

R Programming Project- R Markdown

Alyssa Tran

2023-11-13

Hotel Booking Demand Dataset

This document focuses on **cleaning** and **filtering** a hotel dataset and creating **data visualizations**.

Load Packages

Install the required R packages.

```
install.packages(c("tidyverse", "skimr", "dplyr", "janitor", "ggplot2"))
```

Load the installed packages.

```
library(tidyverse)
library(skimr)
library(dplyr)
library(janitor)
library(ggplot2)
```

Import data

Import data from a .csv in the project folder called “hotel_bookings.csv” and save it as a data frame called hotel_bookings.

Dataset is sourced from the article Hotel Booking Demand Datasets and cleaned by Thomas Mock and Antoine Bichat.

```
hotel_bookings <- read_csv("hotel_bookings.csv")
```

```
## Rows: 119390 Columns: 32
## -- Column specification -----
## Delimiter: ","
## chr  (13): hotel, arrival_date_month, meal, country, market_segment, distrib...
## dbl  (18): is_canceled, lead_time, arrival_date_year, arrival_date_week_numb...
## date  (1): reservation_status_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Preview data

Use various functions to preview the structure of the dataset.

- head()
- str()
- glimpse()

- colnames()

```
colnames(hotel_bookings)
```

```
## [1] "hotel" "is_canceled"
## [3] "lead_time" "arrival_date_year"
## [5] "arrival_date_month" "arrival_date_week_number"
## [7] "arrival_date_day_of_month" "stays_in_weekend_nights"
## [9] "stays_in_week_nights" "adults"
## [11] "children" "babies"
## [13] "meal" "country"
## [15] "market_segment" "distribution_channel"
## [17] "is_repeated_guest" "previous_cancellations"
## [19] "previous_bookings_not_canceled" "reserved_room_type"
## [21] "assigned_room_type" "booking_changes"
## [23] "deposit_type" "agent"
## [25] "company" "days_in_waiting_list"
## [27] "customer_type" "adr"
## [29] "required_car_parking_spaces" "total_of_special_requests"
## [31] "reservation_status" "reservation_status_date"
```

Manipulating data

Arrange the data by lead time, focusing on bookings made far in advance.

```
arrange(hotel_bookings, lead_time)
```

```
## # A tibble: 119,390 x 32
##   hotel      is_canceled lead_time arrival_date_year arrival_date_month
##   <chr>      <dbl>      <dbl>      <dbl> <chr>
## 1 Resort Hotel      0      0      2015 July
## 2 Resort Hotel      0      0      2015 July
## 3 Resort Hotel      0      0      2015 July
## 4 Resort Hotel      0      0      2015 July
## 5 Resort Hotel      0      0      2015 July
## 6 Resort Hotel      0      0      2015 July
## 7 Resort Hotel      0      0      2015 July
## 8 Resort Hotel      0      0      2015 July
## 9 Resort Hotel      0      0      2015 July
## 10 Resort Hotel      0      0      2015 July
## # i 119,380 more rows
## # i 27 more variables: arrival_date_week_number <dbl>,
## #   arrival_date_day_of_month <dbl>, stays_in_weekend_nights <dbl>,
## #   stays_in_week_nights <dbl>, adults <dbl>, children <dbl>, babies <dbl>,
## #   meal <chr>, country <chr>, market_segment <chr>,
## #   distribution_channel <chr>, is_repeated_guest <dbl>,
## #   previous_cancellations <dbl>, previous_bookings_not_canceled <dbl>, ...
```

Calculating data

Calculate the mean, maximum, and minimum lead time.

```
hotel_summary <-
  hotel_bookings %>%
  group_by(hotel) %>%
  summarise(average_lead_time=mean(lead_time),
```

```
min_lead_time=min(lead_time),
max_lead_time=max(lead_time))
```

Check the new data set.

```
head(hotel_summary)
```

```
## # A tibble: 2 x 4
##   hotel          average_lead_time min_lead_time max_lead_time
##   <chr>          <dbl>          <dbl>          <dbl>
## 1 City Hotel      110.              0             629
## 2 Resort Hotel    92.7              0             737
```

Cleaning data

Combine the arrival month and year into one column.

```
new_hotel_bookings_df <- hotel_bookings %>%
  unite(arrival_month_year, c("arrival_date_month", "arrival_date_year"), sep = " ")
```

Rename the variable 'hotel' to be named 'hotel_type' for clarity.

```
new_hotel_bookings_df %>%
  rename(hotel_type = hotel)
```

```
## # A tibble: 119,390 x 31
##   hotel_type  is_canceled lead_time arrival_month_year arrival_date_week_num~1
##   <chr>          <dbl>    <dbl> <chr>                                <dbl>
## 1 Resort Hotel      0      342 July 2015                                27
## 2 Resort Hotel      0      737 July 2015                                27
## 3 Resort Hotel      0       7 July 2015                                27
## 4 Resort Hotel      0      13 July 2015                                27
## 5 Resort Hotel      0      14 July 2015                                27
## 6 Resort Hotel      0      14 July 2015                                27
## 7 Resort Hotel      0       0 July 2015                                27
## 8 Resort Hotel      0       9 July 2015                                27
## 9 Resort Hotel      1      85 July 2015                                27
## 10 Resort Hotel     1      75 July 2015                                27
## # i 119,380 more rows
## # i abbreviated name: 1: arrival_date_week_number
## # i 26 more variables: arrival_date_day_of_month <dbl>,
## #   stays_in_weekend_nights <dbl>, stays_in_week_nights <dbl>, adults <dbl>,
## #   children <dbl>, babies <dbl>, meal <chr>, country <chr>,
## #   market_segment <chr>, distribution_channel <chr>, is_repeated_guest <dbl>,
## #   previous_cancellations <dbl>, previous_bookings_not_canceled <dbl>, ...
```

Bar Charts and Annotating

Save minimum and maximum date variables as annotations to include in bar chart.

```
mindate <- min(new_hotel_bookings_df$arrival_month_year)
maxdate <- max(new_hotel_bookings_df$arrival_month_year)
```

Create a bar chart using ggplot2, this compares market segments for hotel bookings. It is faceted by hotel type, and labels provide additional information about the data and the plot.

```
ggplot(data = new_hotel_bookings_df) +
  geom_bar(mapping = aes(x = market_segment, fill=market_segment)) +
  facet_wrap(~hotel) +
  labs(title="Comparison of market segments by hotel type",
       caption=paste0("Data from: ", mindate, " to ", maxdate),
       x="Market Segment",
       y="Number of Bookings") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

