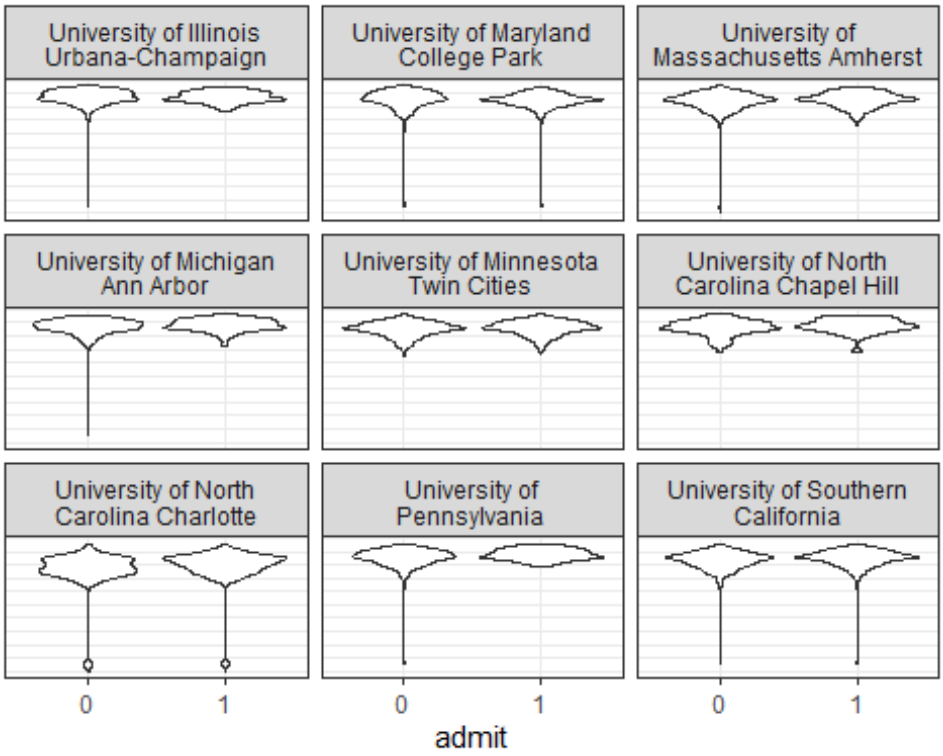

```
ggplot(okvir1, aes(x=admit,y=toeflScore)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



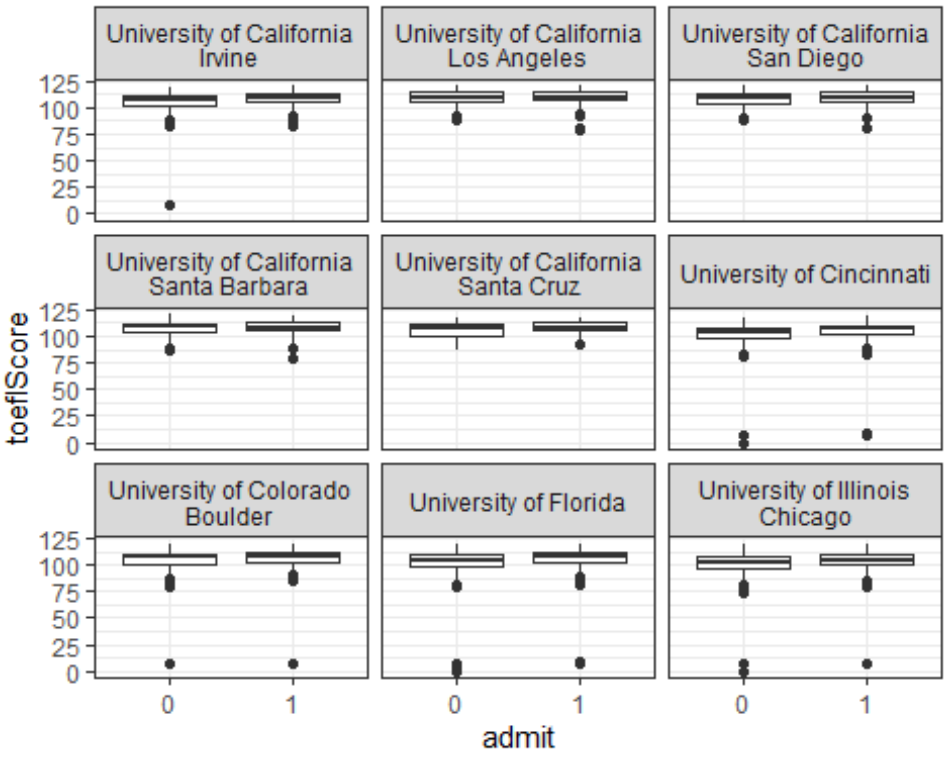

```
ggplot(okvir2, aes(x=admit,y=toeflScore)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



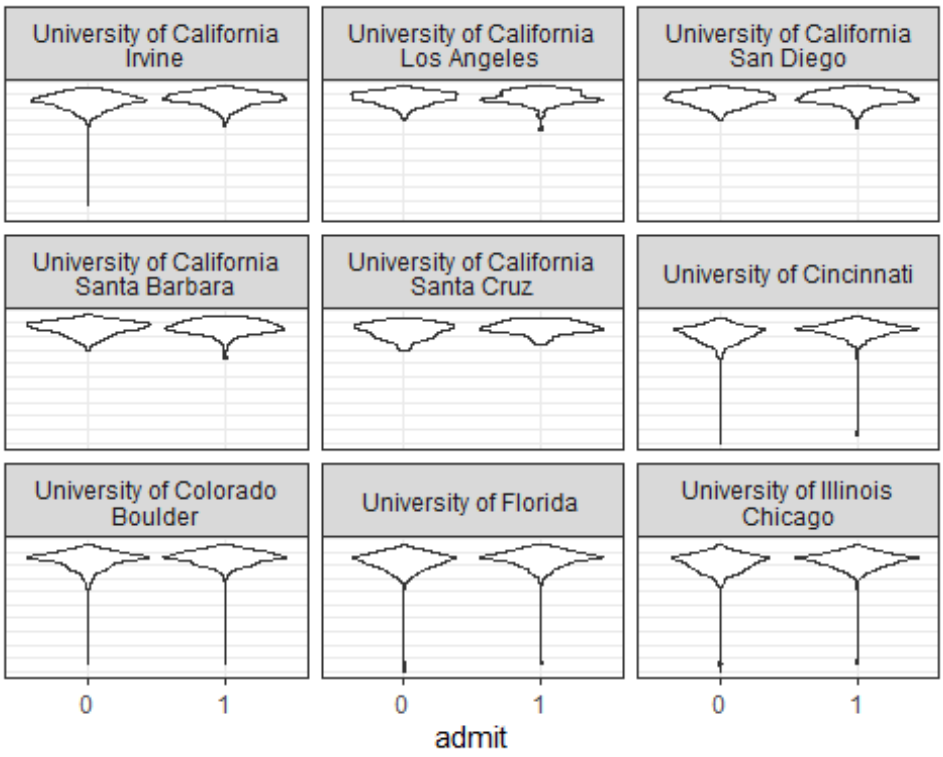
```
ggplot(okvir2, aes(x=admit,y=toeflScore)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



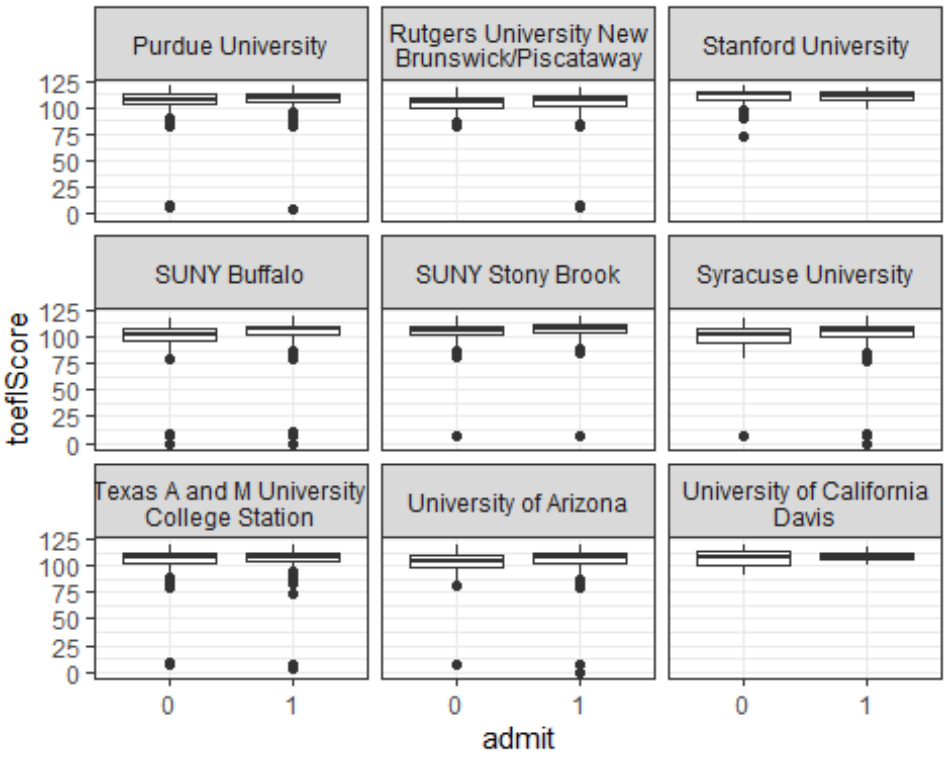

```
ggplot(okvir3, aes(x=admit,y=toeflScore)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



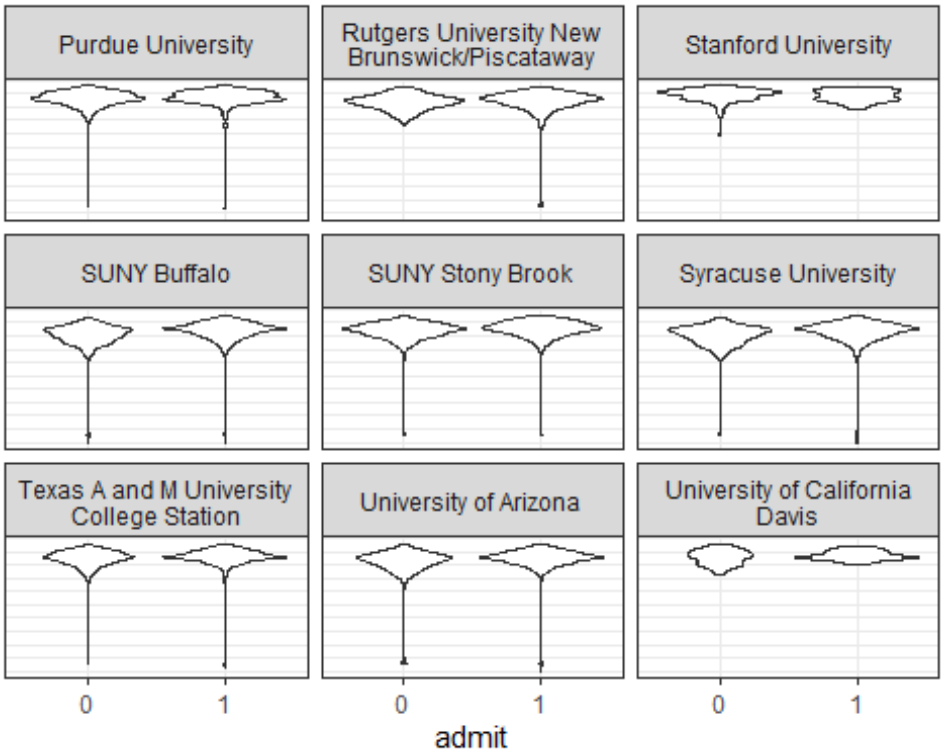

```
ggplot(okvir3, aes(x=admit,y=toeflScore)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



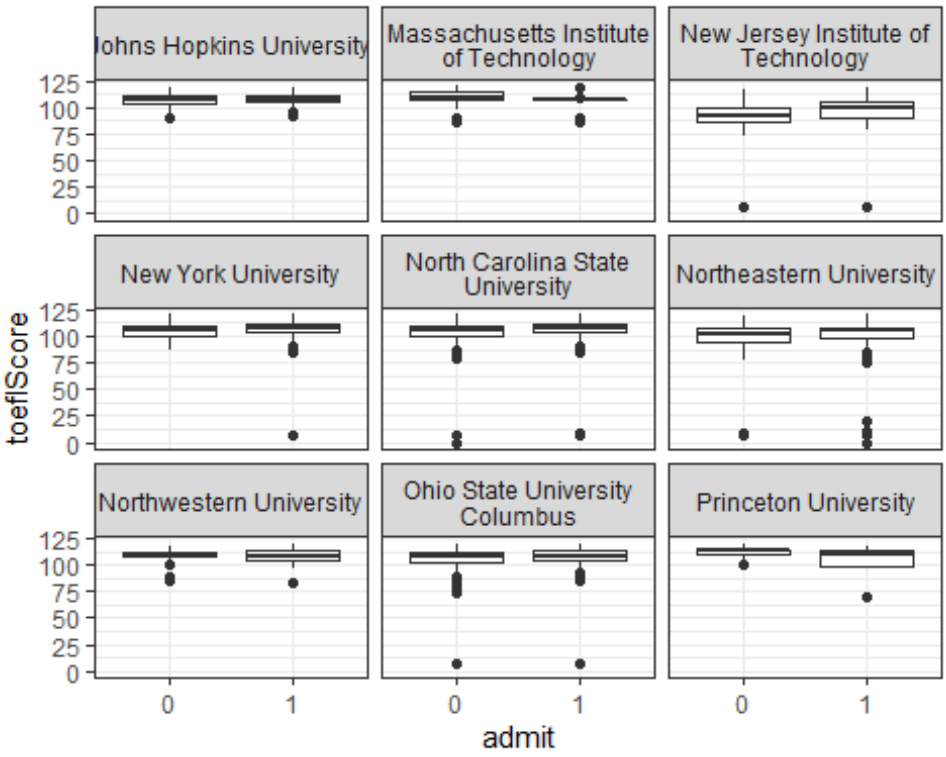

```
ggplot(okvir4, aes(x=admit,y=toeflScore)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



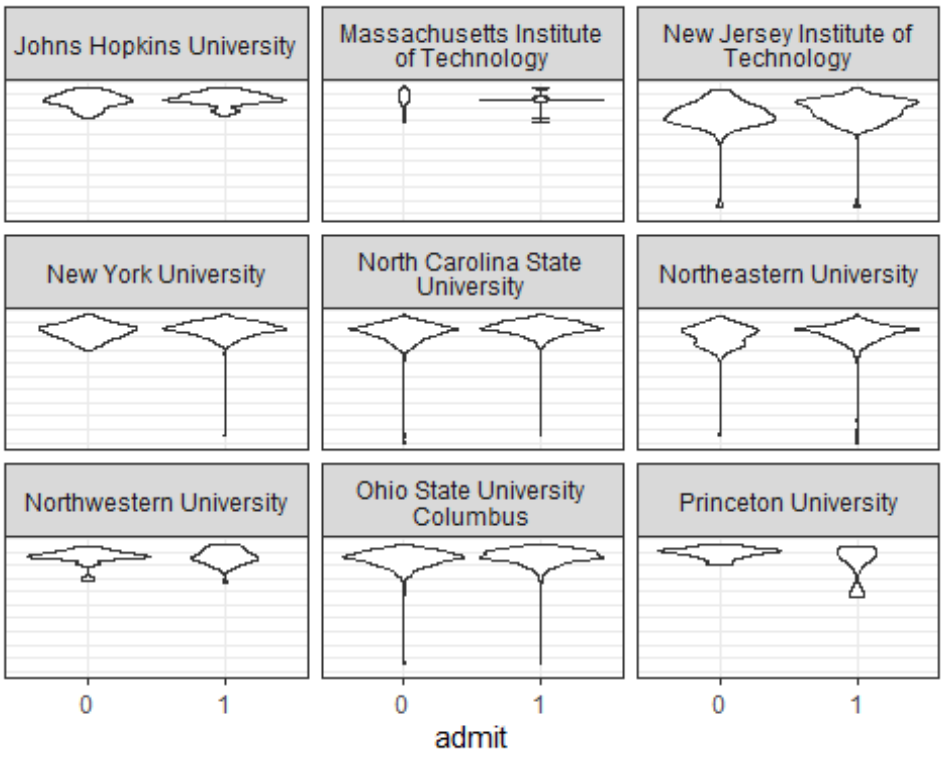
```
ggplot(okvir4, aes(x=admit,y=toeflScore)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



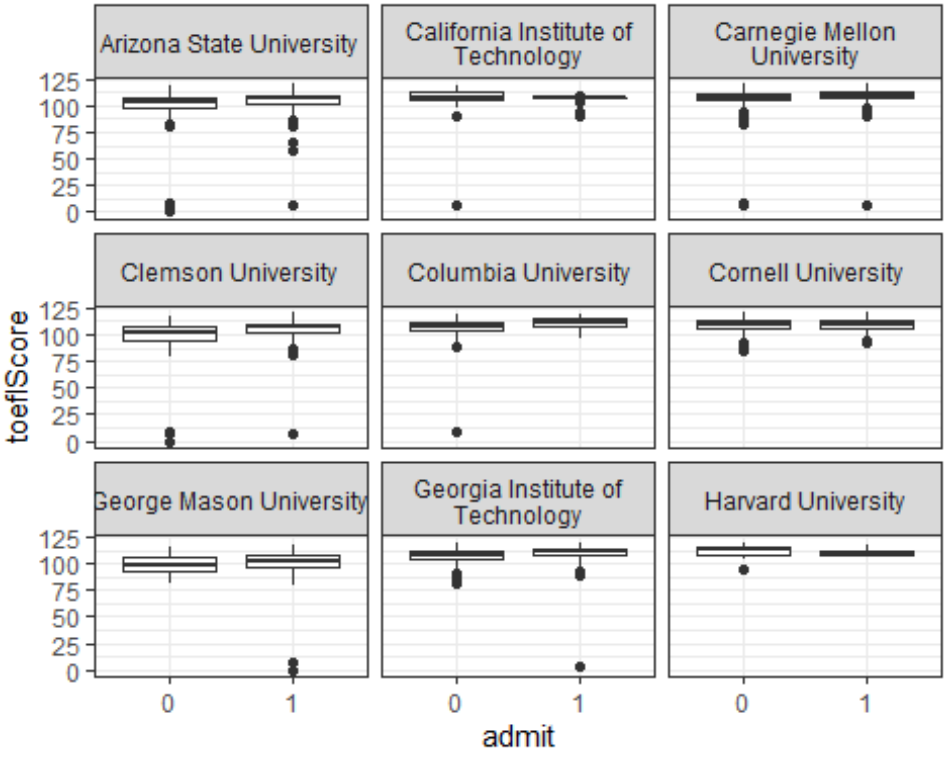

```
ggplot(okvir5, aes(x=admit,y=toeflScore)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



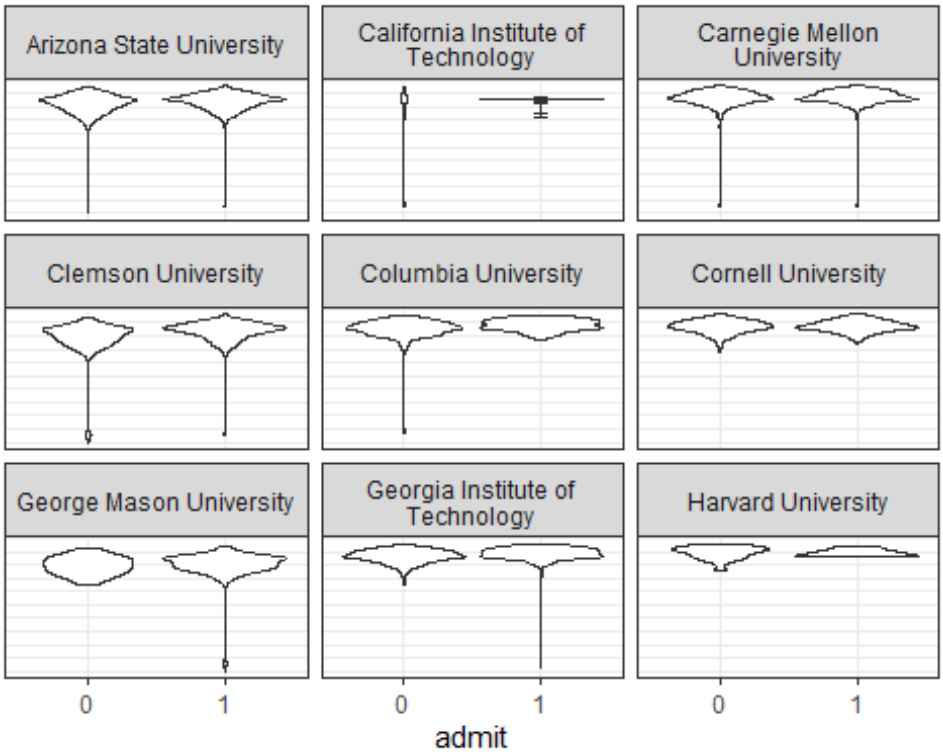

```
ggplot(okvir5, aes(x=admit,y=toeflScore)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```




```
ggplot(okvir6, aes(x=admit,y=toeflScore)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



```
ggplot(okvir6, aes(x=admit,y=toeflScore)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



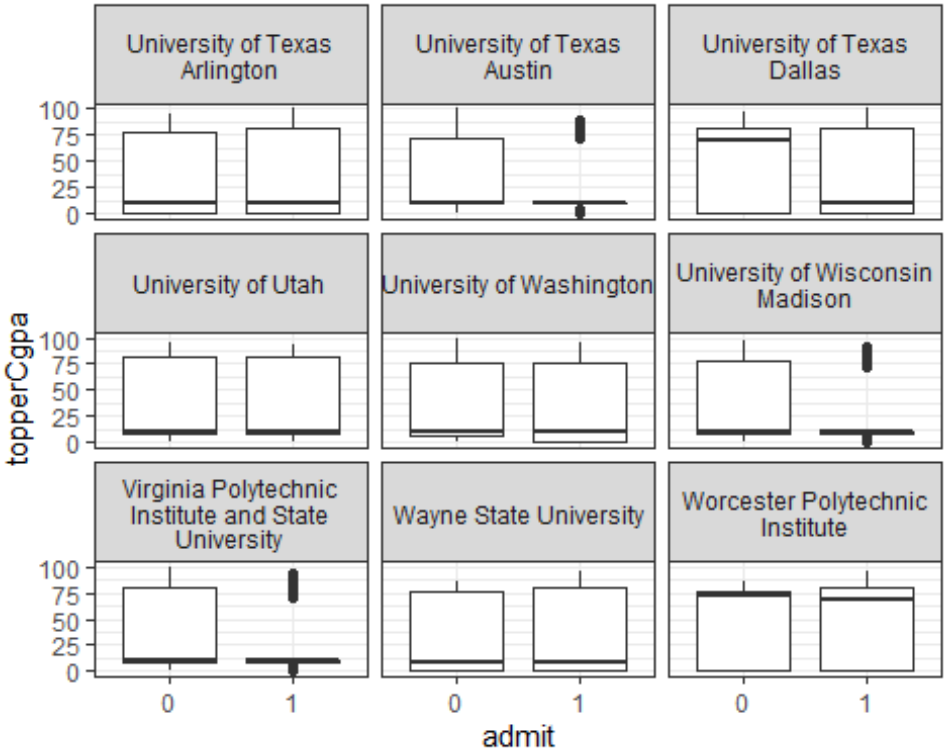

```
#K-S test normalnosti
novi_univerziteti %>% group_by(admit) %>%
summarise(izlaz = list(ks.test(toeflScore, "pnorm", mean=mean(toeflScore,
na.rm = T),
sd=sd(toeflScore, na.rm = T)) %>% tidy), .groups = 'drop') %>%
unnest(c(izlaz))
```

```
## Warning in ks.test(toeflScore, "pnorm", mean = mean(toeflScore, na.rm = T),
:
## ties should not be present for the Kolmogorov-Smirnov test

## Warning in ks.test(toeflScore, "pnorm", mean = mean(toeflScore, na.rm = T),
:
## ties should not be present for the Kolmogorov-Smirnov test

## # A tibble: 2 x 5
##   admit statistic p.value method                                alternative
##   <fct>     <dbl>   <dbl> <chr>                                 <chr>
## 1 0         0.118       0 One-sample Kolmogorov-Smirnov test two-sided
## 2 1         0.135       0 One-sample Kolmogorov-Smirnov test two-sided
```

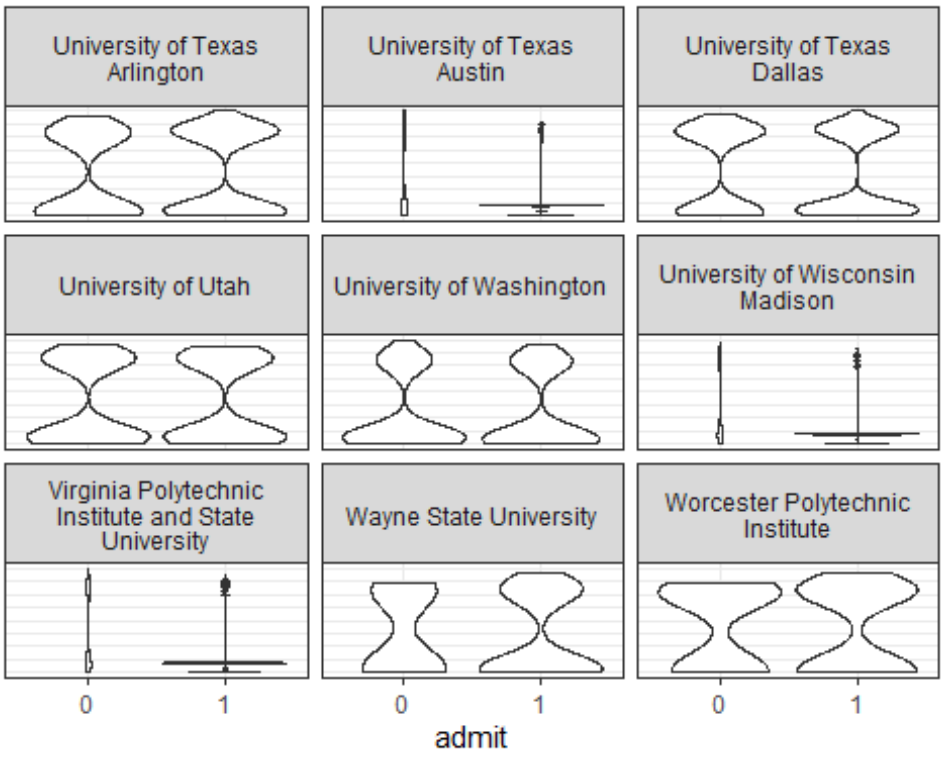
Na osnovu grafika iznad vidimo da toefl skor utiče na to da li je osoba primljena na fakultet ili nije, ali zavisi od fakulteta. Testiranjem normalnosti Kolmogorov-Smirnov testom pokazano je da ne postoji normalnost unutar obe grupe obeležja (p = 0.0 < α=0.05, p = 0.0 < α=0.05).

topperCgpa i admit

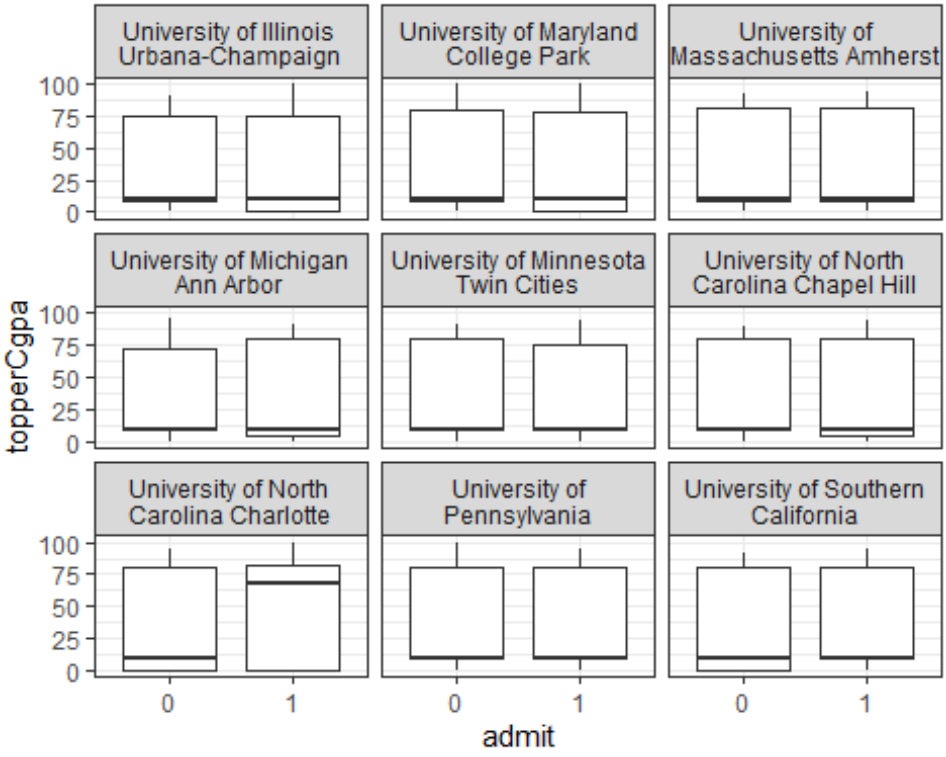
```
ggplot(okvir1, aes(x=admit,y=topperCgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



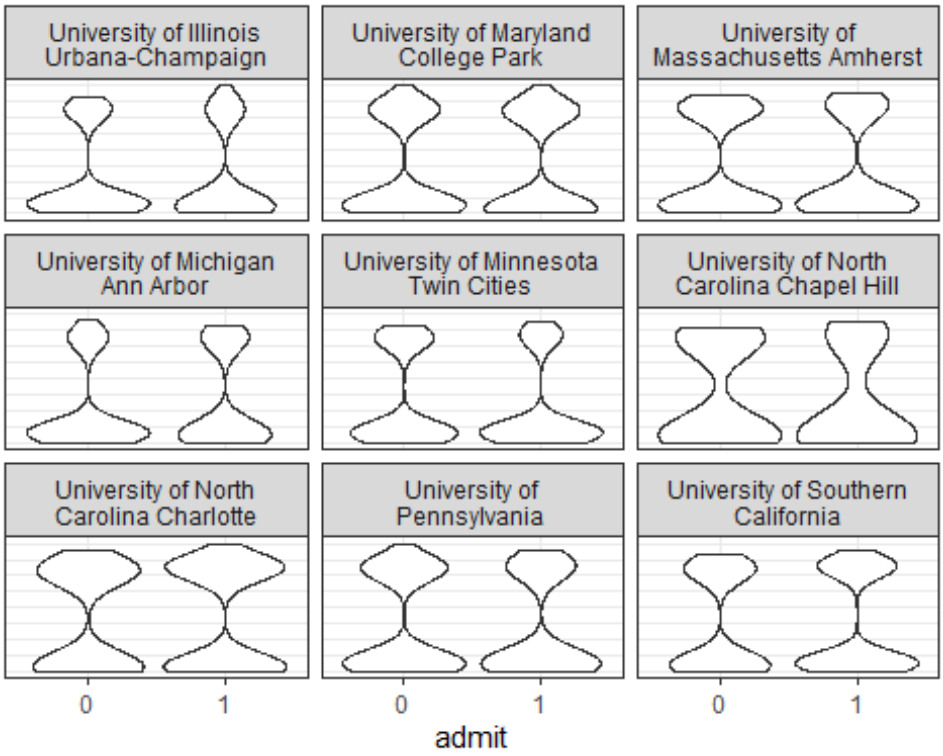

```
ggplot(okvir1, aes(x=admit,y=topperCgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



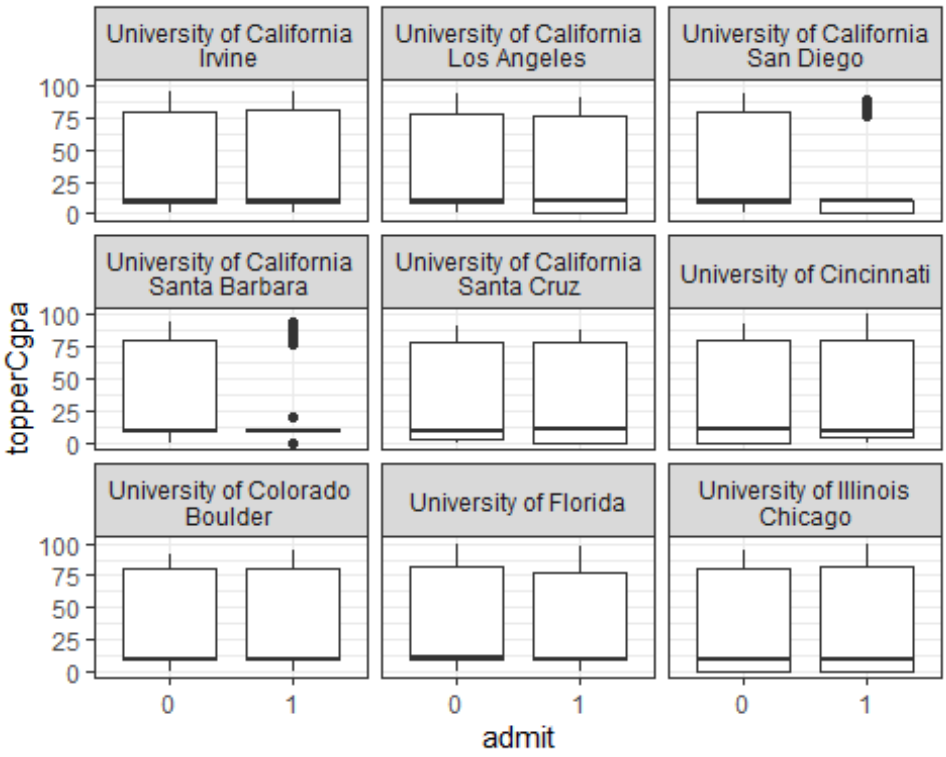

```
ggplot(okvir2, aes(x=admit,y=topperCgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



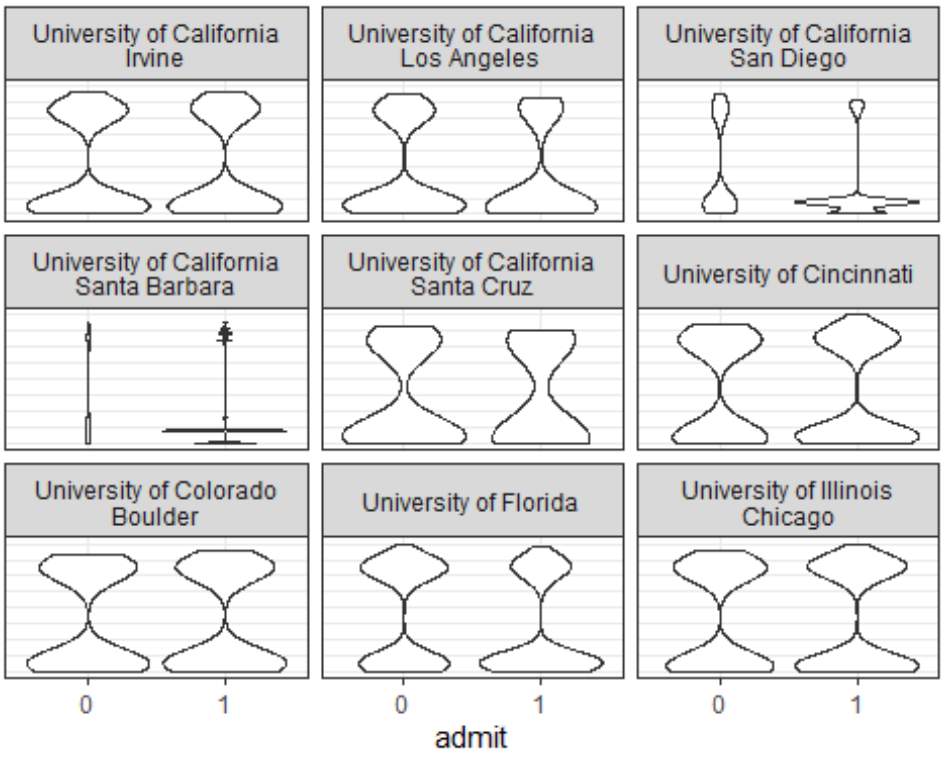
```
ggplot(okvir2, aes(x=admit,y=topperCgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



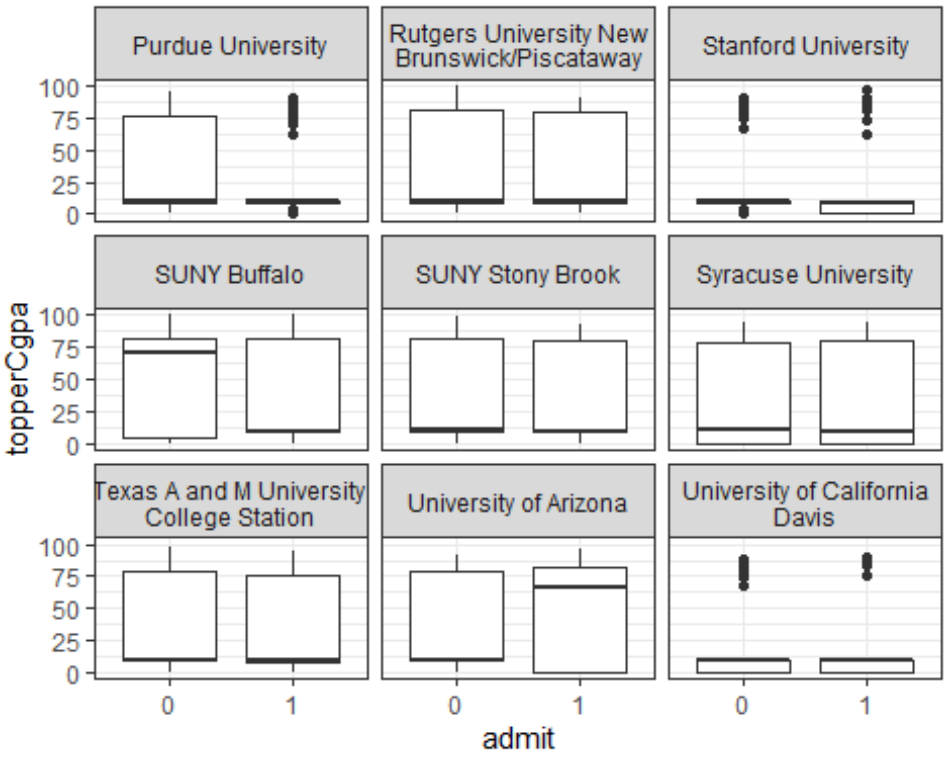

```
ggplot(okvir3, aes(x=admit,y=topperCgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



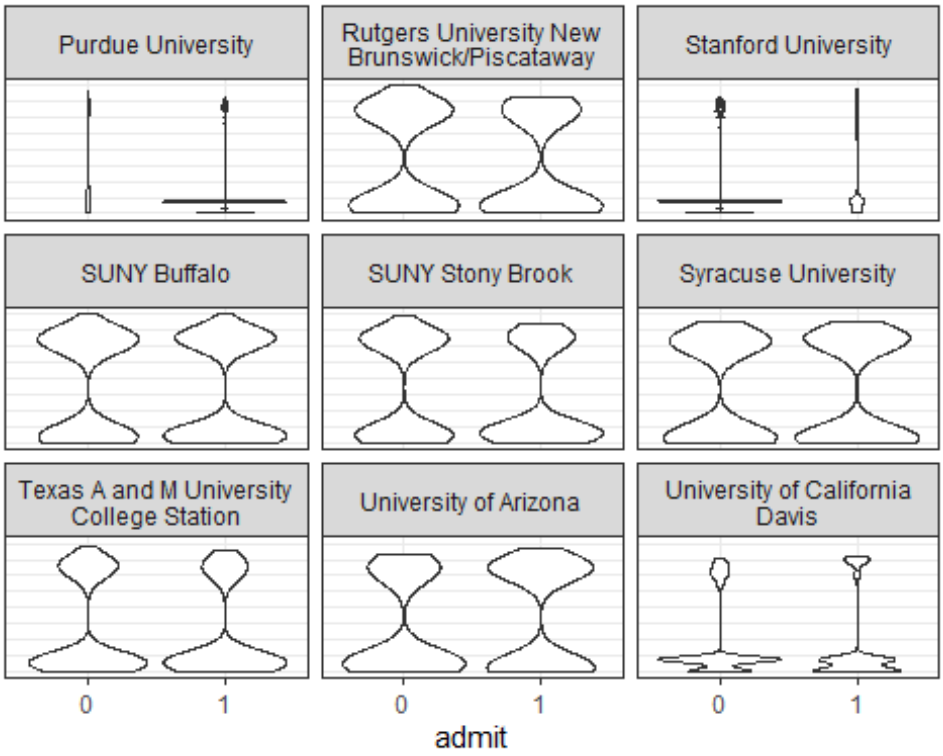

```
ggplot(okvir3, aes(x=admit,y=topperCgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



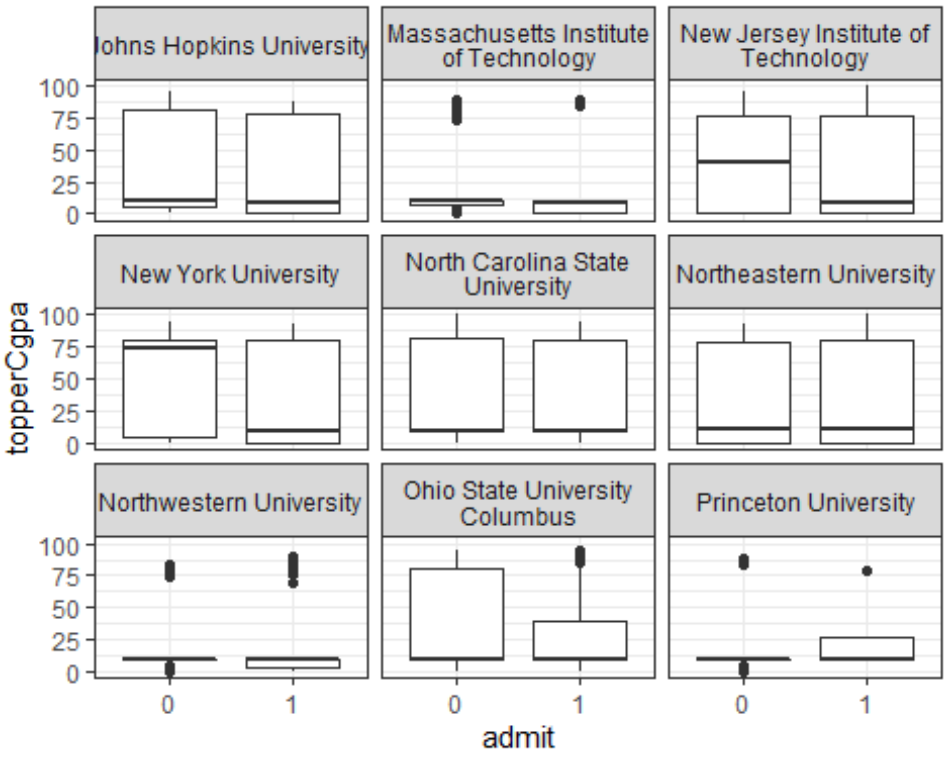

```
ggplot(okvir4, aes(x=admit,y=topperCgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



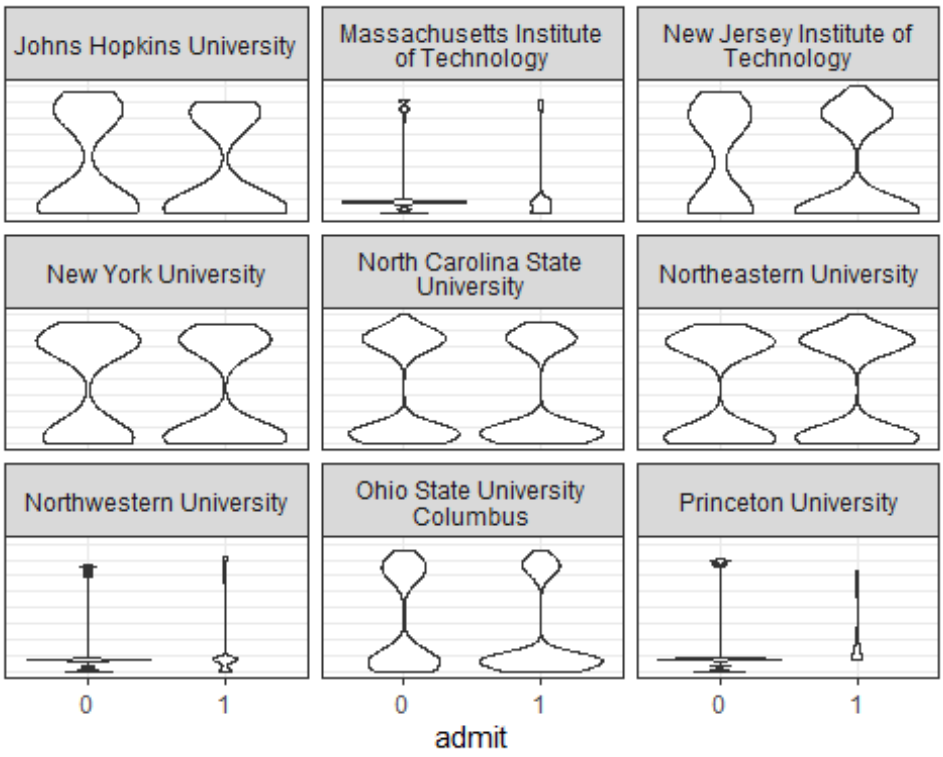
```
ggplot(okvir4, aes(x=admit,y=topperCgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



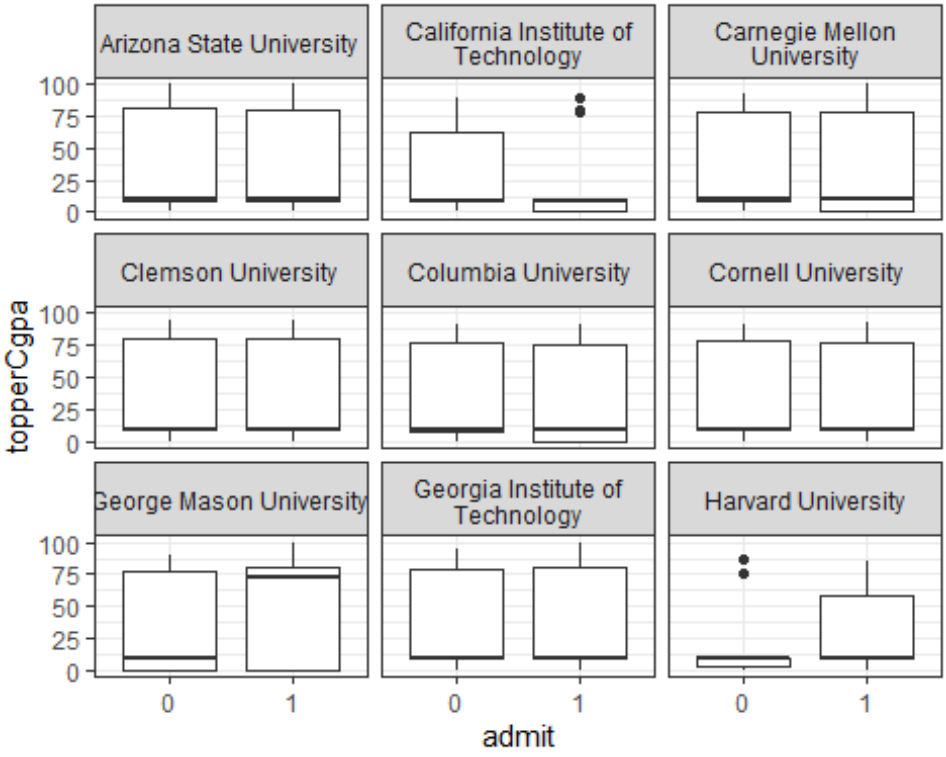

```
ggplot(okvir5, aes(x=admit,y=topperCgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



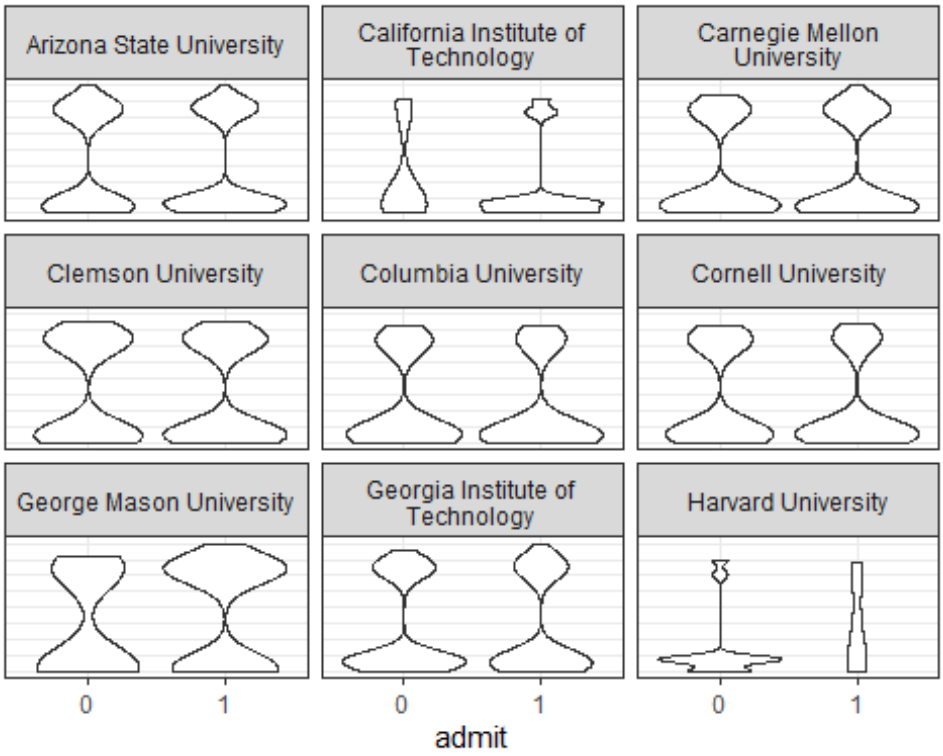

```
ggplot(okvir5, aes(x=admit,y=topperCgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```




```
ggplot(okvir6, aes(x=admit,y=topperCgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



```
ggplot(okvir6, aes(x=admit,y=topperCgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



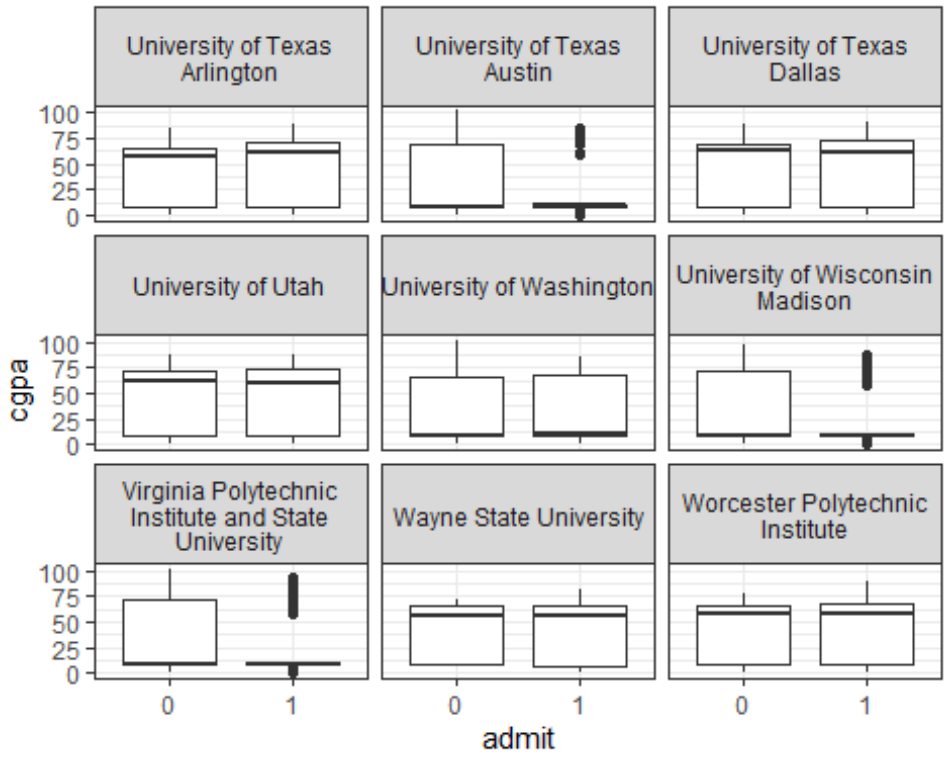

```
#K-S test normalnosti
novi_univerziteti %>% group_by(admit) %>%
summarise(izlaz = list(ks.test(topperCgpa, "pnorm", mean=mean(topperCgpa,
na.rm = T),
sd=sd(topperCgpa, na.rm = T)) %>% tidy), .groups = 'drop') %>%
unnest(c(izlaz))
```

```
## Warning in ks.test(topperCgpa, "pnorm", mean = mean(topperCgpa, na.rm = T),
:
## ties should not be present for the Kolmogorov-Smirnov test

## Warning in ks.test(topperCgpa, "pnorm", mean = mean(topperCgpa, na.rm = T),
:
## ties should not be present for the Kolmogorov-Smirnov test

## # A tibble: 2 x 5
##   admit statistic p.value method                                alternative
##   <fct>     <dbl>   <dbl> <chr>                                 <chr>
## 1 0         0.359       0 One-sample Kolmogorov-Smirnov test two-sided
## 2 1         0.360       0 One-sample Kolmogorov-Smirnov test two-sided
```

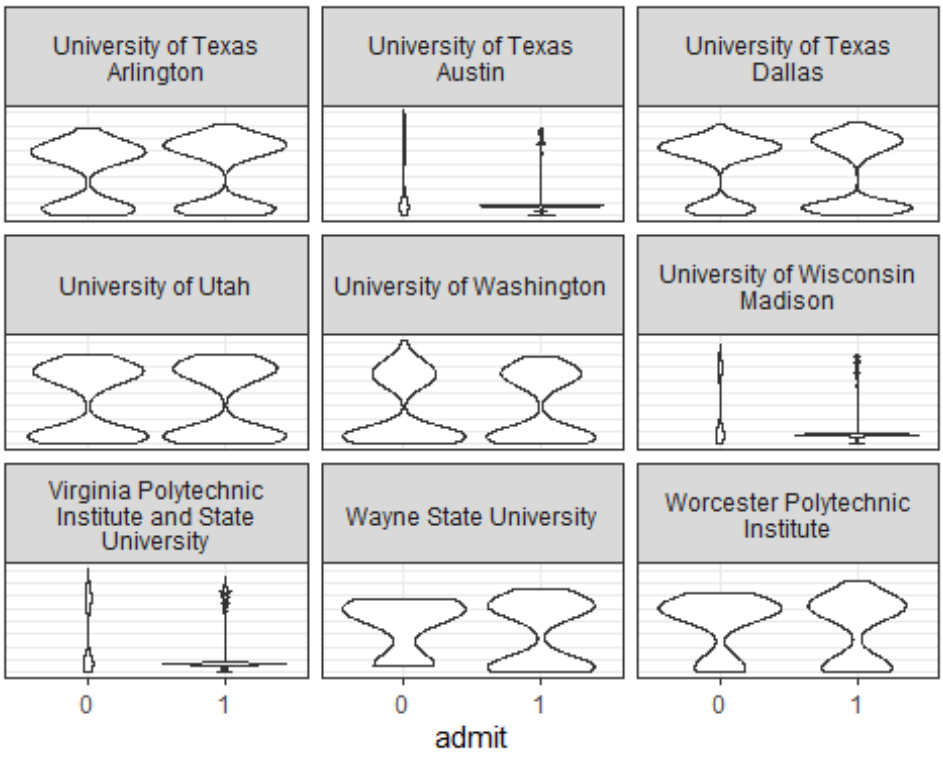
Na osnovu grafika iznad vidimo da srednja prosečna ocena u najvišem delu rang liste utiče na to da li je osoba primljena na fakultet ili nije, ali zavisi od fakulteta. Testiranjem normalnosti Kolmogorov-Smirnov testom pokazano je da ne postoji normalnost unutar obe grupe obeležja ($p = 0.0 < \alpha=0.05$, $p = 0.0 < \alpha=0.05$).

cgpa i admit

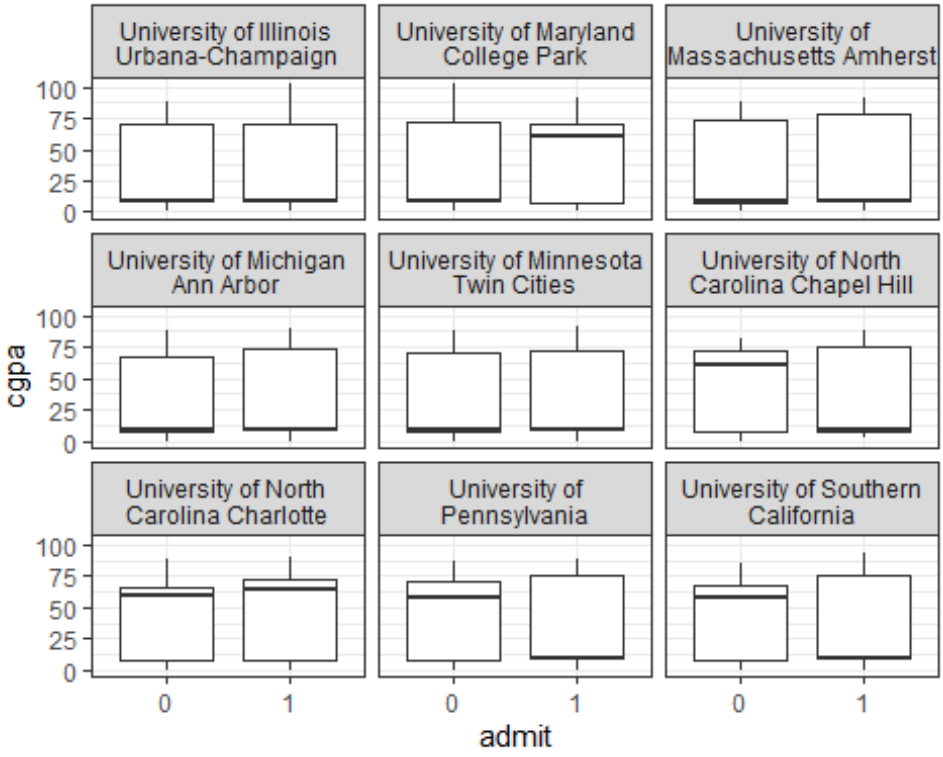
```
ggplot(okvir1, aes(x=admit,y=cgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



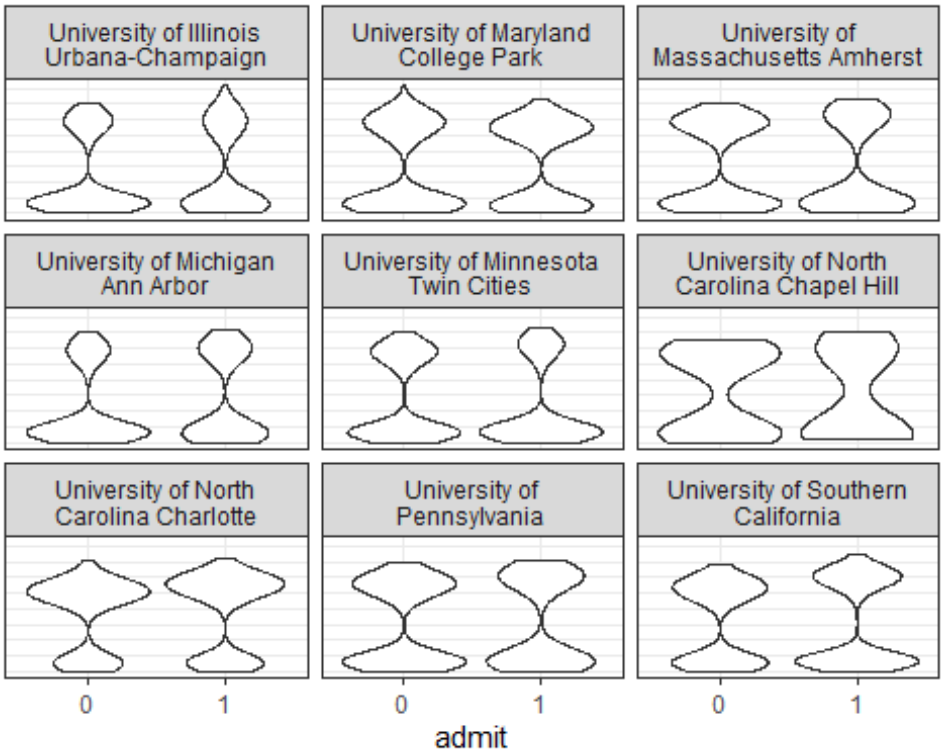

```
ggplot(okvir1, aes(x=admit,y=cgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



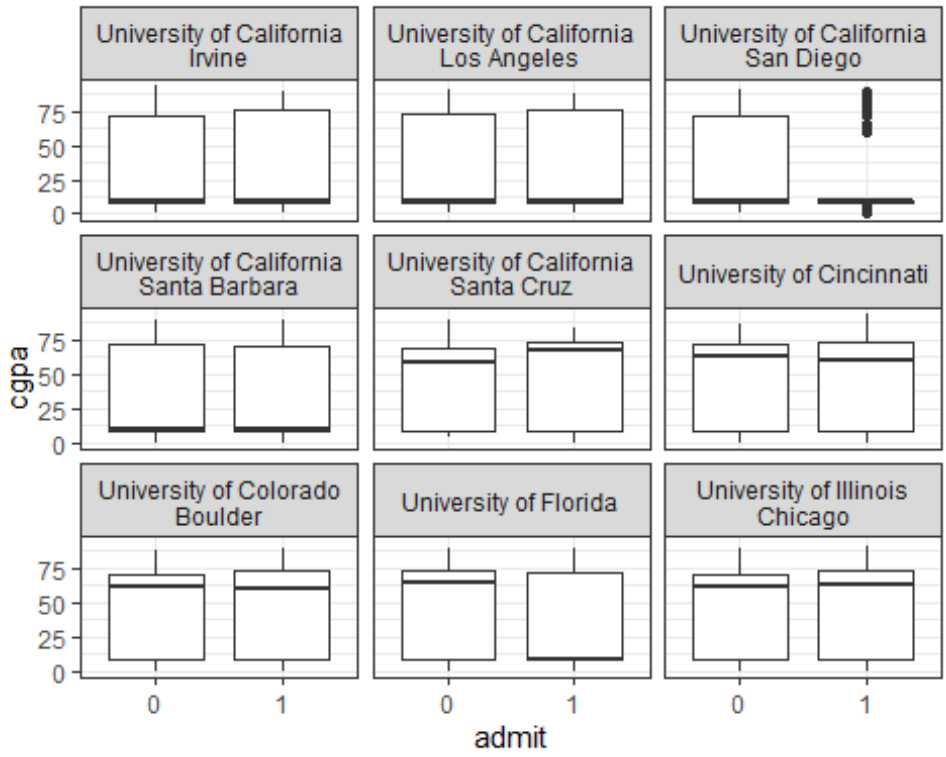

```
ggplot(okvir2, aes(x=admit,y=cgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



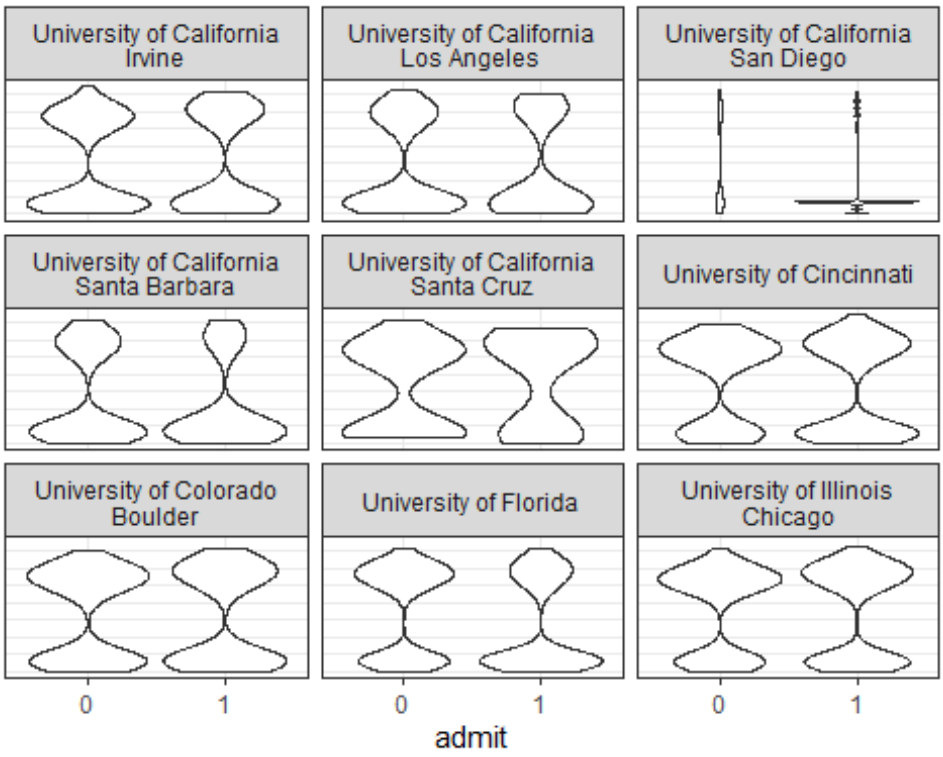
```
ggplot(okvir2, aes(x=admit,y=cgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



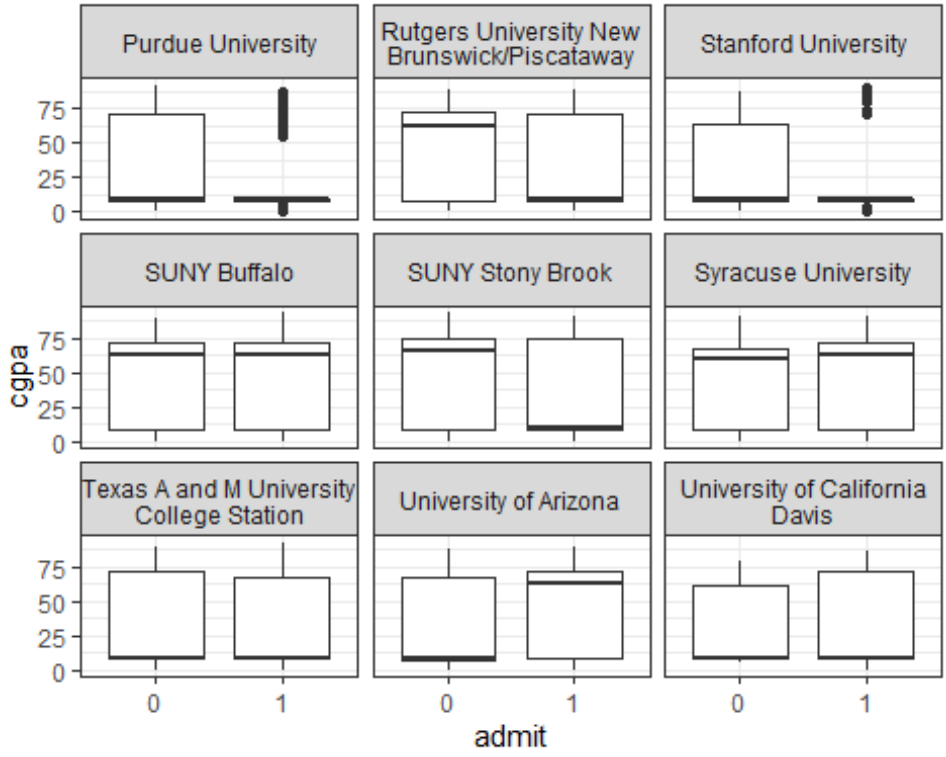

```
ggplot(okvir3, aes(x=admit,y=cgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



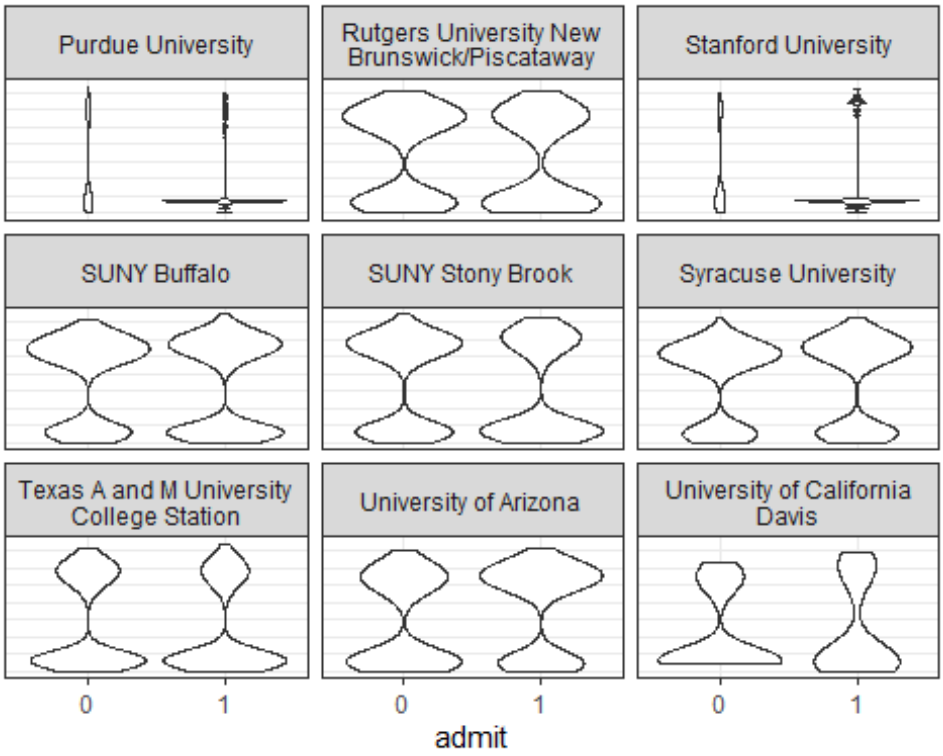

```
ggplot(okvir3, aes(x=admit,y=cgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



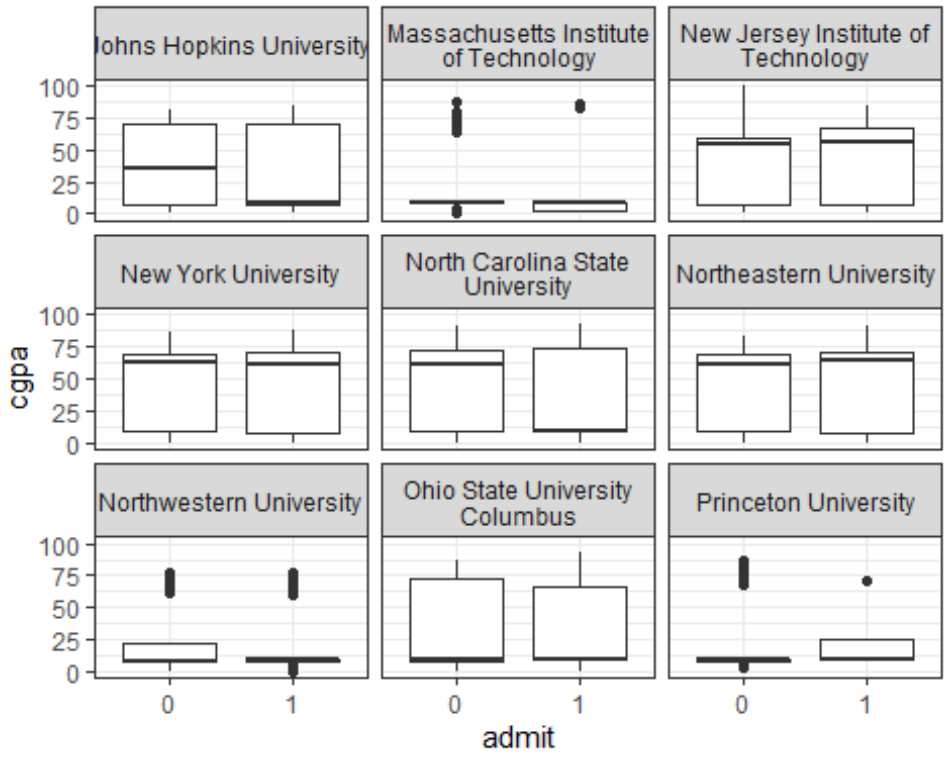

```
ggplot(okvir4, aes(x=admit,y=cgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



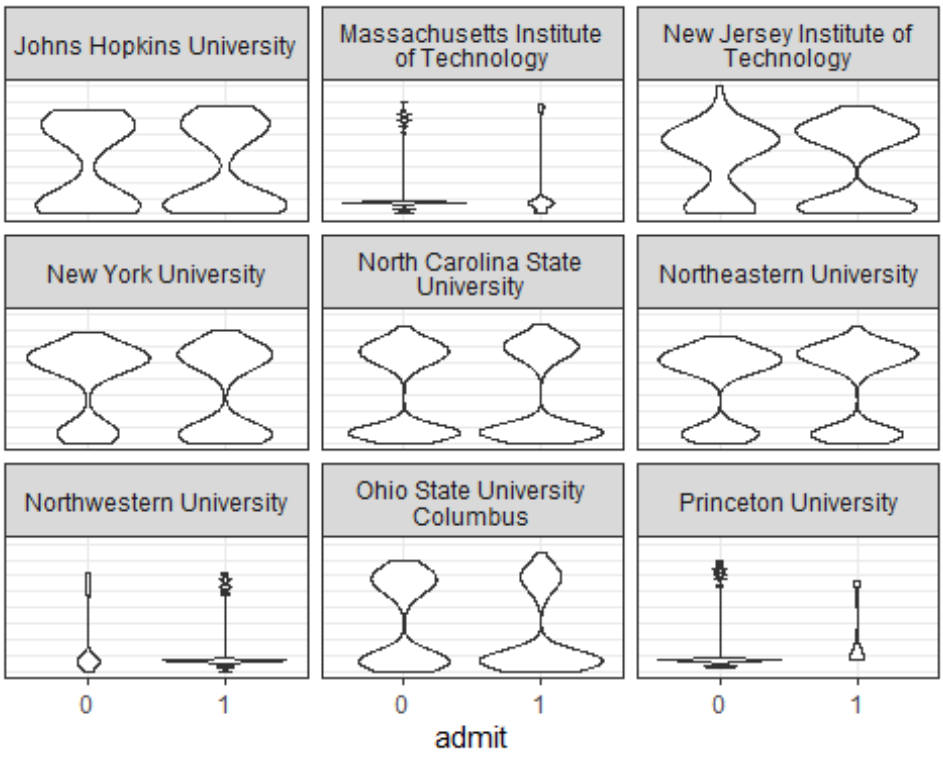
```
ggplot(okvir4, aes(x=admit,y=cgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```



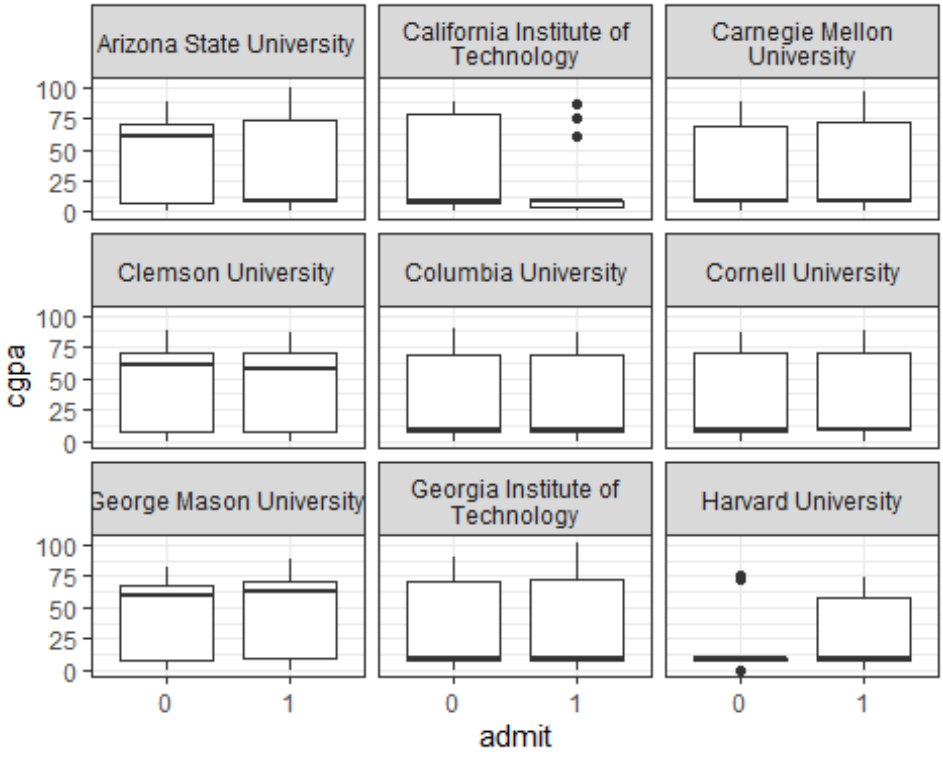

```
ggplot(okvir5, aes(x=admit,y=cgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



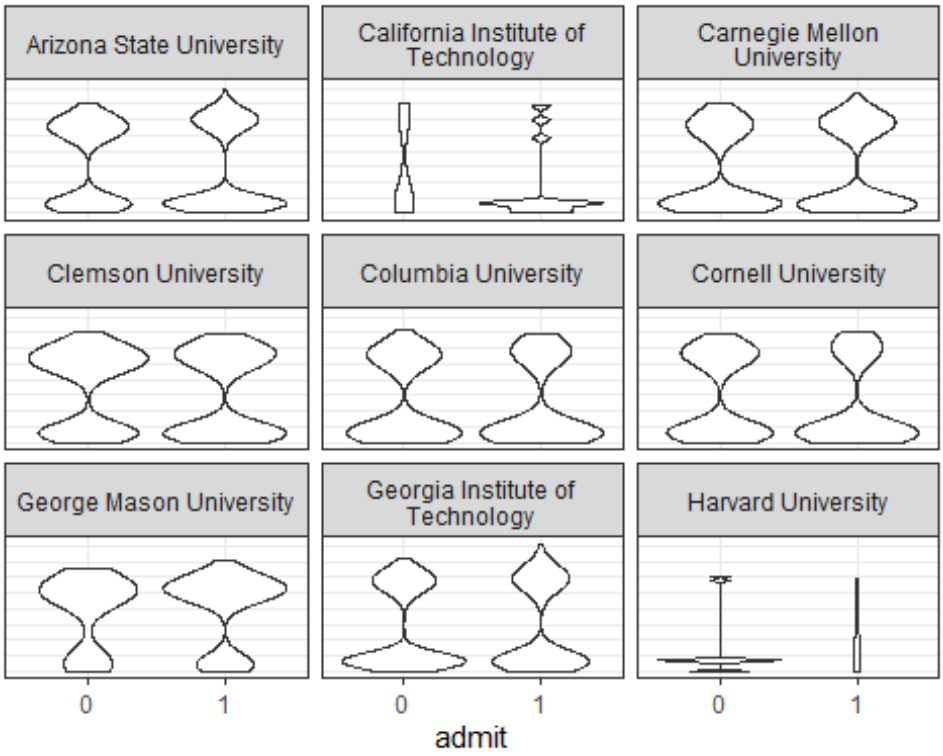

```
ggplot(okvir5, aes(x=admit,y=cgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```




```
ggplot(okvir6, aes(x=admit,y=cgpa)) +
geom_boxplot(alpha=1) + theme_bw()  + facet_wrap(~ univName, nrow = 4,labeller
= label_wrap_gen())
```



```
ggplot(okvir6, aes(x=admit,y=cgpa)) +
geom_violin(alpha=1) +
theme_bw() + ylab(NULL)+theme(axis.text.y = element_blank(),axis.ticks.y =
element_blank()) + facet_wrap(~ univName, nrow = 4,labeller =
label_wrap_gen())
```

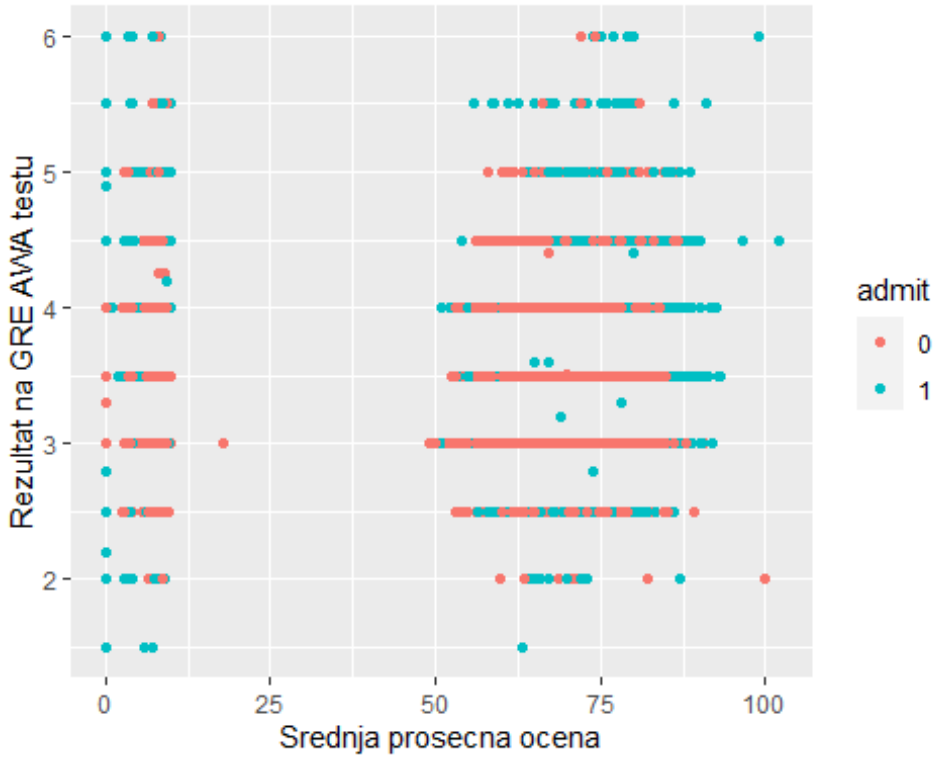



```
#K-S test normalnosti
novi_univerziteti %>% group_by(admit) %>%
summarise(izlaz = list(ks.test(cgpa, "pnorm", mean=mean(cgpa, na.rm = T),
sd=sd(cgpa, na.rm = T)) %>% tidy), .groups = 'drop') %>% unnest(c(izlaz))

## Warning in ks.test(cgpa, "pnorm", mean = mean(cgpa, na.rm = T), sd =
sd(cgpa, :
## ties should not be present for the Kolmogorov-Smirnov test

## Warning in ks.test(cgpa, "pnorm", mean = mean(cgpa, na.rm = T), sd =
sd(cgpa, :
## ties should not be present for the Kolmogorov-Smirnov test

## # A tibble: 2 x 5
##   admit statistic p.value method                                alternative
##   <fct>     <dbl>   <dbl> <chr>                                 <chr>
## 1 0         0.317       0 One-sample Kolmogorov-Smirnov test two-sided
## 2 1         0.320       0 One-sample Kolmogorov-Smirnov test two-sided
```

Na osnovu grafika iznad vidimo da cgpa utiče na to da li je osoba primljena na fakultet ili nije, ali zavisi od fakulteta. Testiranjem normalnosti Kolmogorov-Smirnov testom pokazano je da ne postoji normalnost unutar obe grupe obeležja ($p = 0.0 < \alpha=0.05$, $p = 0.0 < \alpha=0.05$).

cgpa, greA i admit

```
ggplot(novi_univerziteti, aes(x = cgpa, y=greA)) +
 geom_point(aes(colour=admit)) +
 xlab("Srednja prosečna ocena") +
 ylab("Rezultat na GRE AWA testu")
```

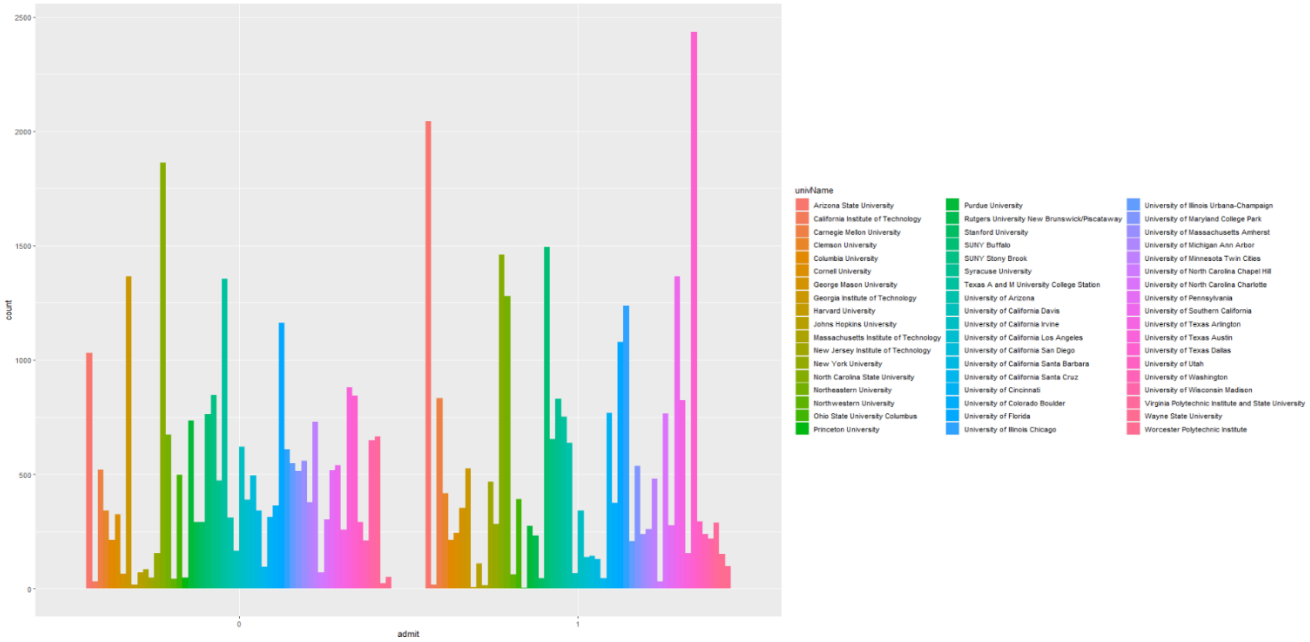


Na osnovu grafika iznad vidimo da ukoliko je student imao veću prosečnu ocenu i ukoliko je imao bolji rezultat na greA testu onda je veća verovatnoća da je primljen na fakultet.

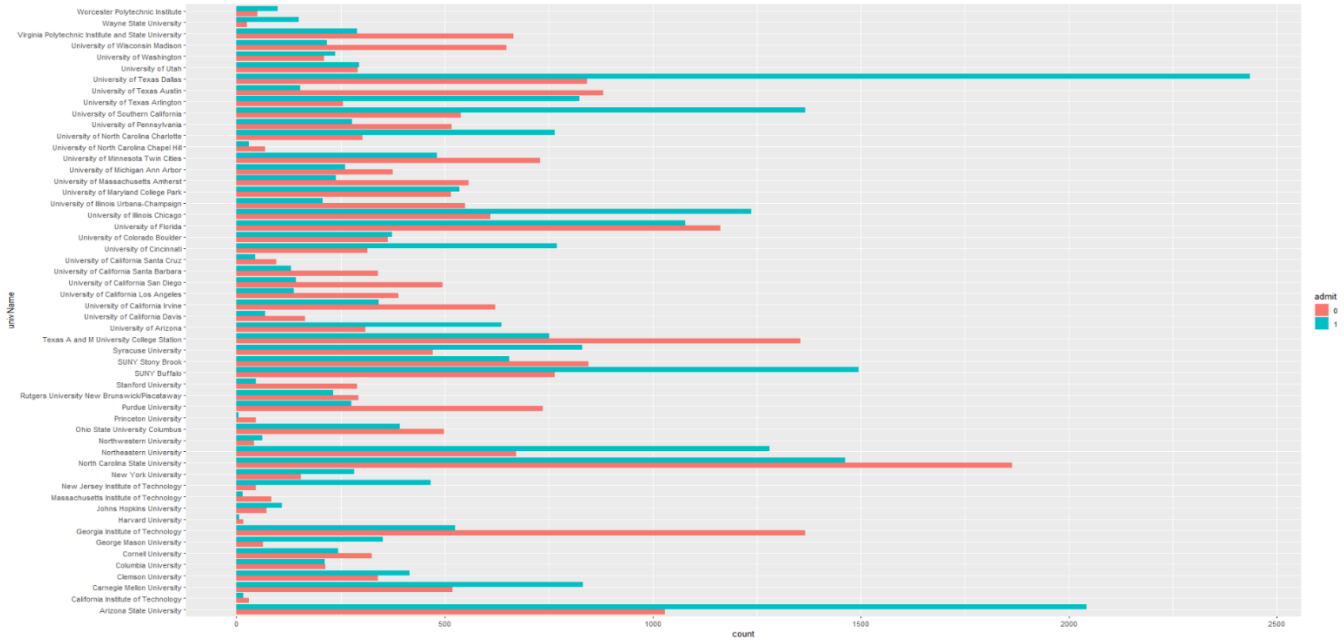
Analiza kategorijskih podataka naspram kategorijskih

univName i admit

```
plt1 <- ggplot(data = novi_univerziteti)+ geom_bar(mapping = aes(x = admit,
fill = univName), position = "dodge")
plt1
```



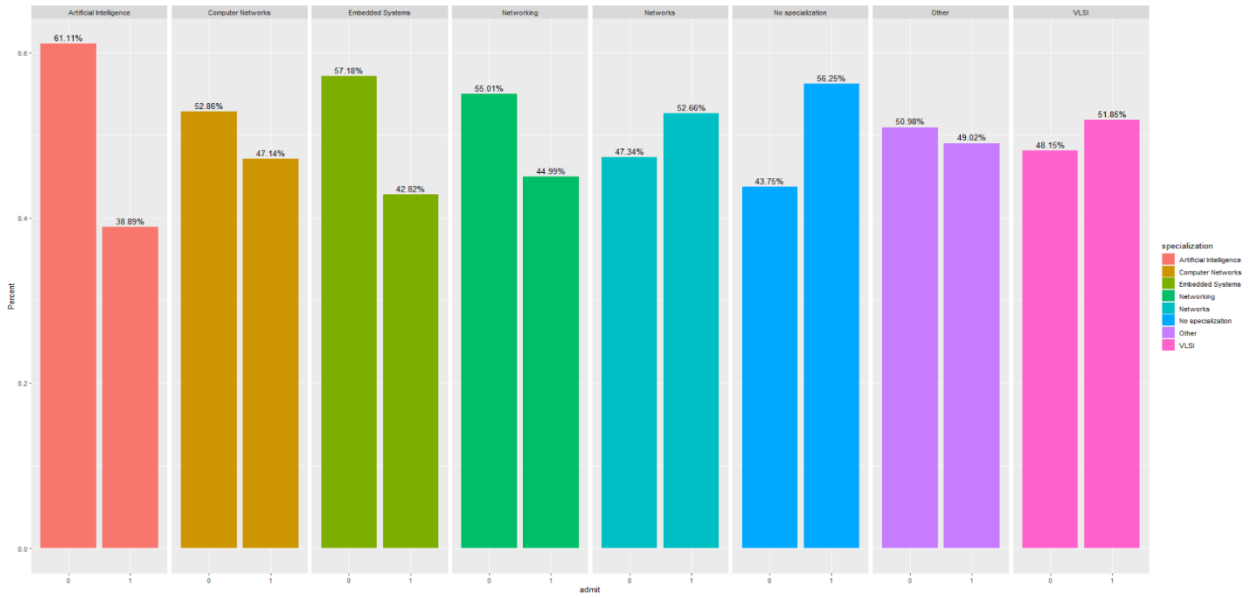

```
plt2 <- ggplot(data = novi_univerziteti)+ geom_bar(mapping = aes(y = univName, fill = admit), position = "dodge")
plt2
```



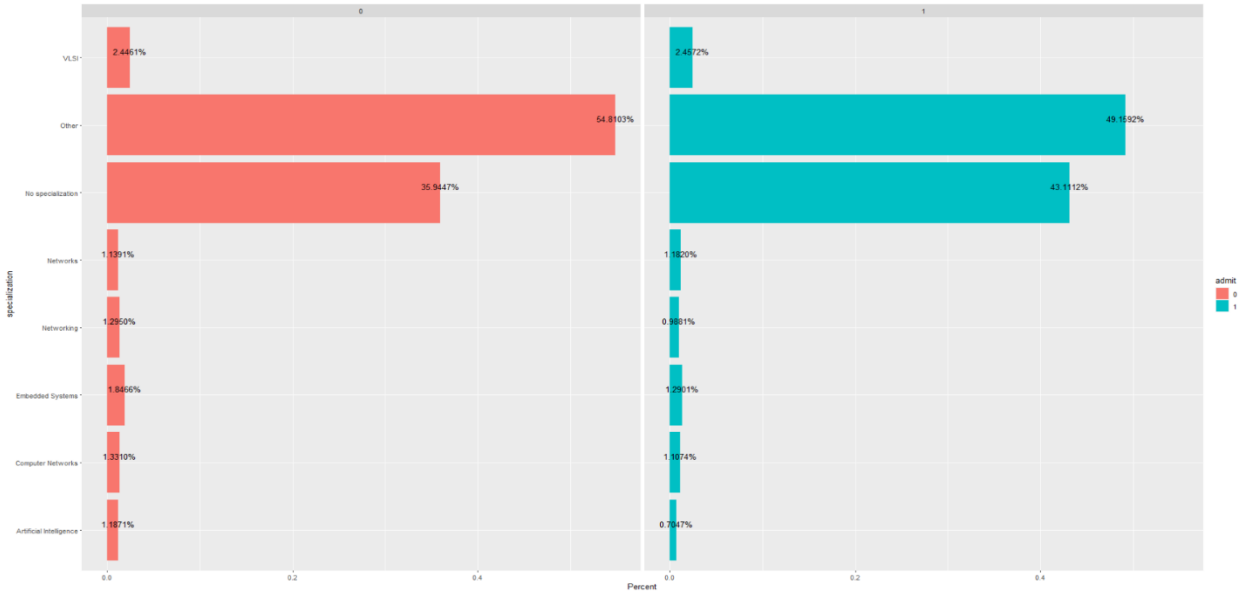
Na osnovu grafika iznad vidimo da je najveći broj kandidata primljen na fakultete *Arizona State University* i *University of Texas Dallas*, dok najveći broj kandidata nije uspelo da se upiše na fakultetima *North Carolina State University* i *Georgia Institute of Technology*. Na drugom grafiku vidimo da je na većini fakulteta većina kandidata koji su konkurisali uspeli su i da se upišu na fakultet, ali postoji manji broj fakulteta gde je zainteresovanost bila velika i većina kandidata nije uspelo da se upiše na željeni fakultet.

specialization i admit

```
plt1 = ggplot(novi_univerziteti, aes(x = admit, group = specialization)) +
  geom_bar(aes(y = ..prop.., fill = specialization), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..), y = ..prop..), stat =
"count", vjust = -.5) +
  labs(y = "Percent", fill = "specialization") +
  facet_grid(~specialization)
plt1
```



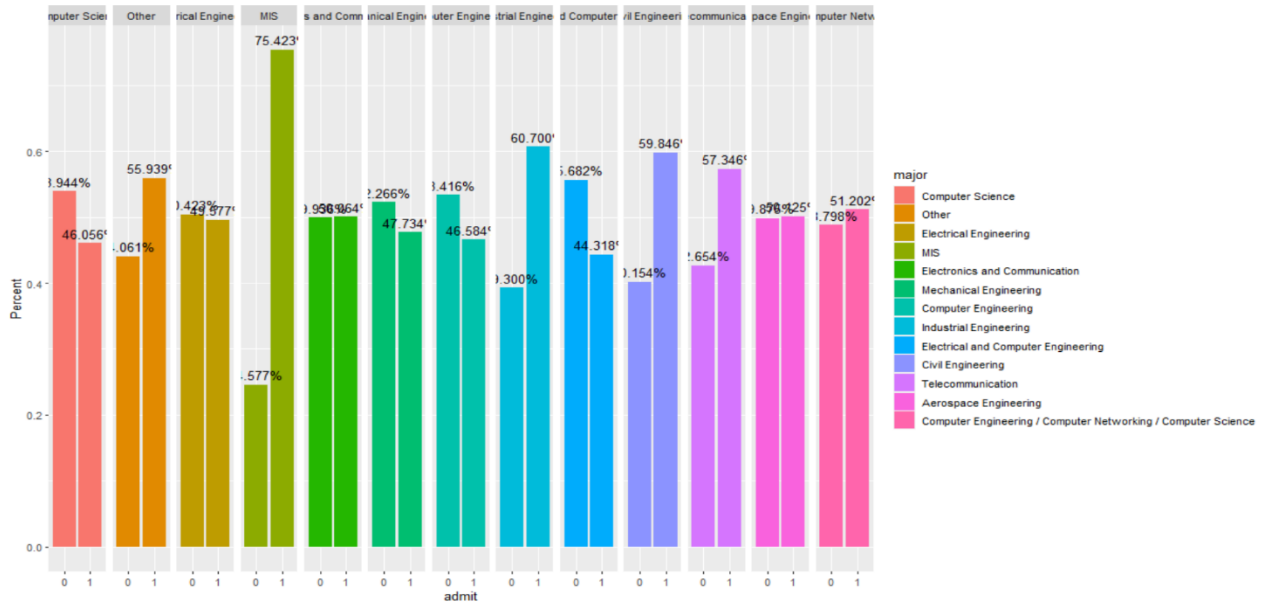

```
plt2 = ggplot(novi_univerziteti, aes(y = specialization, group = admit)) +
  geom_bar(aes(x = ..prop.., fill = admit), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..), x = ..prop..), stat =
"count", vjust = -.5) +
  labs(x = "Percent", fill = "admit") +
  facet_grid(~admit)
plt2
```



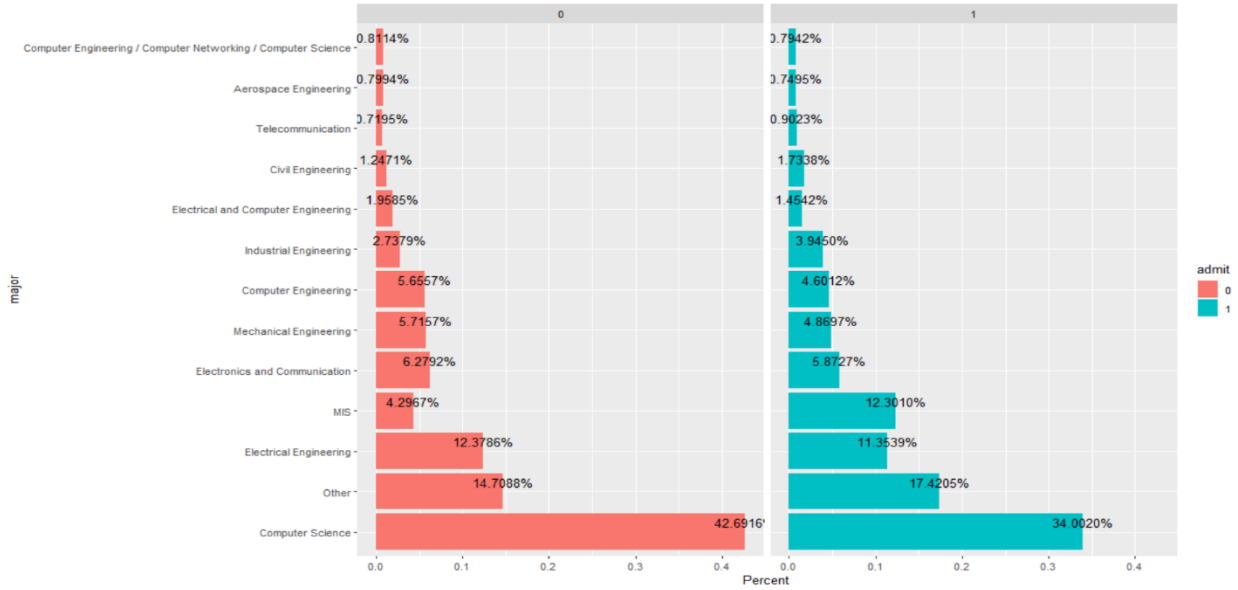
Zanimljiva stvar koju možemo primetiti jeste da đaci koji nemaju specijalizaciju su dosta više primljeni u odnosu na druge đake koji su specijalizovani u neku posebnu oblast.

major i admit

```
plt1 = ggplot(novi_univerziteti, aes(x = admit, group = major)) +
  geom_bar(aes(y = ..prop.., fill = major), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..), y = ..prop..), stat =
"count", vjust = -.5) +
  labs(y = "Percent", fill = "major") +
  facet_grid(~major)
plt1
```



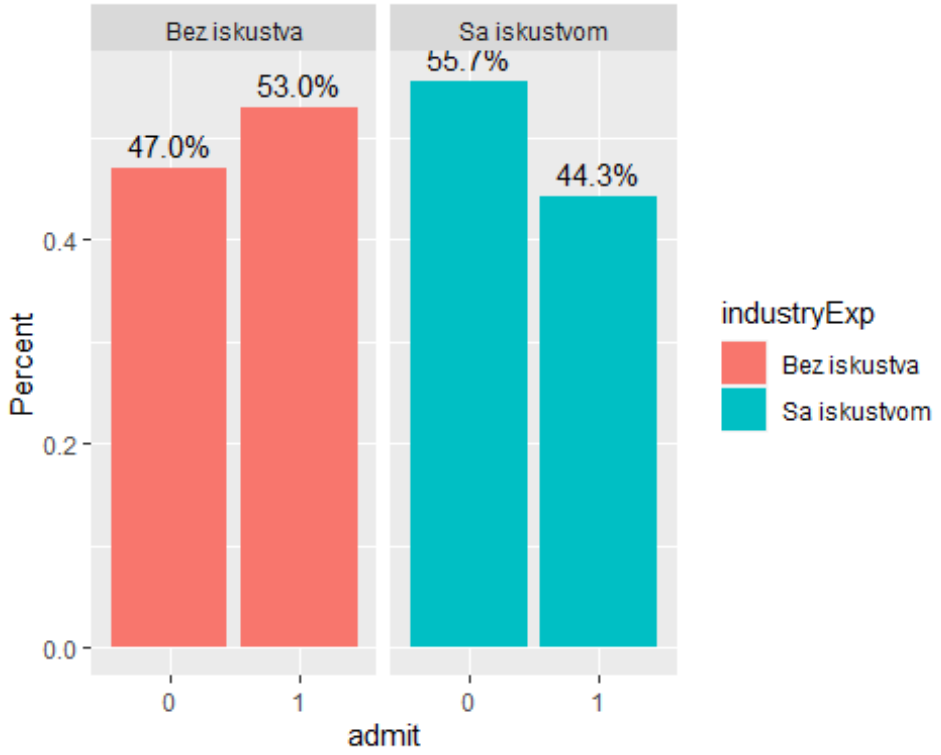
```
plt2 = ggplot(novi_univerziteti, aes(y = major, group = admit)) +
  geom_bar(aes(x = ..prop.., fill = admit), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..), x = ..prop..), stat =
"count", vjust = -.5) +
  labs(x = "Percent", fill = "admit") +
  facet_grid(~admit)
plt2
```



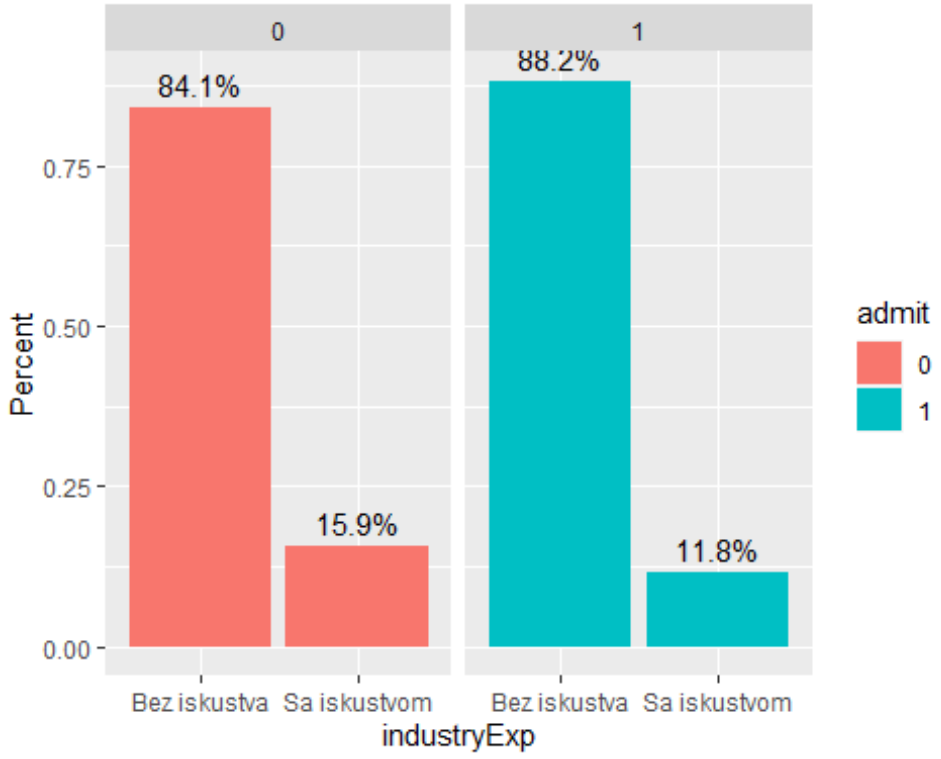
Na grafiku vidimo da je najviše osoba bilo zainteresovano za smer *Computer Science* što je i logično jer se povećava potreba za upotrebom novih tehnologija i tržište rada je veoma veliko za buduće studente. Zbog velike zainteresovanosti za oblast *Computer Science* većina kandidata nisu uspeli da upadnu na fakultet i za sva zanimanja vezana za računare većina kandidata nije uspela da upiše željeni fakultet. Kandidati koji su konkurisali za smer *MIS*(Management Information Systems), *Industrial Engineering* i kandidati sa ostalih smerova, većina njih je upisala željeni fakultet.

industryExp i admit

```
plt1 = ggplot(novi_univerziteti, aes(x = admit, group = industryExp)) +
  geom_bar(aes(y = ..prop.., fill = industryExp), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..), y = ..prop..), stat =
"count", vjust = -.5) +
  labs(y = "Percent", fill = "industryExp") +
  facet_grid(~industryExp)
plt1
```

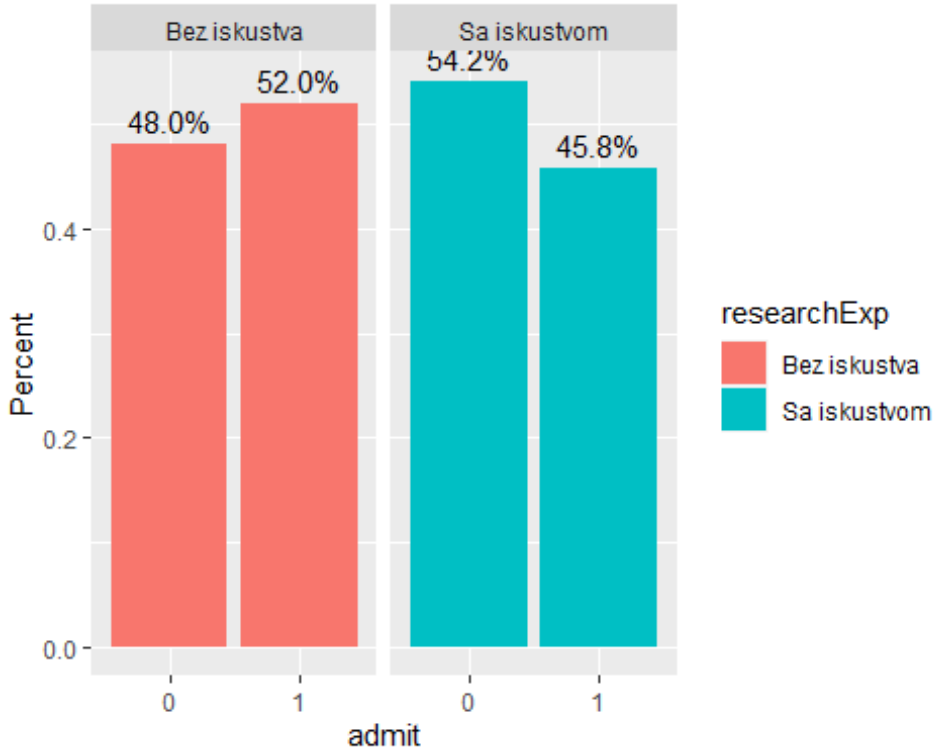



```
plt2 = ggplot(novi_univerziteti, aes(x = industryExp, group = admit)) +
  geom_bar(aes(y = ..prop.., fill = admit), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..), y = ..prop..), stat =
"count", vjust = -.5) +
  labs(y = "Percent", fill = "admit") +
  facet_grid(~admit)
plt2
```

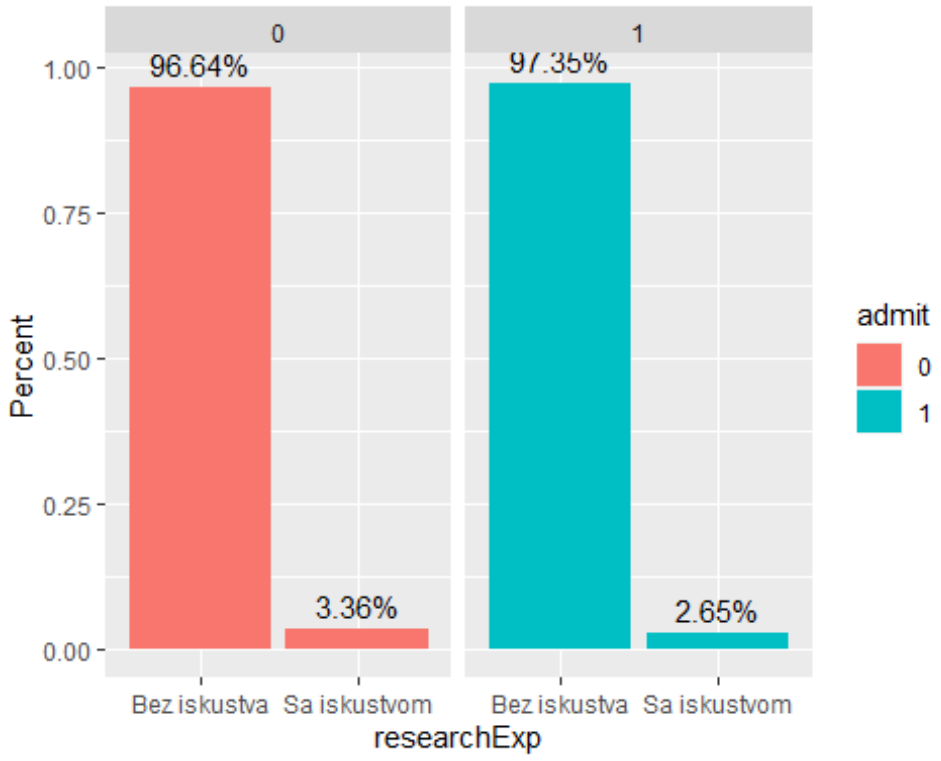


ResearchExp i admit

```
plt1 = ggplot(novi_univerziteti, aes(x = admit, group = researchExp)) +
  geom_bar(aes(y = ..prop.., fill = researchExp), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..), y = ..prop..), stat =
"count", vjust = -.5) +
  labs(y = "Percent", fill = "researchExp") +
  facet_grid(~researchExp)
plt1
```

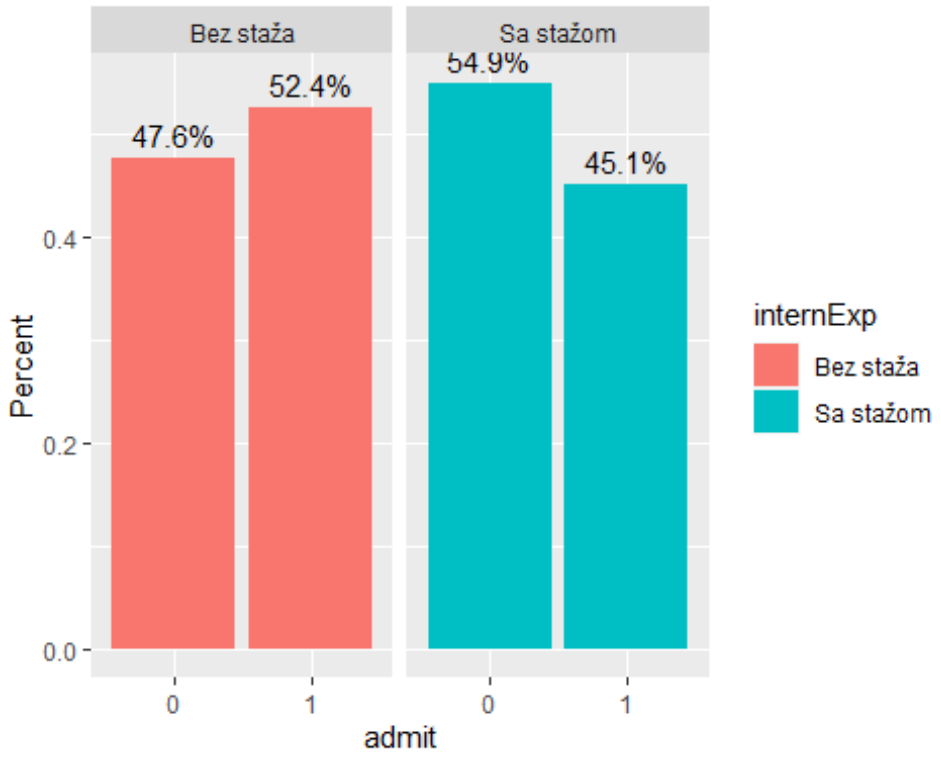



```
plt2 = ggplot(novi_univerziteti, aes(x = researchExp, group = admit)) +
  geom_bar(aes(y = ..prop.., fill = admit), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..), y = ..prop..), stat =
"count", vjust = -.5) +
  labs(y = "Percent", fill = "admit") +
  facet_grid(~admit)
plt2
```

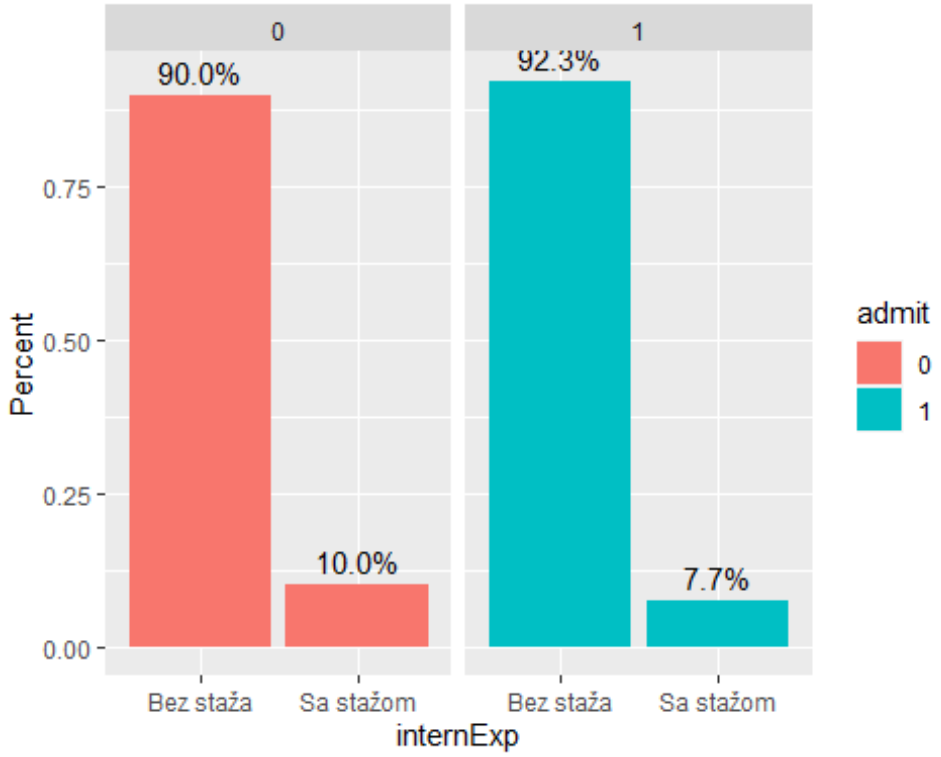


internExp i admit

```
plt1 = ggplot(novi_univerziteti, aes(x = admit, group = internExp)) +
  geom_bar(aes(y = ..prop.., fill = internExp), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..), y = ..prop..), stat =
"count", vjust = -.5) +
  labs(y = "Percent", fill = "internExp") +
  facet_grid(~internExp)
plt1
```




```
plt2 = ggplot(novi_univerziteti, aes(x = internExp, group = admit)) +
  geom_bar(aes(y = ..prop.., fill = admit), stat = "count") +
  geom_text(aes(label = scales::percent(..prop..), y = ..prop..), stat =
"count", vjust = -.5) +
  labs(y = "Percent", fill = "admit") +
  facet_grid(~admit)
plt2
```



Na osnovu prethodnih grafika vidimo da većina kandidata koji su imali bilo kakvo iskustvo nisu uspeli da upadnu na fakultet, dok je većina kandidata bez bilo kakvog iskustva uspela da upadne na fakultet i time zaključujemo da iskustvo kandidata nije bitno pri upisu fakulteta.

Podela na trening i test skupove

Delimo podatke na *Train* i *Test* kao pripremu za klasifikacione modele. U ovom slučaju će *Train* podataka bili 75% iz čitavog skupa.

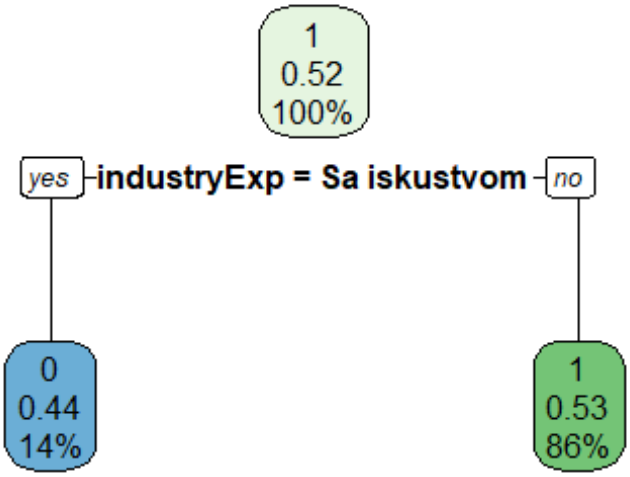
```
smp_siz = floor(0.75*nrow(novi_univerziteti))
smp_siz

## [1] 37977

set.seed(123)
# na slucajan nacin uzimamo uzorak 75% rednih brojeva redova
train_ind = sample(seq_len(nrow(novi_univerziteti)),size = smp_siz)
# kreiramo train sa rednim brojevima smestenim u train_ind
train = novi_univerziteti[train_ind,]
# kreiramo test sa preostalim podacima
test = novi_univerziteti[-train_ind,]
```

Stablo odlučivanja

Stablo odlučivanja je dijagram oblika stabla koji se koristi za utvrđivanje toka akcija u procesu odlučivanja. Svaka grana predstavlja moguće odluke, pojave ili reakcije. Stabla odlučivanja su najjednostavniji mehanizam za klasifikaciju i regresiju. Na osnovu stabla odlučivanja mogu se generisati pravila, koja ljudi mogu da razumeju i koja mogu biti upotrebljena za formiranje baze znanja.

Kreiranje modela

```
tree_model = rpart(admit ~ ., data=subset(train,select=c(admit, greQ, greV,
greA, toeflScore, internExp, industryExp, researchExp)),method="class")

rpart.plot(tree_model)
```




```
printcp(tree_model)

##
## Classification tree:
## rpart(formula = admit ~ ., data = subset(train, select = c(admit,
##     greQ, greV, greA, toeflScore, internExp, industryExp, researchExp)),
##     method = "class")
##
## Variables actually used in tree construction:
## [1] industryExp
```

```
##
## Root node error: 18347/37977 = 0.48311
##
## n= 37977
##
##         CP nsplit rel error  xerror      xstd
## 1 0.036409      0   1.00000 1.00000 0.0053078
## 2 0.010000      1   0.96359 0.96359 0.0052982
```

Možemo primetiti da je naš kreirani model koristio samo 1 obeležje pri konstrukciji stabla odlučivanja, u pitanju je obeležje **industryExp** . Vidimo da je greška osnovnog čvora velika, malo ispod 50%, kao i rel *error* i *xerror* greske.

```
predTree = predict(tree_model,test, type="class")

conf = table(predTree, test$admit, dnn =c("Prediction","Action"))

conf
```

```
##           Action
## Prediction    0    1
##          0  941  820
##          1 5124 5775
```

Metrike

Metrike koje ćemo pratiti u narednom radu jesu metrike preciznosti, odziva, kao i F1 score koja nam ujedno pokazuje više različitih metrika.

Preciznost:

```
(precision = diag(conf) / sum(conf))
```

```
##          0          1
## 0.07432859 0.45616114
```

Odziv:

```
(recall = (diag(conf) / colSums(conf)))
```

```
##         0         1
## 0.1551525 0.8756634
```

F1 score:

```
(F1 = 2*precision*recall/(precision+recall))
```

```
##         0         1
## 0.1005073 0.5998442
```

Preciznost u odnosu na celo obeležje.

```
sum(diag(conf)) / sum(conf)
```

```
## [1] 0.5304897
```

Kao što možemo videti, preciznost nije na zadovoljavajućem nivou.

Random Forest

Random forest je metoda mašinskog učenja za klasifikaciju, regresiju ili druge modele tako što za vreme obuke konstruiše mnoštvo stabala odlučivanja. Za izlaznu vrednost modela klasifikacije random forest bira klasu koja je izabrana od vecine stabala.Za izlaznu vrednost regresionog modela vraća medijanu ili srednju vrednost predikcija svih stabala. Da bismo kreirali model potrebno da je sva kategorijska obeležja imaju najviše 53 različite kategorije, ali naše kategorijsko obeležje *univName* koje je veoma bitno za naše modele na osnovu detaljne prethodne analize ima 54 različite kategorije, tako da smo rešili da izbacimo jednu kategoriju koja se najmanje pojavljuje. Zbog velikog broja podataka u datom skupu neće biti problem ako uklonimo sve redove koji sadrže ovu kategoriju, pogotovo kada uzmemo u obzir da nekoliko kategorija imaju nesto manje i od 100 uzoraka, što predstavlja dosta manje od 1% ukupnog broja uzoraka.

Ponovno ćemo podeliti ceo okvir podataka na train i test, po istom *set.seed-u* da bismo bili sigurni da da ćemo dobiti iste rezultate za randomizaciju, ali ovog puta samo bez uzoraka koji predstavljaju kategoriju sa najmanje uzoraka, a u pitanju je univerzizet **California Institute of Technology**.

```
xtabs(~novi_univerziteti$univName)
```

```
## novi_univerziteti$univName
##                             Arizona State University
##                                                 3003
##                   California Institute of Technology
##                                                   46
##                           Carnegie Mellon University
##                                                 1343
##                                  Clemson University
##                                                  746
##                                Columbia University
##                                                  412
##                                 Cornell University
##                                                  559
##                              George Mason University
##                                                  405
##                      Georgia Institute of Technology
##                                                 1871
##                                   Harvard University
##                                                   20
##                              Johns Hopkins University
##                                                  167
##                 Massachusetts Institute of Technology
##                                                   94
##                     New Jersey Institute of Technology
```

```
##                                                 486
##                                 New York University
##                                                 429
##                     North Carolina State University
##                                                3261
##                             Northeastern University
##                                                1913
##                             Northwestern University
##                                                 100
##                      Ohio State University Columbus
##                                                 874
##                                 Princeton University
##                                                  51
##                                   Purdue University
##                                                 996
##         Rutgers University New Brunswick/Piscataway
##                                                 504
##                                 Stanford University
##                                                 333
##                                       SUNY Buffalo
##                                                2211
##                                    SUNY Stony Brook
##                                                1438
##                                 Syracuse University
##                                                1237
##            Texas A and M University College Station
##                                                2046
##                               University of Arizona
##                                                 923
##                      University of California Davis
##                                                 220
##                     University of California Irvine
##                                                 947
##                 University of California Los Angeles
##                                                 522
##                  University of California San Diego
##                                                 626
##               University of California Santa Barbara
##                                                 458
##                  University of California Santa Cruz
##                                                 131
##                             University of Cincinnati
##                                                1057
##                      University of Colorado Boulder
##                                                 715
##                                University of Florida
##                                                2193
##                      University of Illinois Chicago
##                                                1802
##              University of Illinois Urbana-Champaign
```

```
##                                                 745
##                 University of Maryland College Park
##                                                1020
##                 University of Massachusetts Amherst
##                                                 760
##                    University of Michigan Ann Arbor
##                                                 626
##                 University of Minnesota Twin Cities
##                                                1186
##            University of North Carolina Chapel Hill
##                                                  93
##              University of North Carolina Charlotte
##                                                1050
##                          University of Pennsylvania
##                                                 771
##                   University of Southern California
##                                                1852
##                       University of Texas Arlington
##                                                1054
##                          University of Texas Austin
##                                                1008
##                          University of Texas Dallas
##                                                3236
##                                  University of Utah
##                                                 562
##                            University of Washington
##                                                 443
##                      University of Wisconsin Madison
##                                                 858
## Virginia Polytechnic Institute and State University
##                                                 918
##                               Wayne State University
##                                                 169
##                      Worcester Polytechnic Institute
##                                                 147
```

```
novi_univerziteti = novi_univerziteti[novi_univerziteti$univName !=
"California Institute of Technology",]

smp_siz = floor(0.75*nrow(novi_univerziteti))
smp_siz
```

```
## [1] 37943
```

```
set.seed(123)
# na slucajan nacin uzimamo uzorak 75% rednih brojeva redova
train_ind = sample(seq_len(nrow(novi_univerziteti)),size = smp_siz)
# kreiramo train sa rednim brojevima smestenim u train_ind
train = novi_univerziteti[train_ind,]
# kreiramo test sa preostalim podacima
test = novi_univerziteti[-train_ind,]
```

Nakon što smo izbrisali sve uzorke te kategorije, moramo i da sklonimo samu kategoriju sa liste postojećih kategorija.

```
train$univName = droplevels(train$univName)
test$univName = droplevels(test$univName)
```

Kreiranje modela

```
rf <- randomForest(admit ~ greA + greV + greQ + major + cgpa + toeflScore +
univName, data = train)
rf.pred <- predict(rf,newdata = test)

rf

##
## Call:
##  randomForest(formula = admit ~ greA + greV + greQ + major + cgpa +
toeflScore + univName, data = train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 26.36%
## Confusion matrix:
##       0     1 class.error
## 0 12490  5825   0.3180453
## 1  4178 15450   0.2128592
```

Primećujemo da je greška znatno manja i da je ovaj model znatno bolji.

Kreiraćemo i konfuzionu matricu na osnovu testnog skupa:

```
conf = table(rf.pred, test$admit, dnn =c("Prediction","Action"))

conf

##           Action
## Prediction    0    1
##          0 4134 1339
##          1 1933 5242
```

Metrike

Preciznost:

```
(precision = diag(conf) / sum(conf))

##         0         1
## 0.3268501 0.4144529
```

Odziv:

```
(recall = (diag(conf) / colSums(conf)))
```

```
##         0         1
## 0.6813911 0.7965355
```

F1 score:

```
(F1 = 2*precision*recall/(precision+recall))
```

```
##         0         1
## 0.4417847 0.5452182
```

Preciznost u odnosu na celo obeležje:

```
sum(diag(conf)) / sum(conf)
```

```
## [1] 0.741303
```

Logistička regresija

Logistička regresija je nadgledani algoritam mašinskog učenja koji ispunjava zadatke binarne klasifikacije predviđanjem verovatnoće ishoda, događaja ili posmatranja. Model daje binarni ishod ograničen na dva moguća ishoda: da/ne, 0/1 ili tačno/netačno.

S obzirom da mi rešavamo jedan od navedenih problema, da li je đak primljen na fakultet ili nije, odnosno da li je vrednost 0 ili 1, moći ćemo da koristimo binarnu logističku regresiju za kreiranje modela.

Kreiranje modela

```
logistic_model <- glm(admit ~ univName + major + industryExp + cgpa + greV +
greA + toeflScore,
                      data = train,
                      family = "binomial")

summary(logistic_model)
```

```
##
## Call:
## glm(formula = admit ~ univName + major + industryExp + cgpa +
##     greV + greA + toeflScore, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5537  -1.0040   0.4866   0.9835   2.6082
##
## Coefficients:
##
Estimate
```

```
## (Intercept)                                                    -
2.4209816
## univNameCarnegie Mellon University                             -
0.6368118
## univNameClemson University                                     -
0.5907468
## univNameColumbia University                                    -
1.0085751
## univNameCornell University                                     -
1.2914083
## univNameGeorge Mason University
1.2329013
## univNameGeorgia Institute of Technology                        -
1.9736164
## univNameHarvard University                                     -
2.0168312
## univNameJohns Hopkins University                               -
0.5470842
## univNameMassachusetts Institute of Technology                  -
3.1155198
## univNameNew Jersey Institute of Technology
1.8330369
## univNameNew York University                                    -
0.2687596
## univNameNorth Carolina State University                        -
1.0246984
## univNameNortheastern University
0.0621101
## univNameNorthwestern University                                -
0.7396348
## univNameOhio State University Columbus                         -
1.0826188
## univNamePrinceton University                                   -
3.5891788
## univNamePurdue University                                      -
2.1820307
## univNameRutgers University New Brunswick/Piscataway            -
1.0187007
## univNameStanford University                                    -
3.1231803
## univNameSUNY Buffalo                                           -
0.0729910
## univNameSUNY Stony Brook                                       -
0.8277732
## univNameSyracuse University                                    -
0.2718378
## univNameTexas A and M University College Station               -
1.7990833
## univNameUniversity of Arizona                                  -
0.4445988
```

```
## univNameUniversity of California Davis                              -
1.7758636
## univNameUniversity of California Irvine                             -
1.4508004
## univNameUniversity of California Los Angeles                        -
1.9388123
## univNameUniversity of California San Diego                          -
2.1399768
## univNameUniversity of California Santa Barbara                      -
1.9149290
## univNameUniversity of California Santa Cruz                         -
1.7247727
## univNameUniversity of Cincinnati                                    -
0.1647641
## univNameUniversity of Colorado Boulder                              -
0.8227432
## univNameUniversity of Florida                                       -
0.8724876
## univNameUniversity of Illinois Chicago                              -
0.1731187
## univNameUniversity of Illinois Urbana-Champaign                     -
2.1125078
## univNameUniversity of Maryland College Park                         -
1.0074167
## univNameUniversity of Massachusetts Amherst                         -
1.6077151
## univNameUniversity of Michigan Ann Arbor                            -
1.5897667
## univNameUniversity of Minnesota Twin Cities                         -
1.2573130
## univNameUniversity of North Carolina Chapel Hill                    -
1.8644166
## univNameUniversity of North Carolina Charlotte
0.4033182
## univNameUniversity of Pennsylvania                                  -
1.5464319
## univNameUniversity of Southern California
0.1550507
## univNameUniversity of Texas Arlington
0.6477192
## univNameUniversity of Texas Austin                                  -
2.8999586
## univNameUniversity of Texas Dallas
0.4572441
## univNameUniversity of Utah                                          -
0.7564550
## univNameUniversity of Washington                                    -
1.1607986
## univNameUniversity of Wisconsin Madison                             -
2.2136380
```

```
## univNameVirginia Polytechnic Institute and State University        -
1.8527342
## univNameWayne State University
1.2883118
## univNameWorcester Polytechnic Institute                            -
0.2093789
## majorOther
0.5167175
## majorElectrical Engineering
0.2758930
## majorMIS
1.1828373
## majorElectronics and Communication
0.3035763
## majorMechanical Engineering
0.3344273
## majorComputer Engineering
0.1435364
## majorIndustrial Engineering
0.9184151
## majorElectrical and Computer Engineering
0.0960563
## majorCivil Engineering
1.2625027
## majorTelecommunication
0.2028336
## majorAerospace Engineering
0.3774552
## majorComputer Engineering / Computer Networking / Computer Science
0.1777009
## industryExpSa iskustvom                                            -
0.3832064
## cgpa                                                               -
0.0038675
## greV
0.0006565
## greA
0.3668763
## toeflScore
0.0156640
##                                                                   Std.
Error
## (Intercept)
0.1396330
## univNameCarnegie Mellon University
0.0841960
## univNameClemson University
0.0989436
## univNameColumbia University
0.1252221
```

```
## univNameCornell University
0.1130317
## univNameGeorge Mason University
0.1727326
## univNameGeorgia Institute of Technology
0.0773719
## univNameHarvard University
0.5106972
## univNameJohns Hopkins University
0.1920754
## univNameMassachusetts Institute of Technology
0.3501907
## univNameNew Jersey Institute of Technology
0.1907496
## univNameNew York University
0.1326369
## univNameNorth Carolina State University
0.0624356
## univNameNortheastern University
0.0742123
## univNameNorthwestern University
0.2471453
## univNameOhio State University Columbus
0.0920824
## univNamePrinceton University
0.5316953
## univNamePurdue University
0.0978098
## univNameRutgers University New Brunswick/Piscataway
0.1156426
## univNameStanford University
0.1896848
## univNameSUNY Buffalo
0.0711680
## univNameSUNY Stony Brook
0.0780418
## univNameSyracuse University
0.0853006
## univNameTexas A and M University College Station
0.0735749
## univNameUniversity of Arizona
0.0989375
## univNameUniversity of California Davis
0.1791558
## univNameUniversity of California Irvine
0.0937685
## univNameUniversity of California Los Angeles
0.1241043
## univNameUniversity of California San Diego
0.1219989
```

```
## univNameUniversity of California Santa Barbara
0.1336101
## univNameUniversity of California Santa Cruz
0.2404870
## univNameUniversity of Cincinnati
0.0964000
## univNameUniversity of Colorado Boulder
0.1016890
## univNameUniversity of Florida
0.0687810
## univNameUniversity of Illinois Chicago
0.0767708
## univNameUniversity of Illinois Urbana-Champaign
0.1091913
## univNameUniversity of Maryland College Park
0.0890143
## univNameUniversity of Massachusetts Amherst
0.1032475
## univNameUniversity of Michigan Ann Arbor
0.1070963
## univNameUniversity of Minnesota Twin Cities
0.0840291
## univNameUniversity of North Carolina Chapel Hill
0.2843368
## univNameUniversity of North Carolina Charlotte
0.0925618
## univNameUniversity of Pennsylvania
0.1004210
## univNameUniversity of Southern California
0.0755127
## univNameUniversity of Texas Arlington
0.0972819
## univNameUniversity of Texas Austin
0.1141244
## univNameUniversity of Texas Dallas
0.0670877
## univNameUniversity of Utah
0.1101498
## univNameUniversity of Washington
0.1234894
## univNameUniversity of Wisconsin Madison
0.1049221
## univNameVirginia Polytechnic Institute and State University
0.0966393
## univNameWayne State University
0.2627968
## univNameWorcester Polytechnic Institute
0.2177981
## majorOther
0.0341936
```

```
## majorElectrical Engineering
0.0383940
## majorMIS
0.0504786
## majorElectronics and Communication
0.0497150
## majorMechanical Engineering
0.0531456
## majorComputer Engineering
0.0532318
## majorIndustrial Engineering
0.0660589
## majorElectrical and Computer Engineering
0.0889506
## majorCivil Engineering
0.0945646
## majorTelecommunication
0.1315766
## majorAerospace Engineering
0.1291945
## majorComputer Engineering / Computer Networking / Computer Science
0.1240633
## industryExpSa iskustvom
0.0347021
## cgpa
0.0003599
## greV
0.0000583
## greA
0.0227209
## toeflScore
0.0012889
##                                                                 z value
## (Intercept)                                                    -17.338
## univNameCarnegie Mellon University                              -7.563
## univNameClemson University                                      -5.971
## univNameColumbia University                                     -8.054
## univNameCornell University                                     -11.425
## univNameGeorge Mason University                                  7.138
## univNameGeorgia Institute of Technology                        -25.508
## univNameHarvard University                                      -3.949
## univNameJohns Hopkins University                                -2.848
## univNameMassachusetts Institute of Technology                   -8.897
## univNameNew Jersey Institute of Technology                       9.610
## univNameNew York University                                     -2.026
## univNameNorth Carolina State University                        -16.412
## univNameNortheastern University                                  0.837
## univNameNorthwestern University                                 -2.993
## univNameOhio State University Columbus                         -11.757
## univNamePrinceton University                                    -6.750
```

```
## univNamePurdue University                                          -22.309
## univNameRutgers University New Brunswick/Piscataway                 -8.809
## univNameStanford University                                        -16.465
## univNameSUNY Buffalo                                                -1.026
## univNameSUNY Stony Brook                                           -10.607
## univNameSyracuse University                                         -3.187
## univNameTexas A and M University College Station                   -24.452
## univNameUniversity of Arizona                                       -4.494
## univNameUniversity of California Davis                              -9.912
## univNameUniversity of California Irvine                            -15.472
## univNameUniversity of California Los Angeles                       -15.622
## univNameUniversity of California San Diego                         -17.541
## univNameUniversity of California Santa Barbara                     -14.332
## univNameUniversity of California Santa Cruz                         -7.172
## univNameUniversity of Cincinnati                                    -1.709
## univNameUniversity of Colorado Boulder                              -8.091
## univNameUniversity of Florida                                      -12.685
## univNameUniversity of Illinois Chicago                              -2.255
## univNameUniversity of Illinois Urbana-Champaign                    -19.347
## univNameUniversity of Maryland College Park                        -11.317
## univNameUniversity of Massachusetts Amherst                        -15.571
## univNameUniversity of Michigan Ann Arbor                           -14.844
## univNameUniversity of Minnesota Twin Cities                        -14.963
## univNameUniversity of North Carolina Chapel Hill                    -6.557
## univNameUniversity of North Carolina Charlotte                       4.357
## univNameUniversity of Pennsylvania                                 -15.399
## univNameUniversity of Southern California                            2.053
## univNameUniversity of Texas Arlington                                6.658
## univNameUniversity of Texas Austin                                 -25.411
## univNameUniversity of Texas Dallas                                   6.816
## univNameUniversity of Utah                                          -6.868
## univNameUniversity of Washington                                    -9.400
## univNameUniversity of Wisconsin Madison                            -21.098
## univNameVirginia Polytechnic Institute and State University        -19.172
## univNameWayne State University                                       4.902
## univNameWorcester Polytechnic Institute                             -0.961
## majorOther                                                          15.112
## majorElectrical Engineering                                          7.186
## majorMIS                                                            23.432
## majorElectronics and Communication                                   6.106
## majorMechanical Engineering                                          6.293
## majorComputer Engineering                                            2.696
## majorIndustrial Engineering                                         13.903
## majorElectrical and Computer Engineering                             1.080
## majorCivil Engineering                                              13.351
## majorTelecommunication                                               1.542
## majorAerospace Engineering                                           2.922
## majorComputer Engineering / Computer Networking / Computer Science   1.432
## industryExpSa iskustvom                                            -11.043
## cgpa                                                               -10.745
```

```
## greV                                                    11.261
## greA                                                    16.147
## toeflScore                                              12.153
##                                                        Pr(>|z|)
## (Intercept)                                             < 2e-16
***
## univNameCarnegie Mellon University                     3.93e-14
***
## univNameClemson University                             2.36e-09
***
## univNameColumbia University                            7.99e-16
***
## univNameCornell University                              < 2e-16
***
## univNameGeorge Mason University                        9.50e-13
***
## univNameGeorgia Institute of Technology                 < 2e-16
***
## univNameHarvard University                             7.84e-05
***
## univNameJohns Hopkins University                        0.00440
**
## univNameMassachusetts Institute of Technology           < 2e-16
***
## univNameNew Jersey Institute of Technology              < 2e-16
***
## univNameNew York University                             0.04274
*
## univNameNorth Carolina State University                 < 2e-16
***
## univNameNortheastern University                         0.40263
## univNameNorthwestern University                         0.00277
**
## univNameOhio State University Columbus                  < 2e-16
***
## univNamePrinceton University                           1.47e-11
***
## univNamePurdue University                               < 2e-16
***
## univNameRutgers University New Brunswick/Piscataway     < 2e-16
***
## univNameStanford University                             < 2e-16
***
## univNameSUNY Buffalo                                    0.30507
## univNameSUNY Stony Brook                                < 2e-16
***
## univNameSyracuse University                             0.00144
**
## univNameTexas A and M University College Station        < 2e-16
***
```

```
## univNameUniversity of Arizona                                  7.00e-06
***
## univNameUniversity of California Davis                          < 2e-16
***
## univNameUniversity of California Irvine                         < 2e-16
***
## univNameUniversity of California Los Angeles                    < 2e-16
***
## univNameUniversity of California San Diego                      < 2e-16
***
## univNameUniversity of California Santa Barbara                  < 2e-16
***
## univNameUniversity of California Santa Cruz                    7.39e-13
***
## univNameUniversity of Cincinnati                                0.08742
.
## univNameUniversity of Colorado Boulder                         5.93e-16
***
## univNameUniversity of Florida                                   < 2e-16
***
## univNameUniversity of Illinois Chicago                          0.02413
*
## univNameUniversity of Illinois Urbana-Champaign                 < 2e-16
***
## univNameUniversity of Maryland College Park                     < 2e-16
***
## univNameUniversity of Massachusetts Amherst                     < 2e-16
***
## univNameUniversity of Michigan Ann Arbor                        < 2e-16
***
## univNameUniversity of Minnesota Twin Cities                     < 2e-16
***
## univNameUniversity of North Carolina Chapel Hill               5.49e-11
***
## univNameUniversity of North Carolina Charlotte                 1.32e-05
***
## univNameUniversity of Pennsylvania                              < 2e-16
***
## univNameUniversity of Southern California                       0.04004
*
## univNameUniversity of Texas Arlington                          2.77e-11
***
## univNameUniversity of Texas Austin                              < 2e-16
***
## univNameUniversity of Texas Dallas                             9.39e-12
***
## univNameUniversity of Utah                                     6.53e-12
***
## univNameUniversity of Washington                                < 2e-16
***
```

```
## univNameUniversity of Wisconsin Madison                                  < 2e-16
***
## univNameVirginia Polytechnic Institute and State University          < 2e-16
***
## univNameWayne State University                                        9.47e-07
***
## univNameWorcester Polytechnic Institute                               0.33638
## majorOther                                                            < 2e-16
***
## majorElectrical Engineering                                           6.68e-13
***
## majorMIS                                                              < 2e-16
***
## majorElectronics and Communication                                    1.02e-09
***
## majorMechanical Engineering                                           3.12e-10
***
## majorComputer Engineering                                             0.00701
**
## majorIndustrial Engineering                                           < 2e-16
***
## majorElectrical and Computer Engineering                              0.28019
## majorCivil Engineering                                                < 2e-16
***
## majorTelecommunication                                                0.12318
## majorAerospace Engineering                                            0.00348
**
## majorComputer Engineering / Computer Networking / Computer Science  0.15205
## industryExpSa iskustvom                                               < 2e-16
***
## cgpa                                                                  < 2e-16
***
## greV                                                                  < 2e-16
***
## greA                                                                  < 2e-16
***
## toeflScore                                                            < 2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 52555  on 37942  degrees of freedom
## Residual deviance: 45399  on 37873  degrees of freedom
## AIC: 45539
##
## Number of Fisher Scoring iterations: 4
```

Koristili smo obeležja za koje smo analizom u sekciji **EDA** zaključili da postoji povezanost, i sumarizacijom kreiranog modela vidimo da većina univerziteta su itekako povezani sa obeležjem koji želimo da prediktujemo, zatim i kod obeležja major je slična situacija. Kategorija *sa iskustvom* obeležja industryExp itekako je povezana sa *admit-om*, ali i numerička obeležja cgpa, greV, greA i toeflScore.

Nakon toga ćemo iskoristi kreirani model za prediktovanje nad testnim skupom, i pošto vrednosti mogu biti decimalne i između brojeva 0 i 1, koristićemo prag od 0.5 tako da sve vrednosti manje od vrednosti 0.5 biće okarakterisane kao 0, a veće kao 1.

```
predict_reg <- predict(logistic_model,
                       test, type = "response")

predict_reg <- ifelse(predict_reg >0.5, 1, 0)
```

Zatim ćemo kreirati matricu konfuzije.

```
(conf = table(test$admit, predict_reg, dnn =c("Prediction","Action")))
```

```
##            Action
## Prediction    0    1
##          0 3905 2162
##          1 1881 4700
```

Na prvi pogled vidimo da su rezultati sasvim solidni, ali to ćemo tačnije utvrditi metrikama.

Metrike

Preciznost:

```
(precision = diag(conf) / sum(conf))
```

```
##         0         1
## 0.3087445 0.3716003
```

Odziv:

```
(recall = (diag(conf) / colSums(conf)))
```

```
##         0         1
## 0.6749049 0.6849315
```

F1 score:

```
(F1 = 2*precision*recall/(precision+recall))
```

```
##         0         1
## 0.4236736 0.4818042
```

Preciznost u odnosu na celo obeležje:

```
sum(diag(conf)) / sum(conf)
```

```
## [1] 0.6803447
```

Zaključak

Stablo odluke kreira loše modele sa velikim greškama, i zbog toga se treba izbegavati pri kreiranju modela. Metoda Random Forest najbolje opisuje podatke, međutim potrebno je previše vremena za kreiranje modela zato što Random Forest pri kreiranju najboljeg mogućeg modela kreira 500 različitih stabala, i zbog toga je sam proces veoma dug. Logistička regresija u odnosu na stablo odluke daje znatno bolje rezultate, i kada je potreba brzina za kreiranje modela, logistička regresija itekako može dobro poslužiti. Može se reći kada se uporedi odnos brzine i kvaliteta, logistička regresija daje najbolje rezultate, tako da na osnovu potreba, može se birati neki od ova 2 načina.

Smatramo da smo ovim dokumentom ispunili 3 glavna cilja kojim smo se vodili kreiranjem dokumenta, a to su:

1. Da se izvrši adekvatan opis obeležja, i detaljna analiza uticaja/veza/zavisnosti između obeležja i u skupu podataka
2. Da se na principijalan način izvrši formiranje, odabir i tumačenje najadekvatnijeg modela mašinskog učenja za predviđanje prijema studenata na fakultetima u SAD-a.
3. Da se sirovi skup podataka dovede do nivoa kvaliteta koji omogućava dovoljno pouzdano statističko zaključivanje o vezama između obeležja. kao i formiranje adekvatnih modela mašinskog učenja za predviđanje prijema studenata na fakultetima u SAD-a.