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), Google Sheets, and Tableau, I uncover trends in bookings, cancellations, and customer behavior, providing data-driven business recommendations to optimize hotel operations.

```
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. I then used BigQuery SQL to retrieve the desired information into several CSV files, which were plotted in R, forming a data pipeline. The queries used to retrieve the corresponding sub-datasets are shown commented throughout the report.

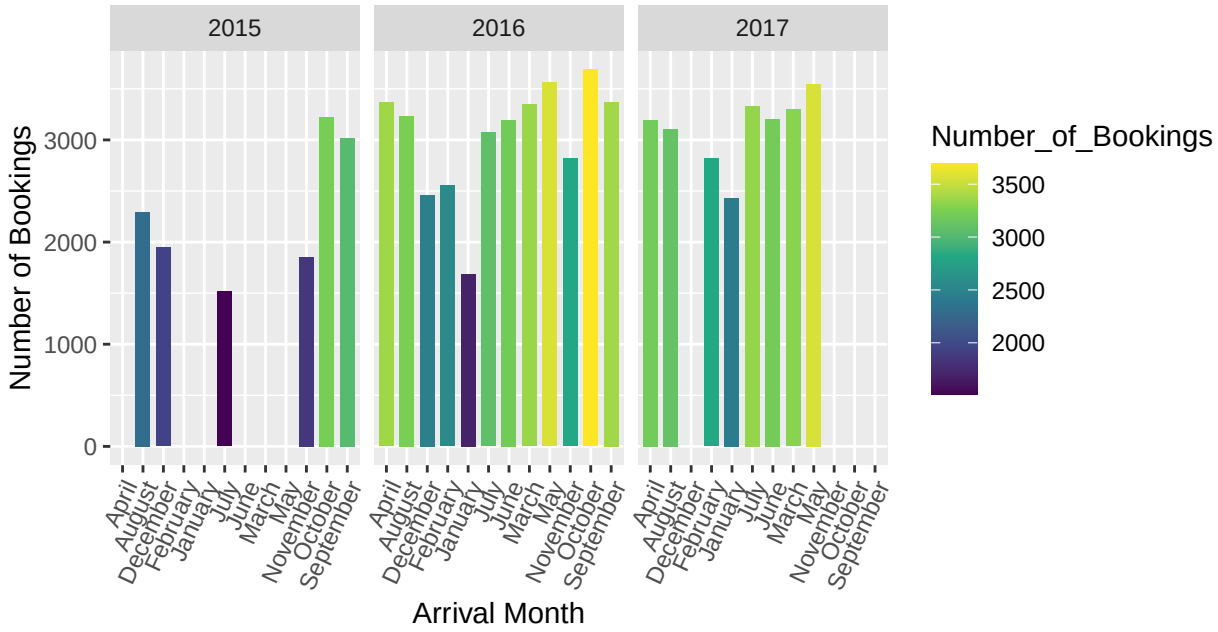
The analysis begins by examining monthly booking trends for the available time period, covering 2015, 2016, and 2017. The bookings per year per month are visualized in Figure 1. As expected, bookings exhibit minimums during the winter months of December–February. Peaks in bookings are observed in 2016 and 2017 during May, while for 2015, a conclusion cannot be made as data are missing for several months.

```
hb1 <- read.csv("HOTEL_BOOKINGS_PER_YEAR_PER_MONTH.csv")
ggplot(data = hb1) +
geom_col(mapping = aes(x = Month_of_Arrival, y = Number_of_Bookings, fill = Number_of_Bookings),
          width = 0.7) +
  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)
```

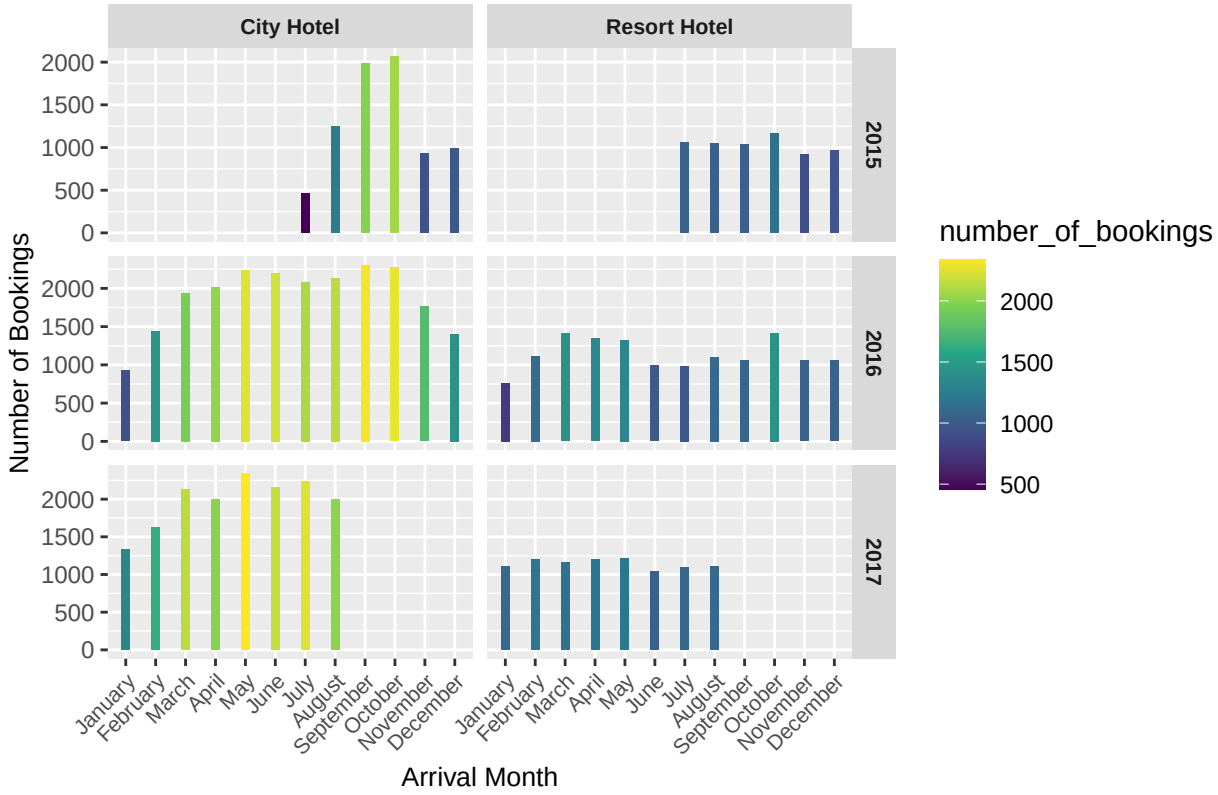
Figure 1: Monthly trends in Bookings



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 explained by the ease of access to the city center and main city attractions. Figure 2 shows that City Hotels experienced peaks in bookings in 2015 and 2016 during October and November, while Resort Hotels do not appear to have any dependency on the month of booking. The 2017 data do not show this peak, possibly due to incomplete data or seasonal shifts.

```
hb2 <- read.csv("BOOKINGS_HOTEL_TYPE_PER_MONTH_PER_YEAR.csv")
ggplot(data = hb2, aes(x = reorder(Month_of_Arrival, match(Month_of_Arrival, month.name)),
                       y = number_of_bookings, fill = number_of_bookings)) +
  geom_col(width = 0.3) +  # Reduced bar width for spacing
 scale_fill_viridis_c() +  # Simple color gradient
  labs(x = "Arrival Month", y = "Number of Bookings", title = "Figure 2: Bookings per type of hotel") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 8),  # Rotate and align x-axis labels, sma
    axis.title = element_text(size = 10),  # Uniform axis title size
    strip.text = element_text(size = 8, face = "bold")  # Style facet labels
  ) +
  facet_grid(Year ~ Type_of_Hotel, scales = "free_y")
```

Figure 2: Bookings per type of hotel




```
# SELECT
# arrival_date_month AS Month_of_Arrival,
# arrival_date_year AS Year_of_Arrival,
# COUNT(*) AS Number_of_Bookings
# hotel
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE
# is_canceled = 0
# GROUP BY
# arrival_date_year,
# arrival_date_month,
# hotel
```

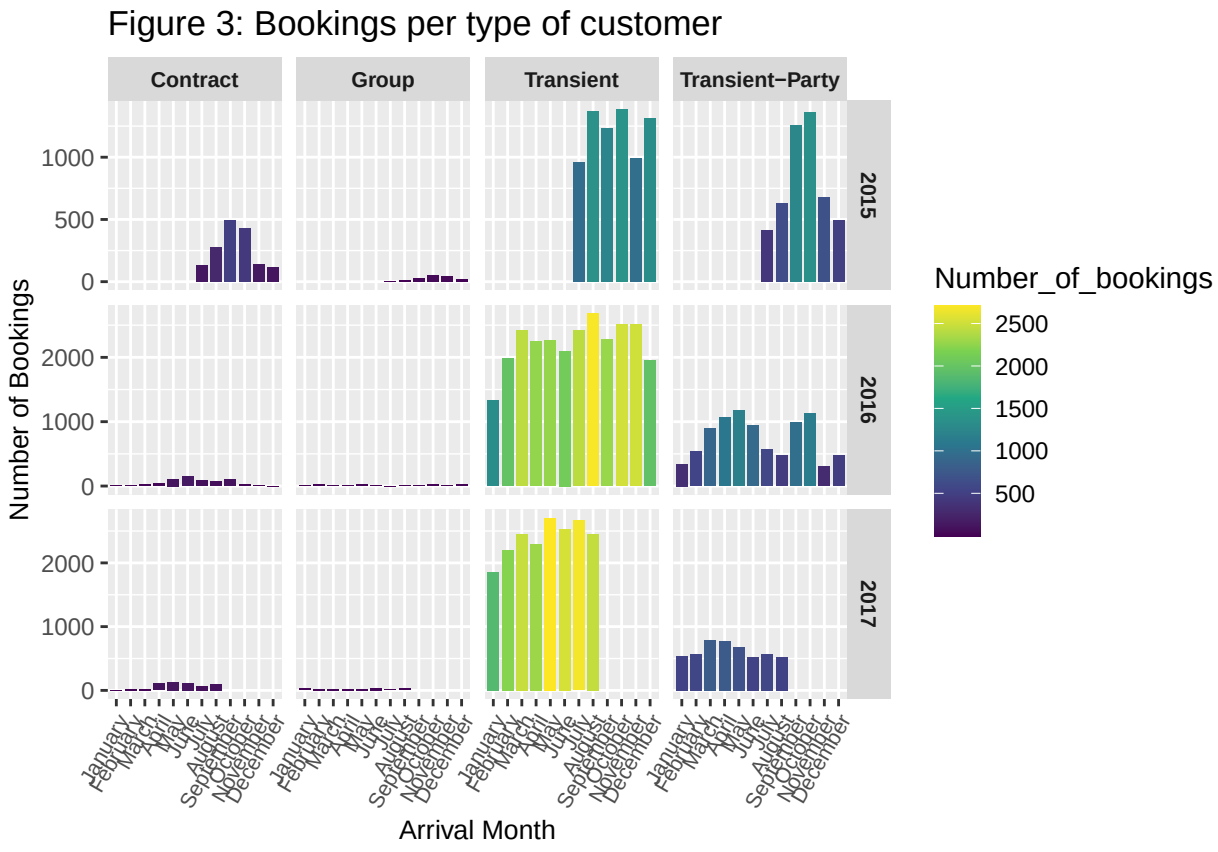
Bookings were then categorized by days spent in the hotel, divided into stays of up to 5 days, 5–10 days, or more than 10 days. Of the bookings, 86.3% stayed up to 5 days, 12.5% stayed for 5–10 days, and 1.2% stayed more than 10 days. This suggests that the hotel could introduce a package offer with cost incentives or added amenities to encourage more customers to stay for 5–10 days.

```
# SELECT
# arrival_date_year,
# COUNT(is_canceled) AS Number_of_bookings
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE
# (stays_in_week_nights < 5)
# stays_in_week_nights >= 5 AND stays_in_week_nights <10
# (stays_in_week_nights >= 10)
```

```
# AND
# is_canceled = 0
# GROUP BY
# arrival_date_year
```

Next, bookings were analyzed by customer type: Transient, Transient-Party, Group, and Contract. Of the total non-canceled bookings, 70.6% were made by Transient customers, while Transient-Party, Group, and Contract bookings accounted for 25%, 0.7%, and 3.7%, respectively. The monthly booking trends in Figure 3 show no prominent peak for Transient customers, while a peak for Transient-Party customers appears in September–October 2015. This peak doesn't appear in 2016 or 2017, so no conclusion can be drawn due to potential data inconsistencies or a small sample size for Transient-Party bookings.

```
hb3 <- read.csv("BOOKINGS_CUSTOMER_TYPE_PER_MONTH_PER_YEAR.csv")
ggplot(data = hb3, aes(x = reorder(Month_of_arrival, match(Month_of_arrival, month.name)),
                       y = Number_of_bookings, fill = Number_of_bookings)) +
  geom_col(width = 0.8) +  # Reduced bar width for spacing
 scale_fill_viridis_c() +  # Simple color gradient
  labs(x = "Arrival Month", y = "Number of Bookings", title = "Figure 3: Bookings per type of customer")
  theme(
    axis.text.x = element_text(angle = 60, hjust = 1, size = 8),  # Rotate and align x-axis labels, sma
    axis.title = element_text(size = 10),  # Uniform axis title size
    strip.text = element_text(size = 8, face = "bold")  # Style facet labels
  ) +
  facet_grid(Year ~ Type_of_customer, scales = "free_y")
```



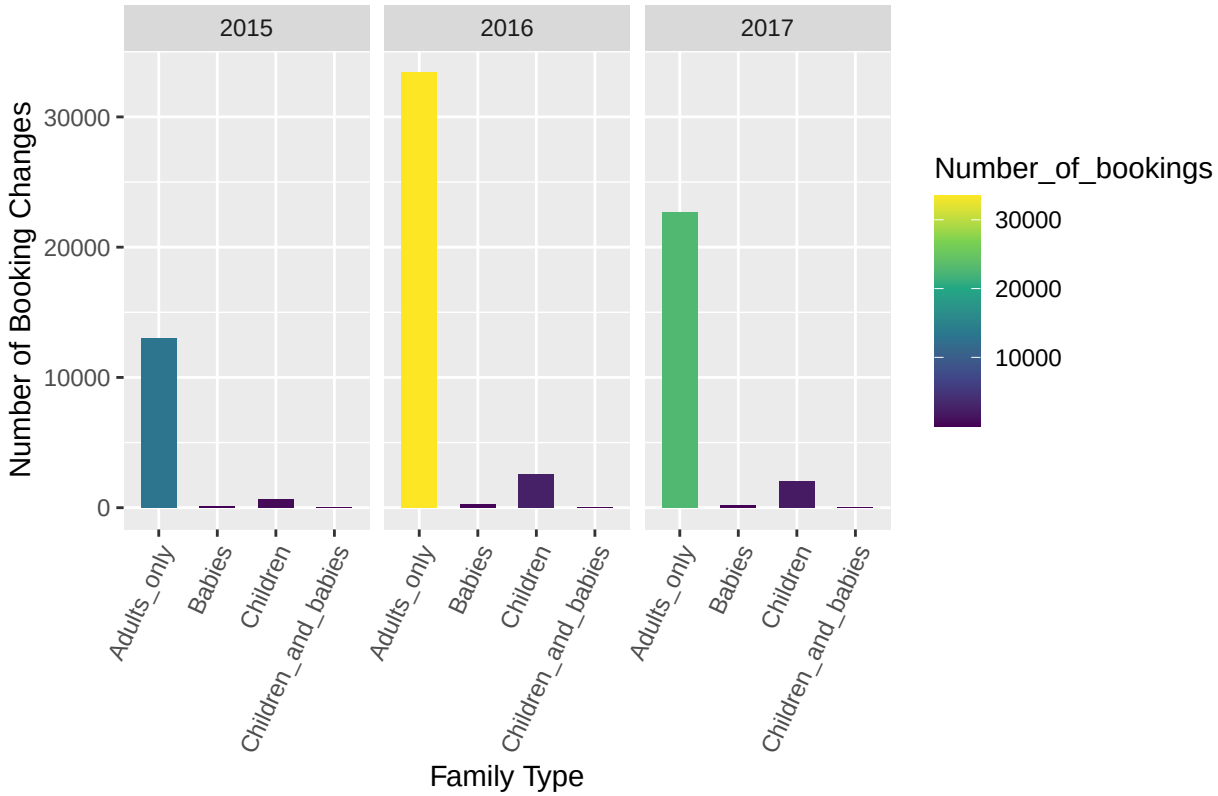

```
# SELECT
# arrival_date_month AS Month_of_Arrival,
```

```
# arrival_date_year AS Year,
# customer_type AS Type_of_Customer,
# COUNT(*)  AS Number_of_bookings
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE
# is_canceled = 0
# GROUP BY
# customer_type,
# arrival_date_year
```

```
hb6 <- read_xlsx("BOOKINGS_MADE_CHILDREN_BABIES_OR_BOTH.xlsx")
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 Booking Changes",
       title = "Figure 4: Bookings made by families with children and/or babies")+

  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()
```

Figure 4: Bookings made by families with children and/or babies



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 only, and 0.2% were families with both children and babies. Finally, 91.9% of bookings were from adults. The hotel could focus marketing campaigns on adults, as they represent the largest segment, simplifying marketing efforts and targeting higher-volume customers.

```
# SELECT
# arrival_date_year AS Year,
# COUNT(*) AS Number_of_Babies,
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE is_canceled = 0 AND babies >= 1 AND children = 0  #BOOKINGS WITH BABIES ONLY
# (is canceled = 0 AND babies = 0 AND Children >0  #BOOKINGS WITH BABIES ONLY )
#   (is_canceled = 0 AND Children = 0 AND babies = 0 #ADULTS ONLY)
#   (is_canceled = 0 AND  Children >0 AND babies >0 #BOTH CHILDREN AND BABIES)
# GROUP BY
# arrival_date_year
```

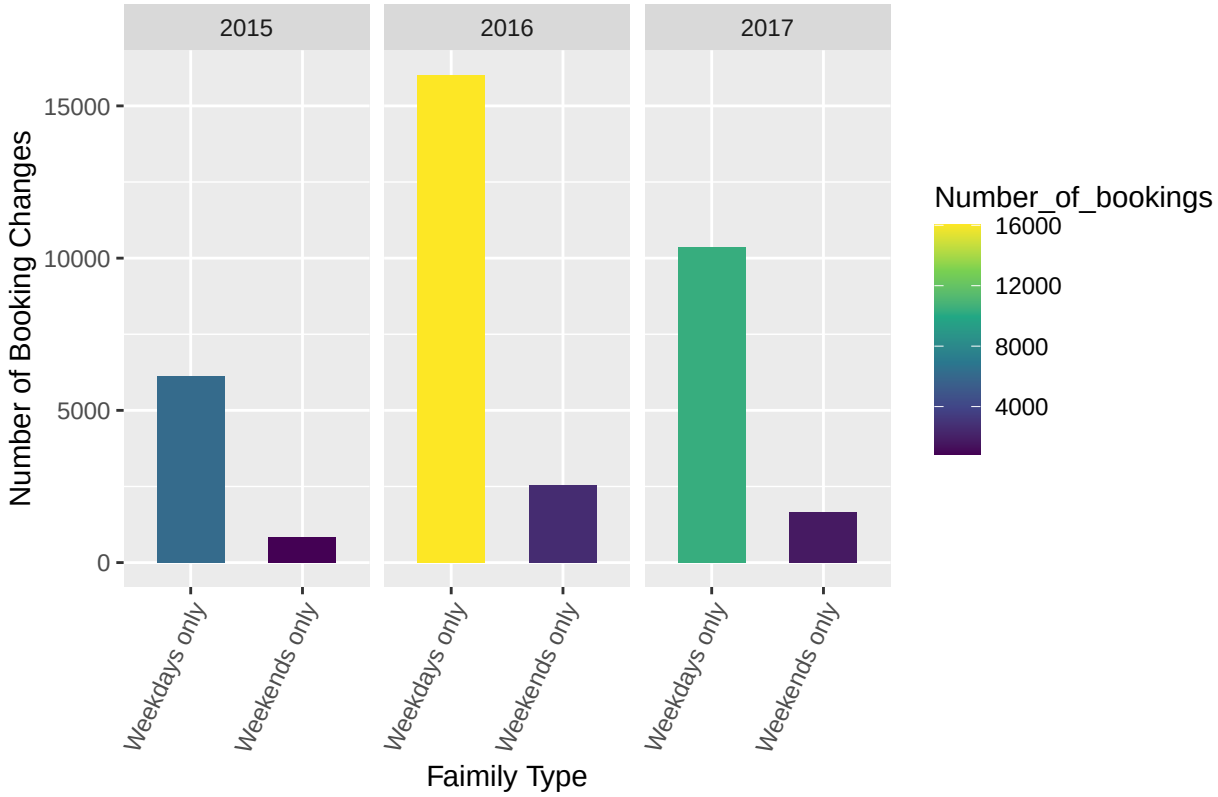
I then analyzed whether bookings involved stays on weekdays only, weekends only, or both weekdays and weekends, with results shown in Figure 5. Of the bookings, 42.7% stayed on weekdays only, 50.5% included both weekdays and weekends, and 6.8% stayed only on weekends. The hotel could promote weekend packages to increase weekend-only stays.

```
hb8 <- read_xlsx("BOOKINGS_STAYED_WEEKEND_ONLY_OR_WEEKDAYS_ONLY.xlsx")

ggplot(data = hb8) +
geom_col(mapping = aes(x = Booking_type, y = Number_of_bookings, fill = Number_of_bookings),
          width = 0.6) +
  facet_wrap(~Year)+
  labs(x = "Faimily Type", y = "Number of Booking Changes",
       title = "Figure 5: Number of Bookings stayed only in weekends or weekdays")+

  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()
```

Figure 5: Number of Bookings stayed only in weekends or weekdays




```
# SELECT
# arrival_date_year,
# COUNT(*) AS Number_of_bookings
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE is_canceled = 0
# AND stays_in_weekend_nights > 0 AND stays_in_week_nights = 0 FOR WEEKEND STAYS
#( AND stays_in_weekend_nights > 0 AND stays_in_week_nights = 0 for WEEKDAY STAYS)
# GROUP BY
# arrival_date_year
```

Most bookings were made through travel agencies, with 83.6% of bookings attributed to them. The hotel could enhance its marketing strategy and advertisements through these partnered agencies, leveraging their reach to boost bookings further.

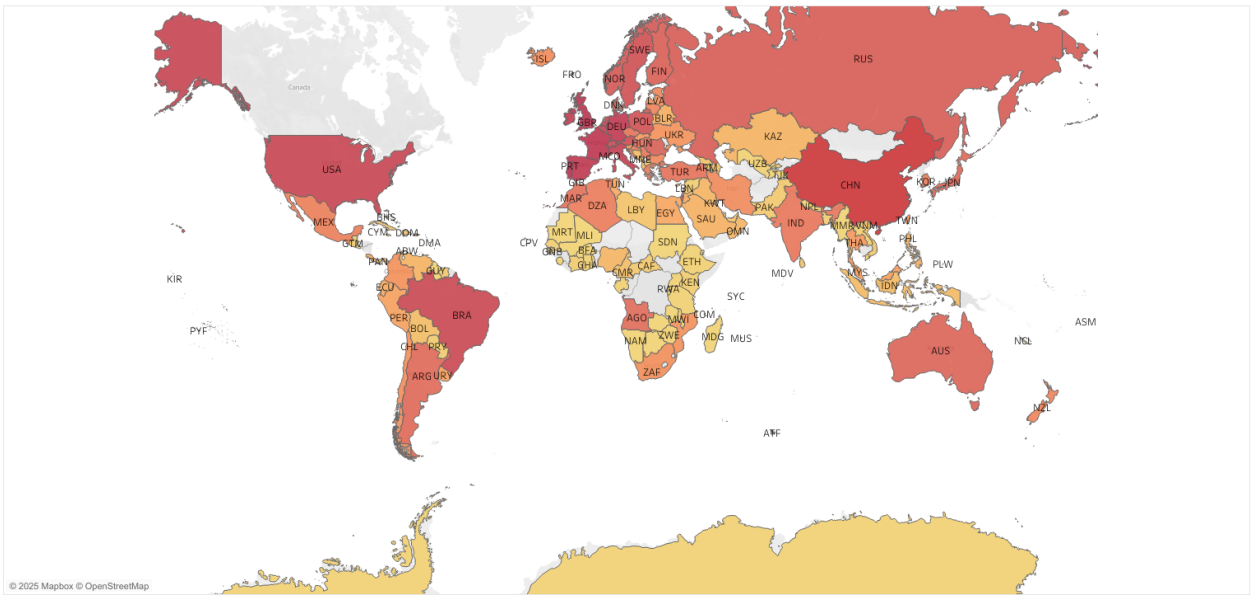
```
# SELECT
# arrival_date_year,
# COUNT(*) AS Number_of_bookings
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE is_canceled = 0
# AND agent != 'NULL'
# GROUP BY
# arrival_date_year
```

Next, I determined the top five visiting countries: 1) Portugal (28.2%), 2) Great Britain (13%), 3) France (11.4%), 4) Spain (8.6%), and 5) Germany (8.1%). Additional marketing campaigns or packages could target these countries, leveraging their high booking volumes to attract more guests. An interactive map visualizing

worldwide bookings, created in Tableau, is shown below.

```
# SELECT
#    COUNT(*) AS Booking_number,
#    country
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE
# is_canceled = 0
# GROUP BY
# country
# ORDER BY
# Booking_number DESC
```

Bookings per Country

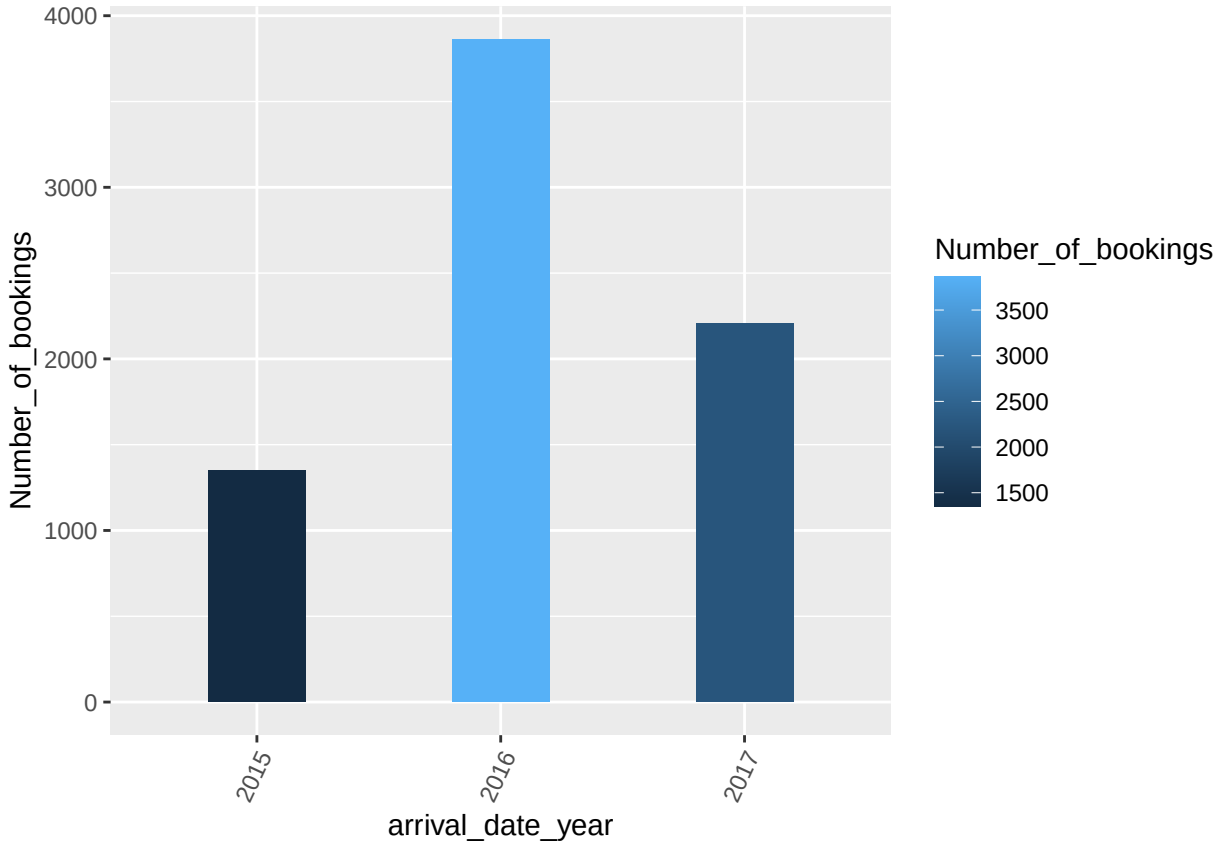



```
hb9 <- read_csv("BOOKINGS_REQUESTED_CAR_PARKING_SPACE.csv")
```

```
## Rows: 3 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): arrival_date_year, Number_of_bookings
##
## 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.
```

```
hb9$arrival_date_year <- as.factor(hb9$arrival_date_year)
#View(hb9)
ggplot(data = hb9) +
geom_bar(mapping = aes(x = arrival_date_year, y = Number_of_bookings, fill = Number_of_bookings),
           stat = "identity", width = 0.4) +

  theme(axis.text.x = element_text(angle = 65, hjust = 1))
```




```
labs(x = "Year", y = "Number of Bookings",
       title = "Figure 6: Bookings Requesting Car Parking Space")
```

```
## $x
## [1] "Year"
##
## $y
## [1] "Number of Bookings"
##
## $title
## [1] "Figure 6: Bookings Requesting Car Parking Space"
##
## attr(,"class")
## [1] "labels"
```

I then analyzed bookings requesting car parking spaces, as shown in Figure 6. The results indicate a steady demand for parking across 2015–2017, suggesting the hotel should maintain sufficient parking facilities to accommodate guest needs.

```
# SELECT
# arrival_date_year,
# COUNT(*) AS Number_of_bookings
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE is_canceled = 0
# AND required_car_parking_spaces >0
# GROUP BY
# arrival_date_year
```

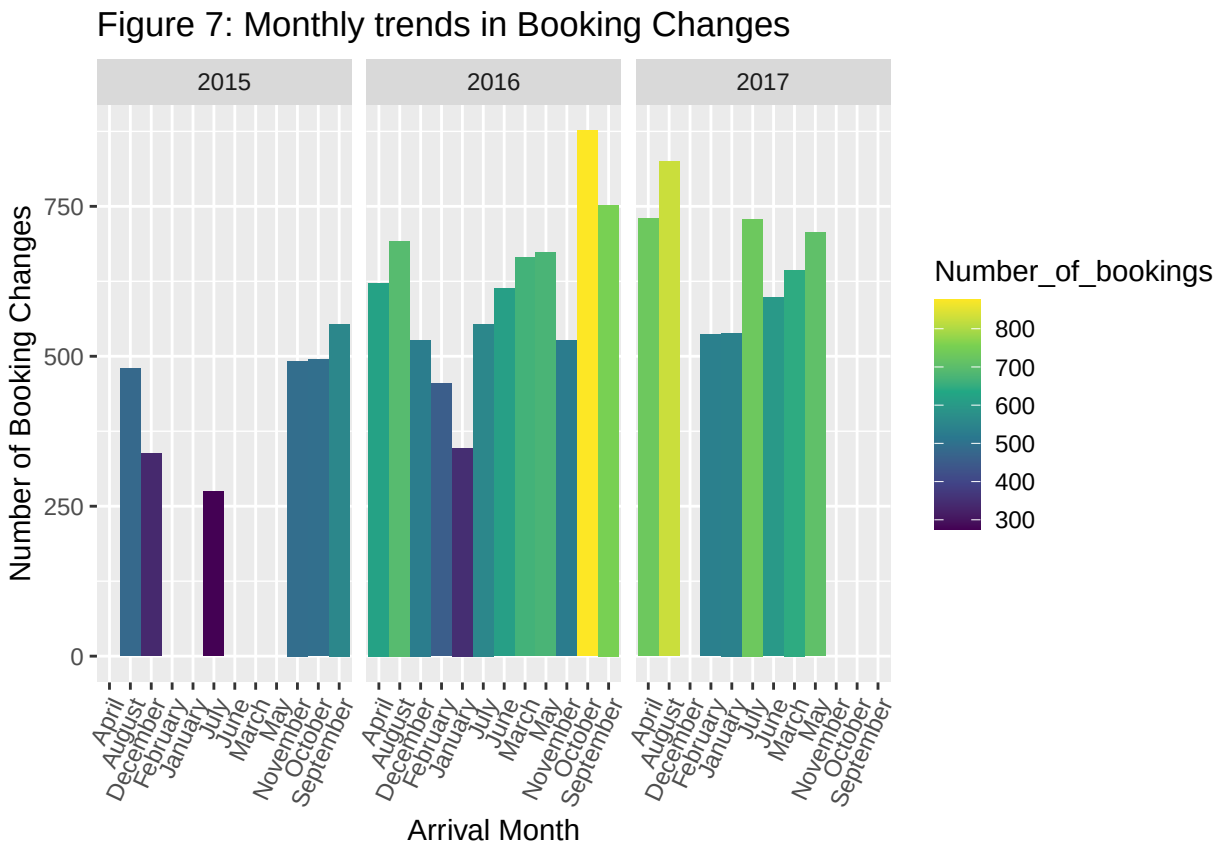
II. Booking changes

In this section, I analyzed booking changes, starting by examining available data to observe any monthly trends, if present. Figure 7 shows no significant trends (sharp peaks or minimums). This suggests booking changes are consistent across months, allowing the hotel to maintain stable change policies.

```
hb4 <- read.csv("BOOKING_CHANGES_PER_YEAR_PER_MONTH.csv")
#View(hb4)

ggplot(data = hb4) +
geom_col(mapping = aes(x = arrival_date_month, y = Number_of_bookings, fill = Number_of_bookings),
          width = 1) +
  facet_wrap(~arrival_date_year)+
  labs(x = "Arrival Month", y = "Number of Booking Changes",
       title = "Figure 7: Monthly trends in Booking Changes")+

  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()
```




```
# SELECT
# arrival_date_month AS Month_of_Arrival,
# arrival_date_year AS Year_of_Arrival,
# COUNT(*) AS Number_of_Bookings


# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
```

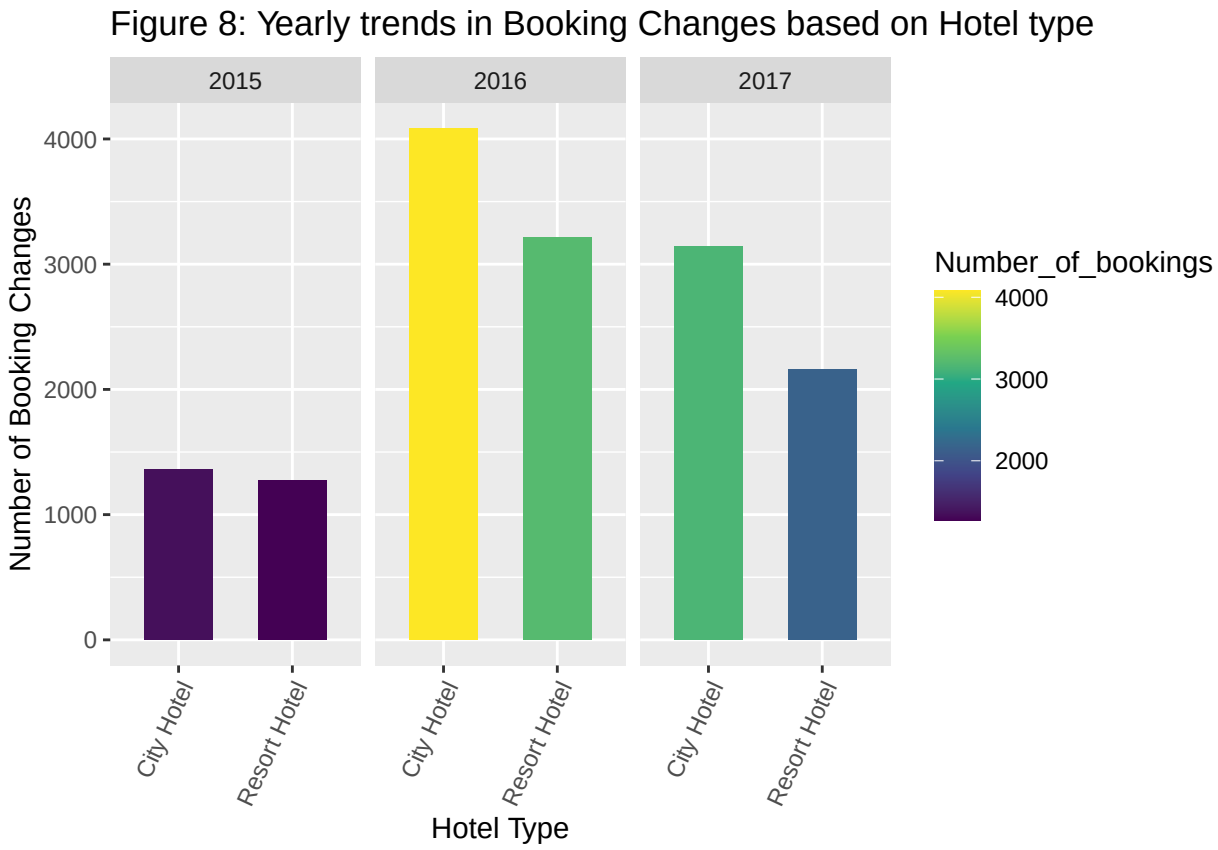
```
# WHERE
# is_canceled = 0
# AND
# booking_changes > 0


# GROUP BY
# arrival_date_year,
# arrival_date_month
```

When examining booking changes by hotel type (Figure 8), 56.3% were from City Hotels, which is expected since most bookings originated from City Hotels. This may reflect their higher guest volume and urban location, leading to more frequent adjustments.

```
hb5 <- read.csv("BOOKING_CHANGES_HOTEL_TYPE.csv")
ggplot(data = hb5) +
geom_col(mapping = aes(x = hotel, y = Number_of_bookings, fill = Number_of_bookings),
          width = 0.6) +
  facet_wrap(~arrival_date_year)+
  labs(x = "Hotel Type", y = "Number of Booking Changes",
       title = "Figure 8: Yearly trends in Booking Changes based on Hotel type")+

  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()
```



Furthermore, booking changes were examined by customer type. Of the changes, 60.8% were from Transient customers, as expected since they made most of the bookings, followed by Transient-Party customers (36.2%),

Contract (2.3%), and Group (0.7%).

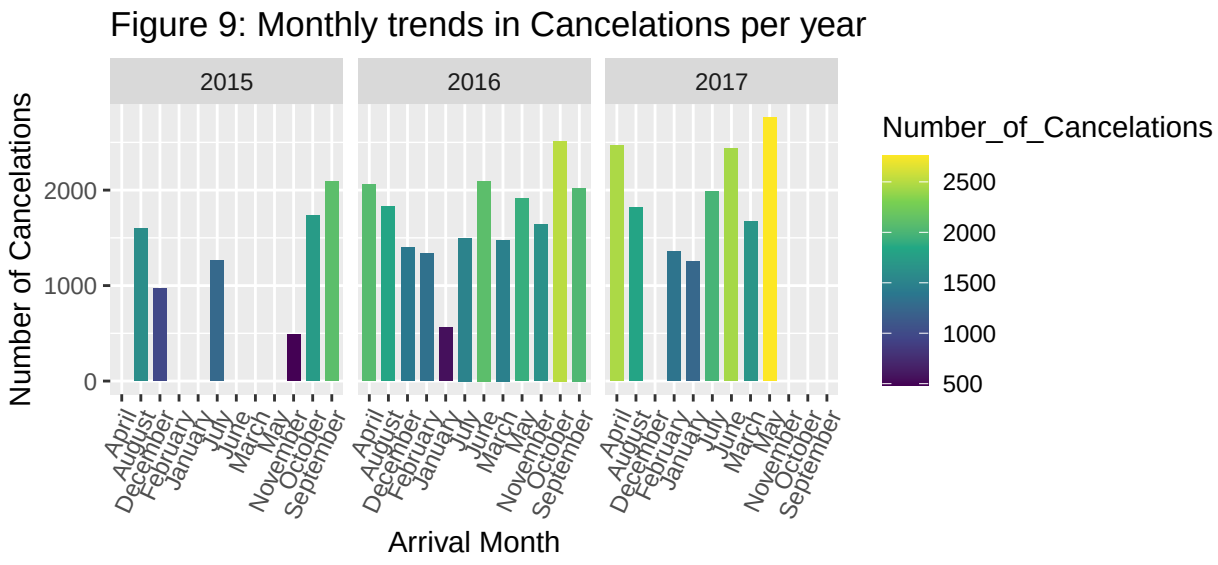
```
# SELECT
# customer_type,
# arrival_date_year,
# COUNT(*) AS Number_of_bookings,
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE is_canceled = 0
# AND
# booking_changes >0
# GROUP BY
# customer_type,
# arrival_date_year
```

III.Cancelations

Finally, I investigated cancellations. As seen in Figure 9, there is no apparent trend in monthly cancellations (no sharp peaks or minimums). This consistency suggests the hotel can maintain steady cancellation policies without seasonal adjustments.

```
hb10 <- read.csv("CANCELATIONS_PER_MONTH_PER_YEAR.csv")
#View(hb10)
ggplot(data = hb10) +
geom_col(mapping = aes(x = Month_of_Arrival, y = Number_of_Cancelations, fill = Number_of_Cancelations)
          width = 0.7) +
  facet_wrap(~arrival_date_year)+
  labs(x = "Arrival Month", y = "Number of Cancelations",
       title = "Figure 9: Monthly trends in Cancelations per year")+

  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()+
  coord_fixed(ratio = 0.005)
```




```
# SELECT
# arrival_date_month AS Month_of_Arrival,
# COUNT(is_canceled) AS Number_of_Cancelations,
```

```
# arrival_date_year
# FROM
# `[Project_Name].hotel_booking_dataset`
# WHERE
# is_canceled = 1
# GROUP BY
# arrival_date_month,
# arrival_date_year
```

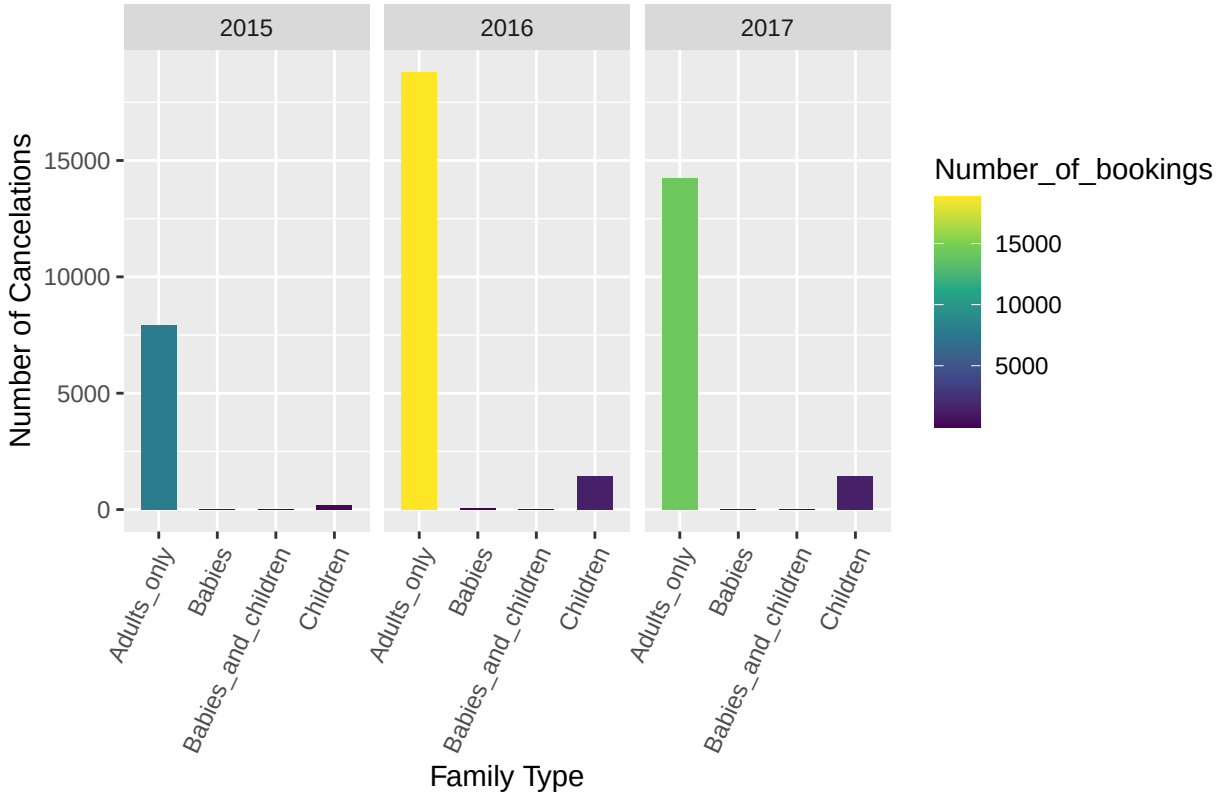
In terms of cancellations made by families (Figure 10), the trend mirrors bookings: 1) Adults (92.6%), 2) Children (7%), 3) Babies (0.3%), and 4) Children and Babies (0.1%). This reflects the higher volume of adult bookings, leading to more cancellations from this group.

```
hb11 <- read_xlsx("CANCELATIONS_CHILDREN_BABIES.xlsx")

#View(hb11)
ggplot(data = hb11) +
geom_col(mapping = aes(x = Family_Type, y = Number_of_bookings, fill = Number_of_bookings),
          width = 0.6) +
  facet_wrap(~Year)+
  labs(x = "Family Type", y = "Number of Cancelations",
       title = "Figure 10: Cancellations made by families with children and/or babies")+

  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()
```

Figure 10: Cancellations made by families with children and/or babies



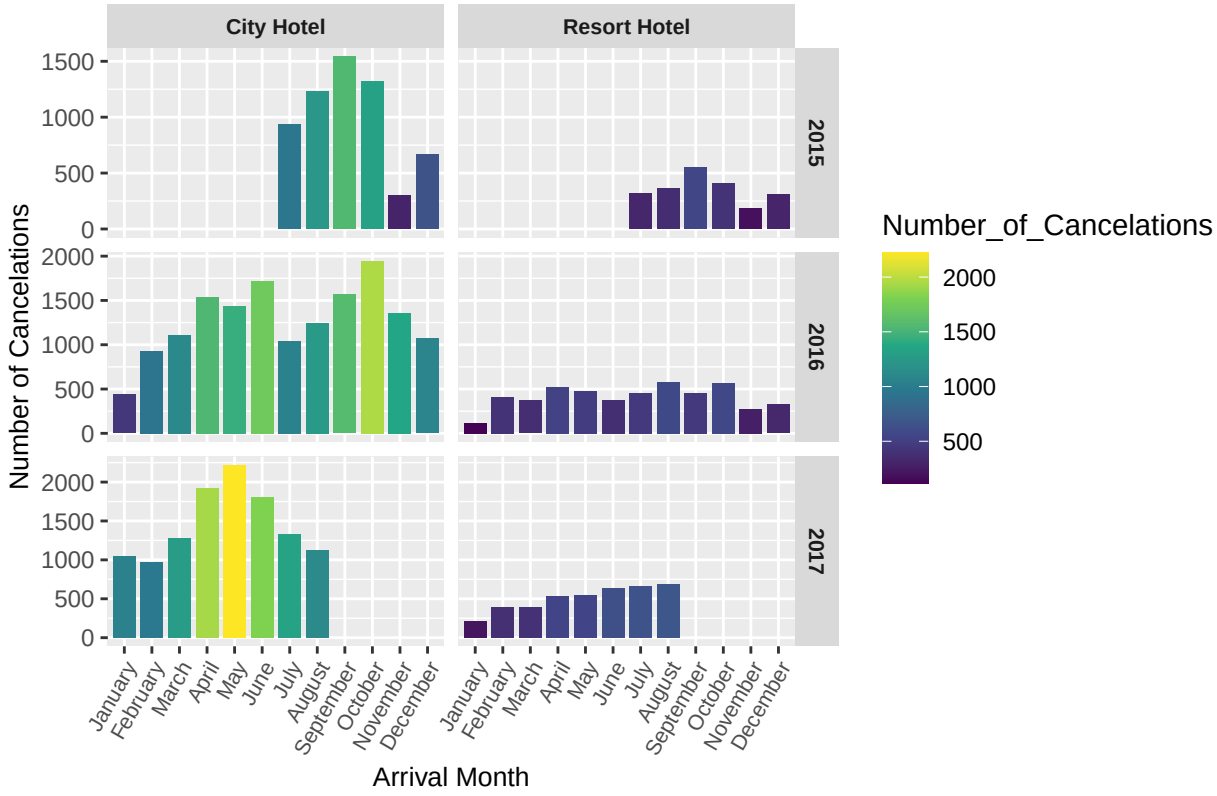

```
# SELECT
# arrival_date_year AS Year,
```

```
# COUNT(*) AS Number_of_Babies,
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE is_canceled = 1 AND babies >= 1 AND children = 0  #BOOKINGS WITH BABIES ONLY
# (is canceled = 1 AND babies = 0 AND Children >0  #BOOKINGS WITH BABIES ONLY )
#   (is_canceled = 1 AND Children = 0 AND babies = 0 #ADULTS ONLY)
#   (is_canceled = 1 AND  Children >0 AND babies >0 #BOTH CHILDREN AND BABIES)
# GROUP BY
# arrival_date_year
```

In terms of hotel type (Figure 11), most cancellations were from City Hotels (75%), while the remaining 25% were for Resort Hotels. This aligns with the higher booking volume of City Hotels, likely leading to more cancellations.

```
hb12 <- read.csv("CANCELATIONS_HOTEL_TYPE.csv")
ggplot(data = hb12, aes(x = reorder(Month_of_Arrival, match(Month_of_Arrival, month.name)),
                        y = Number_of_Cancelations, fill = Number_of_Cancelations)) +
  geom_col(width = 0.8) +  # Reduced bar width for spacing
scale_fill_viridis_c() +  # Simple color gradient
  labs(x = "Arrival Month", y = "Number of Cancelations", title = "Figure 11: Cancelations from hotel ty
  theme(
    axis.text.x = element_text(angle = 60, hjust = 1, size = 8),  # Rotate and align x-axis labels, sma
    axis.title = element_text(size = 10),  # Uniform axis title size
    strip.text = element_text(size = 8, face = "bold")  # Style facet labels
  ) +
  facet_grid(Year ~ hotel, scales = "free_y")
```

Figure 11: Cancelations from hotel type



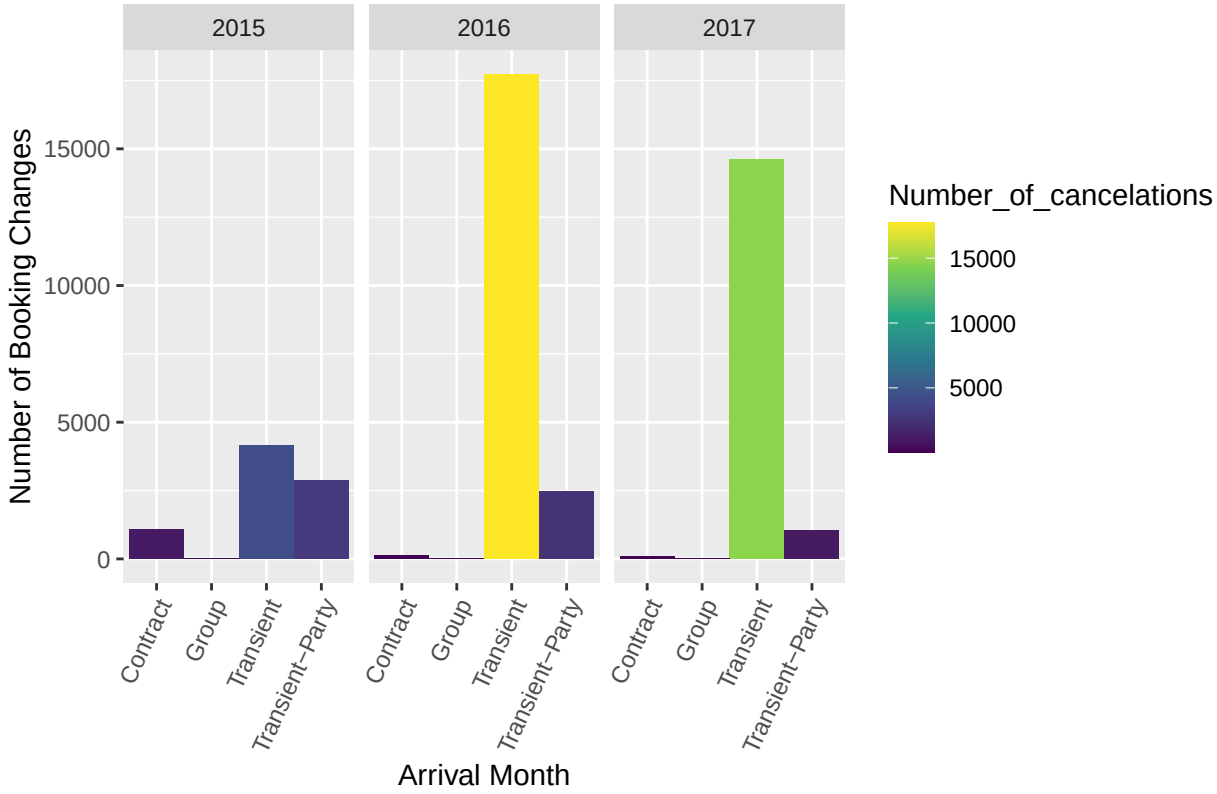
```
# SELECT
# arrival_date_month AS Month_of_Arrival,
# arrival_date_year AS Year,
# COUNT(*) AS Number_of_Cancelations,
# hotel
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE is_canceled = 1
# GROUP BY
# hotel,
# arrival_date_year
```

Regarding customer type (Figure 12), 82.6% of cancellations were from Transient Customers, 14.4% from Transient-Party, 2.8% from Contracts, and 0.1% from Groups, respectively. This reflects the higher volume of Transient bookings, which may be more prone to changes in travel plans.

```
hb13 <- read.csv("CANCELATIONS_CUSTOMER_TYPE.csv")
ggplot(data = hb13) +
geom_col(mapping = aes(x = customer_type, y = Number_of_cancelations, fill = Number_of_cancelations),
          width = 1) +
  facet_wrap(~arrival_date_year)+
  labs(x = "Arrival Month", y = "Number of Booking Changes",
       title = "Figure 12: Cancelations based on customer type")+

  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()
```

Figure 12: Cancelations based on customer type



```
# SELECT
# arrival_date_month AS Month_of_Arrival,
# arrival_date_year AS Year,
# COUNT(*) AS Number_of_Bookings,
# customer_type
# FROM
# `[Project_name].Hotel_Bookings.hotel_booking_dataset`
# WHERE is_canceled = 1
# GROUP BY
# customer_type,
# arrival_date_year,
# arrival_date_month
```

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, during the day categorization step (0 to 5 days, etc.), there was a deficit of 680 entries that weren't in any category. Upon inspection, these were regular entries, likely due to same-day checkouts. Since the rest of the analysis does not use days stayed, this discrepancy can be disregarded, though it would be relevant if revenue data were involved. Future work could address these gaps by incorporating complete datasets or revenue metrics.

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, a dominance of City Hotels (61.5% of bookings), and a high cancellation rate among Transient Customers (82.6%). Recommendations include targeting marketing campaigns toward adults (91.9% of bookings), promoting 5–10 day stay packages, and enhancing partnerships with travel agencies (83.6% of bookings). The Tableau visualization of top countries (e.g., Portugal at 28.2%) further supports targeted marketing strategies toward these countries.

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 business insights. Future work could explore revenue impacts or seasonal pricing strategies to further optimize hotel operations.