# Analiza rezervacija hotelskog smještaja

Grupa Prosječni

Mario Mrvčić, Rujana Perić, Dorjan Štrbac, Sven Winkler

2023-12-29

## OPIS ZADATKA

U današnje vrijeme turistička djelatnost jako je popularna, a tržište puno konkurencije. Zanima nas kako hoteli mogu prilagoditi svoje usluge kako bi poboljšali zadovoljnost gostiju i povećali profitabilnost.

Posjedujući skup podataka koji detaljno dokumentira rezervacije u dva različita hotela želimo istražiti, bolje razumjeti i ispitati veze između boravka u hotelu, godišnjeg doba, socio-ekonomskog statusa gosta, cijena, tipa rezervacija i sličnih parametara danih skupom podataka.

### Učitavanje podataka o rezervacijama

Skup podataka sastoji se od 119390 podataka o rezervacijama u dva hotela. Svaki podatak o rezervaciji sastoji se od 31 varijable koje pružaju uvid u trendove dolaska, demografske podatke gostiju, njihove financijske obrasce...

U nastavku je prikazan:

- podatak o broju zapisa rezervacija

```
h1_data = read.csv("Rezervacije/Rezervacije_/H1.csv")
h2_data = read.csv("Rezervacije/Rezervacije_/H2.csv")
data <- rbind(h1_data, h2_data)
dim(data)
```

```
## [1] 119390     31
```

- primjer nekoliko zapisa

```
head(data)
```

```
##   IsCanceled LeadTime ArrivalDateYear ArrivalDateMonth ArrivalDateWeekNumber
## 1          0      342            2015             July                    27
## 2          0      737            2015             July                    27
## 3          0        7            2015             July                    27
## 4          0       13            2015             July                    27
## 5          0       14            2015             July                    27
## 6          0       14            2015             July                    27
##   ArrivalDateDayOfMonth StaysInWeekendNights StaysInWeekNights Adults Children
```

```
## 1                     1                   0                0      2        0
## 2                     1                   0                0      2        0
## 3                     1                   0                1      1        0
## 4                     1                   0                1      1        0
## 5                     1                   0                2      2        0
## 6                     1                   0                2      2        0
##   Babies      Meal Country MarketSegment DistributionChannel IsRepeatedGuest
## 1      0 BB            PRT        Direct              Direct               0
## 2      0 BB            PRT        Direct              Direct               0
## 3      0 BB            GBR        Direct              Direct               0
## 4      0 BB            GBR     Corporate           Corporate               0
## 5      0 BB            GBR     Online TA               TA/TO               0
## 6      0 BB            GBR     Online TA               TA/TO               0
##   PreviousCancellations PreviousBookingsNotCanceled ReservedRoomType
## 1                     0                           0 C
## 2                     0                           0 C
## 3                     0                           0 A
## 4                     0                           0 A
## 5                     0                           0 A
## 6                     0                           0 A
##   AssignedRoomType BookingChanges     DepositType       Agent     Company
## 1 C                             3 No Deposit             NULL        NULL
## 2 C                             4 No Deposit             NULL        NULL
## 3 C                             0 No Deposit             NULL        NULL
## 4 A                             0 No Deposit              304        NULL
## 5 A                             0 No Deposit              240        NULL
## 6 A                             0 No Deposit              240        NULL
##   DaysInWaitingList CustomerType ADR RequiredCarParkingSpaces
## 1                 0    Transient   0                       0
## 2                 0    Transient   0                       0
## 3                 0    Transient  75                       0
## 4                 0    Transient  75                       0
## 5                 0    Transient  98                       0
## 6                 0    Transient  98                       0
##   TotalOfSpecialRequests ReservationStatus ReservationStatusDate
## 1                      0         Check-Out            2015-07-01
## 2                      0         Check-Out            2015-07-01
## 3                      0         Check-Out            2015-07-02
## 4                      0         Check-Out            2015-07-02
## 5                      1         Check-Out            2015-07-03
## 6                      1         Check-Out            2015-07-03
```

- svi parametri rezervacija

```
names(data)
```

```
##  [1] "IsCanceled"                "LeadTime"
##  [3] "ArrivalDateYear"           "ArrivalDateMonth"
##  [5] "ArrivalDateWeekNumber"     "ArrivalDateDayOfMonth"
##  [7] "StaysInWeekendNights"      "StaysInWeekNights"
##  [9] "Adults"                    "Children"
## [11] "Babies"                    "Meal"
## [13] "Country"                   "MarketSegment"
```

```
## [15] "DistributionChannel"          "IsRepeatedGuest"
## [17] "PreviousCancellations"        "PreviousBookingsNotCanceled"
## [19] "ReservedRoomType"             "AssignedRoomType"
## [21] "BookingChanges"               "DepositType"
## [23] "Agent"                        "Company"
## [25] "DaysInWaitingList"            "CustomerType"
## [27] "ADR"                          "RequiredCarParkingSpaces"
## [29] "TotalOfSpecialRequests"       "ReservationStatus"
## [31] "ReservationStatusDate"
```

# Case study: TRAJANJE BORAVKA GOSTA

Slijedi detaljna analiza trajanja boravka gosta s obrizom na različite parametre.

### Trajanje prosječnog boravka tjekom različitih godišnjih doba

Dogovorno su godišnja doba definirana na idući način:
ljeto - lipanj, srpanj, kolovoz
jesen - rujan, listopad, studeni
zima - prosinac, siječanj, veljača
proljeće -ožujak, travanj, svibanj

Definiramo čeriti pomoćna skupa podataka za svako godišnje doba grupirajući po paramteru "ArrivalDateMonth".

```
summer = rbind(data[data$ArrivalDateMonth == "June",],
               data[data$ArrivalDateMonth == "July",],
               data[data$ArrivalDateMonth == "August",])

autumn = rbind(data[data$ArrivalDateMonth == "September",],
               data[data$ArrivalDateMonth == "October",],
               data[data$ArrivalDateMonth == "November",])

winter = rbind(data[data$ArrivalDateMonth == "December",],
               data[data$ArrivalDateMonth == "January",],
               data[data$ArrivalDateMonth == "February",])

spring = rbind(data[data$ArrivalDateMonth == "March",],
               data[data$ArrivalDateMonth == "April",],
               data[data$ArrivalDateMonth == "May",])
```

Pomoćnim skupovima dodajemo i parametar "TotalStays" koji predstavlja sumu vrijednosti parametara "StaysInWeekendNights" i "StaysInWeekNights" kako bismo dobili uvid o trajanju boravka gosta neovisno o danima u tjednu.

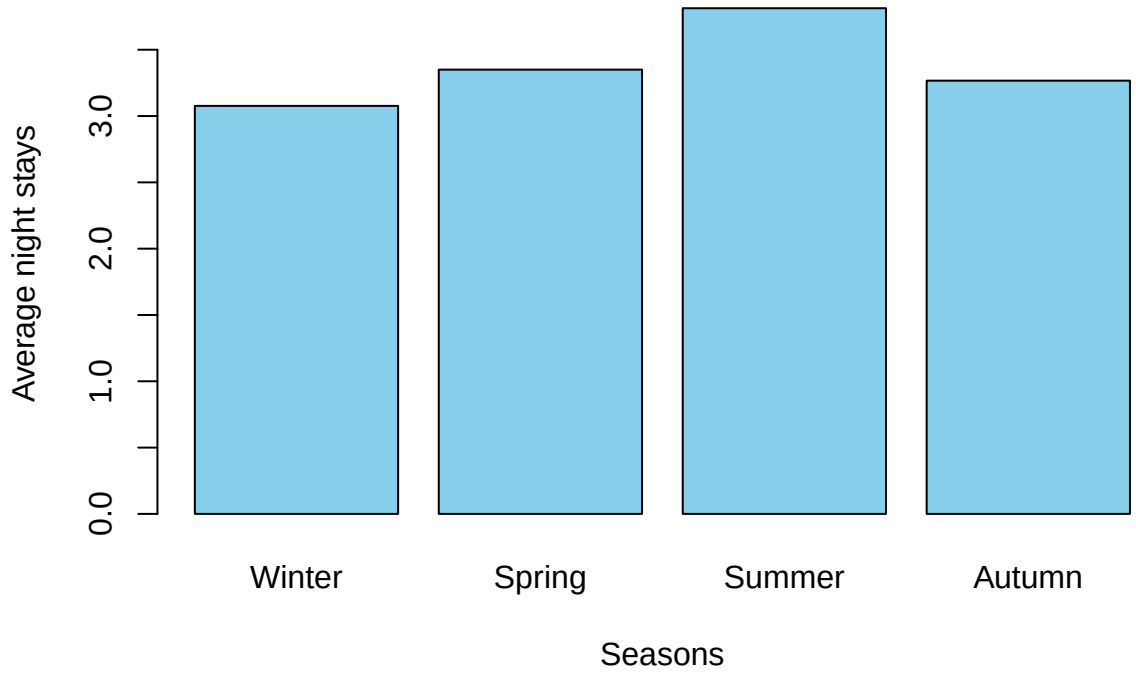
```
library(magrittr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
summer_total <- summer %>%
  mutate(TotalStays = StaysInWeekendNights + StaysInWeekNights)

autumn_total <- autumn %>%
  mutate(TotalStays = StaysInWeekendNights + StaysInWeekNights)

winter_total <- winter %>%
  mutate(TotalStays = StaysInWeekendNights + StaysInWeekNights)

spring_total <- spring %>%
  mutate(TotalStays = StaysInWeekendNights + StaysInWeekNights)
```

Računamo prosječnu duljinu trajanja boravka za svaki pomoćni skup podataka, rezultat prikazujemo stupičastim dijagramom.

```r
seasons_means_vector = data.frame(
  Season = c("Winter", "Spring", "Summer", "Autumn"),
  Means = c(mean(winter_total$TotalStays),
            mean(spring_total$TotalStays),
            mean(summer_total$TotalStays),
            mean(autumn_total$TotalStays))
)

barplot(seasons_means_vector$Means, names.arg = seasons_means_vector$Season, col = "skyblue", xlab = "S
```



```r
#binding adjusted data
data_total_stays = rbind(winter_total, spring_total, summer_total, autumn_total)
```

Iz histograma možemo zaključiti da je prosječni boravak gostiju najduži ljeti, a najkraći zimi.

### Trajanje prosječnog boravka i tip iznajmljene sobe

Naše istraživačko pitanje je slijedeće:
Možemo li trajanje boravka povezati s tipom iznajmljene sobe?

Pogledajmo najprije koje sve kategorije soba za iznajmljivanje postoje.

```r
#all different values of "ReservedRoomType" variable
room_types = unique(data_total_stays$AssignedRoomType)
room_types
```

```
##  [1] "I               " "A               " "E               " "D               "
##  [5] "F               " "C               " "H               " "G               "
##  [9] "B               " "P               " "K               " "L               "
```

Iz priloženog možemo vidjeti da postoji 12 različitih kategorija soba, prema tome, uzimamo 12 podskupova podataka i uspoređujemo njihove aritmetičke sredine duljine trajanja boravka (varijabla "TotayStays").

Za dati odgovor na istraživačko pitanje koristit ćemo se jednofaktorskim ANOVA modelom.

Predpostavke ANOVE su: - nezavisnost pojedinih podataka u uzorcima - normalna razdioba podataka - homogenost varijanci među populacijama

Kako nam veličine grupa nisu podjednake, radimo provjeru normalnosti Lillieforsovom inačicom Kolmogorov-Smirnovljenog testa. Razmatrat ćemo tip sobe kao varijablu koja određuje grupe i ukupnu duljinu boravka kao zavisnu varijablu.

```r
require(nortest)
```

```
## Loading required package: nortest
```

```r
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "I               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "I               ",     ]$TotalStays
## D = 0.26446, p-value < 2.2e-16
```

```r
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "A               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "A               ",     ]$TotalStays
## D = 0.21952, p-value < 2.2e-16
```

```
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "E               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "E               ",     ]$TotalStays
## D = 0.17111, p-value < 2.2e-16
```

```
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "D               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "D               ",     ]$TotalStays
## D = 0.19488, p-value < 2.2e-16
```

```
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "F               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "F               ",     ]$TotalStays
## D = 0.18765, p-value < 2.2e-16
```

```
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "C               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "C               ",     ]$TotalStays
## D = 0.162, p-value < 2.2e-16
```

```
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "H               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "H               ",     ]$TotalStays
## D = 0.15848, p-value < 2.2e-16
```

```
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "G               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "G               ",     ]$TotalStays
## D = 0.16133, p-value < 2.2e-16
```

```
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "P               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "P               ",     ]$TotalStays
## D = 0.53025, p-value = 1.337e-10
```

```
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "B               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "B               ",     ]$TotalStays
## D = 0.18077, p-value < 2.2e-16
```

```
lillie.test(data_total_stays[data_total_stays$AssignedRoomType == "K               ", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$AssignedRoomType == "K               ",     ]$TotalStays
## D = 0.30595, p-value < 2.2e-16
```

Iznimno malene p-vrijednosti nam govore kako podaci nisu normalno distribuirani. Usprkos tome, jednofaktorska ANOVA je robusna statistika s obzirom na predpostavku normalnosti. Dobro tolerira odstupanja od normalne distribucije.

Što se tiče homogenosti varijanci različitih populacija, potrebno je testirati:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$
$$H_1 : \text{barem dvije varijance nisu iste.}$$

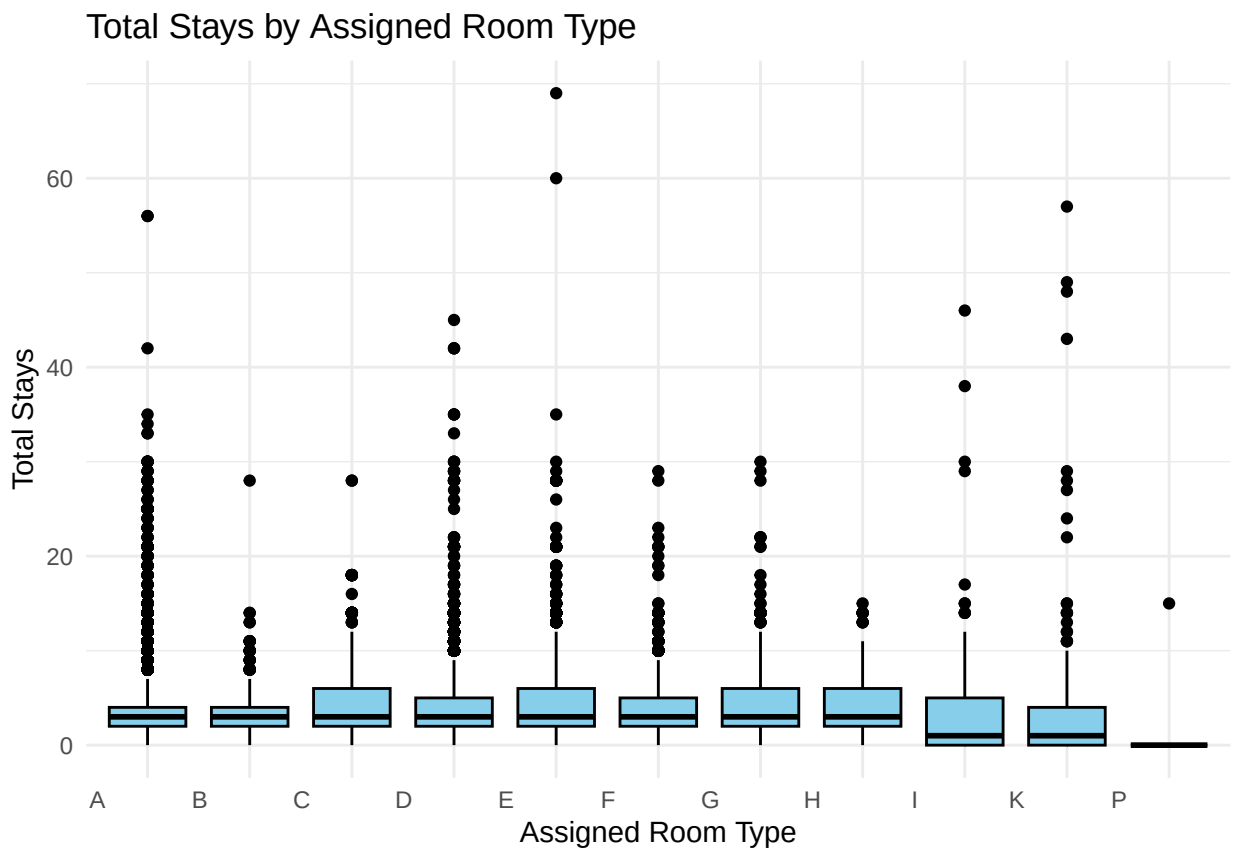
Navedenu hipotezu možemo testirati Bartlettovim testom.

```
data_total_stays$AssignedRoomType <- factor(data_total_stays$AssignedRoomType)

# Check the number of observations in each group
group_counts <- table(data_total_stays$AssignedRoomType)
valid_groups <- names(group_counts[group_counts >= 2])

#If there is at least two oservations in the group, proceed
filtered_data <- subset(data_total_stays, AssignedRoomType %in% valid_groups)

#Bartlettovim test
bartlett.test(filtered_data$TotalStays ~ filtered_data$AssignedRoomType)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  filtered_data$TotalStays by filtered_data$AssignedRoomType
## Bartlett's K-squared = 6714.5, df = 10, p-value < 2.2e-16
```

Varijance varijable "TotalStays" nisu podjednake, što je grafički prikazano u nastavku.

```
#box and whisker plot --> mean, rangs, median, variances....
library(ggplot2)

# Assuming filtered_data$AssignedRoomType is a factor
filtered_data$AssignedRoomType <- factor(filtered_data$AssignedRoomType)

# Create a boxplot using ggplot2
ggplot(filtered_data, aes(x = AssignedRoomType, y = TotalStays)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Total Stays by Assigned Room Type",
       x = "Assigned Room Type",
       y = "Total Stays") +
  theme_minimal()
```




Dani podaci o distibuciji koja nije normalna i nehomogenosti varijanci navodi nas na zaključak da prosječna duljina boravka varira ovisno o tipu iznajmljene sobe, a to ćemo sada testirati.

Naša testne hipoteze biti će:

$H_0$ : prosječna duljina boravka jednaka je za sve tipove iznajmljenih soba

$H_1$ : za barem jedan tip iznajmljene sobe prosječna duljina boravka različita je no u ostalih

```
#ANOVA
a = aov(filtered_data$TotalStays ~ filtered_data$AssignedRoomType)
summary(a)
```

```
##                                   Df Sum Sq Mean Sq F value Pr(>F)
## filtered_data$AssignedRoomType     10  14597  1459.7   227.4 <2e-16 ***
## Residuals                      119378 766260     6.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Zaključak:
Na temelju malene p-vrijednosti odbacujemo nultu hipotezu u korist altervativne hipoteze: **postoje razlike u duljini trajanja boravka ovisno o tipu iznajmljene sobe.**

### Trajanje prosječnog boravka i kategorija gosta

Naše iduće istraživačko pitanje je slijedeće:
Možemo li trajanje boravka povezati s kategorijom gosta?

Procedura zaključivanja i odgovaranja na ovo istraživačko pitanje ista je prethodnom. Pogledajmo koje sve kategorije gostiju postoje.

```
#all different values of "CustomerType" variable
customer_types = unique(data_total_stays$CustomerType)
customer_types
```

```
## [1] "Transient"       "Contract"        "Transient-Party" "Group"
```

Iz priloženog možemo vidjeti da postoji 4 različitih kategorija gostiju, prema tome, uzimamo 4 podskupova podataka i uspoređujemo njihove aritmetičke sredine duljine trajanja boravka (varijabla "TotayStays").

Za dati odgovor na istraživačko pitanje koristit ćemo se jednofaktorskim ANOVA modelom.

Provjeravamo jesu li zadovoljene predpostavke ANOVE (za provjeru predpostavke normalnosti koristimo se Lillieforsovom inačicom KS testa, a za provjeru predpostavke homogensoti varijanca koristimo se Bartlettovim testom).

```
#normality check
require(nortest)

lillie.test(data_total_stays[data_total_stays$CustomerType == "Transient", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$CustomerType == "Transient",     ]$TotalStays
## D = 0.19487, p-value < 2.2e-16
```

```
lillie.test(data_total_stays[data_total_stays$CustomerType == "Contract", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$CustomerType == "Contract",     ]$TotalStays
## D = 0.17507, p-value < 2.2e-16
```

```r
lillie.test(data_total_stays[data_total_stays$CustomerType == "Transient-Party", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$CustomerType == "Transient-Party",     ]$TotalStays
## D = 0.23459, p-value < 2.2e-16
```

```r
lillie.test(data_total_stays[data_total_stays$CustomerType == "Group", ]$TotalStays)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_total_stays[data_total_stays$CustomerType == "Group", ]$TotalStays
## D = 0.21935, p-value < 2.2e-16
```

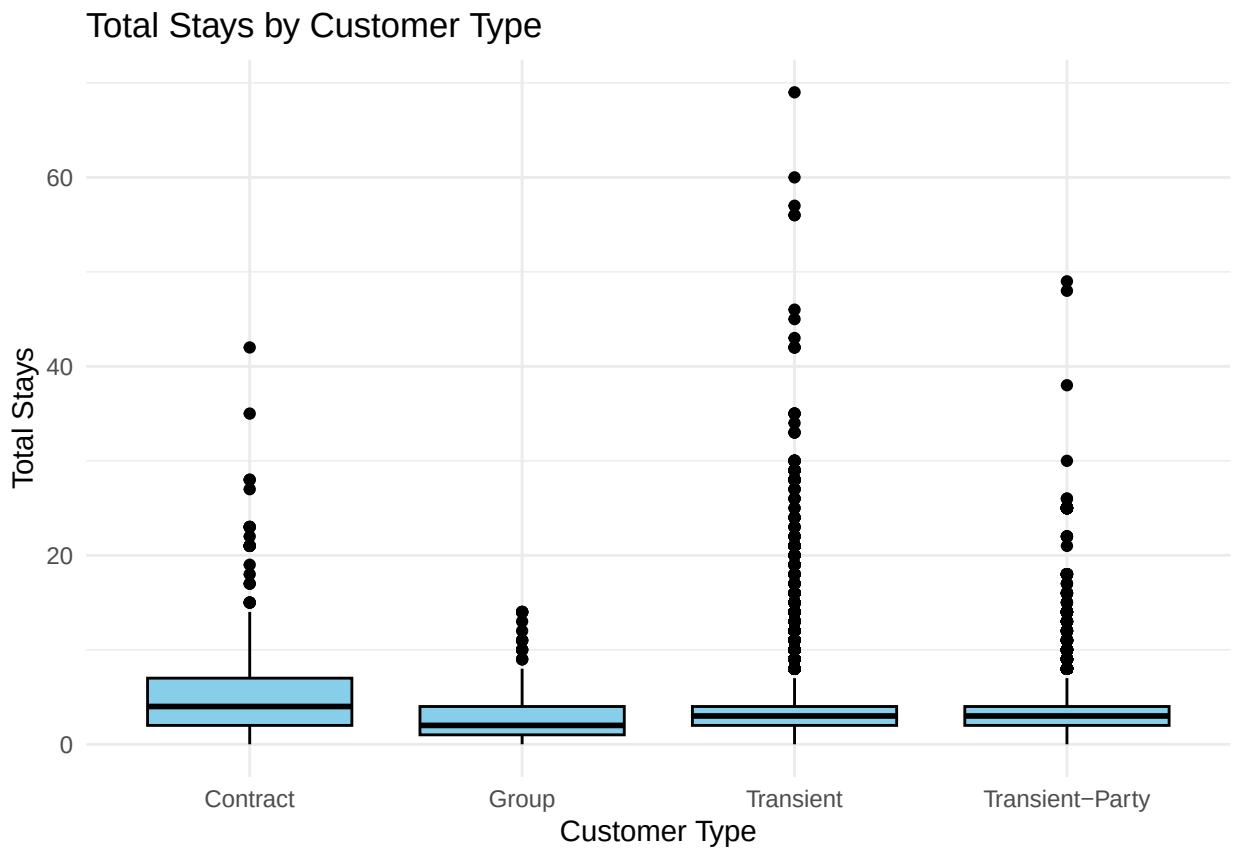
Iznimno malene p-vrijednosti nam govore kako podaci nisu normalno distribuirani. Usprkos tome, jednofaktorska ANOVA je robusna statistika s obzirom na predpostavku normalnosti.

```r
# variance homogenity check
data_total_stays$CustomerType <- factor(data_total_stays$CustomerType)

# Check the number of observations in each group
group_counts <- table(data_total_stays$CustomerType)
valid_groups <- names(group_counts[group_counts >= 2])

#If there is at least two oservations in the group, proceed
filtered_data <- subset(data_total_stays, CustomerType %in% valid_groups)

#Bartlettovim test
bartlett.test(filtered_data$TotalStays ~ filtered_data$CustomerType)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  filtered_data$TotalStays by filtered_data$CustomerType
## Bartlett's K-squared = 3615, df = 3, p-value < 2.2e-16
```

Zbog malene p-vrijednosti odbacujemo homogenost varijanci, što prikazujemo i grafički.

```r
#box and whisker plot --> mean, rangs, median, variances....
library(ggplot2)

# Assuming filtered_data$CustomerType is a factor
filtered_data$CustomerType <- factor(filtered_data$CustomerType)

# Create a boxplot using ggplot2
ggplot(filtered_data, aes(x = CustomerType, y = TotalStays)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Total Stays by Customer Type",
       x = "Customer Type",
       y = "Total Stays") +
  theme_minimal()
```




Dani podaci o distibuciji koja nije normalna i nehomogenosti varijanci navodi nas na zaključak da prosječna duljina boravka varira ovisno o tipu gosta, a to ćemo sada i testirati jednoparametarskom ANOVOM.

Naša testne hipoteze biti će:

$H_0$ : prosječna duljina boravka jednaka je za sve tipove gostiju

$H_1$ : za barem jedan tip gosta prosječna duljina boravka različita je no u ostalih

```
#ANOVA
a = aov(filtered_data$TotalStays ~ filtered_data$CustomerType)
summary(a)
```

```
##                               Df Sum Sq Mean Sq F value Pr(>F)
## filtered_data$CustomerType     3  18121    6040   945.5 <2e-16 ***
## Residuals                 119386 762742       6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Zaključak:
Na temelju malene p-vrijednosti odbacujemo nultu hipotezu u korist altervativne hipoteze: **postoje razlike u duljini trajanja boravka ovisno o tipu gosta.**

Diskusija:
Kao što smo i mogli naslutiti, duljina boravka gosta varira s obzirom na različite parametre. Najdulji prosječni boravak jest ljeti, a duljina boravka varira i zavisno od tipa iznajmljene sobe te kategorije gosta.

## Case study: CIJENE

## Case study: PANSION, POLUPANSION

## Case study: OTKAZIVANJE REZERVACIJA

## Case study: SKUPINE GOSTIJU IZ RAZLIČITIH ZEMALJA