Visual analytics of hotel bookings data

Xavier de Moner

2024-11-29

Carreguem Llibreries i dades

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## 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
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.2.3
```

```
##
## Attaching package: 'lubridate'
```

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

```
# Carreguem les dades
x <- read.csv("hotel_bookings.csv", stringsAsFactors = FALSE)

# Convertim columnes de dates
x$dia <- as_date(paste0(x$arrival_date_year, "-", x$arrival_date_month, "-", x$arrival_date_d
ay_of_month))
```

2. Netegem dades

```
# eliminem outliers i registres inconsistents
x <- x %>%
  filter(adults < 10, children < 10, babies < 10, adr >= 0 & adr < 1000) %>%  # Outliers
  filter((adults + children + babies) > 0) %>%  # Almenys una persona
  mutate(children = ifelse(is.na(children), 0, children))
```

3 Preparació de dades per visualitzacions

3.1 ADR i cancel.lació x països

```
# ADR mitjà i taxa de cancel.lacions per país
dades_paisos <- x %>%
  group_by(country) %>%
  summarise(
    adr_avg = mean(adr, na.rm = TRUE),  # ADR mitjà
    cancel_rate = mean(as.numeric(is_canceled)) * 100,  # ½  % de cancel·lacions
    total_reserves = n()  # Total reserves
  ) %>%
  filter(total_reserves >= 100)  # països >= 100 reserves

# Exportem dades x país
write.csv(dades_paisos, "dades_paisos.csv", row.names = FALSE)
```

3.2 Comparem locals vs estranger

```
# Comparem amb portuguesos
dades_origen <- x %>%
  mutate(origin = ifelse(country == "PRT", "Resident", "Foreigner")) %>%
  group_by(origin) %>%
  summarise(
    adr_avg = mean(adr, na.rm = TRUE),  # ADR mitjà
    cancel_rate = mean(as.numeric(is_canceled)) * 100,  # % de cancel·lacions
    total_reserves = n(),  # Total reserves
    cancel_losses = sum(as.numeric(is_canceled) * adr, na.rm = TRUE)  # Pèrdues per cancel·la
cions
  )


# Exportem
write.csv(dades_origen, "dades_origen.csv", row.names = FALSE)
```

3.3 TEndències temporals reserves

```
# Reserves agrupades x dia
dades_temps <- x %>%
  group_by(dia) %>%
  summarise(
    reserves = n()
  )

# Exportem
write.csv(dades_temps, "dades_temps.csv", row.names = FALSE)
```

3.4 Tipologia de viatges

```
# Categoritzem viatgers
x <- x %>%
  mutate(
    tipo_viatge = case_when(
      stays_in_weekend_nights == 0 ~ "Work",
      stays_in_week_nights == 0 ~ "Weekend",
      stays_in_week_nights == 5 & stays_in_weekend_nights >= 3 ~ "Package",
      TRUE ~ "Leisure"
    )
  )

# Exportem
write.csv(x %>% select(dia, tipo_viatge, adr, is_canceled), "dades_viatges.csv", row.names =
FALSE)
```

4. Simulació de beneficis

```
# Escenaris de beneficis
projeccio_beneficis <- data.frame(
  Escenari = c("Actual", "Reduir cancel.lacions en 10%", "Augmentar ADR en 10%"),
  Benefici = c(
    sum(x$adr * (1 - as.numeric(x$is_canceled)), na.rm = TRUE),
    sum(x$adr * (1 - as.numeric(x$is_canceled) * 0.9), na.rm = TRUE),
    sum((x$adr * 1.1) * (1 - as.numeric(x$is_canceled)), na.rm = TRUE)
  )
)

# Exportem
write.csv(projeccio_beneficis, "projeccio_beneficis.csv", row.names = FALSE)
```

```
# Llegir els fitxers
dades_paisos <- read.csv("dades_paisos.csv", stringsAsFactors = FALSE)
countries_codes <- read.csv("countries-codes.csv", sep = ";", stringsAsFactors = FALSE)

# Combinar les dades basant-nos en el codi ISO3
dades_combined <- dades_paisos %>%
  left_join(countries_codes, by = c("country" = "ISO3.CODE")) %>%
  filter(!is.na(ONU.CODE)) %>%  # Filtrar països sense codi ONU.CODE
  mutate(ONU.CODE = as.integer(ONU.CODE)) %>%  # Assegurar que ONU.CODE és numèric
  select(country, adr_avg, cancel_rate, total_reserves, ONU.CODE) # Seleccionar columnes útil
s

# Escriure el fitxer final
write.csv(dades_combined, "dades_paisos_with_codes.csv", row.names = FALSE)
```

```
# Llibreries
library(dplyr)

# Carreguem dades originals
dades_paisos <- read.csv("dades_paisos.csv", stringsAsFactors = FALSE)
hotel_bookings <- read.csv("hotel_bookings.csv", stringsAsFactors = FALSE)

# Agreguem nits totals per país a partir de `hotel_bookings`
nits_totals <- hotel_bookings %>%
  group_by(country) %>%
  summarise(
    stays_week_nights = sum(stays_in_week_nights, na.rm = TRUE),  # Nits entre setmana
    stays_weekend_nights = sum(stays_in_weekend_nights, na.rm = TRUE)  # Nits de cap de setma
na
  )

# Unim dades agregades amb `dades_paisos`
dades_paisos_enriquides <- dades_paisos %>%
  left_join(nits_totals, by = "country") %>%
  mutate(
    total_nights = stays_week_nights + stays_weekend_nights,  # Total nits
    benefit_total = adr_avg * total_nights  # Benefici total
  ) %>%
  select(country, adr_avg, cancel_rate, total_reserves, stays_week_nights, stays_weekend_nigh
ts, total_nights, benefit_total)

# Fitxer final
write.csv(dades_paisos_enriquides, "dades_paisos_enriquides.csv", row.names = FALSE)
```

```
# Llegim el fitxer hotel_bookings.csv
hotel_data <- read.csv("hotel_bookings.csv", stringsAsFactors = FALSE)

# Convertim les dates en un format adequat
library(dplyr)
library(lubridate)

hotel_data <- hotel_data %>%
  mutate(date = ymd(paste(arrival_date_year, arrival_date_month, arrival_date_day_of_month, s
ep = "-"))) %>%
  filter(!is.na(country), adr > 0)  # Filtrar registres sense país o ADR invàlid

# Calculem  despesa total (ADR * dies d'estada)
hotel_data <- hotel_data %>%
  mutate(total_spent = adr * (stays_in_week_nights + stays_in_weekend_nights))

# Agrupem per país, any i mes, i sumar la despesa total
monthly_data <- hotel_data %>%
  mutate(year = year(date), month = month(date)) %>%
  group_by(country, year, month) %>%
  summarise(total_spent = sum(total_spent, na.rm = TRUE)) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'country', 'year'. You can override using
## the `.groups` argument.
```

```
# Filtrem països amb  taxa de cancel·lació < 26%
country_cancel_rates <- hotel_data %>%
  group_by(country) %>%
  summarise(cancel_rate = mean(is_canceled) * 100) %>%
  filter(cancel_rate < 26)

selected_countries <- country_cancel_rates$country

# Filtrem dades només per  països seleccionats
filtered_monthly_data <- monthly_data %>%
  filter(country %in% selected_countries)

# Fitxer final
write.csv(filtered_monthly_data, "filtered_monthly_data.csv", row.names = FALSE)
```

```
# Llibreries necessàries
library(dplyr)

# Carregar dades
hotel_data <- read.csv("hotel_bookings.csv", stringsAsFactors = FALSE)

# Convertim columnes de data en un format comprensible
library(lubridate)
hotel_data <- hotel_data %>%
  mutate(date = ymd(paste(arrival_date_year, arrival_date_month, arrival_date_day_of_month, s
ep = "-")))

# Filtrar per 2016 i Portugal
portugal_2016 <- hotel_data %>%
  filter(arrival_date_year == 2016, country == "PRT") %>%
  mutate(total_days = stays_in_weekend_nights + stays_in_week_nights, # Total dies
         average_daily_cost = adr) # Cost mitjà per dia

# Calcular pèrdues
portugal_2016_resum <- portugal_2016 %>%
  summarise(
    total_days = sum(total_days, na.rm = TRUE),
    average_daily_cost = mean(average_daily_cost, na.rm = TRUE),
    cancel_rate = mean(is_canceled) * 100, # Percentatge cancel·lacions
    total_loss = total_days * (cancel_rate / 100) * average_daily_cost # Pèrdues totals
  )

# Guardar les dades resultants en un CSV
write.csv(portugal_2016_resum, "portugal_2016_resum.csv", row.names = FALSE)

# Mostra el resultat a la consola
print(portugal_2016_resum)
```

```
##   total_days average_daily_cost cancel_rate total_loss
## 1      62805           89.28307    55.57547    3116351
```

```
# Filtrem països taxacancel·lació < 26% menys Portugal
countries_under_26 <- hotel_data %>%
  group_by(country) %>%
  summarise(cancel_rate = mean(is_canceled) * 100) %>%
  filter(cancel_rate < 26 & country != "PRT") %>%
  pull(country)

# Filtrem dades per a l'any 2016 i països
selected_2016 <- hotel_data %>%
  filter(arrival_date_year == 2016, country %in% countries_under_26) %>%
  mutate(
    total_days = stays_in_weekend_nights + stays_in_week_nights,  # Total de dies d'estada
    average_daily_cost = adr,  # Cost mitjà diari
    total_spent = adr * total_days,  # Despesa total
    actual_benefit = total_spent * (1 - is_canceled),  # Benefici actual
    improved_benefit = total_spent * (1 - is_canceled * 0.9),  # Benefici amb cancel·lacions
reduïdes un 10%
    improved_adr_benefit = (adr * 1.1) * total_days * (1 - is_canceled)  # Benefici amb un au
gment del 10% en ADR
  )

# Agregar resultats per resumir
selected_2016_summary <- selected_2016 %>%
  summarise(
    total_days = sum(total_days, na.rm = TRUE),  # Total de dies
    average_daily_cost = mean(adr, na.rm = TRUE),  # Cost mitjà diari
    actual_benefit = sum(actual_benefit, na.rm = TRUE),  # Benefici actual
    improved_benefit = sum(improved_benefit, na.rm = TRUE),  # Benefici millorat (cancel·laci
ons reduïdes)
    improved_adr_benefit = sum(improved_adr_benefit, na.rm = TRUE),  # Benefici millorat (ADR
augmentat)
    additional_benefit = improved_benefit - actual_benefit,  # Benefici addicional (cancel·la
cions)
    additional_benefit_adr = improved_adr_benefit - actual_benefit  # Benefici addicional (AD
R)
  )

# Exportem
write.csv(selected_2016_summary, "selected_countries_2016_summary.csv", row.names = FALSE)

print(selected_2016_summary)
```

```
##   total_days average_daily_cost actual_benefit improved_benefit
## 1     105773           103.0357        8144620          8388807
##   improved_adr_benefit additional_benefit additional_benefit_adr
## 1              8959082           244187.3                 814462
```

Resum dels arxius resultants

dades_paisos.csv: ADR mitjà, taxa de cancel·lacions i reserves totals per país. dades_origen.csv: Comparació locals vs estrangers (ADR, cancel·lacions, pèrdues). dades_temps.csv: Tendència temporal de reserves. dades_viatges.csv: Tipologia de viatges (work, weekend, package, leisure). projeccio_beneficis.csv: Beneficis projectats per diferents escenaris. “dades_paisos_with_codes.csv” “dades_paisos_enriquides.csv”