Hotel Booking Analysis

Michalis Loizos

2025-05-09

Introduction

This independent project, developed to showcase my data analytics skills, analyzes 119,390 hotel bookings from 2015 to 2017 to deliver hospitality insights. Using BigQuery SQL, R (tidyverse, ggplot2), and other tools, the project provides actionable insights.

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.4     v tidyr     1.3.1
## -- 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
```

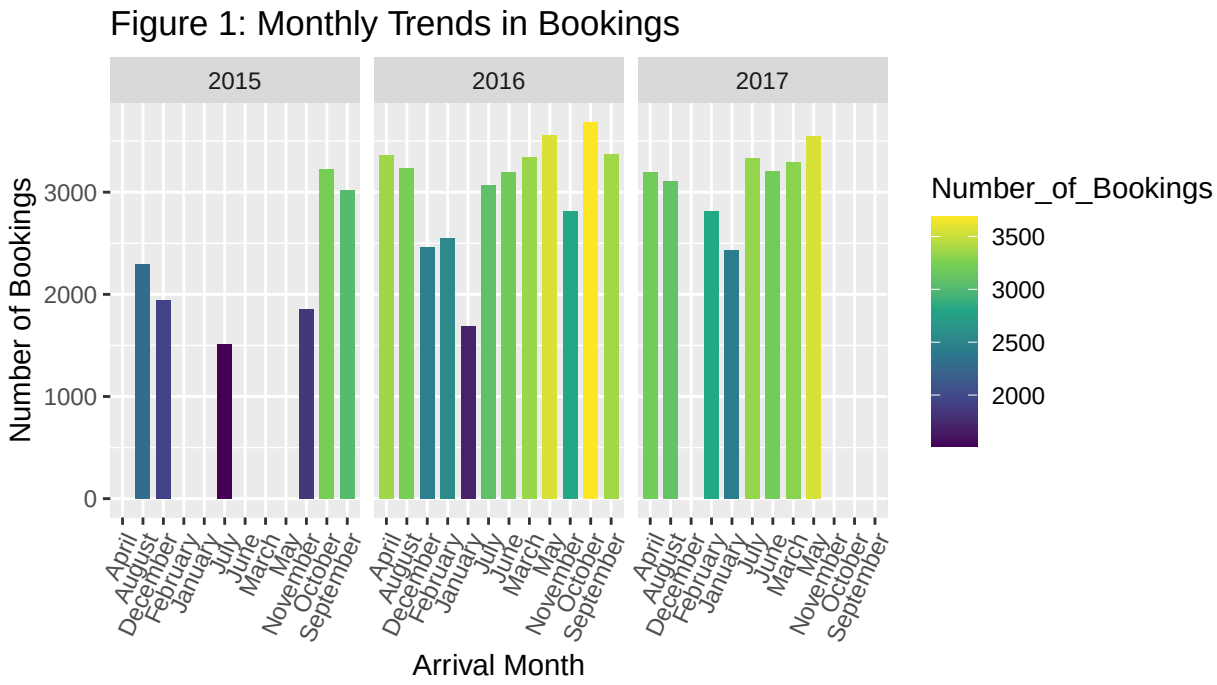
```
library(readxl)
```

I. Results on Bookings

Before analysis, I inspected the dataset to determine if cleaning was required, checking for inconsistencies or missing values. No major cleaning was necessary beyond handling missing data for 2015.

The analysis begins by examining monthly booking trends for the available time period, covering 2015, 2016, and 2017. The bookings per year per month are visualized in **Figure 1**. As expected, bookings peak during specific months.

```
hb1 <- read.csv("HOTEL_BOOKINGS_PER_YEAR_PER_MONTH.csv")
figure_1 <- ggplot(data = hb1) +
  geom_col(mapping = aes(x = Month_of_Arrival, y = Number_of_Bookings, fill = Number_of_Bookings), widt
  facet_wrap(~Year_of_Arrival) +
  labs(x = "Arrival Month", y = "Number of Bookings", title = "Figure 1: Monthly Trends in Bookings") +
  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c() +
  coord_fixed(ratio = 0.005)
figure_1
```

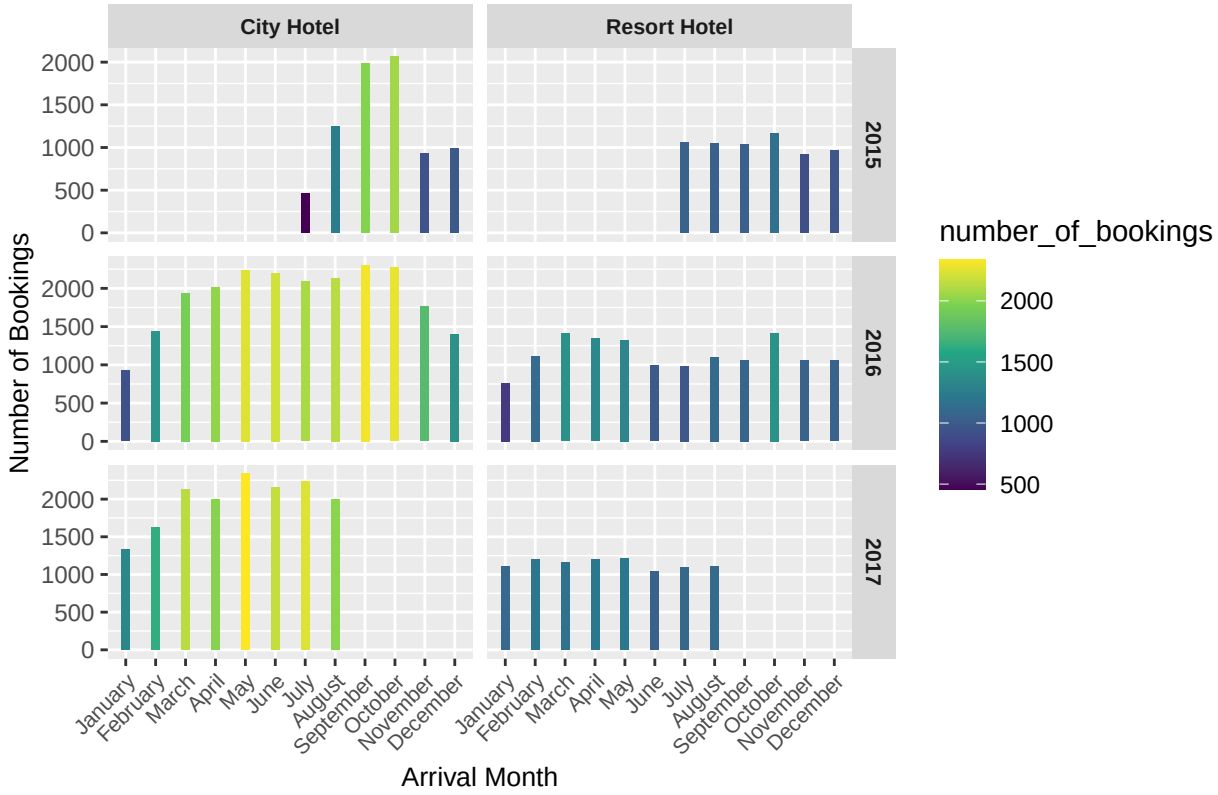



```
ggsave("Figure_1.png", plot = figure_1, width = 8, height = 5)
```

Next, the bookings were evaluated in terms of hotel type. Specifically, during the recorded data period, 38.5% of bookings were made in Resort Hotels, while 61.5% were for a City Hotel. This can be visualized in **Figure 2**.

```
hb2 <- read.csv("BOOKINGS_HOTEL_TYPE_PER_MONTH_PER_YEAR.csv")
figure_2 <- ggplot(data = hb2, aes(x = reorder(Month_of_Arrival, match(Month_of_Arrival, month.name)), y
  geom_col(width = 0.3) +
  scale_fill_viridis_c() +
  labs(x = "Arrival Month", y = "Number of Bookings", title = "Figure 2: Bookings per Hotel Type") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 8),
    axis.title = element_text(size = 10),
    strip.text = element_text(size = 8, face = "bold")
  ) +
  facet_grid(Year ~ Type_of_Hotel, scales = "free_y")
figure_2
```

Figure 2: Bookings per Hotel Type

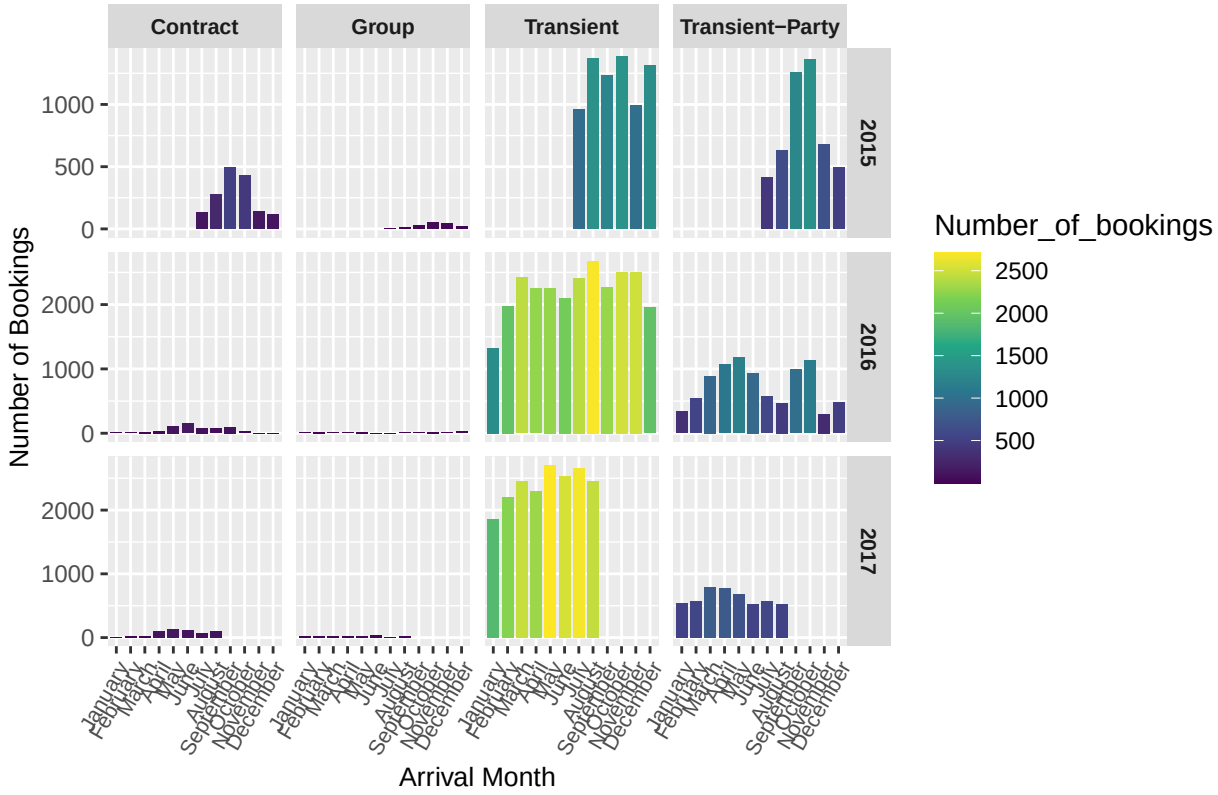



```
ggsave("Figure_2.png", plot = figure_2, width = 8, height = 5)
```

Bookings were then categorized by customer type: Transient, Transient-Party, Group, and Contract. Of the total non-canceled bookings, 70.6% were made by Transient customers (**Figure 3**).

```
hb3 <- read.csv("BOOKINGS_CUSTOMER_TYPE_PER_MONTH_PER_YEAR.csv")
figure_3 <- ggplot(data = hb3, aes(x = reorder(Month_of_arrival, match(Month_of_arrival, month.name)), 
  geom_col(width = 0.8) +
  scale_fill_viridis_c() +
  labs(x = "Arrival Month", y = "Number of Bookings", title = "Figure 3: Bookings by Customer Type") +
  theme(
    axis.text.x = element_text(angle = 60, hjust = 1, size = 8),
    axis.title = element_text(size = 10),
    strip.text = element_text(size = 8, face = "bold")
  ) +
  facet_grid(Year ~ Type_of_customer, scales = "free_y")
figure_3
```

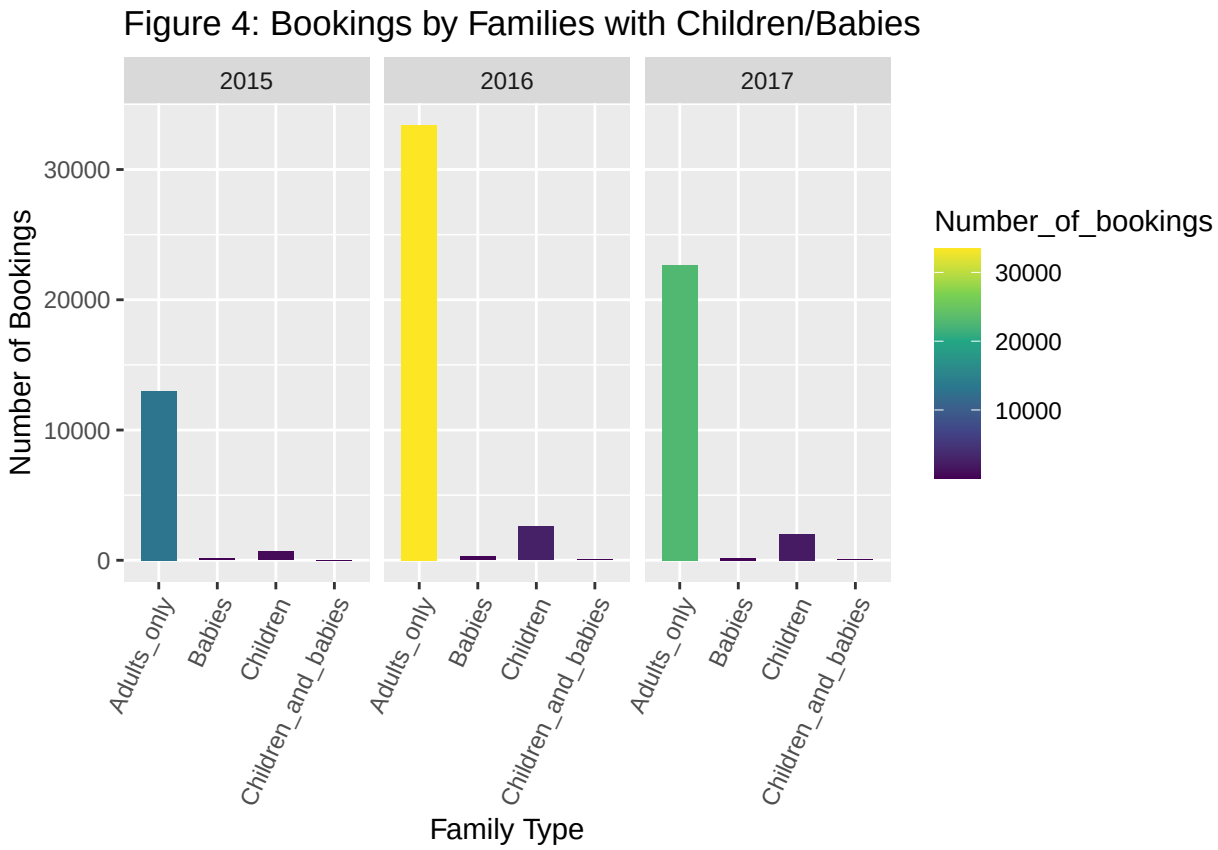
Figure 3: Bookings by Customer Type




```
ggsave("Figure_3.png", plot = figure_3, width = 8, height = 5)
```

I then determined what percentage of bookings were made by families with children and/or babies (**Figure 4**). Of the total bookings, 7.1% were from families with children, 0.8% were families with babies, and 0.1% were families with both children and babies.

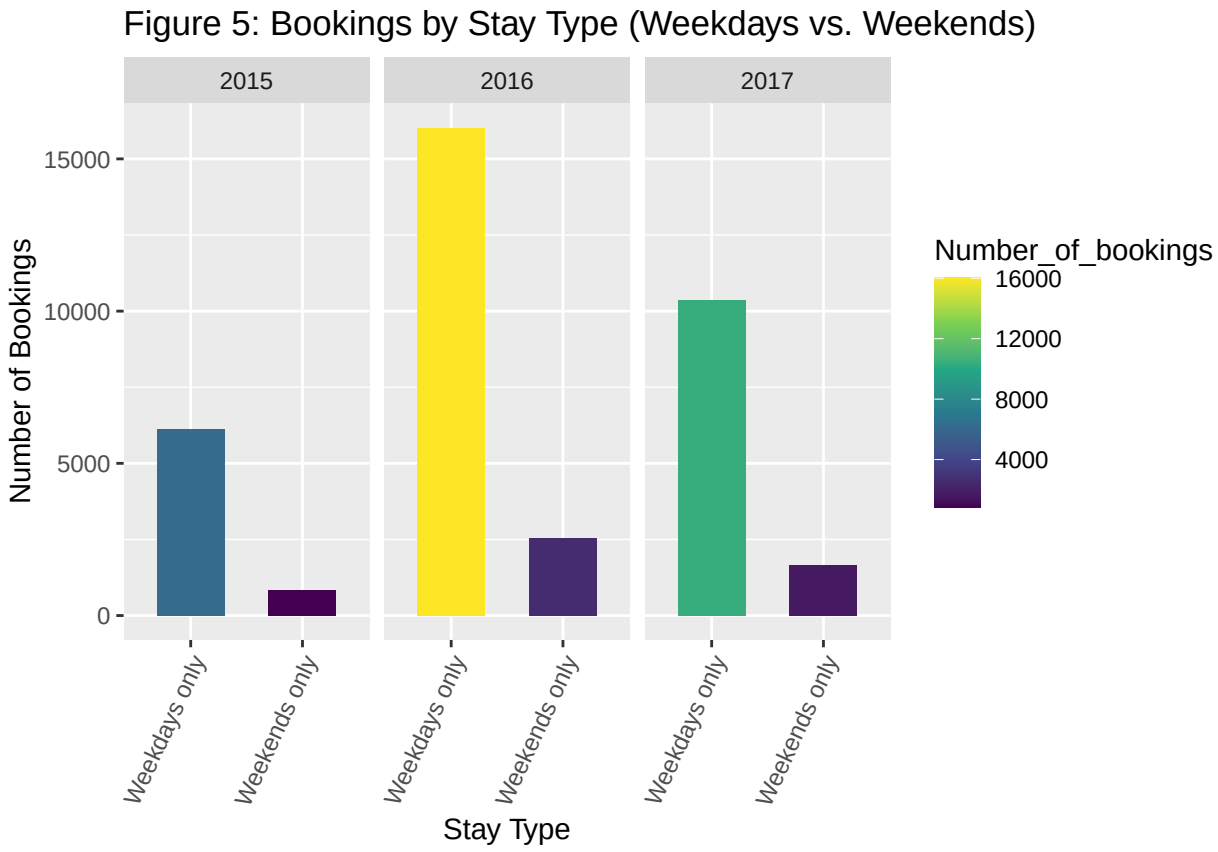
```
hb6 <- read_xlsx("BOOKINGS_MADE_CHILDREN_BABIES_OR_BOTH.xlsx")
figure_4 <- ggplot(data = hb6) +
  geom_col(mapping = aes(x = Category, y = Number_of_bookings, fill = Number_of_bookings), width = 0.6)
  facet_wrap(~Year) +
  labs(x = "Family Type", y = "Number of Bookings", title = "Figure 4: Bookings by Families with Childre
  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()
figure_4
```




```
ggsave("Figure_4.png", plot = figure_4, width = 8, height = 5)
```

I then analyzed whether bookings involved stays only on weekdays, weekends, or both, with results shown in **Figure 5**. Of the bookings, 42.7% stayed on weekdays only, 50.5% stayed during weekends, and 6.8% stayed during both.

```
hb8 <- read_xlsx("BOOKINGS_STAYED_WEEKEND_ONLY_OR_WEEKDAYS_ONLY.xlsx")
figure_5 <- ggplot(data = hb8) +
  geom_col(mapping = aes(x = Booking_type, y = Number_of_bookings, fill = Number_of_bookings), width = 
  facet_wrap(~Year) +
  labs(x = "Stay Type", y = "Number of Bookings", title = "Figure 5: Bookings by Stay Type (Weekdays vs
  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()
figure_5
```



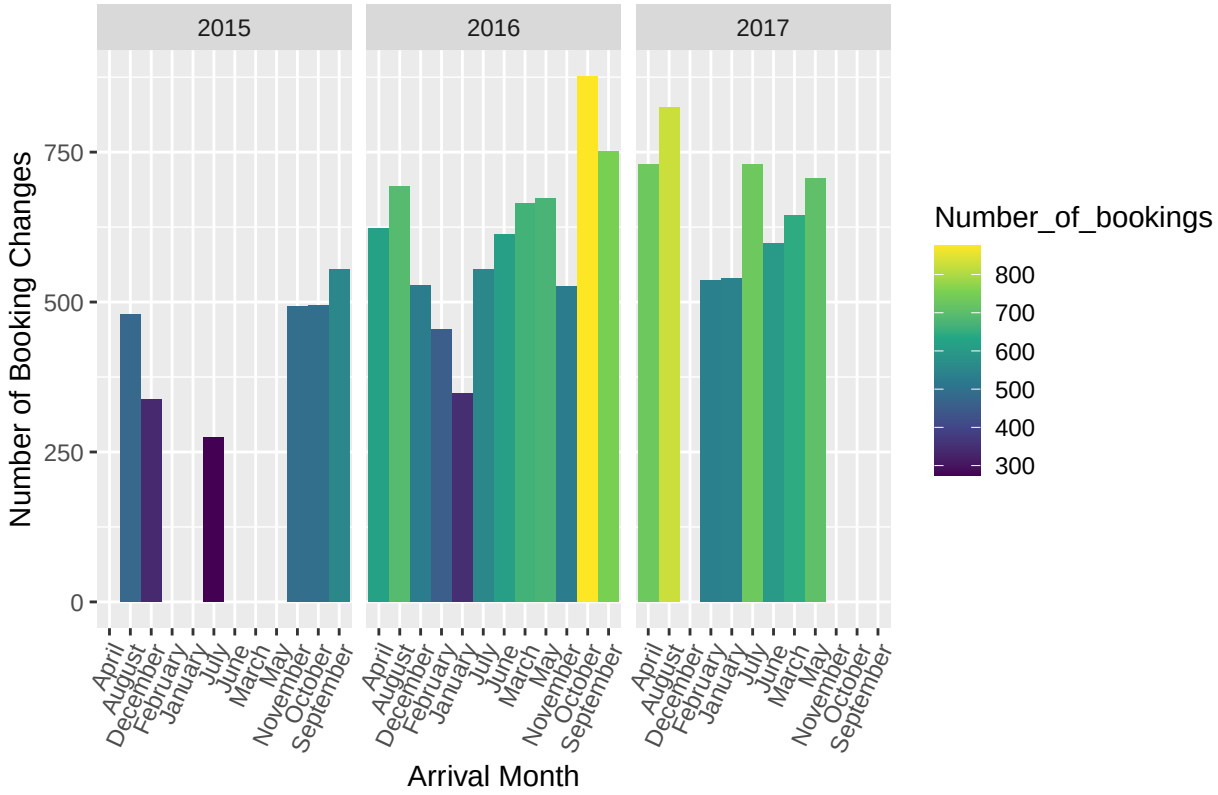

```
ggsave("Figure_5.png", plot = figure_5, width = 8, height = 5)
```

II. Booking Changes

In this section, I analyzed booking changes, starting by examining available data to observe any monthly trends. **Figure 6** shows no significant trends (sharp peaks or minimums). This suggests that booking changes were consistent throughout the years.

```
hb4 <- read.csv("BOOKING_CHANGES_PER_YEAR_PER_MONTH.csv")
figure_6 <- ggplot(data = hb4) +
  geom_col(mapping = aes(x = arrival_date_month, y = Number_of_bookings, fill = Number_of_bookings), wi
  facet_wrap(~arrival_date_year) +
  labs(x = "Arrival Month", y = "Number of Booking Changes", title = "Figure 6: Monthly Trends in Booki
  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()
figure_6
```

Figure 6: Monthly Trends in Booking Changes

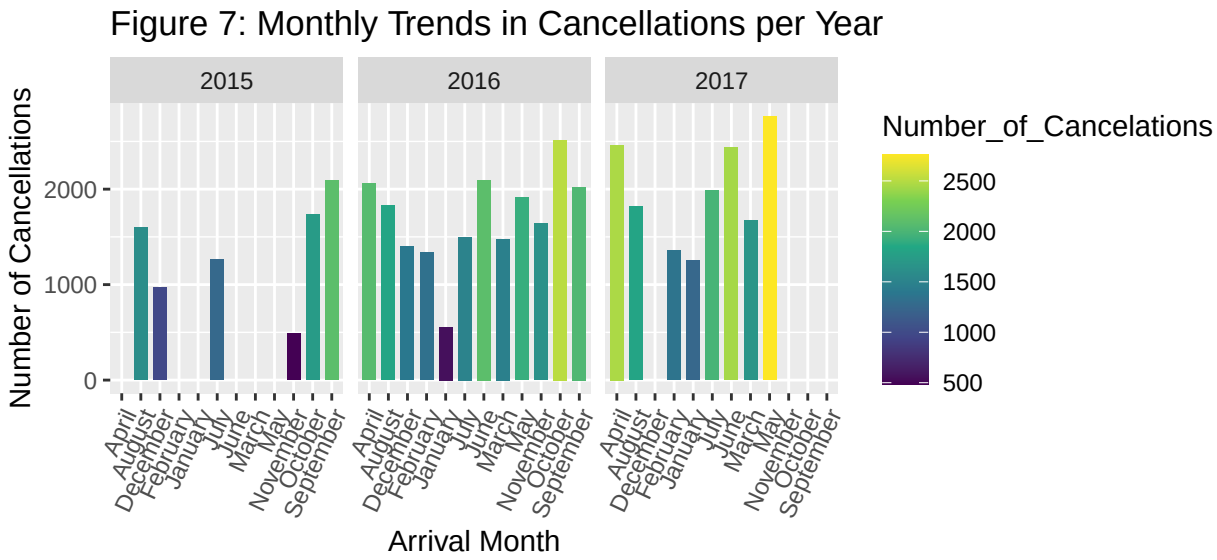



```
ggsave("Figure_6.png", plot = figure_6, width = 8, height = 5)
```

III. Cancellations

Finally, I investigated cancellations. As seen in **Figure 7**, there is no apparent trend in monthly cancellations (no sharp peaks or minimums). This consistency suggests the hotel can maintain steady cancellation rates.

```
hb10 <- read.csv("CANCELATIONS_PER_MONTH_PER_YEAR.csv")
figure_7 <- ggplot(data = hb10) +
  geom_col(mapping = aes(x = Month_of_Arrival, y = Number_of_Cancelations, fill = Number_of_Cancelations
  facet_wrap(~arrival_date_year) +
  labs(x = "Arrival Month", y = "Number of Cancellations", title = "Figure 7: Monthly Trends in Cancella
  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c() +
  coord_fixed(ratio = 0.005)
figure_7
```




```
ggsave("Figure_7.png", plot = figure_7, width = 8, height = 5)
```

IV. Limitations

Data from 2015 and 2017 are included in the report; however, since data for several months are missing for 2015 and some for 2017, no definite conclusions can be drawn for 2015 especially. Additionally, external factors (e.g., marketing campaigns, local events) were not considered in the analysis.

V. Conclusions

This project analyzed 119,390 hotel bookings from 2015 to 2017, uncovering key trends in bookings, cancellations, and customer behavior. Key findings include a peak in bookings during May for 2016 and 2017 and higher bookings for City Hotels compared to Resort Hotels.

This project showcases my skills in BigQuery SQL (data querying), R (tidyverse, ggplot2 for visualizations), Google Sheets (data cleaning), and Tableau (interactive maps), as well as my ability to deliver actionable insights.