

Hotel Bookings Dataset Report

Marko Dananić [JMBAG: 0036541487, e-mail: md54148@fer.hr]

Dario Huić [JMBAG: 0036538469, e-mail: dario.huic@fer.hr]

Matija Alojz Stuhne [JMBAG: 0036540079, e-mail: ms54007@fer.hr]

2024-01-26

Uvod

Tema ovog dokumenta provedba je eksploatorne analize podataka nad skupom podataka o rezervacijama hotela. Skup podataka preuzet je s Kaggle-a, a može se pronaći na sljedećoj poveznici: [Hotel Bookings data set](#).

U podatkovnom skupu vidljivi su podaci o dva hotela. Oba hotela nalaze se u Portugalu. Jedan je gradski hotel (City Hotel), koji se nalazi u centru grada Lisabona, a drugi je hotel u sklopu odmarališta (Resort Hotel), koji se nalazi u obližnjoj turističkoj regiji Algrave.

Uređivanje csv datoteke

Preuzeta *hotel_bookings.csv* datoteka sadrži dva stupaca koja je lakše razumjeti kada se podaci iz tih stupaca pretvore u logičke vrijednosti ("is_canceled", "is_repeated_guest"). Nule (0) su pretvorene u FALSE, a jedinice (1) u TRUE.

Učitavanje podataka

```
hotel_bookings <- read_csv("../data/hotel_bookings_02.csv", show_col_types = FALSE)

hotel_bookings$arrival_date_month <- custom_order_months(hotel_bookings$arrival_date_month)
```

Opis podatkovnog skupa

Skup podataka sadrži informacije o 119,390 rezervacija hotela između 1. srpnja 2015. i 31. kolovoza 2017. godine. Svaki redak csv datoteke predstavlja jednu zabilježenu rezervaciju u hotelskom sustavu.

```
glimpse(hotel_bookings)
```

```
## Rows: 119,390
## Columns: 33
## $ index          <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 1~
## $ hotel          <chr> "Resort Hotel", "Resort Hotel", "Resort~
## $ is_canceled    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALS~
## $ lead_time      <dbl> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, ~
## $ arrival_date_year <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 201~
## $ arrival_date_month <ord> July, July, July, July, July, July, Jul~
## $ arrival_date_week_number <dbl> 27, 27, 27, 27, 27, 27, 27, 27, 27, ~
## $ arrival_date_day_of_month <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ stays_in_weekend_nights <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ stays_in_week_nights <dbl> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, ~
```

```

## $ adults          <dbl> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ children        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ babies          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ meal            <chr> "BB", "BB", "BB", "BB", "BB", "BB", "BB", "BB~
## $ country         <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "GBR~
## $ market_segment <chr> "Direct", "Direct", "Direct", "Corporat~
## $ distribution_channel <chr> "Direct", "Direct", "Direct", "Corporat~
## $ is_repeated_guest <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALS~
## $ previous_cancellations <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_bookings_not_canceled <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ reserved_room_type <chr> "C", "C", "A", "A", "A", "A", "C", "C", ~
## $ assigned_room_type <chr> "C", "C", "C", "A", "A", "A", "C", "C", ~
## $ booking_changes  <dbl> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ deposit_type     <chr> "No Deposit", "No Deposit", "No Deposit~
## $ agent            <dbl> NA, NA, NA, 304, 240, 240, NA, 303, 240~
## $ company          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ days_in_waiting_list <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ customer_type    <chr> "Transient", "Transient", "Transient", ~
## $ adr              <dbl> 0.00, 0.00, 75.00, 75.00, 98.00, 98.00, ~
## $ required_car_parking_spaces <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ total_of_special_requests <dbl> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 3, ~
## $ reservation_status <chr> "Check-Out", "Check-Out", "Check-Out", ~
## $ reservation_status_date <chr> "01-07-15", "01-07-15", "02-07-15", "02~

```

Svaka rezervacija sadrži sljedeće informacije:

- **hotel** - hotel u kojem je rezervacija napravljena (H1 = Resort Hotel ili H2 = City Hotel)
- **is_canceled** - da li je rezervacija otkazana ili nije (1 = otkazana, 0 = nije otkazana)
- **lead_time** - broj dana između datuma rezervacije i datuma dolaska
- **arrival_date_year, arrival_date_month, arrival_date_week_number, arrival_date_day_of_month** - datum dolaska u hotel
- **stays_in_weekend_nights** - broj noćenja (subota ili nedjelja) koje je gost ostao u hotelu
- **stays_in_week_nights** - broj noćenja (ponedjeljak do petak) koje je gost ostao u hotelu
- **adults** - broj odraslih osoba
- **children** - broj djece
- **babies** - broj beba
- **meal** - tip rezervacije (BB - Bed and Breakfast, HB - half board, FB - full board, SC - self catering)
- **country** - država iz koje je gost došao
- **market_segment** - segment tržišta (Direct - direktan, Online TA - online putnička agencija, Offline TA/TO - offline putnička agencija, Corporate - korporativno, Groups - grupe)
- **distribution_channel** - kanal distribucije (TA/TO - putnička agencija, Direct - direktan, Corporate - korporativno, GDS - globalna distribucijska usluga)
- **is_repeated_guest** - da li je gost već boravio u hotelu (1 = da, 0 = ne)
- **previous_cancellations** - broj prethodnih otkazivanja rezervacija
- **previous_bookings_not_canceled** - broj prethodnih rezervacija koje nisu otkazane
- **reserved_room_type** - tip rezervirane sobe

- **assigned_room_type** - tip dodijeljene sobe
- **booking_changes** - broj promjena koje su napravljene u rezervaciji
- **deposit_type** - tip depozita (No Deposit - nema depozita, Non Refund - nepovratni depozit, Refundable - povratni depozit)
- **agent** - ID agenta koji je napravio rezervaciju
- **company** - ID kompanije koja je napravila rezervaciju
- **days_in_waiting_list** - broj dana koje je rezervacija bila na listi čekanja
- **customer_type** - tip rezervacije (Contract - ugovor, Group - grupa, Transient - prolazan, Transient-Party - prolazna grupa)
- **adr** - prosječna dnevna stopa
- **required_car_parking_spaces** - broj potrebnih parkirnih mjesta
- **total_of_special_requests** - broj posebnih zahtjeva
- **reservation_status** - status rezervacije (Canceled - otkazana, Check-Out - odjava, No-Show - gost nije došao)
- **reservation_status_date** - datum posljednje promjene statusa rezervacije

Eksploratorna analiza podataka

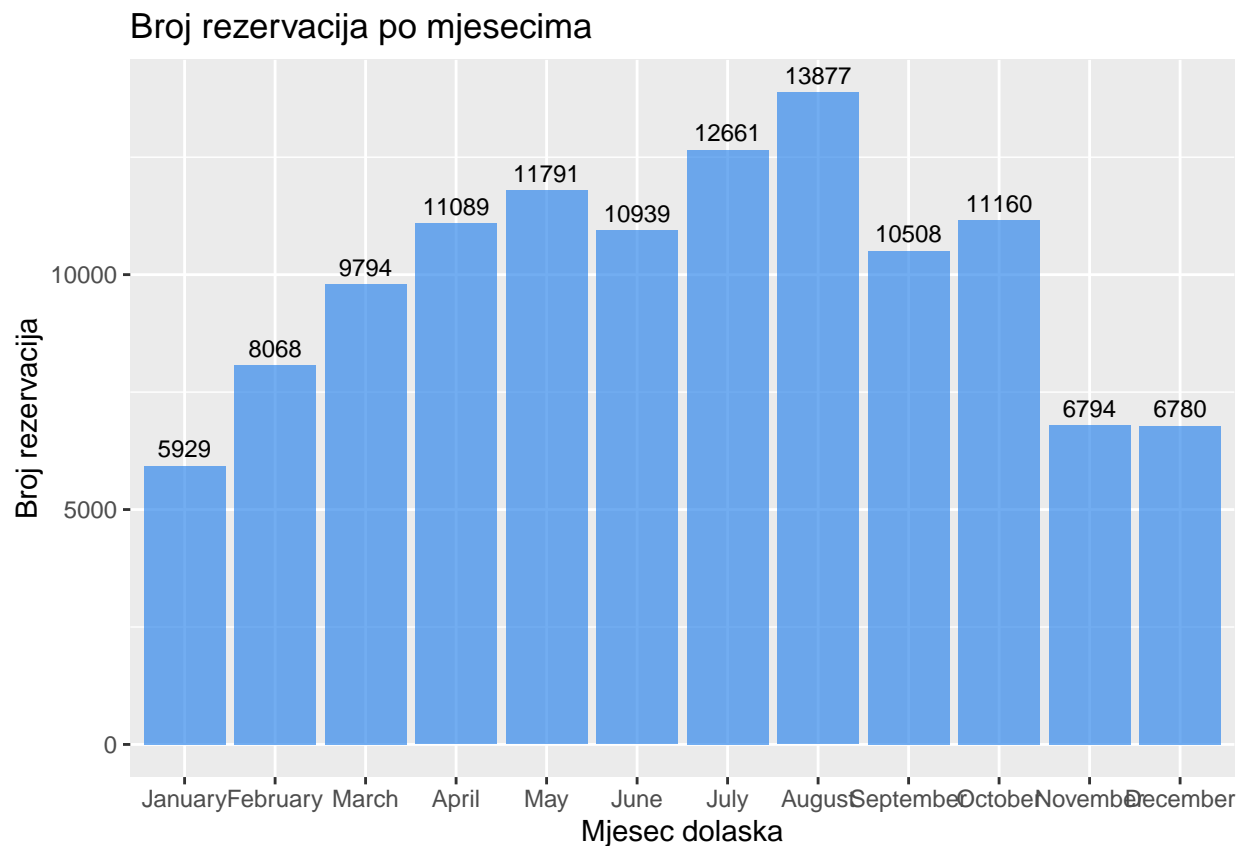
Provest ćemo eksploratornu analizu nad skupom podataka o rezervacijama hotelskih smještaja. Cilj je istražiti podatke i otkriti neke zanimljive činjenice o ponašanju gostiju hotela.

Broj rezervacija po mjesecima

Zanima nas kako se broj rezervacija mijenja tijekom godine. Kako bismo to saznali, sumirat ćemo broj rezervacija po mjesecima i prikazati ih u obliku histograma.

```
num_per_month <- hotel_bookings %>% count(arrival_date_month)

ggplot(num_per_month, aes(x = arrival_date_month, y = n)) +
  geom_col(fill = "#3489eb", alpha = 0.7) +
  geom_text(aes(label = n), vjust = -0.5, size = 3) +
  labs(x = "Mjesec dolaska", y = "Broj rezervacija", title = "Broj rezervacija po mjesecima")
```

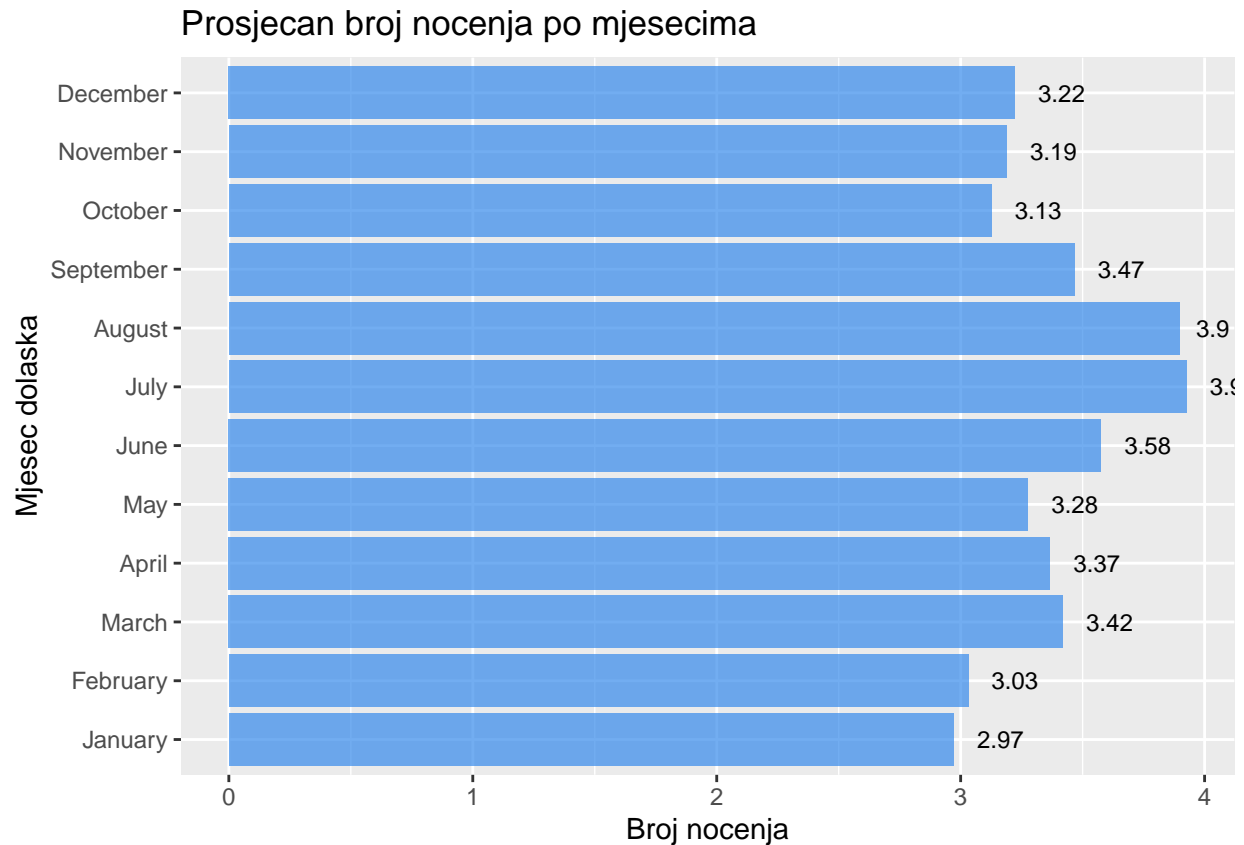


Iz grafa možemo iščitati kako je najveći broj rezervacija u ljetnim mjesecima, posebno u srpnju i kolovozu. Najmanji broj rezervacija je u prosincu i siječnju, što je i očekivano, s obzirom da je to najhladniji period godine u Europi.

Prosječan broj noćenja po mjesecima

Analizom podataka o prosječnom boravku gostiju po mjesecima, dobivamo uvid u to koliko su dugo gosti ostajali u hotelima tijekom godine. Istražit ćemo postoji li razlika u prosječnoj duljini boravka gostiju u ljetnim i zimskim mjesecima.

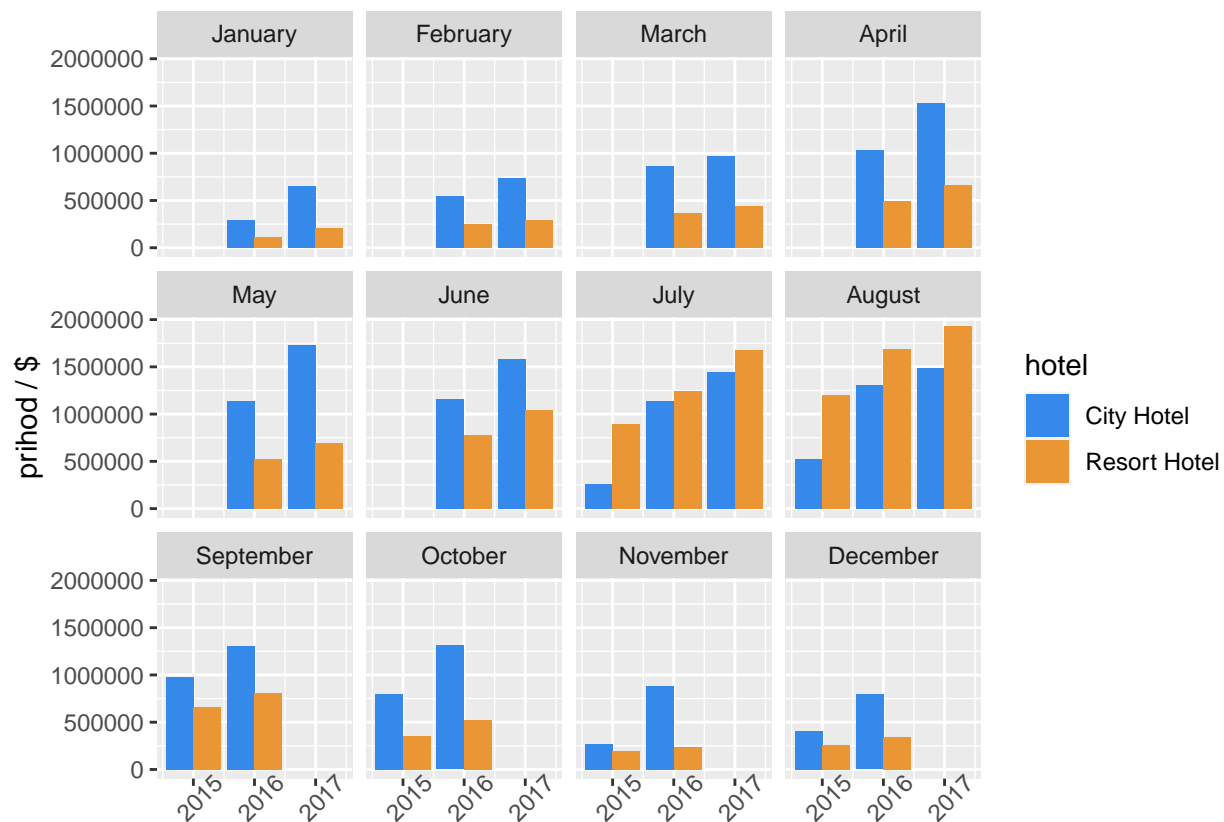
```
avg_stay_per_month <- hotel_bookings %>%
  group_by(arrival_date_month) %>%
  summarise(average_stay = mean(stays_in_week_nights + stays_in_weekend_nights)) %>%
  arrange(average_stay)
```



Možemo primjetiti kako su gosti najduže boravili u hotelima tijekom ljetnih mjeseci, vjerojatno zbog povećanog broja godišnjih odmora i toplijeg vremena. U mjesecu srpnju vidimo vrhunac prosječnog boravka od skoro 4 noćenja po gostu. S druge strane, prosječni boravak u siječnju bio je najkraći, s manje od 3 noćenja po gostu.

Prihod po mjesecima

Koristit ćemo slične filtere kao i do sada kako bismo izračunali mjesečni prihod svakog hotela. Očekujemo sličan rezultat prošlim analizama, s obzirom da je prihod direktno povezan s brojem rezervacija i brojem noćenja.



Uočavamo postepeni rast mjesečnog prihoda tijekom godine, s vrhuncem u ljetnim mjesecima.

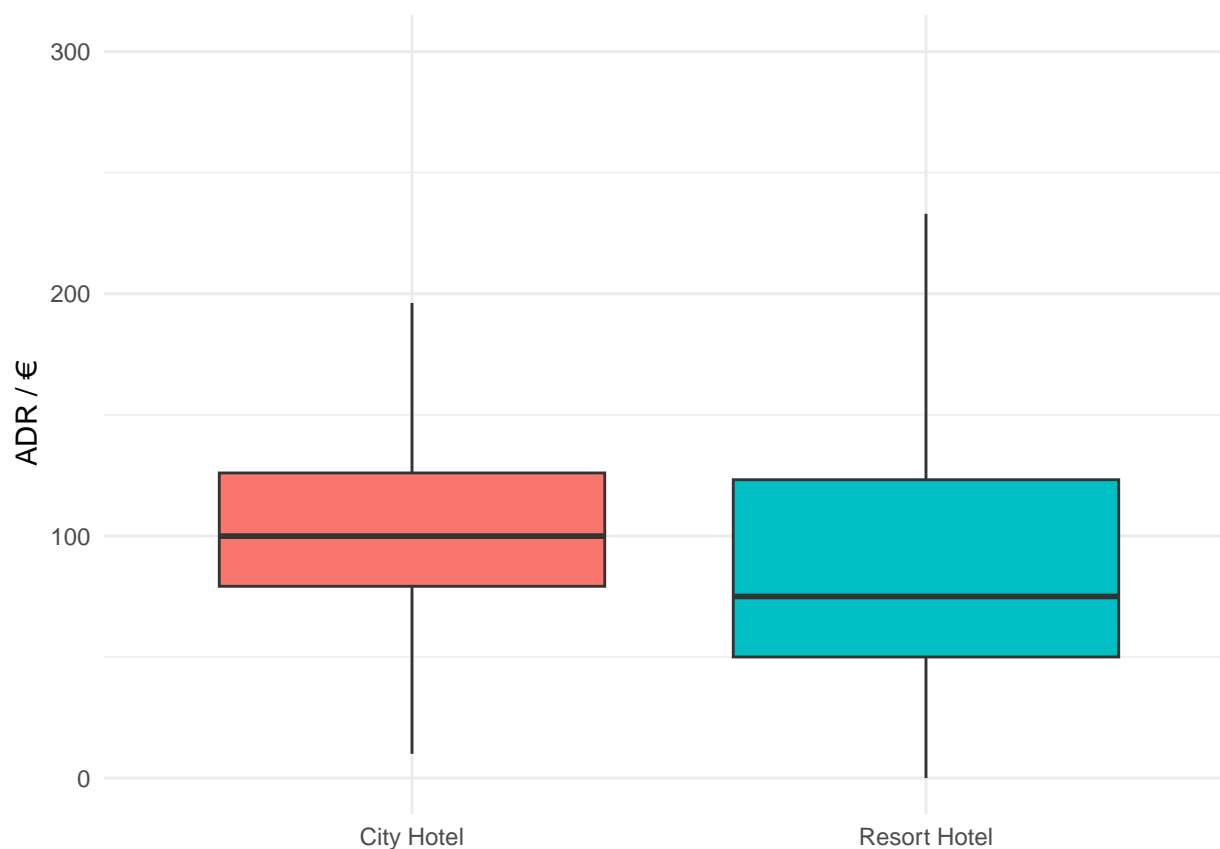
Promjene ADR-a s različitim faktorima

ADR (Average Daily Rate) predstavlja prosječnu cijenu sobe po noćenju. U ovom dijelu analizirat ćemo kako se mijenja ADR s različitim faktorima; kao što su: tip hotela, država iz koje dolaze gosti te tržišni segment.

Usporedba ADR-a i tipa hotela

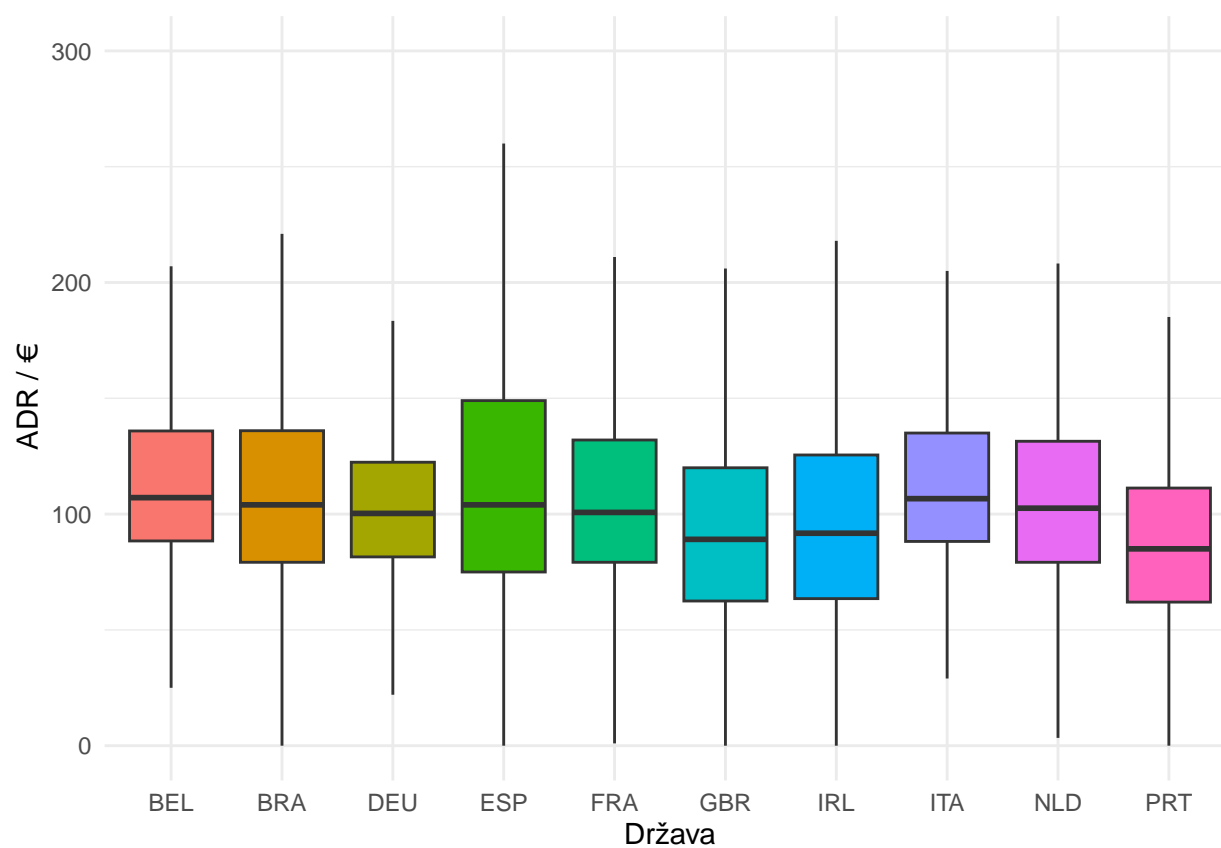
Zanima nas razlika u prosječnoj dnevnoj stopi (ADR) između gradskih i resortnih hotela.

```
ggplot(hotel_bookings, aes(x = hotel, y = adr, fill = hotel)) +  
  geom_boxplot(outlier.shape = NA) +  
  labs(x = NULL, y = "ADR / €") +  
  theme_minimal() +  
  theme(legend.position = "none") +  
  ylim(0,300)
```



Iz ovog boxplot dijagrama, možemo vidjeti da je medijan cijene sobe u gradskom hotelu (City Hotel) viši nego u resortnom hotelu (Resort Hotel). Ljudi koji odsjedaju u gradskom hotelu, u prosjeku izdvajaju više za sobu nego ljudi koji borave u resortnom hotelu. Međutim, postoji veći raspon cijena u resortnom hotelu, što znači da u odnosu na onaj gradski, resortni nudi veći izbor soba (što jeftinijih, što skupljih).

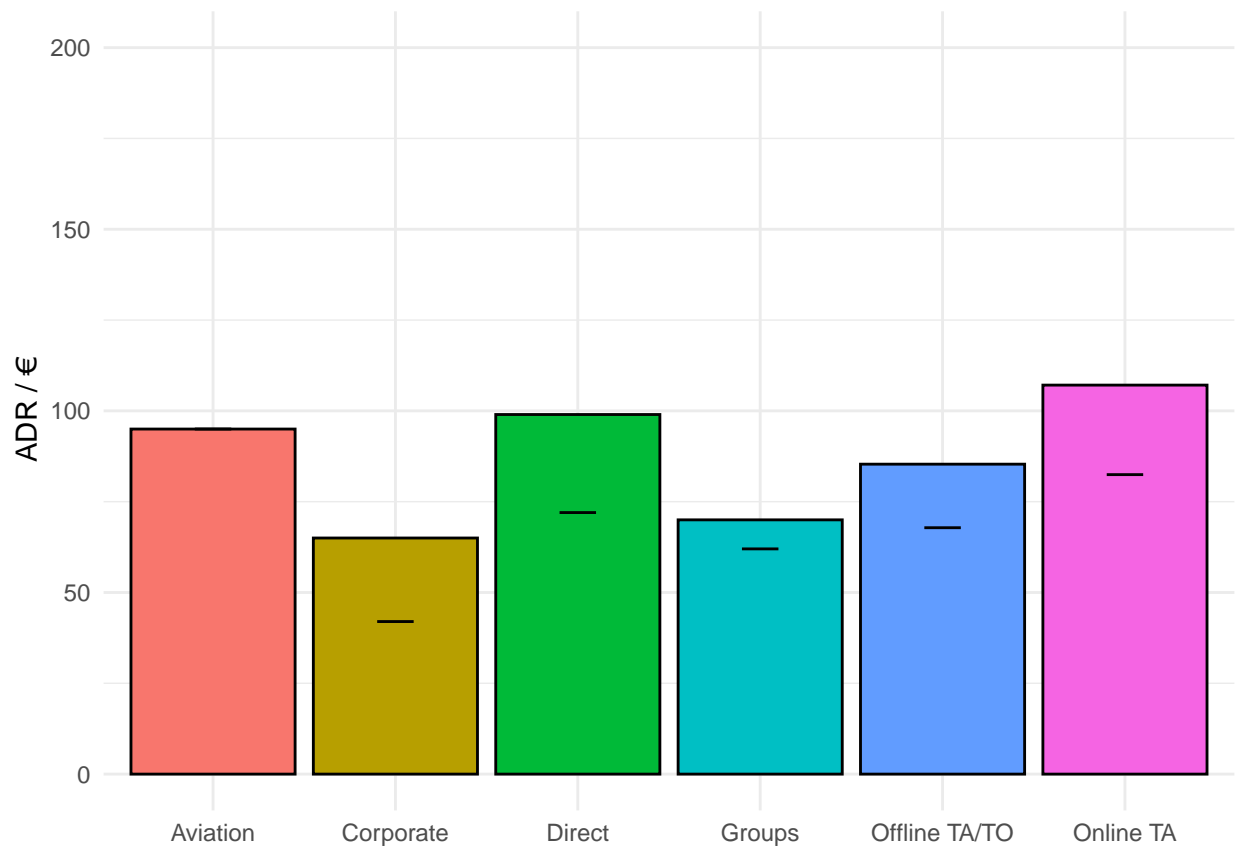
Usporedba ADR-a i zemlje podrijetla



Vidimo da je medijan cijene sobe najviši za goste iz Italije, dok je najniži za goste iz Portugala. Gosti iz Španjolske imaju najveći raspon cijena.

Usporedba ADR-a i tržišnog segmenta

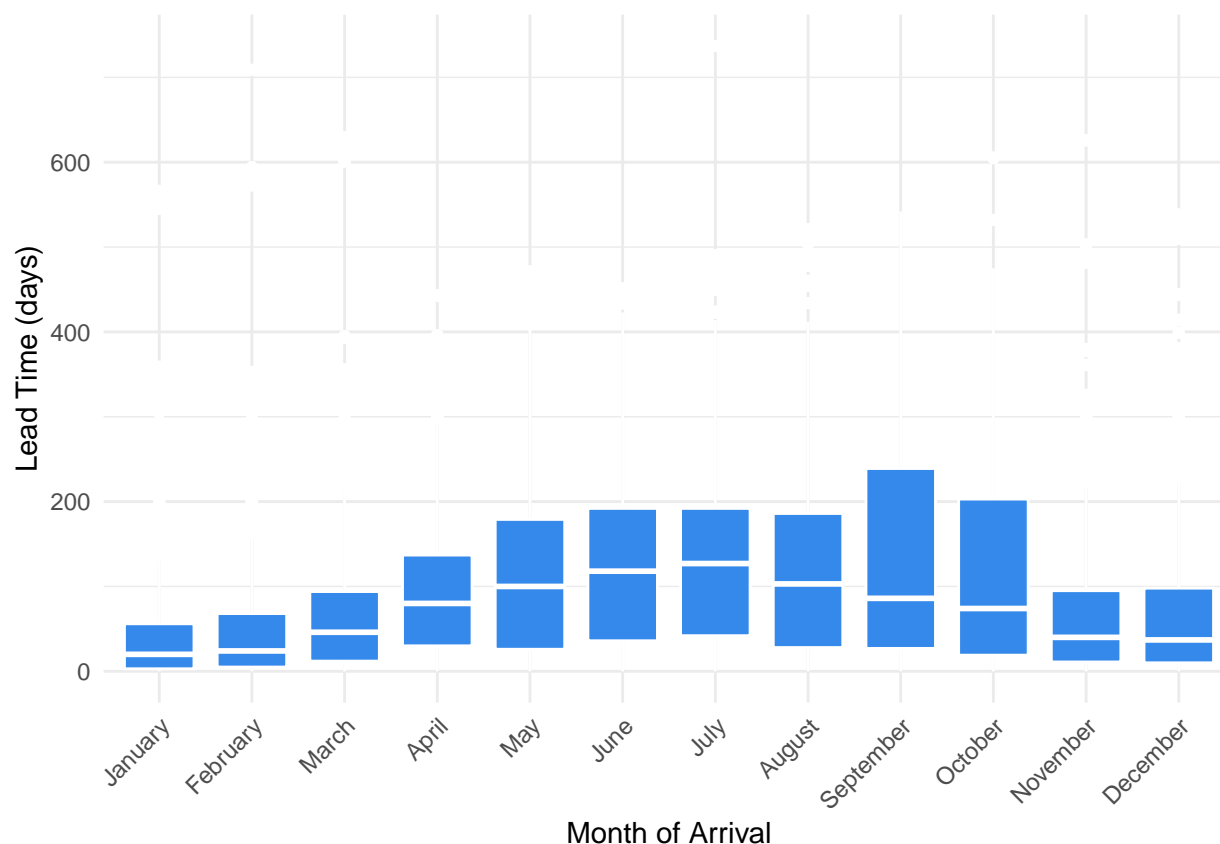
```
filtered_data <- hotel_bookings %>%  
  filter((market_segment %in% c("Aviation", "Corporate", "Direct", "Groups", "Offline TA/TO", "Online TA"))  
  
ggplot(filtered_data, aes(x = market_segment, y = adr, fill = market_segment)) +  
  geom_bar(stat = "summary", fun = "median", position = "dodge", color = "black") +  
  geom_errorbar(  
    stat = "summary", fun.min = function(x) quantile(x, 0.25),  
    fun = function(x) quantile(x, 0.75), position = "dodge", width = 0.2  
  ) +  
  labs(x = NULL, y = "ADR / €") +  
  theme_minimal() +  
  theme(legend.position = "none") +  
  ylim(0, 200)
```



Najmanje cijene soba su za one u kojima odsjedaju gosti čiji su smještaj platile tvrtke, dok su najskuplje cijene soba pri onim gostima koji su sobe rezervirali preko online turističkih agencija.

Analiza značajke lead_time po mjesecu dolaska

Značajka lead_time predstavlja broj dana koji je prošao između datuma rezervacije i datuma dolaska osobe/a u hotel.



Uočavamo da je medijan broja dana između rezervacije i dolaska najveći u srpnju i kolovozu, što može značiti da ljudi unaprijed rezerviraju svoje ljetne odmore. Medijan je najniži u studenom. U zimskim mjesecima ljudi rezerviraju svoje odmore u kraćem vremenskom razdoblju.

Korelacija između otkazivanja i 3 potencijalna faktora

Tablica nam prikazuje potencijalne razloge zbog kojih ljudi otkazuju svoje rezervacije. Duže vrijeme između rezervacije i datuma dolaska, vrijeme provedeno na listi čekanja te informacija da je osoba već otkazala rezervaciju vodi zaključku da će osoba otkazati svoju rezervaciju.

Zanimljivo je primijetiti da je gotovo 6000 osoba koje su prethodno otkazale rezervaciju otkazalo i svoju trenutnu rezervaciju.

Table 1: Correlation Table

is_canceled	avg_lead_time	avg_waiting_time	count_previous_canceled
FALSE	79.98469	1.589868	542
TRUE	144.84882	3.564083	5942

```
kable(result_table, format = "latex", caption = "Correlation Table", ) %>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE) %>%
  column_spec(1:ncol(result_table), background = "white")
```

Top 10 zemalja iz kojih dolaze gosti - podrijetla posjetitelja

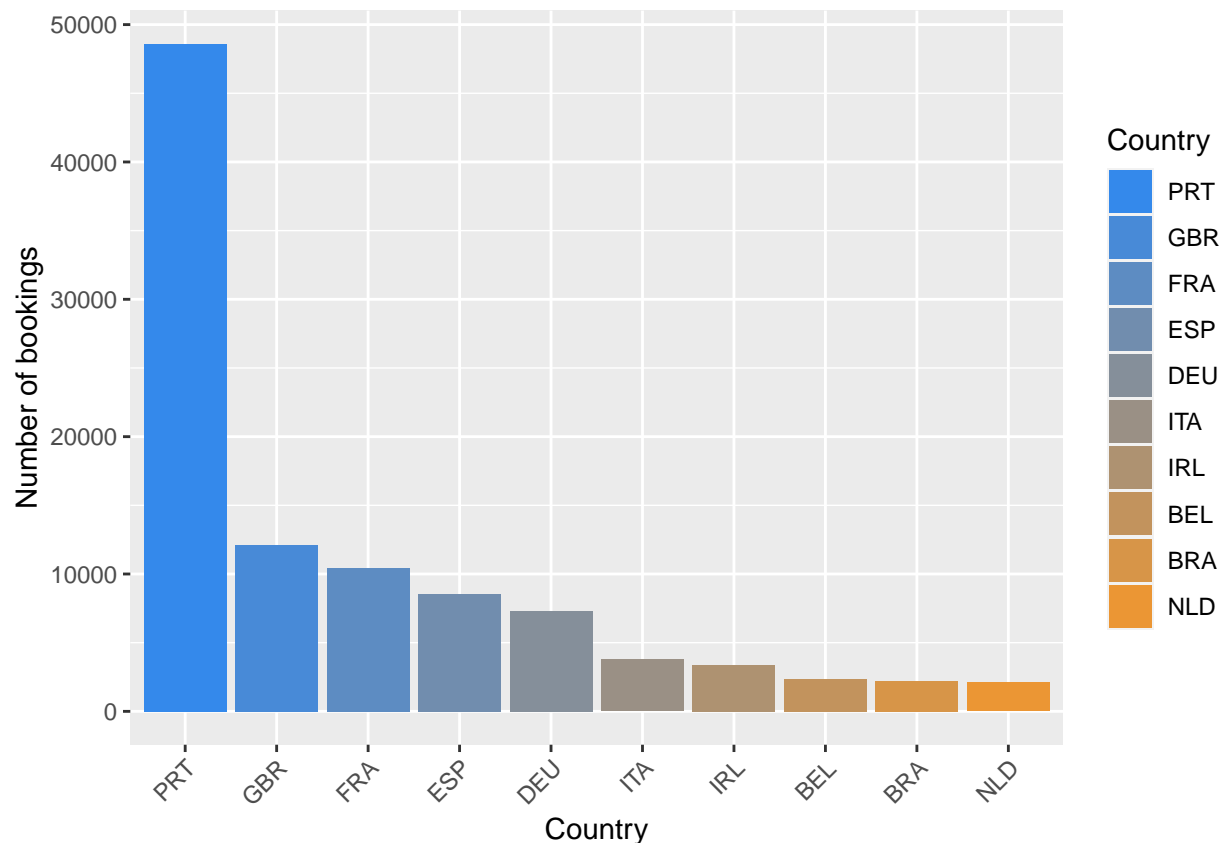
U ovom dijelu istražiti ćemo podrijetlo gostiju, tj. iz kojih zemalja dolaze. S obzirom da se hoteli nalaze u Portugalu, očekujemo da će većina gostiju biti iz Europe. Provedbom ove analize možemo dobiti uvid u to koliko je hotelima važno da se fokusiraju na određene zemlje, odnosno je li potrebno ulagati u marketing u određenim zemljama.

Koristit ćemo funkciju `table()` kako bismo izračunali broj rezervacija po zemljama. Zatim ćemo sortirati podatke kako bismo mogli izdvojiti top 10 zemalja iz kojih dolaze gosti.

```
total_bookings_per_country <- table(hotel_bookings$country)

top_countries <- as.data.frame(total_bookings_per_country)
top_countries <- top_countries[order(-top_countries$Freq), ]
top_10_countries <- head(top_countries, 10)

top_10_countries$Var1 <- factor(top_10_countries$Var1, levels = top_10_countries$Var1)
```



Utvrđili smo da je najveći broj gostiju iz Portugala (što je i očekivano s obzirom da se hoteli nalaze u Portugalu). Zatim slijede gosti iz Velike Britanije, Francuske, Španjolske i Njemačke. Ove zemlje su također u blizini Portugala, što je vjerojatno razlog zašto je broj gostiju iz ovih zemalja poprilično visok.

Karta svijeta s ukupnim brojem gostiju

Na ovom grafu prikazat ćemo raspodjelu broja gostiju u hotelima na karti svijeta. Podaci su grupirani prema zemljama, a boje na karti označavaju ukupan broj gostiju (odrasli + djeca + bebe). Što je boja intenzivnija, broj gostiju koji su došli iz te zemlje je veći.

```
world <- ne_countries(scale = "medium", returnclass = "sf")

hotel_bookings_agg <- hotel_bookings %>%
  group_by(country) %>%
  summarise(total_guests = sum(adults + children + babies, na.rm = TRUE))

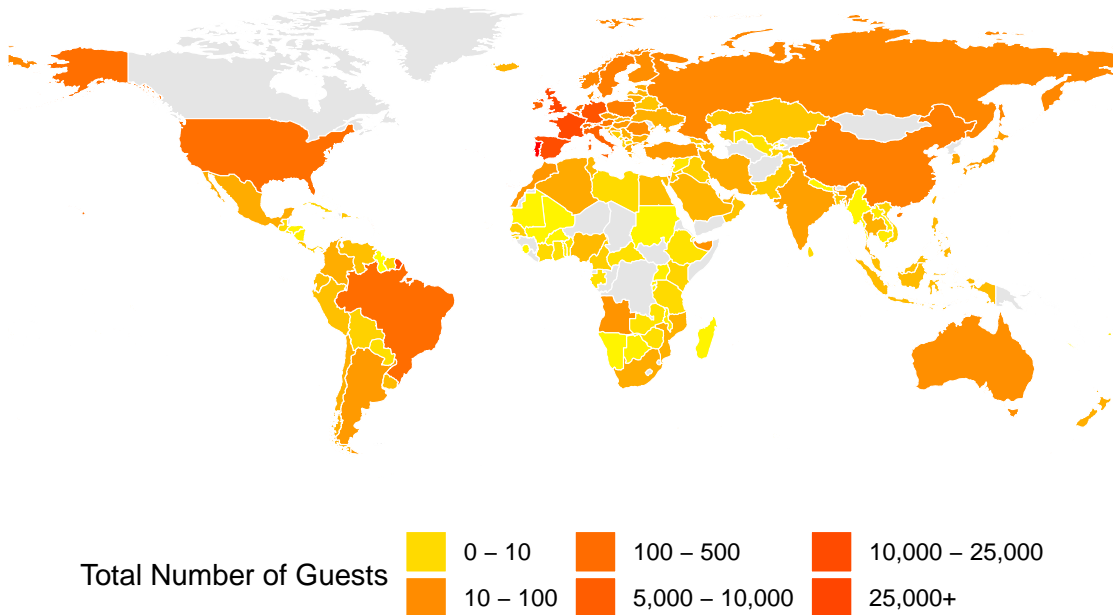
hotel_bookings_map <- left_join(world, hotel_bookings_agg, by = c("iso_a3" = "country"))

ggplot(hotel_bookings_map) +
  geom_sf(aes(fill = total_guests), color = "white") +
  scale_fill_gradient(
    low = "yellow",
    high = "red",
    na.value = "grey90",
    trans = "log",
    breaks = c(0, 10, 1000, 5000, 10000, 20000, 25000),
    labels = c("0", "0 - 10", "10 - 100", "100 - 500", "5,000 - 10,000", "10,000 - 25,000", "25,000+"),
```

```

    guide = guide_legend(title = "Total Number of Guests")
  ) +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    axis.text = element_blank(),
    axis.title = element_blank(),
    axis.ticks = element_blank(),
    panel.border = element_blank(),
    panel.grid = element_blank()
  ) +
  coord_sf(xlim = c(-180, 180), ylim = c(-55, 90))

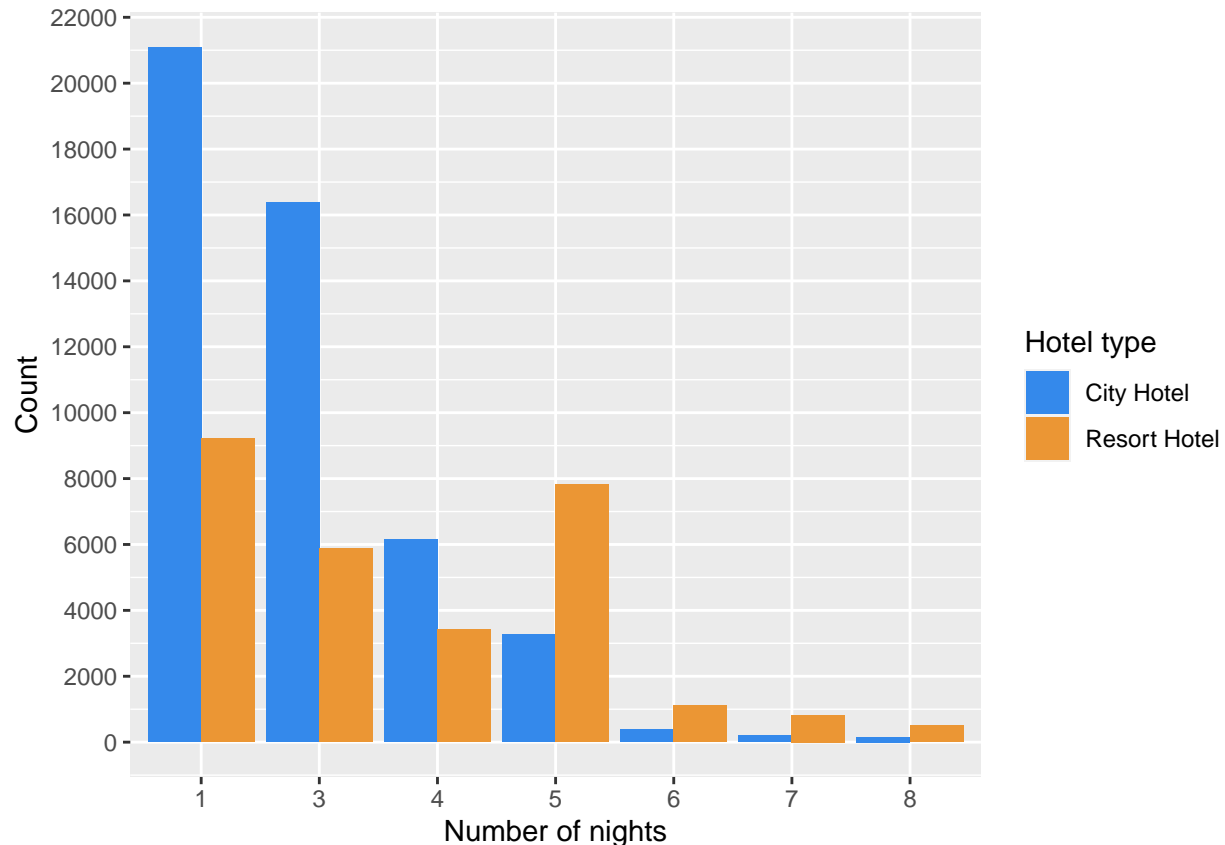
```



Kao što vidimo najintenzivnija boja, crvena, je u Portugalu i okolnim zemljama, što znači da najveći broj gostiju dolazi iz tog područja, kao što smo i već ranije zaključili. Kod zemalja s afričkog kontinenta, boja je žuta ili čak siva, što znači da od tamo dolazi najmanje gostiju. Sva siva područja označavaju zemlje iz kojih hoteli još nisu imali goste.

Trend odsjedanja za svaki tip hotela

Analizom podataka o broju noćenja dobivamo uvid u to koliko su noći gosti boravili u kojem hotelu. Ovaj graf pruža pregled raspodjele broja noćenja u hotelima za određene duljine boravka.



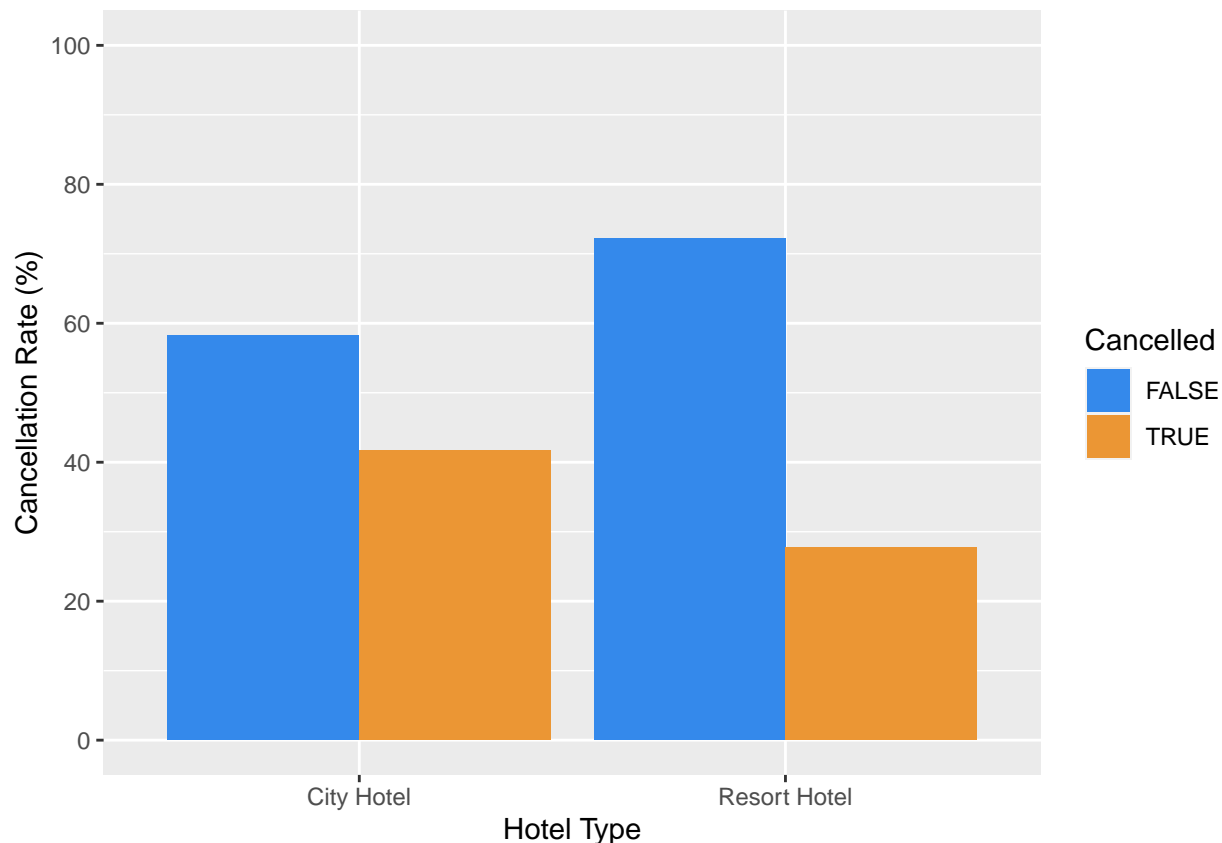
Utvrđili smo kako najviše odsjedanja u jednoj noći ima u gradskom hotelu što je i za očekivati jer su u gradu uvijek prisutni putnici koji traže smještaj, npr. tranzitni putnici ili poslovni ljudi. Također, vidimo da se ističe boravak od 5 noći u resortnom hotelu, što je poprilično razumno, s obzirom da je taj hotel namijenjen svojevrsnim odmorima, a ne samo za jednonoćno odsjedanje.

Postotci otkazivanja po tipu hotela

Na ovom grafu prikazat ćemo stope otkazivanja. Podaci su grupirani po tipu hotela, a zatim su izračunate stope otkazivanja u postotcima.

```
cancellation_rates <- hotel_bookings %>%
  group_by(hotel, is_canceled) %>%
  summarise(count = n()) %>%
  group_by(hotel) %>%
  mutate(cancellation_rate = count / sum(count) * 100)

ggplot(cancellation_rates, aes(x = hotel, y = cancellation_rate, fill = as.factor(is_canceled))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Hotel Type",
       y = "Cancellation Rate (%)",
       fill = "Cancelled") +
  scale_y_continuous(breaks = seq(0, 100, by = 20), limits = c(0, 100)) +
  scale_fill_manual(values = c("#3489eb", "#EB9634"))
```



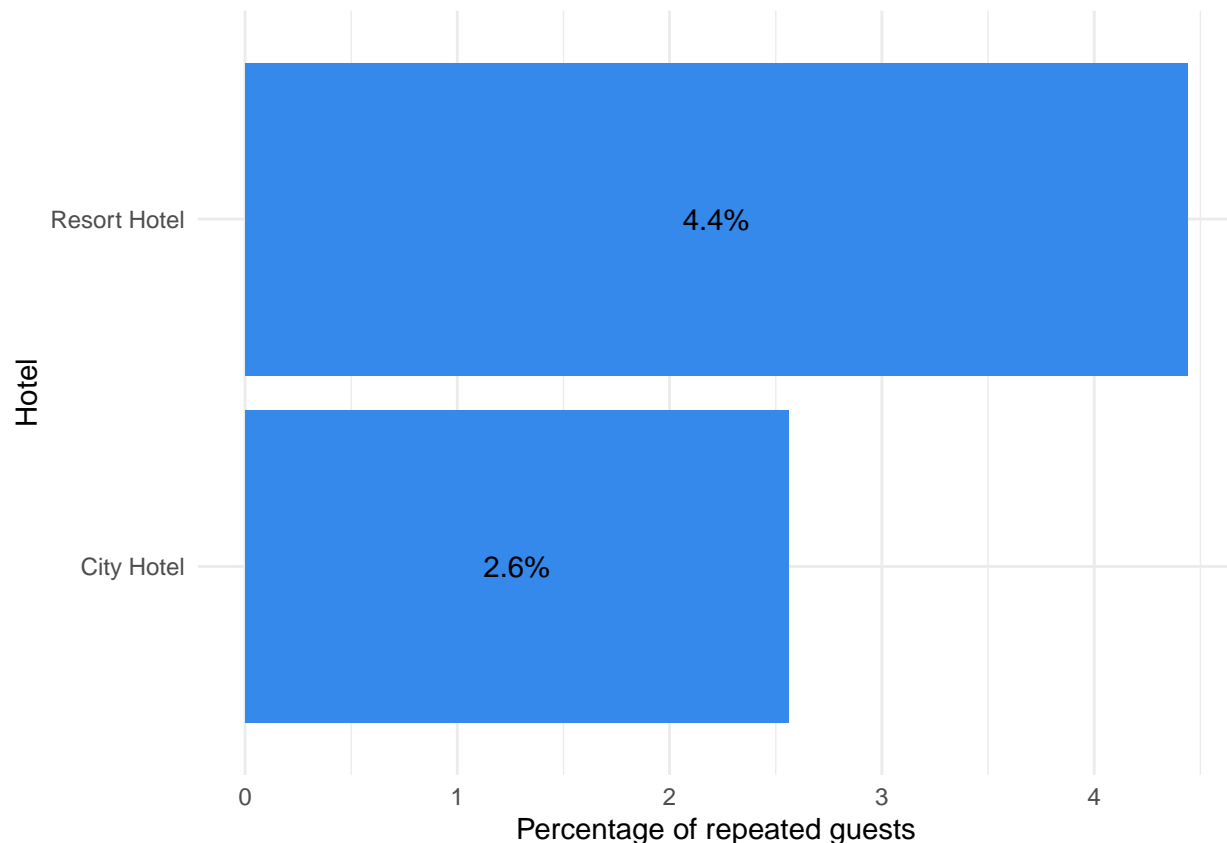
Kao što možemo vidjeti, u gradskom hotelu nije neka prevelika razlika u stopama otkazivanja i neotkazivanja; otprilike 40% ljudi koji rezerviraju hotel ga i otkazu dok oko 60% ljudi zapravo odsjeda u hotelu nakon što ga rezerviraju. Velika razlika može se primijetiti kod otkazivanja hotela u resortu. Manje od 30% ljudi koji rezerviraju hotel ga i otkazu što znači da više od 70% ljudi odsjeda u hotelu nakon što ga rezerviraju, što je i logično jer hotel se u resortu rezervira na više dana zbog odmora te je rezervacija najčešće pomno planirana.

Postotak ponovljenih gostiju po hotelu

Ovdje želimo prikazati postotak ponovljenih gostiju. Podaci su grupirani prema tipu hotela, a zatim su izračunati postotci gostiju koji su odsjedali u hotelu više puta.

```
repeated_guest_percentage <- hotel_bookings %>%
  group_by(hotel, is_repeated_guest) %>%
  summarise(count = n()) %>%
  group_by(hotel) %>%
  mutate(percentage = count / sum(count) * 100) %>%
  filter(is_repeated_guest == 1)
```

```
ggplot(repeated_guest_percentage, aes(x = percentage, y = hotel, fill = as.factor(is_repeated_guest))) +
  geom_bar(stat = "identity") +
  labs(x = "Percentage of repeated guests",
       y = "Hotel") +
  theme_minimal() +
  scale_fill_manual(values = c("#3489eb"), guide = "none") +
  geom_text(aes(label = sprintf("%.1f%%", percentage)),
            position = position_stack(vjust = 0.5))
```



Možemo zaključiti kako je veći postotak ponovljenih gostiju u resortnom hotelu; to je tako upravo zbog teze kako ti hoteli služe za odmor pa se čovjek u njima rado vraća ako mu je boravak bio ugodan, dok su oni gradski tranzitnije prirode.

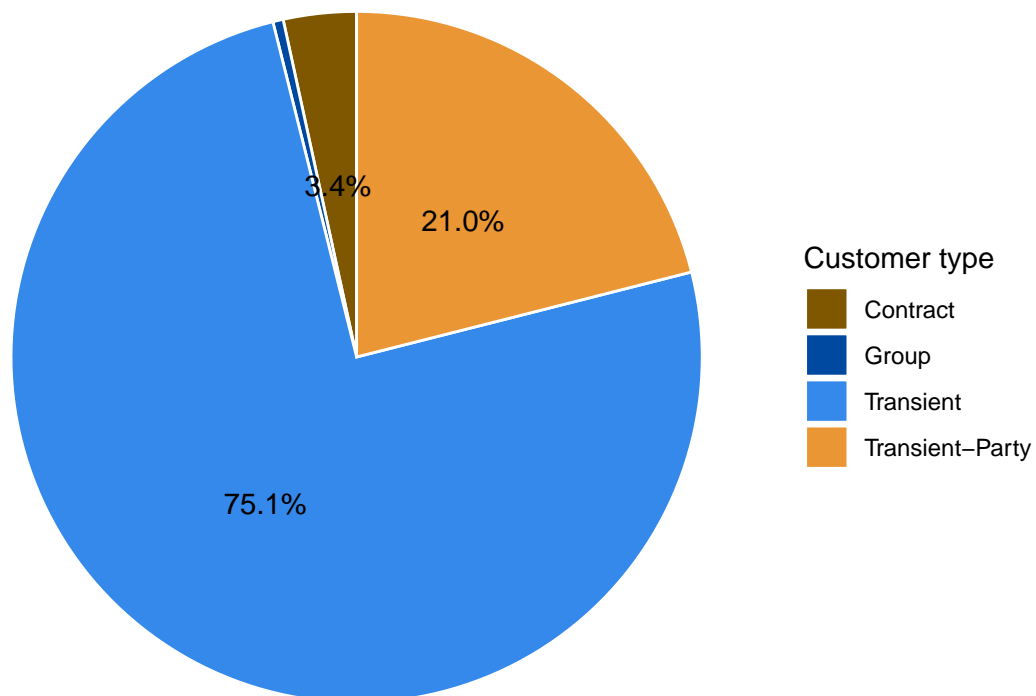
Distribucija tipova gostiju

Želimo saznati kakva je podjela različitih tipova gostiju u hotelima. Podaci su grupirani prema tipu gosta, a zatim su izračunati postotci za svaki tip.

```
customer_type_counts <- hotel_bookings %>%
  group_by(customer_type) %>%
  summarise(count = n()) %>%
  mutate(percentage = (count / sum(count)) * 100)

label_threshold <- 3

ggplot(customer_type_counts, aes(x = "", y = count, fill = customer_type)) +
  geom_bar(stat = "identity", position = "fill", color = "white", width = 1) +
  geom_text(data = filter(customer_type_counts, percentage >= label_threshold),
            aes(label = sprintf("%.1f%%", percentage), y = count / 2),
            position = position_fill(vjust = 0.5)) +
  coord_polar("y") +
  theme_void() +
  theme(legend.position = "right") +
  labs(fill = "Customer type") +
  scale_fill_manual(values = c("#7f5700", "#0049a0", "#3489eb", "#EB9634"))
```

Kao što možemo vidjeti, vodeći su gosti u prolazu (Transient guest), koji čine 75% svih gostiju, zatim sa 21% dolaze gosti koji rezerviraju hotel sami za sebe (Transient - Party) bilo to na recepciji ili online, ali ne preko agencije ili neke treće strane. S manjim postotcima tu su i gosti koji dolaze preko ugovora ili grupno.
