Hotel Bookings

Raja Syam

2023-05-15

Scenario

As a junior data analyst for a hotel booking company, i have been creating visualizations in `R` with the `ggplot2` package to share insights about my data with stakeholders. After creating a series of visualizations using `ggplot()`, `ggplot2` aesthetics, and filters, my stakeholder asks me to add annotations to my visualizations to help explain mine findings in a presentation. Luckily, `ggplot2` has annotation functions built in.

Step 1: Import the data

In this step i import data into R Studio with the code below to read in the file 'hotel_bookings.csv' into a data frame:

```
hotel_bookings <- read.csv("C:/Users/OWNER/Desktop/R/bookings_df/hotel_bookings.csv")
```

Step 2: Look the data

By now, i pretty familiar with this data set. But i can refresh my memory with the `head()` and `colnames()` functions. With two code below to get at a sample of the data and also preview all the column names:

```
head(hotel_bookings)
```

```
##          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
```

```
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"
```

Step 3: Install and load the 'ggplot2' and 'tidyverse' packages (optional)

This step may take a few minutes, before i visualize my data i load and install the packages with `ggplot()` function and `tidyverse`. To install and load `ggplot2` i need run this code:

and also to install and load `tidyverse` i need to run this code:

```
library(tidyverse)
```

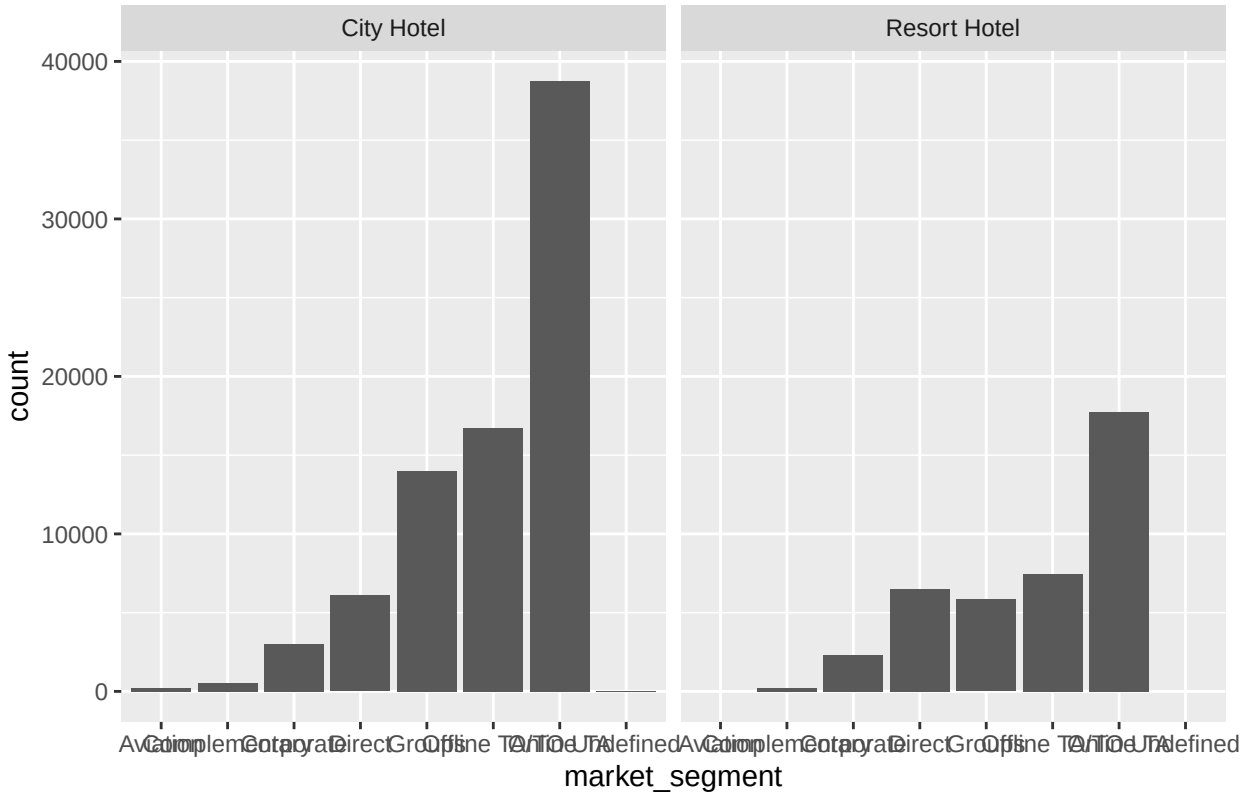
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Step 4: Annotating chart

Stakeholder tells me that they would like me to share they visualization breaking down payment type by city because it will help inform how the company targets promotions in the future. They ask me to create a cleaned and labeled version and save it as a .png file for them to include in a presentation.

Firstly, i make the code like this:

```
ggplot(data = hotel_bookings) +
  geom_bar(mapping = aes(x = market_segment)) +
  facet_wrap(~hotel) +
  labs(title="")
```



The stakeholder also want to add another detail about what time period this data covers. To do this, i need to find out when the data is from. i realize i can use the `min()` function on the year column in the data:

```
min(hotel_bookings$arrival_date_year)
```

```
## [1] 2015
```

And the `max()` function:

```
max(hotel_bookings$arrival_date_year)
```

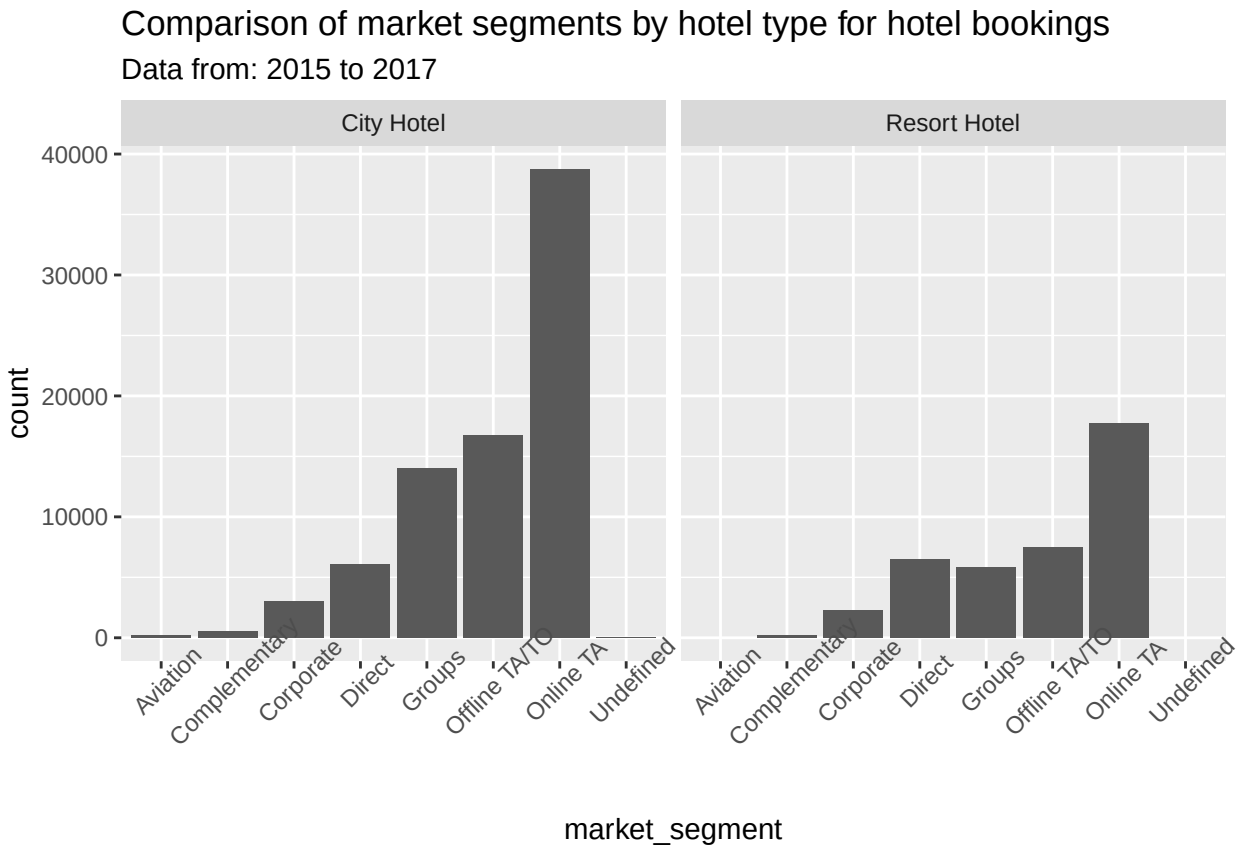
```
## [1] 2017
```

But, beside of that i will need to save them as variables in order to easily use them in my labeling; the following code creates two of those variables:

```
mindate <- min(hotel_bookings$arrival_date_year)
maxdate <- max(hotel_bookings$arrival_date_year)
```

Now, i will add in a subtitle using `subtitle=` in the `labs()` function. Then, i can use the `paste0()` function to use my newly-created variables in my labels. This is really handy, because if the data gets updated and there is more recent data added, i don't have to change the code below because the variables are dynamic:

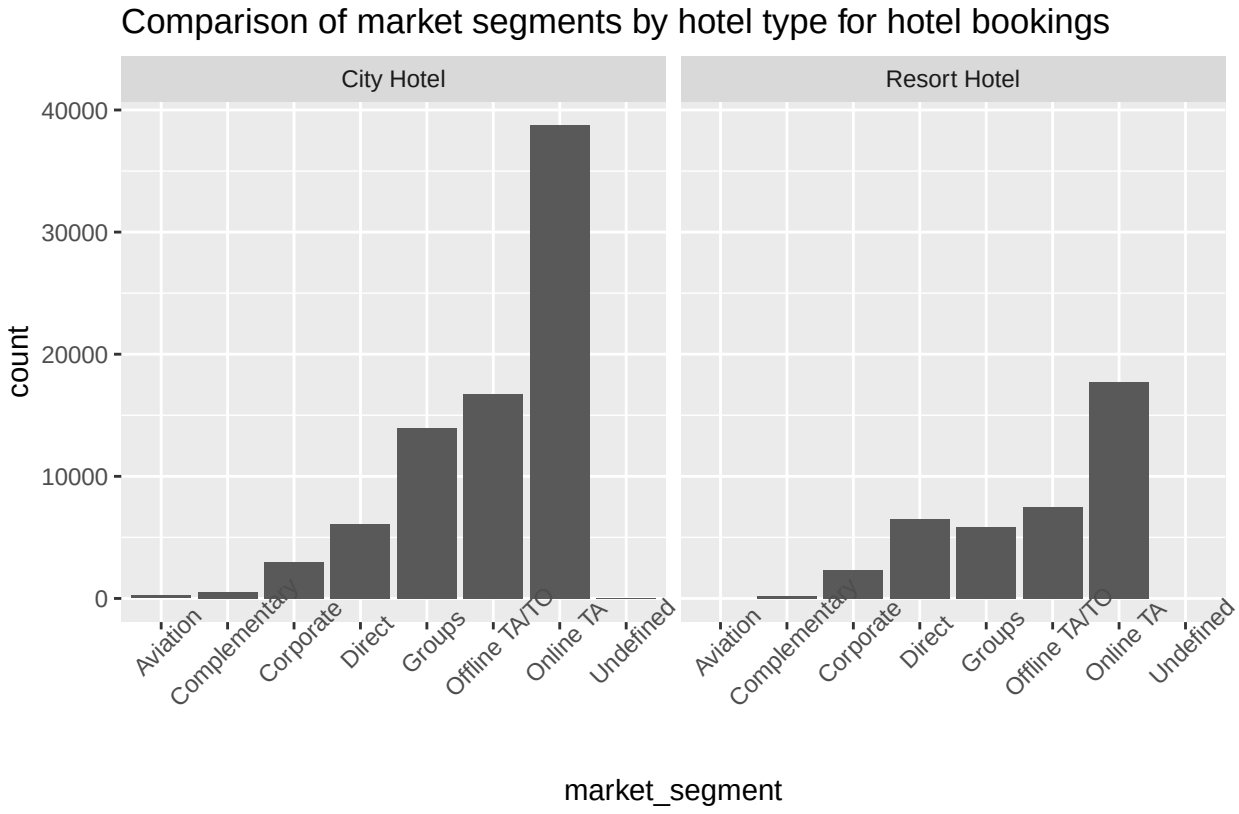
```
ggplot(data = hotel_bookings) +
  geom_bar(mapping = aes(x = market_segment)) +
  facet_wrap(~hotel) +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title="Comparison of market segments by hotel type for hotel bookings",
       subtitle=paste0("Data from: ", mindate, " to ", maxdate))
```



This code will add the subtitle 'Data from: 2015 to 2017' underneath the title i added earlier to the chart.

I realize that this chart is displaying the technical details a little too prominently. I don't want that to be the second thing people notice during the presentation. I decide to switch the `subtitle` to a `caption` which will appear in the bottom right corner instead.

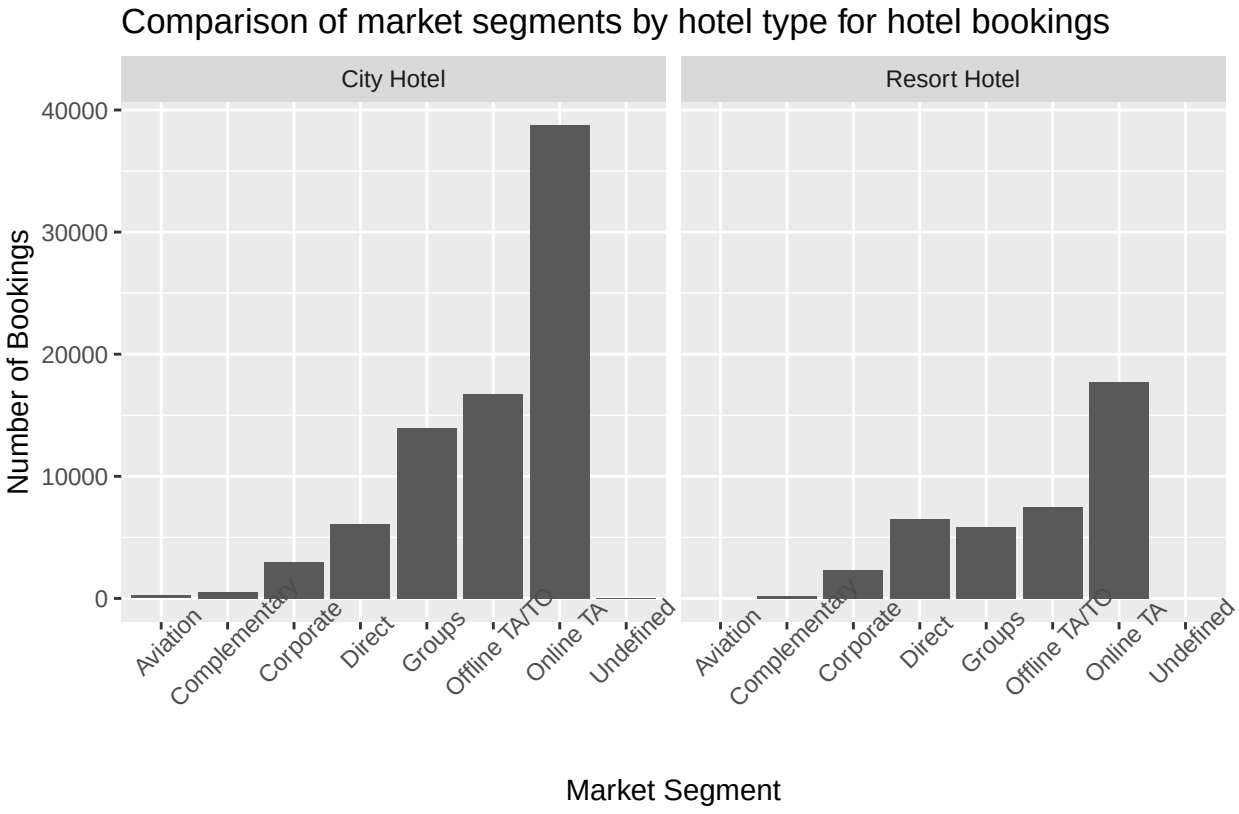
```
ggplot(data = hotel_bookings) +
  geom_bar(mapping = aes(x = market_segment)) +
  facet_wrap(~hotel) +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title="Comparison of market segments by hotel type for hotel bookings",
       caption=paste0("Data from: ", mindate, " to ", maxdate))
```



This code chunk makes a slight change to the visualization i created in the last code; now the "data from: 2015 to 2017" subtitle is in the bottom right corner.

Now i want to clean up the x and y axis labels to make sure they are really clear. To do that, i can add to the `labs()` function and use `x=` and `y=`. Feel free to change the text of the label and play around with it:

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



Now i have the data visualization from earlier, but now the x and y axis labels have been changed from 'market_segment' and 'count' to 'Market Segment' and 'Number of Bookings' so that the chart is clearer.

Step 5: Saving chart

Now, it's time to save what i just created so i can easily share with stakeholders. To save it i use `ggsave()` function, it will save the image as a 7x7 at the file path, and also to save that plot as a .png file named `hotel_booking_chart`, which makes it clear to stakeholders what the .png file contains. So this that code:

```
ggsave('hotel_booking_chart.png')
```

```
## Saving 6.5 x 4.5 in image
```