Exploratory Data Analysis of Hotel Booking Demand

Author: Santiago Boccardo
Date: July 31st 2025
Dataset: The dataset used in this project is publicly available and open for educational and non-commercial use. It was obtained from Kaggle and can be accessed here: Hotel Booking Demand

Executive Summary

This project is part of my portfolio as a data scientist. It explores and analyzes a publicly available dataset on hotel bookings with the aim of identifying patterns in customer behavior, investigating factors associated with reservation cancellations, and extracting insights to inform hotel management decisions. The analysis includes data cleaning, feature engineering, visualization, and interpretation using Python (Google Colab) and Google Looker Studio.

Objectives

- Explore hotel booking behavior by customer type, country of origin, and seasonality.
- Identify key drivers of booking cancellations.
- Detect insights and trends that could support business decisions in the hospitality sector.

Dataset Description

- **Source:** Kaggle
- **Title:** Hotel Booking Demand
- **Number of records:** 119,390
- **Number of columns:** 32
- **Date range:** 2015 to 2017
- **Key features:**
    - Hotel type (City Hotel / Resort Hotel)
    - Arrival date (year, month, `day_of_month`)
    - Stay duration (`stays_in_week_nights, stays_in_weekend_nights`)
    - Number of guests (`adults, children, babies`)
    - Cancellation status (`is_canceled`)
    - Booking channel
    - Country of origin
    - Special requests (`total_of_special_requests`)

Methodology

Tools used:

- **Google Colab** with Python (pandas, matplotlib, seaborn, numpy)
- **Google Looker Studio** for interactive visual dashboards
- **GitHub** for hosting and project presentation

Data preparation steps:

1. **Data loading and initial review**
2. **Data Cleaning and Preprocessing**:
    - Ensured appropriate data types
    - Removed duplicate records
    - Filtered out invalid rows (e.g., bookings with zero total guests or zero adults)
    - Reviewed and handled missing values in key columns (children, country, agent, company)
3. **Feature engineering:**
    - Created new columns like `total_guests` = `adults` + `children` + `babies`

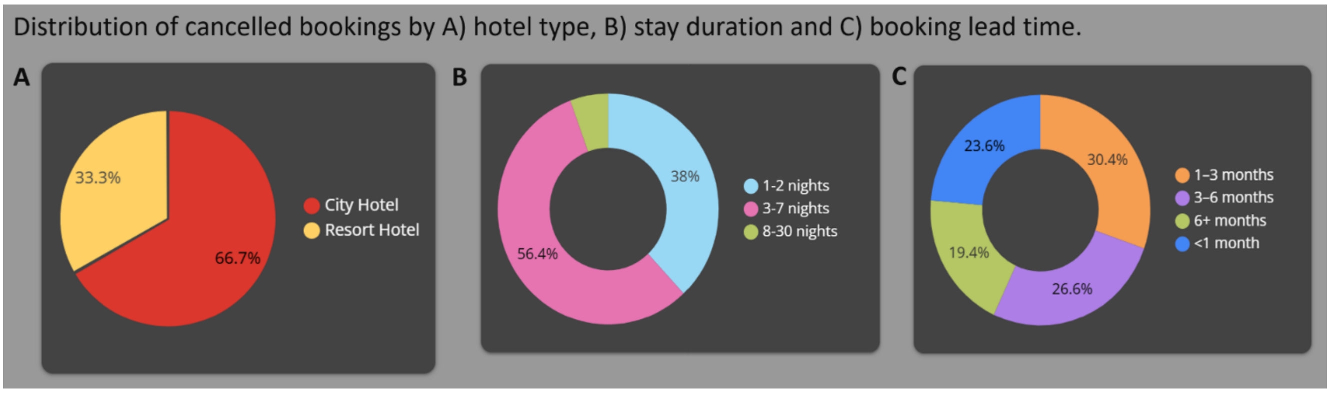
   - Converted month names to numeric values for date ordering
   - Combined arrival day/month/year into a single `arrival_date` field
   - Encoded categorical variables when needed
4. **Exported the cleaned dataset** for visualization and dashboarding

Exploratory Analysis & Visualizations

Key areas of analysis included:

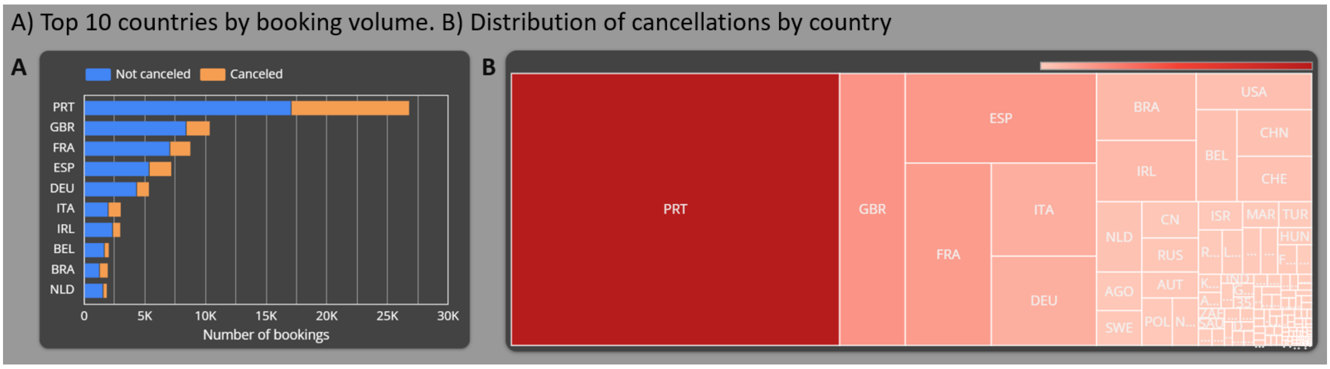
Reservation and Cancellations

- The cleaned dataset includes **86,420** bookings. On average, each reservation includes **2 guests** and corresponds to a **short stay**.
- Canceled bookings account for **27%** of total reservations, most of which are linked to **City Hotels**.
- Higher cancellation rates were observed for **shorter stays**.
- No major differences in cancellation proportion were found regarding **booking lead time**.


Distribution of cancelled bookings by A) hotel type, B) stay duration and C) booking lead time.

Booking Origins

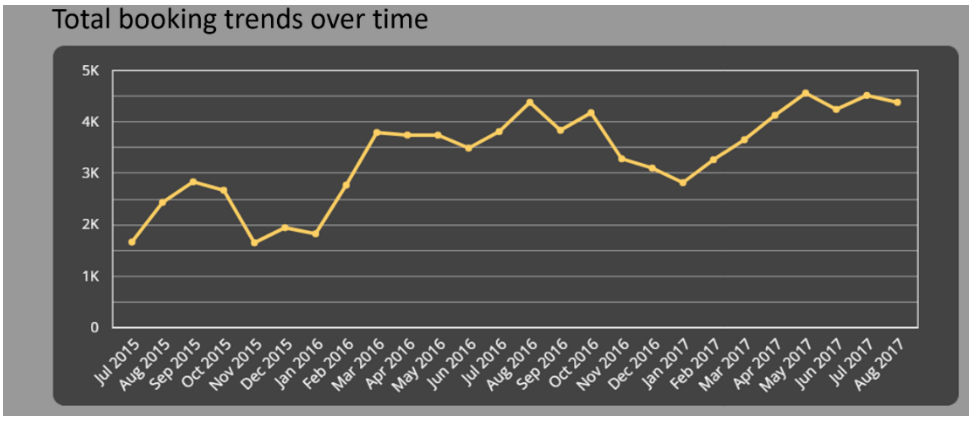
- The top 5 countries by booking volume are:
  1. Portugal
  2. United Kingdom
  3. France
  4. Spain
  5. Germany
- Even though Portugal leads in booking numbers, it also shows a considerably high cancellation rate.

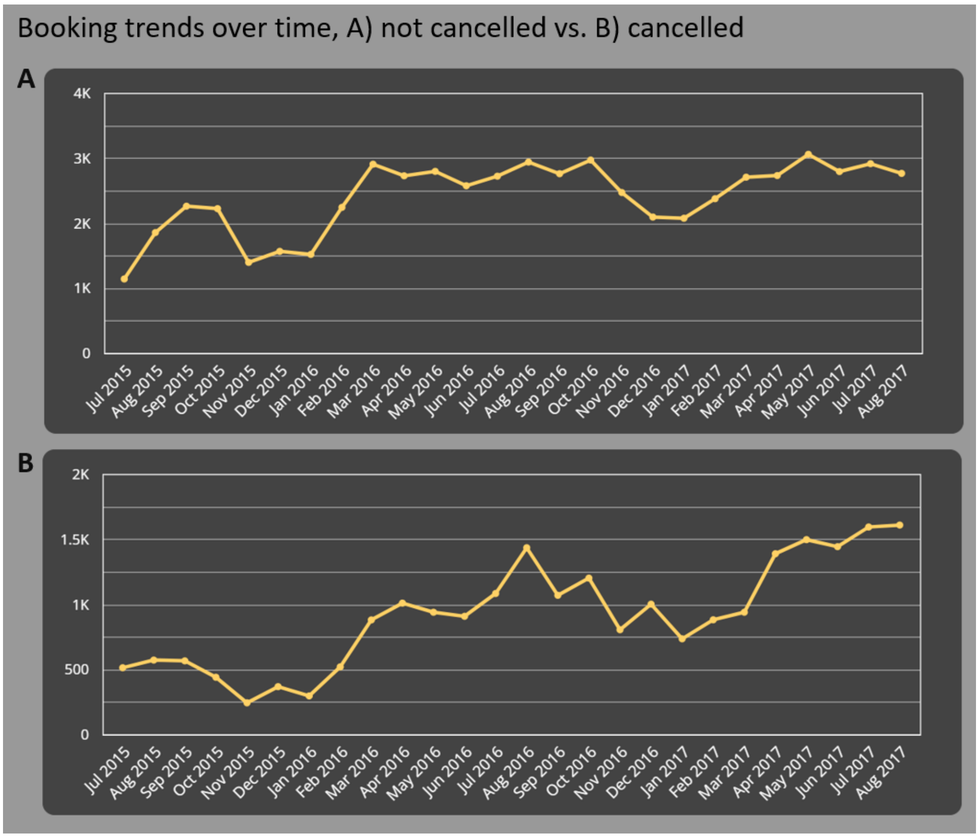

A) Top 10 countries by booking volume. B) Distribution of cancellations by country

Seasonality

An analysis of total bookings over time reveals a recurring increase between July and October, becoming more pronounced in recent years. Conversely, the lowest booking numbers consistently occur between November and January.

However, this trend varies when bookings are split by status. Non-canceled bookings increased at the beginning of 2016 and remained relatively stable, aside from a yearly drop between November and February. Canceled bookings, on the other hand, show higher variability with no clear seasonal pattern, but exhibit a general upward trend throughout the analyzed period.

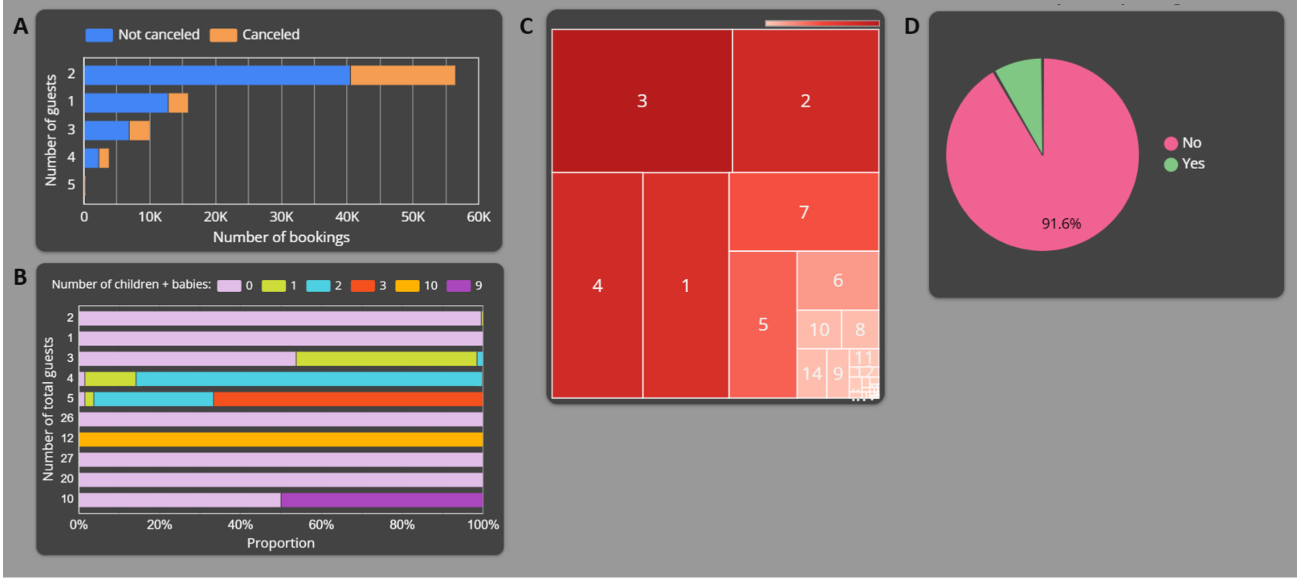




Guest Profiles

- Most bookings were for **two adults with no children**.
- The majority of reservations lasted between **1 and 4 nights**, as shown by the highest frequency in this range. However, the average stay length across all bookings is 4 nights, likely influenced by a smaller number of long-duration stays that skew the mean upward.
- Across the five most common guest group sizes, 75–85% of bookings include bed and breakfast. The remaining reservations are mostly divided between half board and no meal options, while full board is rarely chosen.
- A majority of guests did **not request parking**.

A) Number of bookings by total guest count. B) Proportion of children within bookings, grouped by total guest count. C) Distribution of total bookings by length of stay (nights). D) Parking request distribution in total bookings.



Interactive Dashboard

A dashboard was created in **Google Looker Studio**, organized into three main pages:

1. **General Overview**
   - KPIs such as cancellation rate, average stay duration, and total guest count.
   - Booking distribution by hotel type, country and length of stay.
   - Booking trends over time.
2. **Cancellations' insight**
   - Cancellation behavior by country, number of people per booking, stay duration, lead time and company ID.
3. **Gests' Profile & Stay Behavior**
   - Bookings by number of guests and nights.
   - Proportion of children and meal preferences based on the number of people per booking.
   - Frequency of special requests and parking requests.

Click here to view the interactive dashboard

Key Findings

- **Cancellation rates were high**, especially in City Hotels and for short-duration stays.
- **Portugal led** in both the number of bookings and cancellations.
- **Non-canceled bookings dropped** consistently every year between November and February.
- **Most bookings were for two adults** staying **1–4 nights**.
- **Bed & breakfast** is the most common meal plan across all guest group sizes.

While this report offers a general exploratory overview, the structure of the dataset and the design of the dashboard allow for more granular analysis by cross-referencing variables. This enables further exploration of customer behavior and booking trends across different dimensions.

About the Author of this report

Santiago Boccardo

Data Analyst / Data Scientist
Biochemist & PhD in Chemical Sciences

Health sciences researcher with 8+ years in immunology and data analysis. Currently transitioning into data science to support evidence-based decision-making.

LinkedIn | GitHub