Hilton International Hotels Case Study

Teresiah Karumba

2023-07-07

Loading required packages

```
pacman::p_load(tidyverse, dplyr, openxlsx, janitor, ggplot2, lubridate,latexpdf)
```

Working Directory

```
setwd("C:/Users/teresiah.karumba/[redacted]/BI Automations 20

getwd()
```

```
## [1] "C:/Users/teresiah.karumba/[redacted]/BI Automations 2
```

Loading the Dataset

```
hotel_df <- read.csv("hotel_bookings.csv (1)/hotel_bookings.csv") %>% clean_names()
names(hotel_df)
```

```
##  [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"
```

Data Structure

```
# Data structure

class(hotel_df) #returns the class attribute of our data
```

```
## [1] "data.frame"
```

```
glimpse(hotel_df) #checking the data types of our columns
```

```
## Rows: 119,390
## Columns: 32
## $ hotel                         <chr> "Resort Hotel", "Resort Hotel", "Resort~
## $ is_canceled                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, ~
## $ lead_time                     <int> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, ~
## $ arrival_date_year             <int> 2015, 2015, 2015, 2015, 2015, 2015, 201~
## $ arrival_date_month            <chr> "July", "July", "July", "July", "July",~
## $ arrival_date_week_number      <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,~
## $ arrival_date_day_of_month     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ stays_in_weekend_nights       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ stays_in_week_nights          <int> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, ~
## $ adults                        <int> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ children                      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ babies                        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ meal                          <chr> "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             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_cancellations        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_bookings_not_canceled <int> 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               <int> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ deposit_type                  <chr> "No Deposit", "No Deposit", "No Deposit~
## $ agent                         <chr> "NULL", "NULL", "NULL", "304", "240", "~
## $ company                       <chr> "NULL", "NULL", "NULL", "NULL", "NULL",~
## $ days_in_waiting_list          <int> 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   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ total_of_special_requests     <int> 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> "2015-07-01", "2015-07-01", "2015-07-02~
```

```
head(hotel_df) #Having a view of our data set.
```

```
##          hotel is_canceled lead_time arrival_date_year arrival_date_month
## 1 Resort Hotel           0       342              2015               July
## 2 Resort Hotel           0       737              2015               July
## 3 Resort Hotel           0         7              2015               July
```

```
## 4 Resort Hotel           0        13              2015               July
## 5 Resort Hotel           0        14              2015               July
## 6 Resort Hotel           0        14              2015               July
##   arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights
## 1                       27                         1                      0
## 2                       27                         1                      0
## 3                       27                         1                      0
## 4                       27                         1                      0
## 5                       27                         1                      0
## 6                       27                         1                      0
##   stays_in_week_nights adults children babies meal country market_segment
## 1                    0      2        0      0   BB     PRT         Direct
## 2                    0      2        0      0   BB     PRT         Direct
## 3                    1      1        0      0   BB     GBR         Direct
## 4                    1      1        0      0   BB     GBR      Corporate
## 5                    2      2        0      0   BB     GBR      Online TA
## 6                    2      2        0      0   BB     GBR      Online TA
##   distribution_channel is_repeated_guest previous_cancellations
## 1               Direct                0                       0
## 2               Direct                0                       0
## 3               Direct                0                       0
## 4            Corporate                0                       0
## 5                TA/TO                0                       0
## 6                TA/TO                0                       0
##   previous_bookings_not_canceled reserved_room_type assigned_room_type
## 1                              0                  C                  C
## 2                              0                  C                  C
## 3                              0                  A                  C
## 4                              0                  A                  A
## 5                              0                  A                  A
## 6                              0                  A                  A
##   booking_changes deposit_type agent company days_in_waiting_list customer_type
## 1               3   No Deposit  NULL    NULL                    0     Transient
## 2               4   No Deposit  NULL    NULL                    0     Transient
## 3               0   No Deposit  NULL    NULL                    0     Transient
## 4               0   No Deposit   304    NULL                    0     Transient
## 5               0   No Deposit   240    NULL                    0     Transient
## 6               0   No Deposit   240    NULL                    0     Transient
##   adr required_car_parking_spaces total_of_special_requests reservation_status
## 1   0                           0                         0          Check-Out
## 2   0                           0                         0          Check-Out
## 3  75                           0                         0          Check-Out
## 4  75                           0                         0          Check-Out
## 5  98                           0                         1          Check-Out
## 6  98                           0                         1          Check-Out
##   reservation_status_date
## 1              2015-07-01
## 2              2015-07-01
## 3              2015-07-02
## 4              2015-07-02
## 5              2015-07-03
## 6              2015-07-03
```

```r
dim(hotel_df) #shape of the data #119390 rows of data #32 rows of data
```

```
## [1] 119390     32
```

```r
summary(hotel_df) #Summary of our data set
```

```
##     hotel            is_canceled       lead_time   arrival_date_year
##  Length:119390      Min.   :0.0000   Min.   :  0   Min.   :2015
##  Class :character   1st Qu.:0.0000   1st Qu.: 18   1st Qu.:2016
##  Mode  :character   Median :0.0000   Median : 69   Median :2016
##                     Mean   :0.3704   Mean   :104   Mean   :2016
##                     3rd Qu.:1.0000   3rd Qu.:160   3rd Qu.:2017
##                     Max.   :1.0000   Max.   :737   Max.   :2017
##
##  arrival_date_month arrival_date_week_number arrival_date_day_of_month
##  Length:119390      Min.   : 1.00            Min.   : 1.0
##  Class :character   1st Qu.:16.00            1st Qu.: 8.0
##  Mode  :character   Median :28.00            Median :16.0
##                     Mean   :27.17            Mean   :15.8
##                     3rd Qu.:38.00            3rd Qu.:23.0
##                     Max.   :53.00            Max.   :31.0
##
##  stays_in_weekend_nights stays_in_week_nights     adults
##  Min.   : 0.0000         Min.   : 0.0         Min.   : 0.000
##  1st Qu.: 0.0000         1st Qu.: 1.0         1st Qu.: 2.000
##  Median : 1.0000         Median : 2.0         Median : 2.000
##  Mean   : 0.9276         Mean   : 2.5         Mean   : 1.856
##  3rd Qu.: 2.0000         3rd Qu.: 3.0         3rd Qu.: 2.000
##  Max.   :19.0000         Max.   :50.0         Max.   :55.000
##
##     children           babies              meal             country
##  Min.   : 0.0000   Min.   : 0.000000   Length:119390      Length:119390
##  1st Qu.: 0.0000   1st Qu.: 0.000000   Class :character   Class :character
##  Median : 0.0000   Median : 0.000000   Mode  :character   Mode  :character
##  Mean   : 0.1039   Mean   : 0.007949
##  3rd Qu.: 0.0000   3rd Qu.: 0.000000
##  Max.   :10.0000   Max.   :10.000000
##  NA's   :4
##  market_segment     distribution_channel is_repeated_guest
##  Length:119390      Length:119390        Min.   :0.00000
##  Class :character   Class :character     1st Qu.:0.00000
##  Mode  :character   Mode  :character     Median :0.00000
##                                          Mean   :0.03191
##                                          3rd Qu.:0.00000
##                                          Max.   :1.00000
##
##  previous_cancellations previous_bookings_not_canceled reserved_room_type
##  Min.   : 0.00000       Min.   : 0.0000                Length:119390
##  1st Qu.: 0.00000       1st Qu.: 0.0000                Class :character
##  Median : 0.00000       Median : 0.0000                Mode  :character
##  Mean   : 0.08712       Mean   : 0.1371
##  3rd Qu.: 0.00000       3rd Qu.: 0.0000
```

```
##  Max.   :26.00000       Max.   :72.0000
##
##  assigned_room_type booking_changes   deposit_type          agent
##  Length:119390      Min.   : 0.0000   Length:119390      Length:119390
##  Class :character   1st Qu.: 0.0000   Class :character   Class :character
##  Mode  :character   Median : 0.0000   Mode  :character   Mode  :character
##                     Mean   : 0.2211
##                     3rd Qu.: 0.0000
##                     Max.   :21.0000
##
##    company          days_in_waiting_list customer_type           adr
##  Length:119390      Min.   :  0.000      Length:119390      Min.   :  -6.38
##  Class :character   1st Qu.:  0.000      Class :character   1st Qu.:  69.29
##  Mode  :character   Median :  0.000      Mode  :character   Median :  94.58
##                     Mean   :  2.321                         Mean   : 101.83
##                     3rd Qu.:  0.000                         3rd Qu.: 126.00
##                     Max.   :391.000                         Max.   :5400.00
##
##  required_car_parking_spaces total_of_special_requests reservation_status
##  Min.   :0.00000             Min.   :0.0000            Length:119390
##  1st Qu.:0.00000             1st Qu.:0.0000            Class :character
##  Median :0.00000             Median :0.0000            Mode  :character
##  Mean   :0.06252             Mean   :0.5714
##  3rd Qu.:0.00000             3rd Qu.:1.0000
##  Max.   :8.00000             Max.   :5.0000
##
##  reservation_status_date
##  Length:119390
##  Class :character
##  Mode  :character
##
##
##
##
```

```r
#character columns to factor

hotel_df <- hotel_df %>% mutate(hotel = as.factor(hotel),
                      meal = as.factor(meal),
                      country = as.factor(country),
                      market_segment = as.factor(market_segment),
                      distribution_channel = as.factor(distribution_channel),
                      reserved_room_type = as.factor(reserved_room_type),
                      assigned_room_type = as.factor(assigned_room_type),
                      deposit_type = as.factor(deposit_type),
                      customer_type = as.factor(customer_type),
                      reservation_status = as.factor(reservation_status),
                      is_canceled = as.factor(is_canceled),
                      is_repeated_guest = as.factor(is_repeated_guest),
                      arrival_date_month = as.factor(arrival_date_month)
                     )

#Attaching levels to the factor variables
hotel_df$is_canceled <- factor(hotel_df$is_canceled, labels = c("No","Yes"))
```

```
hotel_df$is_repeated_guest <- factor(hotel_df$is_repeated_guest, labels = c("No","Yes"))


#Merging Columns

# hotel_df <- hotel_df %>% unite(arrival_date, arrival_date_year,arrival_date_month,arrival_date_day_of
```

Data Cleaning

```
listMissingColumns <- colnames(hotel_df)[ apply(hotel_df, 2, anyNA)]
print(listMissingColumns) #Children column has missing data
```

Data Completeness

```
## [1] "children"
```

```
hotel_df %>% filter(is.na(children)) -> miss_children

#omit any data with NA

na.omit(hotel_df)-> hotel_df

dim(hotel_df) #shape of the data #119386 rows of data #32 rows of data
```

```
## [1] 119386     32
```

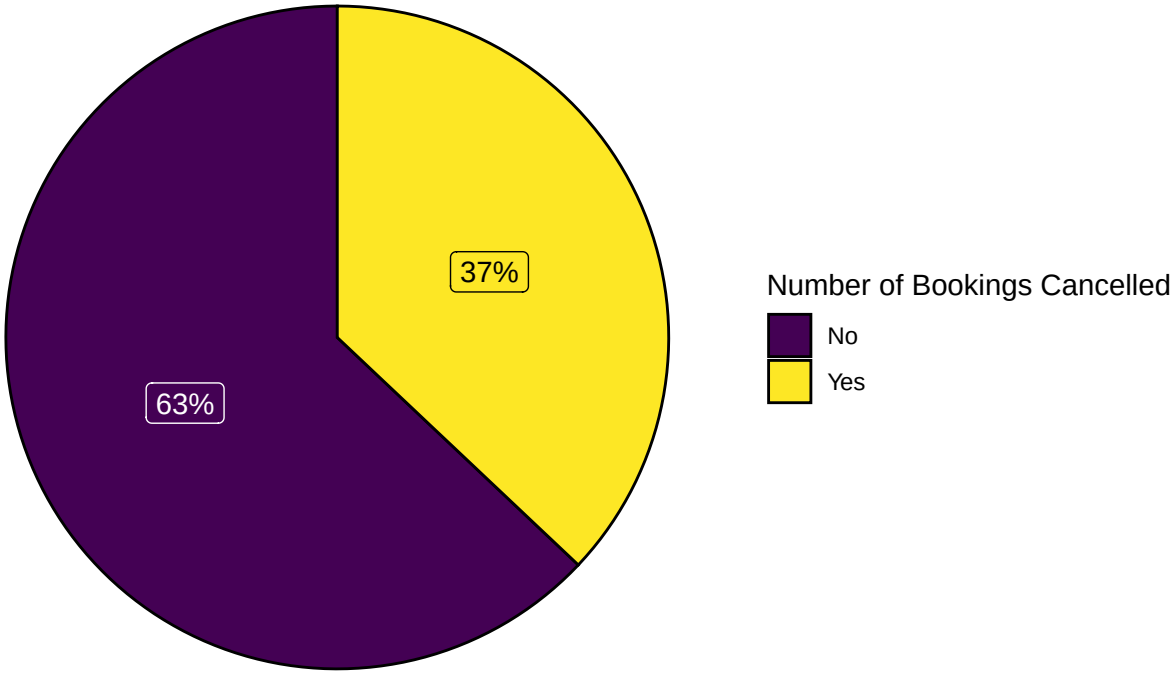
Data Analysis

```
#Creating the proportions

cancelled_bookings <- hotel_df %>%
  group_by(is_canceled) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n`/sum(`n`)) %>%
  mutate(labels = scales::percent(perc))


print(cancelled_bookings)
```

How many bookings were cancelled?

```
## # A tibble: 2 x 4
##   is_canceled     n  perc labels
##   <fct>       <int> <dbl> <chr>
## 1 No          75166 0.630 63%
## 2 Yes         44220 0.370 37%
```

```
plot1 <- ggplot(cancelled_bookings, aes(x = "", y = perc, fill = is_canceled)) +
  geom_col(color = "black") +
  geom_label(aes(label = labels), colour = c("white", 1),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE) +
  guides(fill = guide_legend(title = "Number of Bookings Cancelled")) +
  scale_fill_viridis_d() +
  coord_polar(theta = "y") +
  theme_void()


print(plot1)
```



What was the booking ratio between resort hotels and city hotels?

```
#Creating the proportions

resort <- hotel_df %>% filter(hotel == "Resort Hotel")
city <- hotel_df %>% filter(hotel == "City Hotel")

resort_count <- nrow(resort)
city_count <- nrow(city)
```

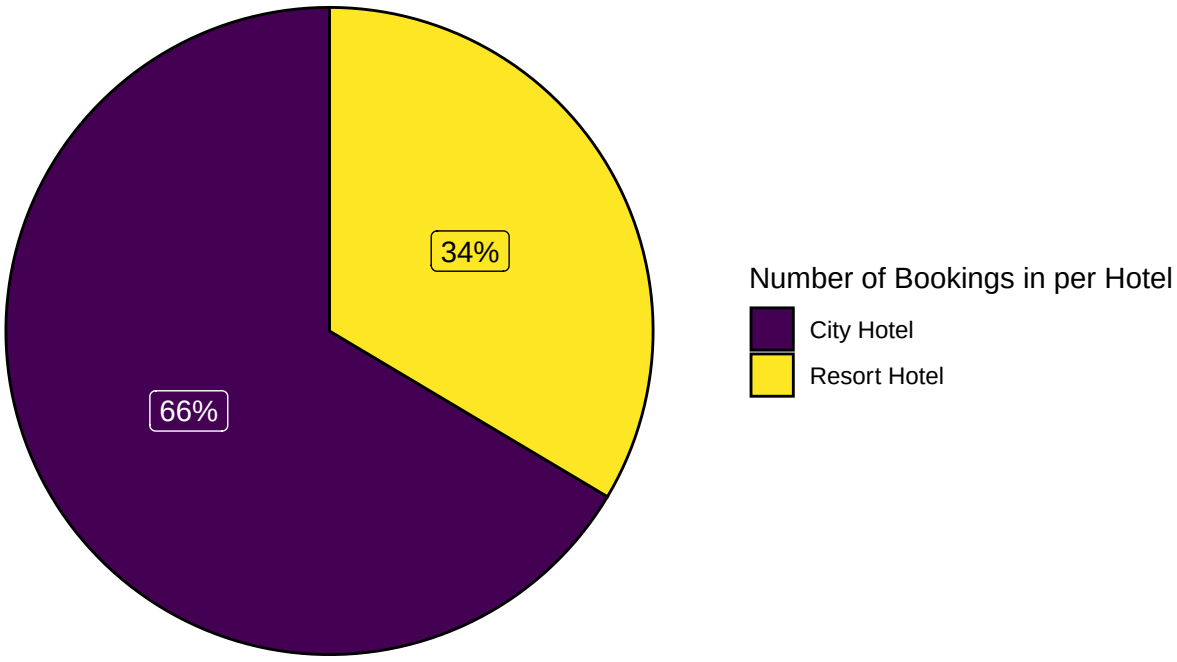
```r
ratio <- resort_count/city_count

# Print the ratio in ratio format
ratio_string <- sprintf("%d:%d", resort_count, city_count)

print(ratio_string)
```

```
## [1] "40060:79326"
```

```r
print(ratio)
```

```
## [1] 0.5050047
```

```r
hotel_ratio <- hotel_df %>%
  group_by(hotel) %>%
  count() %>%
  ungroup() %>%
  mutate(perc = `n`/sum(`n`)) %>%
  mutate(labels = scales::percent(perc))



plot2 <- ggplot(hotel_ratio, aes(x = "", y = perc, fill = hotel)) +
  geom_col(color = "black") +
  geom_label(aes(label = labels), colour = c("white", 1),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE) +
  guides(fill = guide_legend(title = "Number of Bookings in per Hotel")) +
  scale_fill_viridis_d() +
  coord_polar(theta = "y") +
  theme_void()


print(plot2)
```



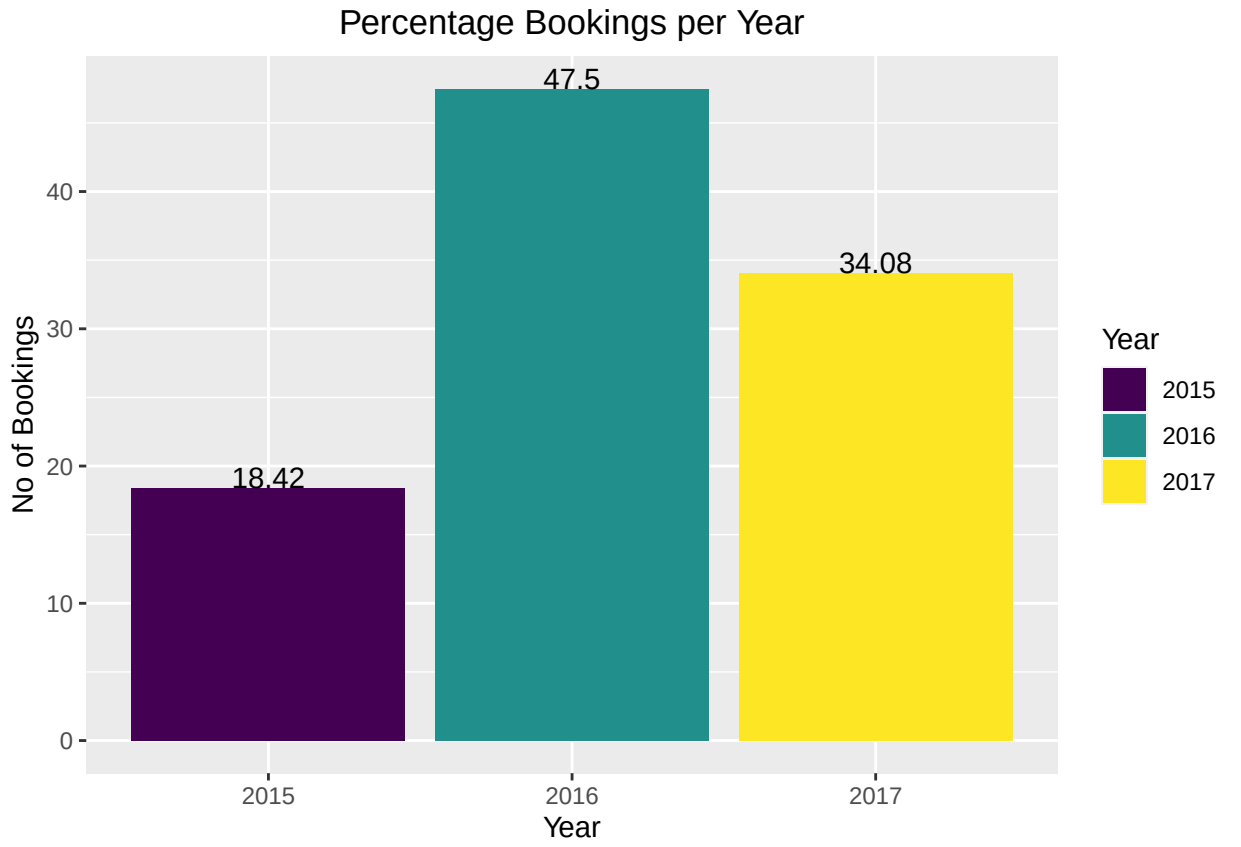
What was the percentage booking for each year?

```
#Creating the proportions

t1 <- table(hotel_df$arrival_date_year)

t2 <- round(prop.table(t1)*100, digits = 2)

print(t2)
```

```
##
##  2015  2016  2017
## 18.42 47.50 34.08
```

```
# Data frame with the percentages

booking_yrdf <- as.data.frame(t2)


names(booking_yrdf) <- c("Year", "Percentage")


booking_yrdf <-  booking_yrdf[order(booking_yrdf$Percentage, decreasing = TRUE),]
```

```r
plot3 <- ggplot(booking_yrdf, aes(x = Year, y = Percentage)) +
  geom_bar(stat = "identity", mapping = aes(x = Year, fill = Year)) +
  geom_text(aes(label = Percentage), vjust = 0, colour = "black") +
  scale_fill_viridis_d()


#adding Chart Title
plot3 <- plot3 + ggtitle("Percentage Bookings per Year") + theme(plot.title = element_text(hjust = 0.5))

plot3 <- plot3 + labs(y="No of Bookings")

plot3 <- plot3 + labs(x="Year")


print(plot3)
```



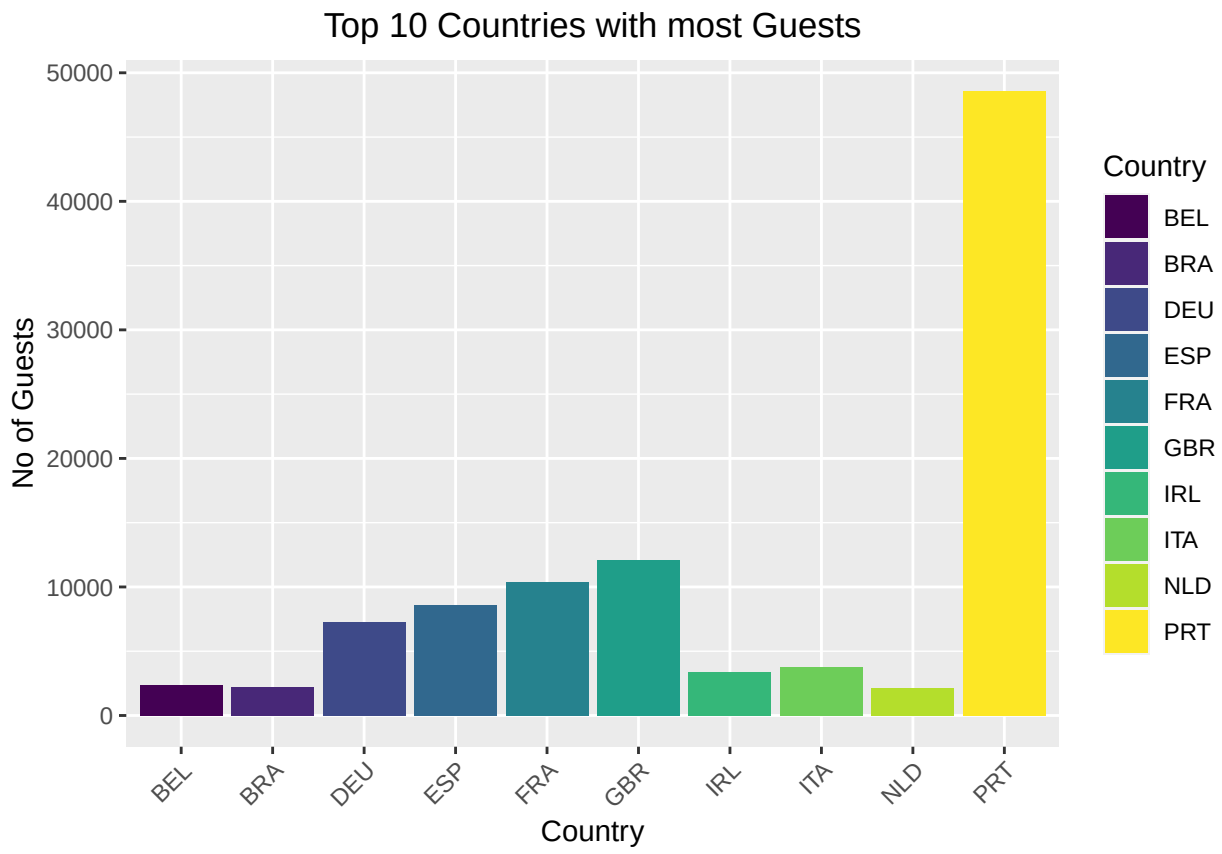
Which was the busiest month for hotels?

```r
#August was the busiest month for the hotels

t3 <- table(hotel_df$arrival_date_month)

# Data frame with the counts
```

```r
busy_month <- as.data.frame(t3)


names(busy_month) <- c("Month", "Tally")


busy_month <-  busy_month[order(busy_month$Tally, decreasing = TRUE),]


plot4 <- ggplot(busy_month, aes(x = Month, y = Tally)) +
  geom_bar(stat = "identity", mapping = aes(x = Month, fill = Month)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_viridis_d()


#adding Chart Title
plot4 <- plot4 + ggtitle("Bookings per Month") + theme(plot.title = element_text(hjust = 0.5))

plot4 <- plot4 + labs(y="No of Bookings")

plot4 <- plot4 + labs(x="Month")



print(plot4)
```

Most Guest come from which Country?

```
#Most guests came from Portugal (PRT)

t4 <- table(hotel_df$country)

guest_df <- as.data.frame(t4)

names(guest_df) <- c("Country", "Count")

guest_df <-  guest_df[order(guest_df$Count, decreasing = TRUE),]

top_10_countries <- head(guest_df, n=10)


plot5 <- ggplot(top_10_countries, aes(x = reorder(Country, - Count) , y = Count)) +
  geom_bar(stat = "identity", mapping = aes(x = Country, fill = Country)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
   scale_fill_viridis_d()

#adding Chart Title
plot5 <- plot5 + ggtitle("Top 10 Countries with most Guests") + theme(plot.title = element_text(hjust =

plot5 <- plot5 + labs(y="No of Guests")

plot5 <- plot5 + labs(x="Country")


print(plot5)
```



How long do most people stay in hotels?

```
#Creating a column to calculate total number of days stayed

hotel_df <- hotel_df %>%
  mutate(total_days = stays_in_weekend_nights + stays_in_week_nights)


avg_stay <- mean(hotel_df$total_days) %>% round(0) %>% as.character()

print(paste("Average stay at the hotel is", avg_stay, "days"))
```

```
## [1] "Average stay at the hotel is 3 days"
```

```
med_stay <- median(hotel_df$total_days) %>% as.character()

print(paste("Most people stay at the hotel for", med_stay, "days"))
```

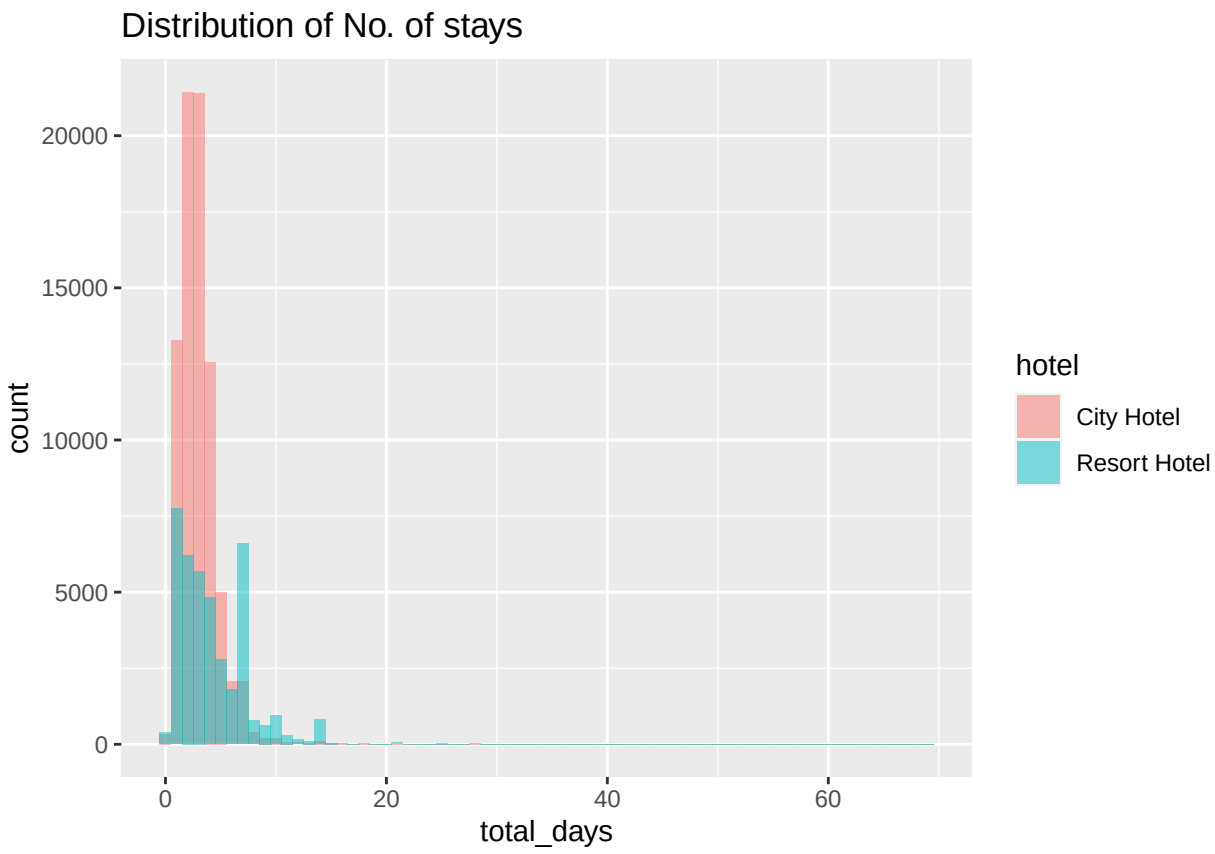
```
## [1] "Most people stay at the hotel for 3 days"
```

```
plot6 <- ggplot(hotel_df, aes(x=total_days,fill=hotel)) +
  geom_histogram(binwidth=1, alpha=0.5, position = 'identity') +
```

```
  labs(title  = "Distribution of No. of stays",
       fill = "hotel")
  scale_fill_viridis_d()
```

```
## <ggproto object: Class ScaleDiscrete, Scale, gg>
##     aesthetics: fill
##     axis_order: function
##     break_info: function
##     break_positions: function
##     breaks: waiver
##     call: call
##     clone: function
##     dimension: function
##     drop: TRUE
##     expand: waiver
##     get_breaks: function
##     get_breaks_minor: function
##     get_labels: function
##     get_limits: function
##     guide: legend
##     is_discrete: function
##     is_empty: function
##     labels: waiver
##     limits: NULL
##     make_sec_title: function
##     make_title: function
##     map: function
##     map_df: function
##     n.breaks.cache: NULL
##     na.translate: TRUE
##     na.value: NA
##     name: waiver
##     palette: function
##     palette.cache: NULL
##     position: left
##     range: environment
##     rescale: function
##     reset: function
##     scale_name: viridis_d
##     train: function
##     train_df: function
##     transform: function
##     transform_df: function
##     super:  <ggproto object: Class ScaleDiscrete, Scale, gg>
```

```
print(plot6)
```



The most booked accomodation type

```
t5 <- table(hotel_df$customer_type)

accomodation_df <- as.data.frame(t5)

names(accomodation_df) <- c("Customer_Type", "Count")

accomodation_df[order(accomodation_df$Count, decreasing = TRUE),]
```

```
##     Customer_Type Count
## 3       Transient 89613
## 4 Transient-Party 25120
## 1        Contract  4076
## 2           Group   577
```

```
plot7 <- ggplot(accomodation_df, aes(x = Customer_Type , y = Count)) +
  geom_bar(stat = "identity", mapping = aes(x = Customer_Type, fill = Customer_Type)) +
   scale_fill_viridis_d()

#adding Chart Title
plot7 <- plot7 + ggtitle("Customer Type Distribution") + theme(plot.title = element_text(hjust = 0.5))
```

```r
plot7 <- plot7 + labs(y="No of Guests")

plot7 <- plot7 + labs(x="Customer Type")


print(plot7)
```