**Data Analysis Using Tableau to Predict Hotel Booking Cancellations**

****

Prepared by

**[Rajeev Yenugula]**

**[016045757]**

**[rajeev.yenugula@sjsu.edu]**

### Department of Applied Data Science, San Jose State University DATA 230: Data Visualization

**Major Professor: Andrew H. Bond December 11, 2022**

**By**

**Rajeev Yenugula 016045757**

**CONTENTS**

[Project Abstract](#_Project_Abstract)…………………………………………………………………………………. 5

Scope of the Project……………………………………………………………………………… 6

Objectives…………………………………………………………………………………………7

Workflow segregation…………………………………………………………………………… 8

Introduction……………………………………………………………………………………… 9

Dataset Description

Column Description

Exploratory Data Analysis……………………………………………………………………... 14

Framing questions

Filtering out the ideas

Cleaning the Data

EDA Analysis

Visualization of Insights

Data Collection and Preparation……………………………………………………………… 16

Dataset

Sample of the raw dataset

Data preparation flow chart from Tableau Software

Summary of Major Achievements and key dissemination activities………………………... 18

Deliverables…………………………………………………………………………………….. 19

Let's find insights from the dataset…………………………………………………………….20

Diagram1: Hotel Type 20

Diagram 2: Number of bookings by cancellations 21

Diagram 3: Booking cancellations on months 22

Cancellations status 23

Average lead time by hotel 24

Number of bookings by market segment 25

Deposit type count 26

ADR by year and hotel 27

Number of Arrivals by Month and Hotel 28

Hotel bookings by country 29

Average waiting days by room type 30

Dashboards……………………………………………………………………………………... 31

Dashboard 1 - Overview of The Hotel Bookings 31

Dashboard 2 - Details 31

Dashboard 3 – Details Continued 32

Story…………………………………………………………………………………………….. 33

Conclusion……………………………………………………………………………………… 36

Strategies to counter higher cancellations at hotel……………………………………………36

References……………………………………………………………………………………… 37

GitHub link: <https://github.com/rajeevyen/Data_Visualization_project>

# **Project Abstract**

Bookings of hotels have increased drastically in recent years. The hotel industry is very volatile, and bookings depend on various factors such as the type of hotel, seasonality, days of the week, and many more. This dataset contains 2 years of reservation information for a city hotel and a resort hotel. Details such as booking time, personal information, parking details, number of adults, children or infants, etc. Using this historical data, Hotels can examine the influence of the past on current economic outcomes. We can use the patterns to predict future bookings. This project is about collecting data about hotel bookings and analyzing the same.

We will be tackling this problem statement in two stages:

1. Extracting and transforming (cleaning the noisy data) the data for analysis like the number of cancellations and Number of bookings on a weekday vs weekends
2. Analyzing the data to bring out meaningful insights.

We will create meaningful visualizations in Tableau and derive useful conclusions. Using the results from the above analysis, a business can make key decisions regarding the customer experience they desire to deliver.

# **[Scope](https://docs.google.com/document/d/1p9LZHDwc9Ejj79b91vycoa8iSd4edP-l/edit" \l "heading%3Dh.2et92p0) of the Project**

The purpose of this project is to use Exploratory Data Analysis (EDA) to get insights from the data set to determine which factors have contributed the most to the prediction of cancellations. This will be accomplished via the use of Tableau data visualization. It is usually a good practice to first analyze the data and then attempt to extract as many insights as possible from it. We can estimate which factors are most likely to cancel or book a reservation using Tableau, which might help us make better projections and minimize uncertainty in business choices

Suppose we were hired as consultants to answer this key question: **when will a customer cancel or not show up for their reservation?**  
This can help a hotel plan for things like food and personal needs. Perhaps some hotels also use this model to offer more rooms than they have to make more money or the like.

This hotel booking dataset can help you explore those questions!

# **Objectives**

* Overview of the Hotel Bookings.
* Details of the Hotel Bookings.
* The arrival of hotels by Month.
* Bookings by Country
* How Many Bookings Were Cancelled?
* What is the booking ratio between Resort Hotel and City Hotel?
* What is the percentage of booking for each year?
* Which is the busiest month for a hotel?
* From which country do most guests come?
* How Long Do People Stay in the hotel?
* This visualization provides insights into hotel bookings and cancellations.

### After that, we made the predictive model to predict whether the booking will be canceled or not

### **We will:**

* Perform Feature Engineering to make new features
* Perform the Data Selection to select only relevant features
* Tranform the Data (Categorial to Numerical)
* Split the data (Train Test Split)
* Model the data (Fit the Data)
* And finally, Evaluate our model

**Work Flow Segregation**

We have divided our workflow into the following 3 Steps:

* Data collection and Understanding
* Data cleaning and Manipulation
* Exploratory data analysis

**Introduction**

**Dataset Description**

The hotel\_bookings.csv file dataset contains reservation information for a city hotel and a resort hotel and includes information such as when the reservation was made, length of stay, the number of adults, children, and/or infants, and the number of parking spaces available, among other things.  
Both hotels are located in Portugal (H1 in the tourist region of the Algarve and H2 in the city of Lisbon). The distance between these two places is almost 280 km by car and both towns border the North Atlantic.

**Column Description**

This data set contains a single file that compares various booking information between two hotels: a city hotel and a resort hotel. This dataset has 32 columns with descriptions as follows:

* **hotel**: Hotel (H1 = Resort Hotel or H2 = City Hotel).
* **Is\_canceled**: Value indicating if the booking was canceled (1) or not (0).
* **lead\_time**: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
* **arrival\_date\_year**: Year of arrival date.
* **arrival\_date\_month**: Month of arrival date.
* **arrival\_date\_week**: Week number of year for arrival date.
* **arrival\_date\_day**: Day of arrival date.
* **stays\_in\_week\_nights**: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel.
* **adult**: Number of adults.
* **children**: Number of children.
* **babies**: Number of babies.
* **meal**: Type of meal booked. Categories are presented in standard hospitality meal packages:
* 1. Undefined/SC – no meal package;
* 2. BB – Bed & Breakfast;
* 3. HB – Half board (breakfast and one other meal – usually dinner);

4. FB – Full board (breakfast, lunch, and dinner)

* **country**: Country of origin. Categories are represented in the ISO 3155–3:2013 format
* **market\_segment**: Market segment designation. In categories, the term **TA** means “Travel Agents” and **to** means “Tour Operators”.
* **distribution\_channel**: Booking distribution channel. The term **TA** means “Travel Agents” and **to** means “Tour Operators”.
* **is\_repeated\_guest**: Value indicating if the booking name was from a repeated guest (1) or not (0).
* **previous\_cancellations**: Number of previous bookings that were canceled by the customer before the current booking.
* **previous\_bookings\_not\_canceled**: Number of previous bookings not canceled by the customer before the current booking.
* **reserved\_room\_type**: Code of room type reserved. Code is presented instead of designation for anonymity reasons.
* **assigned\_room\_type**: Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
* **booking\_changes**: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation.
* **deposit\_type**: Indication if the customer deposited to guarantee the booking. This variable can assume three categories: BO and TR/Value calculated based on the payments identified for the booking in the transaction (TR) table before the bookings arrival or cancellation date.

1. No Deposit – no deposit was made;

2. Non-Refund – a deposit was made in the value of the total stay cost;

3. Refundable – a deposit was made with a value under the total cost of the stay.

* **agent**: ID of the travel agency that made the booking.
* **company**: ID of the company/entity that made the booking or is responsible for paying the booking. ID is presented instead of designation for anonymity reasons.
* **days\_in\_waiting\_list**: Number of days the booking was on the waiting list before it was confirmed to the customer
* **customer\_type**: Type of booking, assuming one of four categories:
  1. Contract - when the booking has an allotment or other type of contract associated with it;
  2. Group – when the booking is associated with a group;
  3. Transient – when the booking isn't part of a group/contract and isn't associated with another transient booking.
  4. Transient-party – when the booking is transient but is associated with at least another transient booking.
* **ADR**: Average daily rate.
* **required\_car\_parking\_spaces**: Number of car parking spaces required by the customer.
* **total\_of\_special\_requests**: Number of special requests made by the customer (e.g. twin bed or high floor).
* **reservation\_status**: Reservation is the last status, assuming one of three categories:

1. Canceled – booking was canceled by the customer;
2. Check-Out – customer has checked in but already departed;
3. No-Show – the customer did not check in and did inform the hotel of the reason why

* **reservation\_status\_date**: The date at which the last status was set. This variable can be used in conjunction with the Reservation status to understand when was the booking canceled or when the customer checked out of the hotel.

# **Exploratory Data analysis**

**Framing the questions:** Before any form of analysis, it is important to frame the questions that we want to know from the data. For this, I performed many creative thinking techniques such as brainstorming to take out all the questions that could be asked related to the dataset.

**Filtering out the ideas:** After getting a long list of questions and assumptions that we want to solve from the dataset. As per the approach, we started filtering out our question list. One approach that we used to filter questions was to think from a customer and business perspective and look for answers to only those questions that mattered to any one of them.

**Cleaning the Data:** As mentioned before, I got this dataset from Kaggle. By clean we mean that it didn’t have any nested lists or dictionaries as row elements or wrong types in data frame columns. However, it had four variables with null values, so we had to take them into account before proceeding further with our analysis.

**EDA analysis:** By EDA we mean exploratory data analysis. In this, we looked at the data frame and decided our target variables (Important Columns) based upon which we were going to conduct further analysis.

**Visualization of Insights:** After we completed the analysis of our data, we used Tableau to present our analysis graphically. We used Pie charts, bar charts, scatterplots, and much more to give complex insights in an eye-catching manner. We learned a lot about the different visualization tools that are available for data analysis.

**Drawing Conclusions and Finding Answers:** Finally, we warped up each analysis by drawing out conclusions from them.

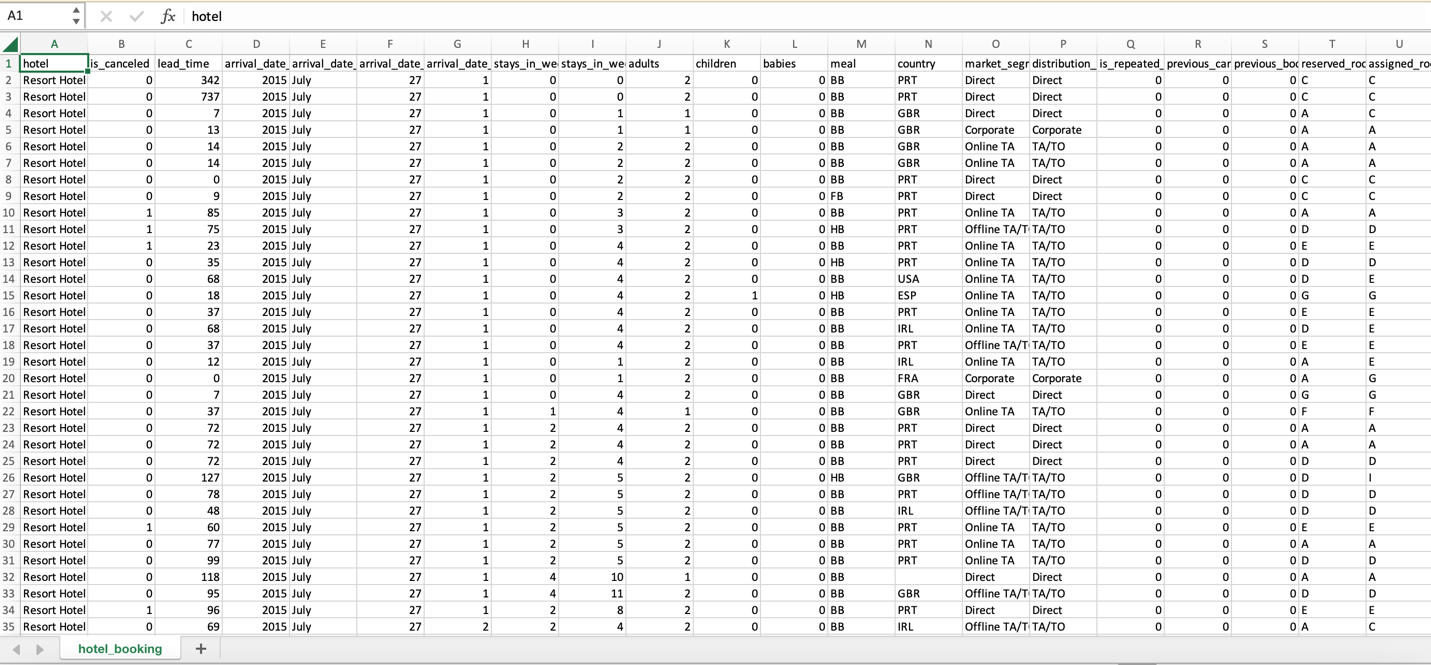
# **Data Collection and Preparation**

**Dataset:**

This dataset contains 119390 observations for a City Hotel and a Resort Hotel. Each observation represents a hotel booking between the 1st of July 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were canceled. Dataset was in .csv format.

**Sample of the raw dataset**

Since this is hotel real data, all data elements about hotel or customer identification were deleted. Four Columns, 'name', 'email', 'phone number', and 'credit card' have been artificially created and added to the dataset.

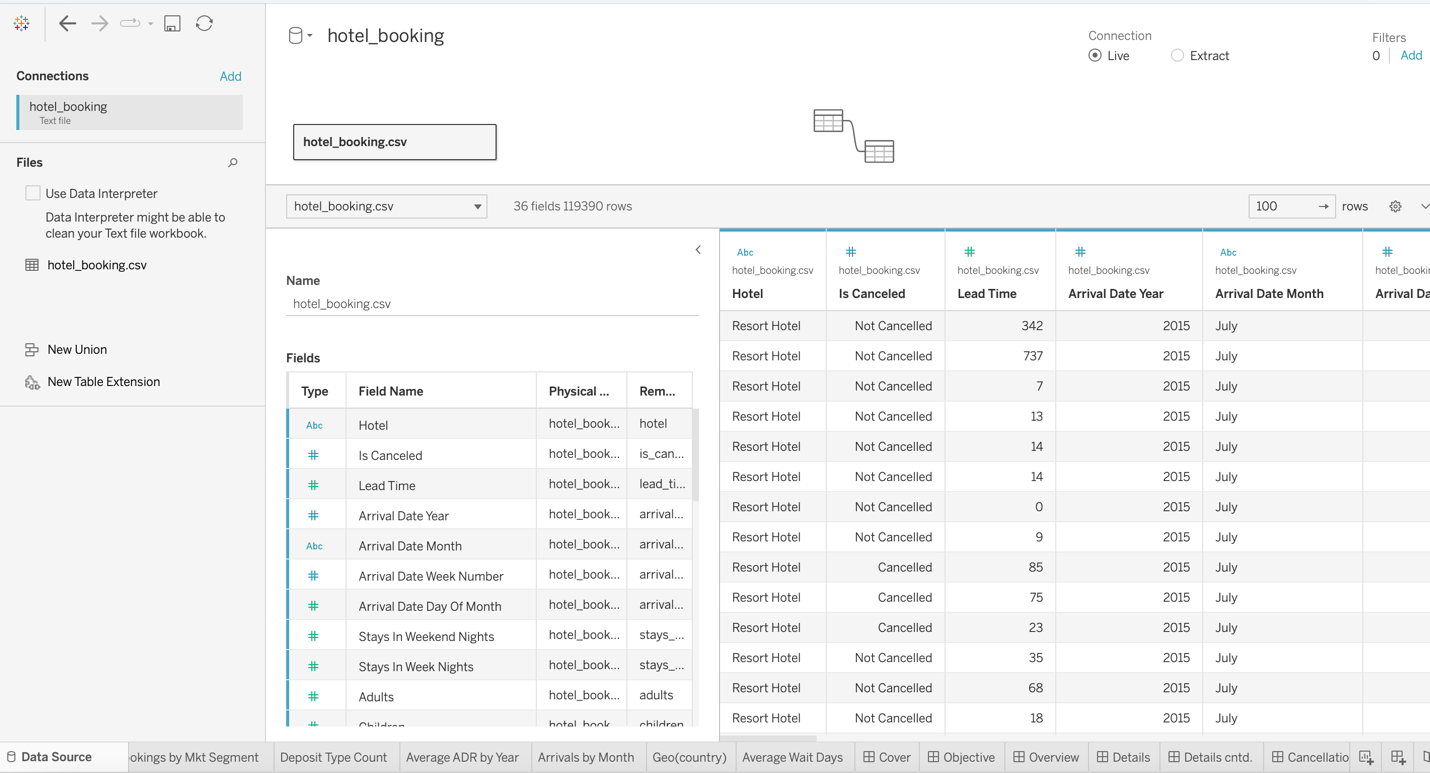


**Data Wrangling**

Based on the initial assessment we found that the data was pretty much clean except for some missing values in a few columns.

1. The dataset has a shape of (119390, 32) which means that it contains approximately 1.2 lakh rows and 32 columns.
2. Our Dataset has 4 columns with float64 type, 16 columns with int64 type, and 12 columns with the object type.
3. In our Dataset, we observed null values in the following columns:
   * 4 null values in the children column
   * 488 null values in the country column
   * 16,340 null values in the agent column
   * 112,593 null values in the company column

**Data preparation flow chart from Tableau Software.**



**Summary of Major Achievements and key dissemination activities**

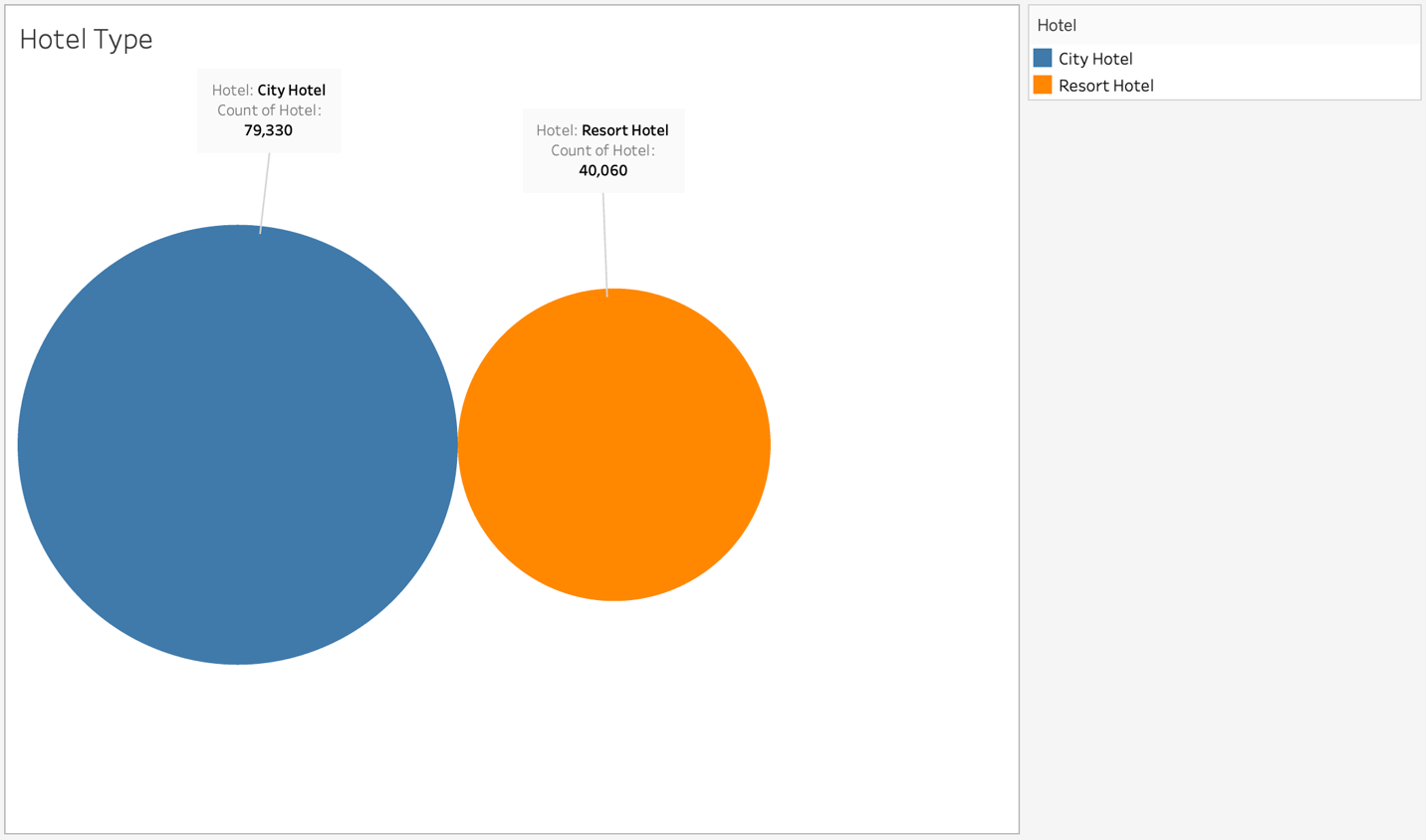
Understanding the data have taken a lot of time since the topic is related to hotels. The dataset has a lot of null values, so dealing with the null values has taken a lot of time.

# **Deliverables**

The most important deliverable of this project is to provide 5 important dashboards for overall Hotel Booking analysis and story. A project presentation is also given. And the Final Project report will be submitted. All the data processing and visualization codes will be provided.

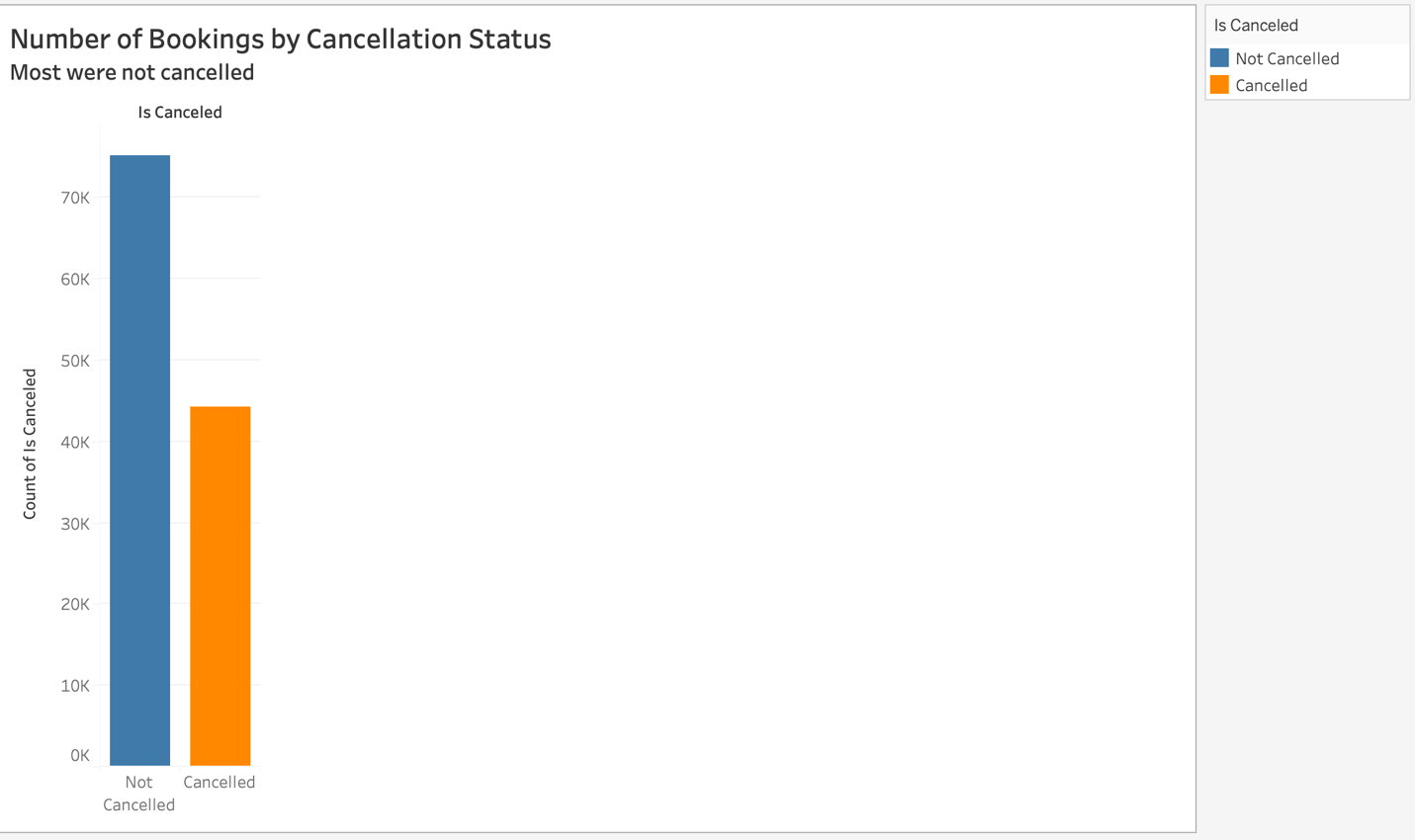
**Let’s find insights from the Dataset**

**Diagram 1**

****

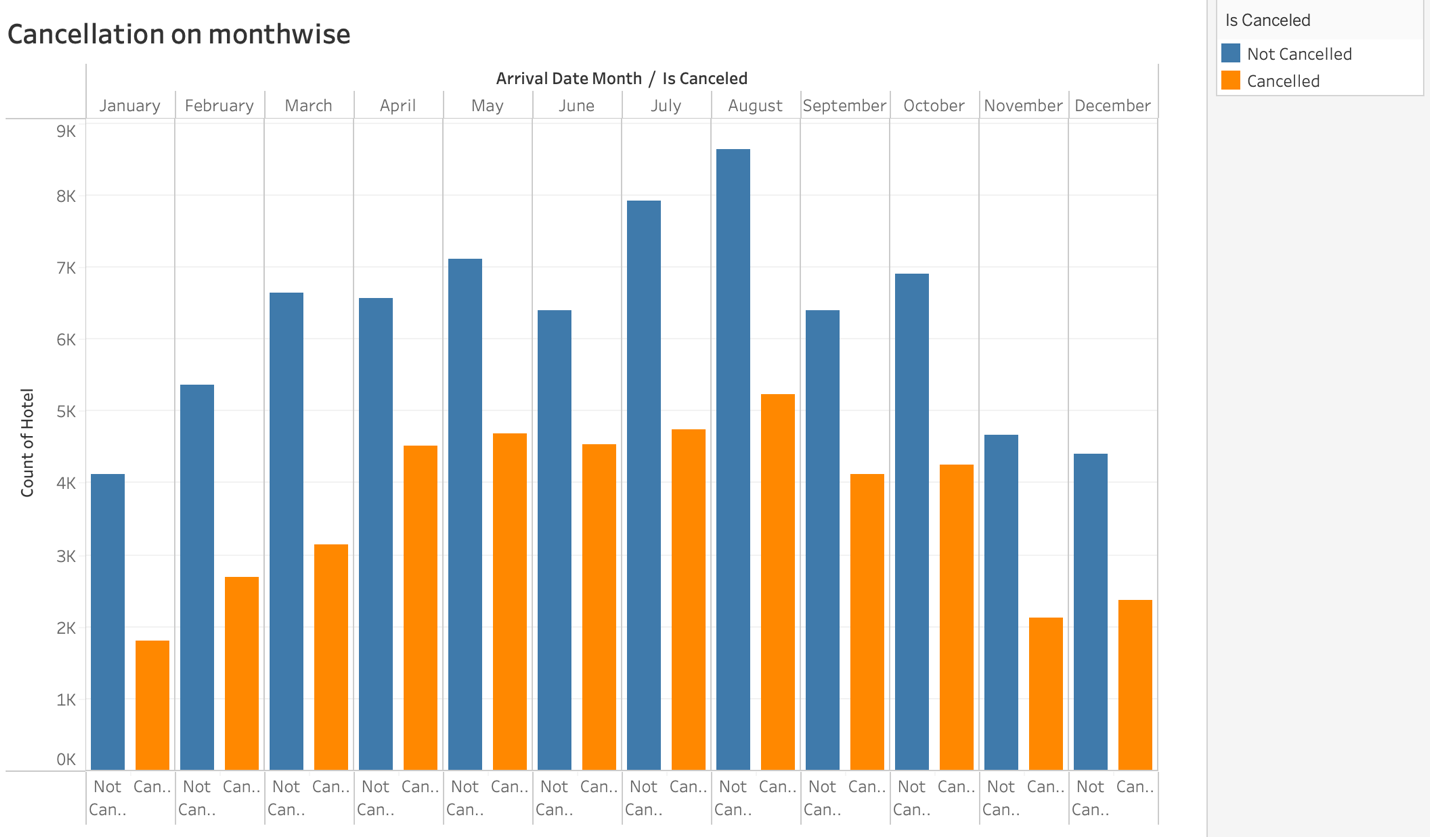
The data covers two datasets that provide information on hotel demand. Two hotels are in the area, the first is mentioned as a resort hotel(H1), while the second is identified as a city hotel(H2). Both sets of data show that the 40,060 observations in H1 are described by the same 31 variables. The 79,330 observations are in H2, which is the same structure as in H1. Each observation corresponds to a hotel reservation.

**Diagram 2**

****

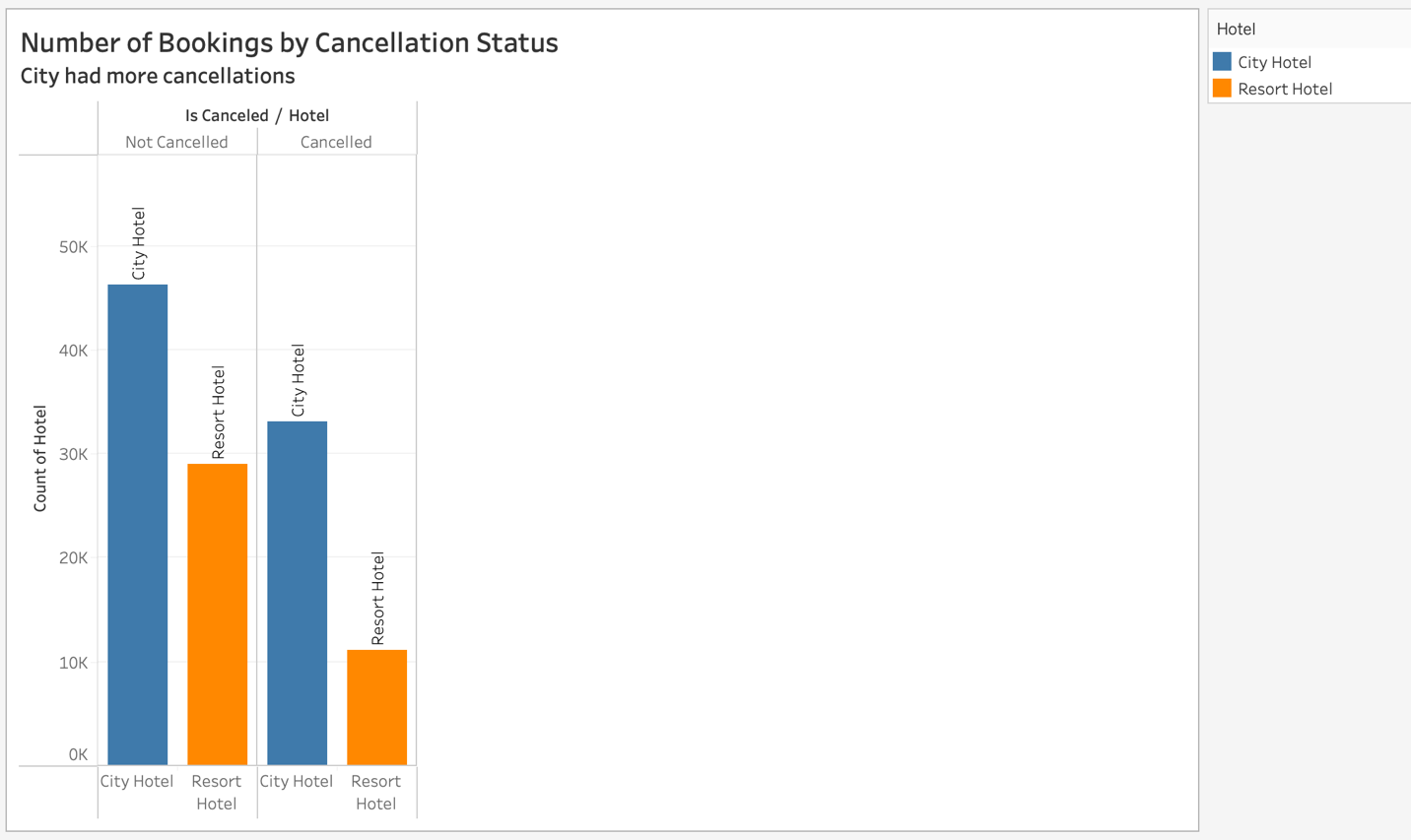
City Hotels have over 33k booking cancellations, resulting in a revenue loss of 3.5M, while the Resort hotels have over 11k cancellations owing to a loss of 1.1M. This includes reservations that were really arrived but later canceled in either of the two databases.

**Diagram 3**

****

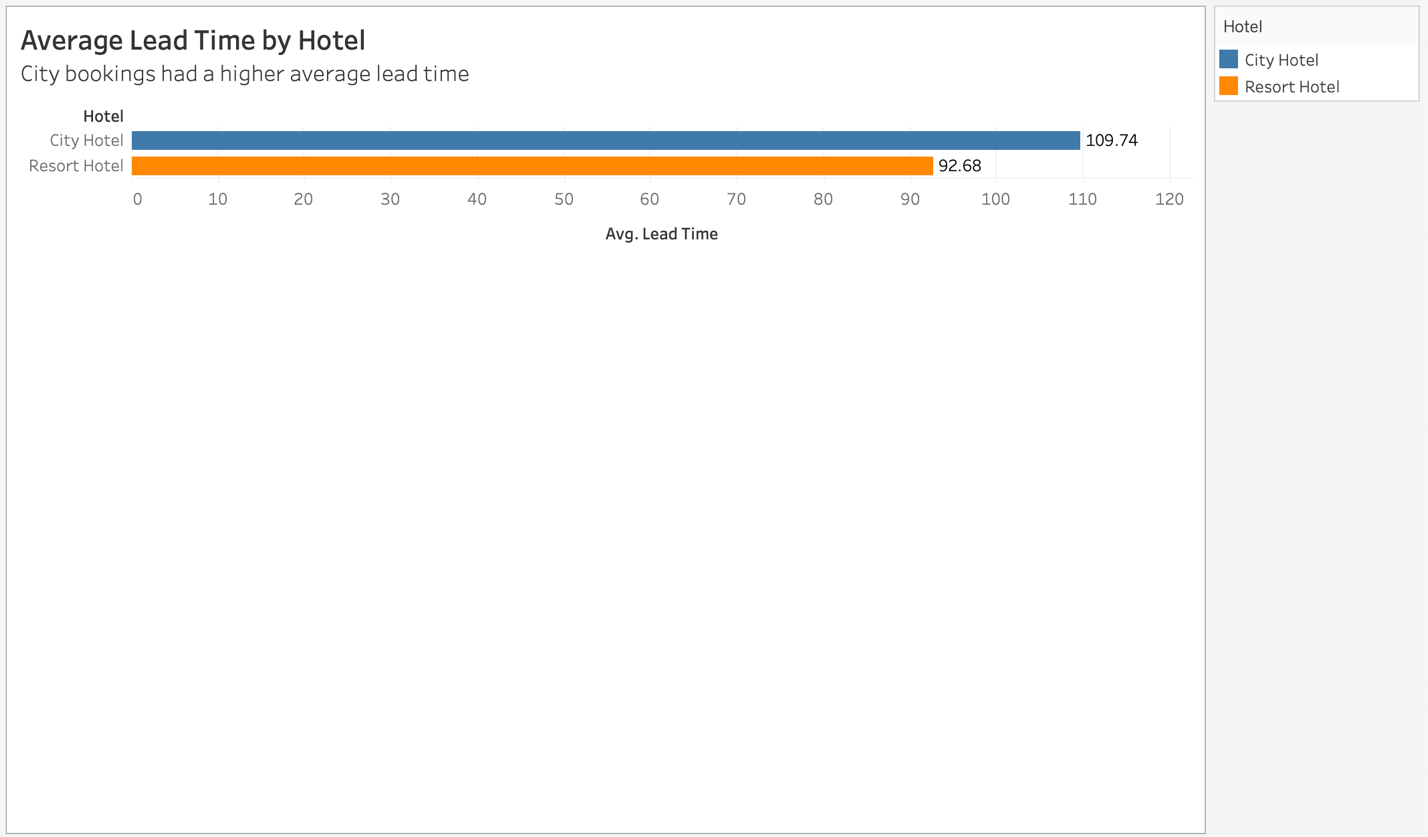
This graph data displays that the highest cancellations rate was seen in the month of August representing around 5,000. The lowest cancellations rate was in the month of January.

**Diagram 4:**

****

City Hotels have over 33k booking cancellations, resulting in a revenue loss of 3.5M, while the Resort hotels have over 11k cancellations owing to a loss of 1.1M. Almost 19 % of people did not cancel their bookings even after not getting the same room that they reserved while booking a hotel. Only 2.5 % of people canceled the booking. It is evident that even if customers are not given the rooms they requested throughout the booking process, there is no appreciable impact on cancellations of reservations.

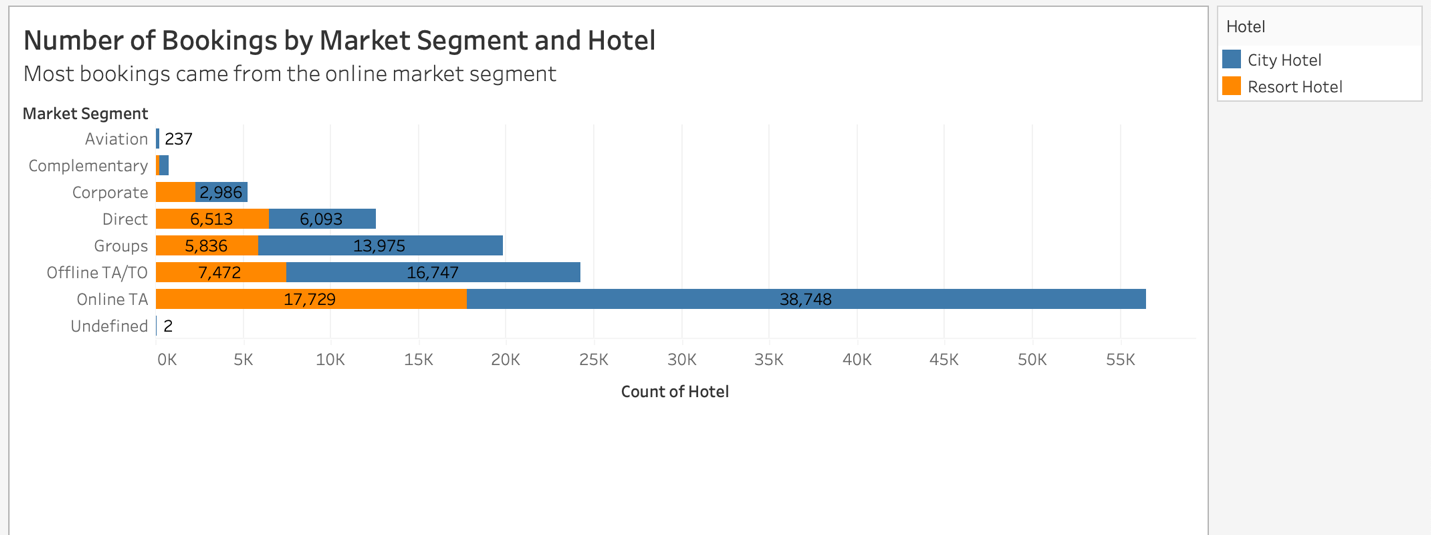
**Diagram 5:**

****

**Average lead time by the hotel:**

City bookings had a higher average lead time compared to resort hotels.

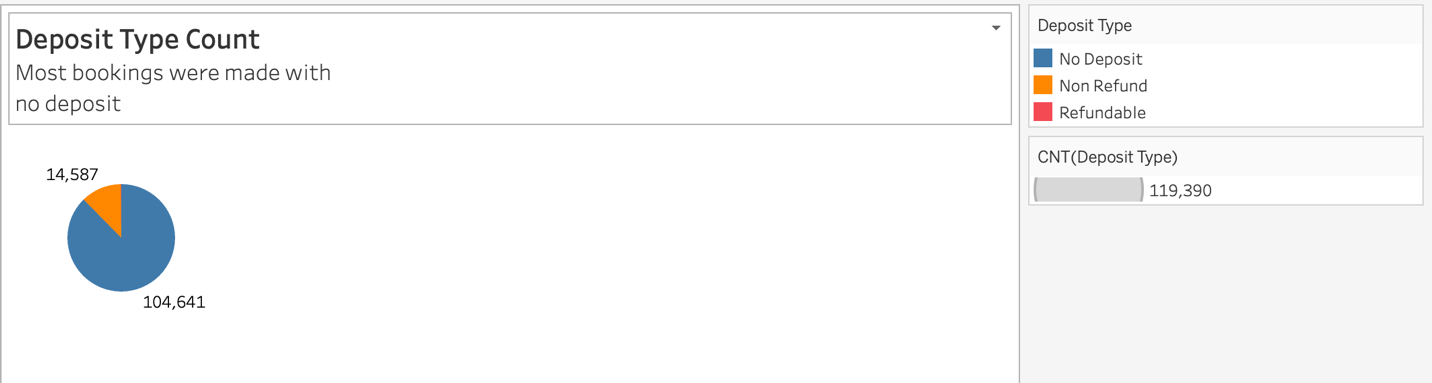
**Diagram 6:**

****

**Number of Bookings by Market Segment:**

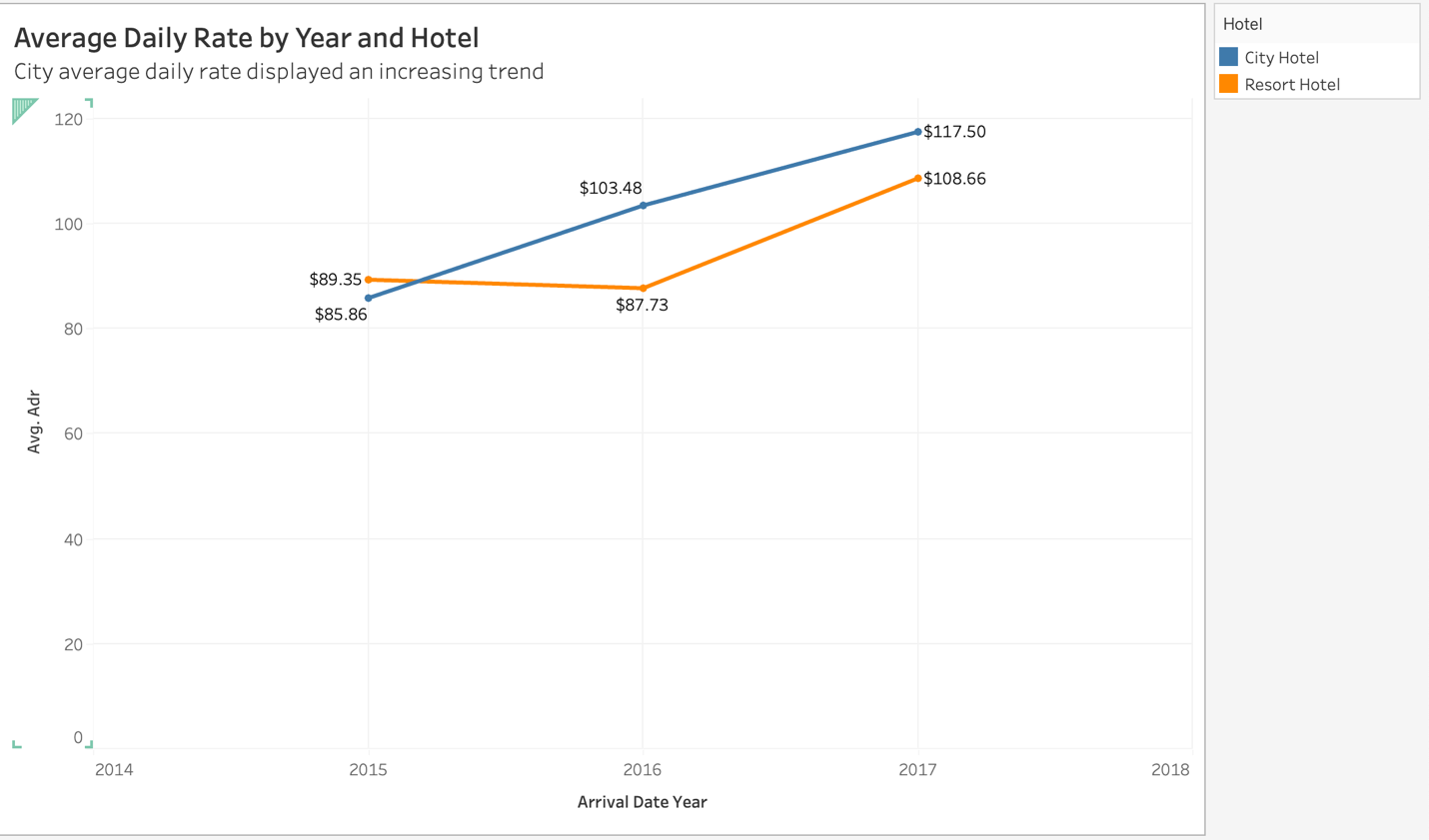
Here the blue color indicates “City Hotel” and orange indicates “Resort Hotel”. Most bookings came from the online market segment for both the resort and city hotels. Over 82% of the bookings come from the guests booking through the distributor's TA/TO (Travel Agent and Operator) Distributors which get revenue of over 3.2M for the City hotel and 0.99M for the Resort Hotel. Of the TA/TO distributor the market segment Online, group and offline bookings result in 92% cancellation representatives both across city hotels and resort hotels.

**Diagram 7:**

****

Most of the bookings were made with no deposit.

**Diagram 8:**

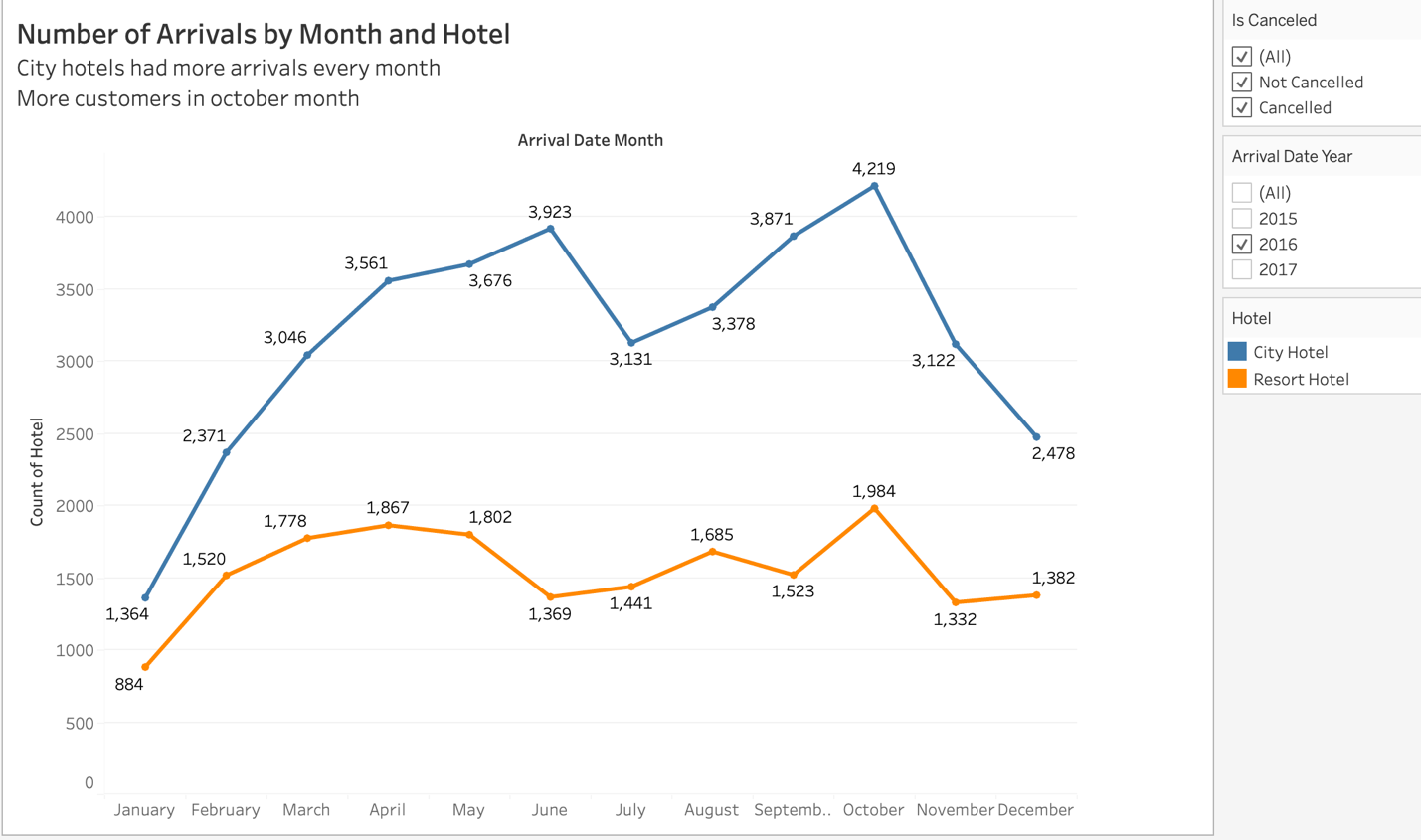
****

In this line graph, the blue line indicates the revenue of city hotels and the orange indicates resort hotels. As we can see in the above graph, City hotel's ADR is an increasing trend compared to resorts.

In comparison to City Hotels, the ADR for Resort Hotel is higher in the months of July and August. Perhaps people wish to vacation in resort hotels this summer.

January, February, March, April, October, November, and December are the ideal months for visitors to resort or city hotels because of the low average daily rate throughout these months.

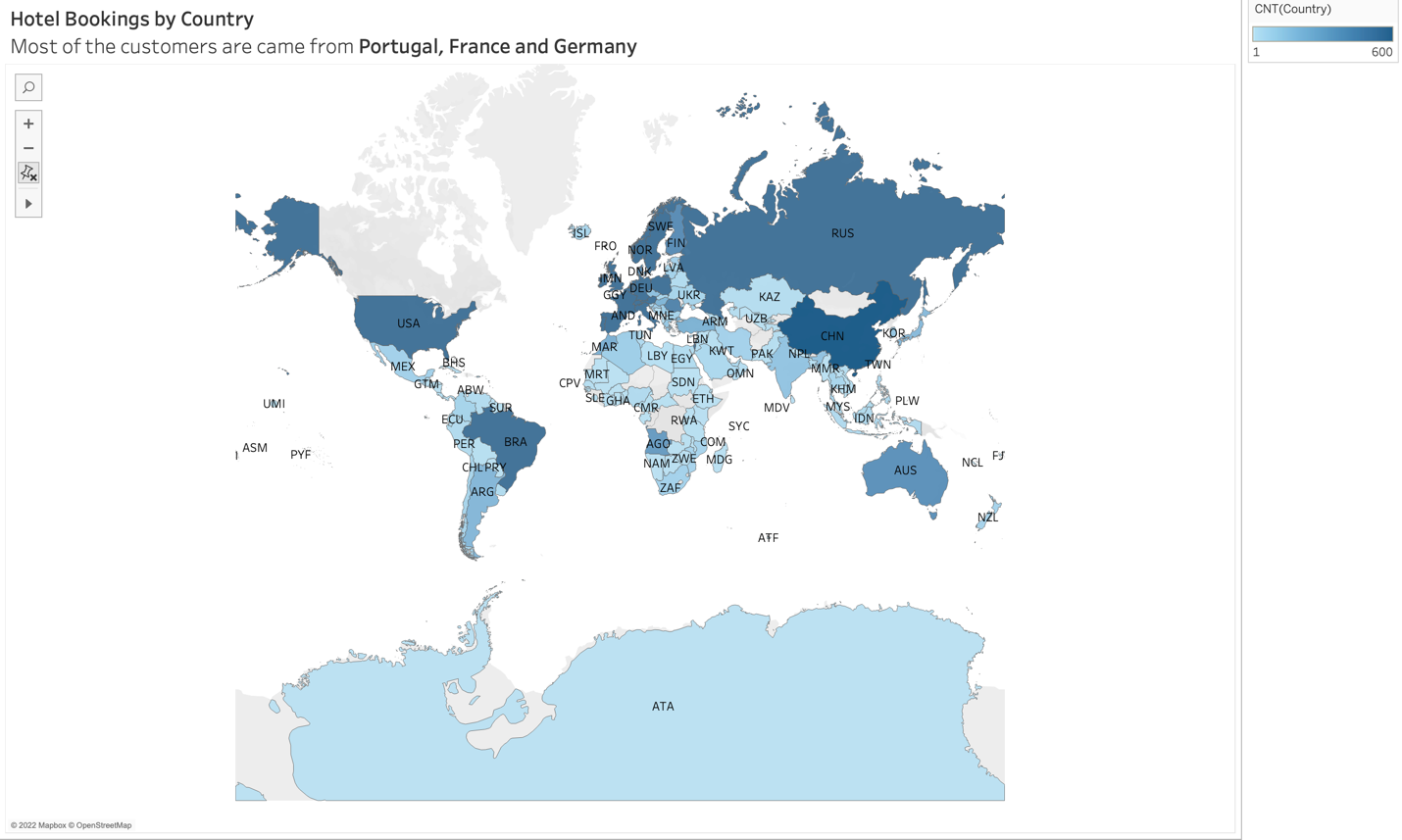
**Diagram 9:**

****

**Which are the peak and last month of booking concerning the number of bookings?**

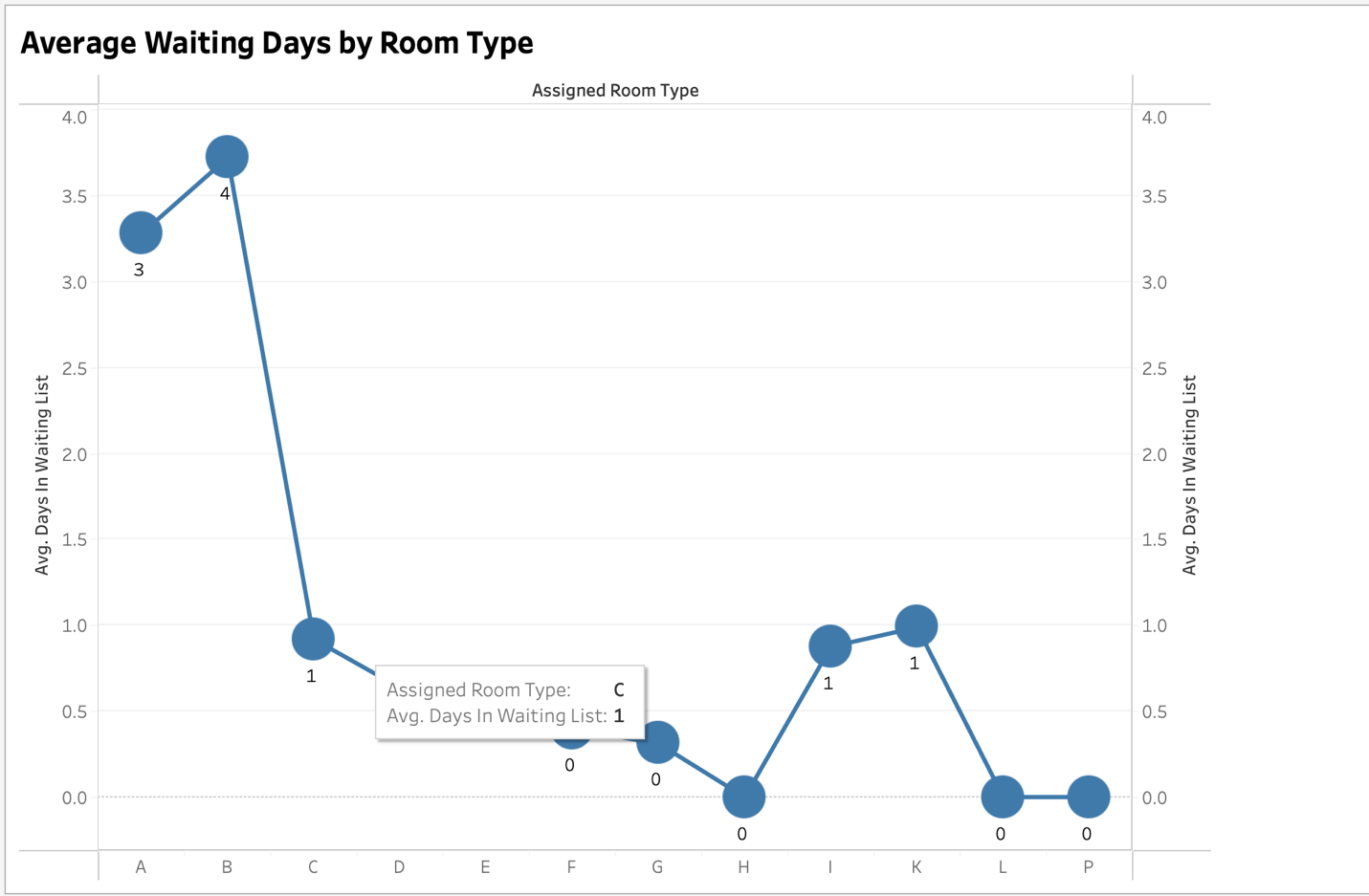
From the graph data, we can see that city hotel had more arrivals every month. In the year 2016, June and October’s months were most occupied by bookings representing 3,923 for June and 4,219 for August. For both sorts of hotels, the busiest periods are during the summer months of June and October, which is to be expected as summer approaches. Moreover, the least bookings were made in the month of January approximately 1,364, November roughly 3,122, and December which is 2,478.

**Diagram 10:**

****

European Countries get the most bookings over 84% (37k), which has higher compared to the other countries. The country Portugal has a percentage of bookings close to 64% (27k).

**Diagram 11:**

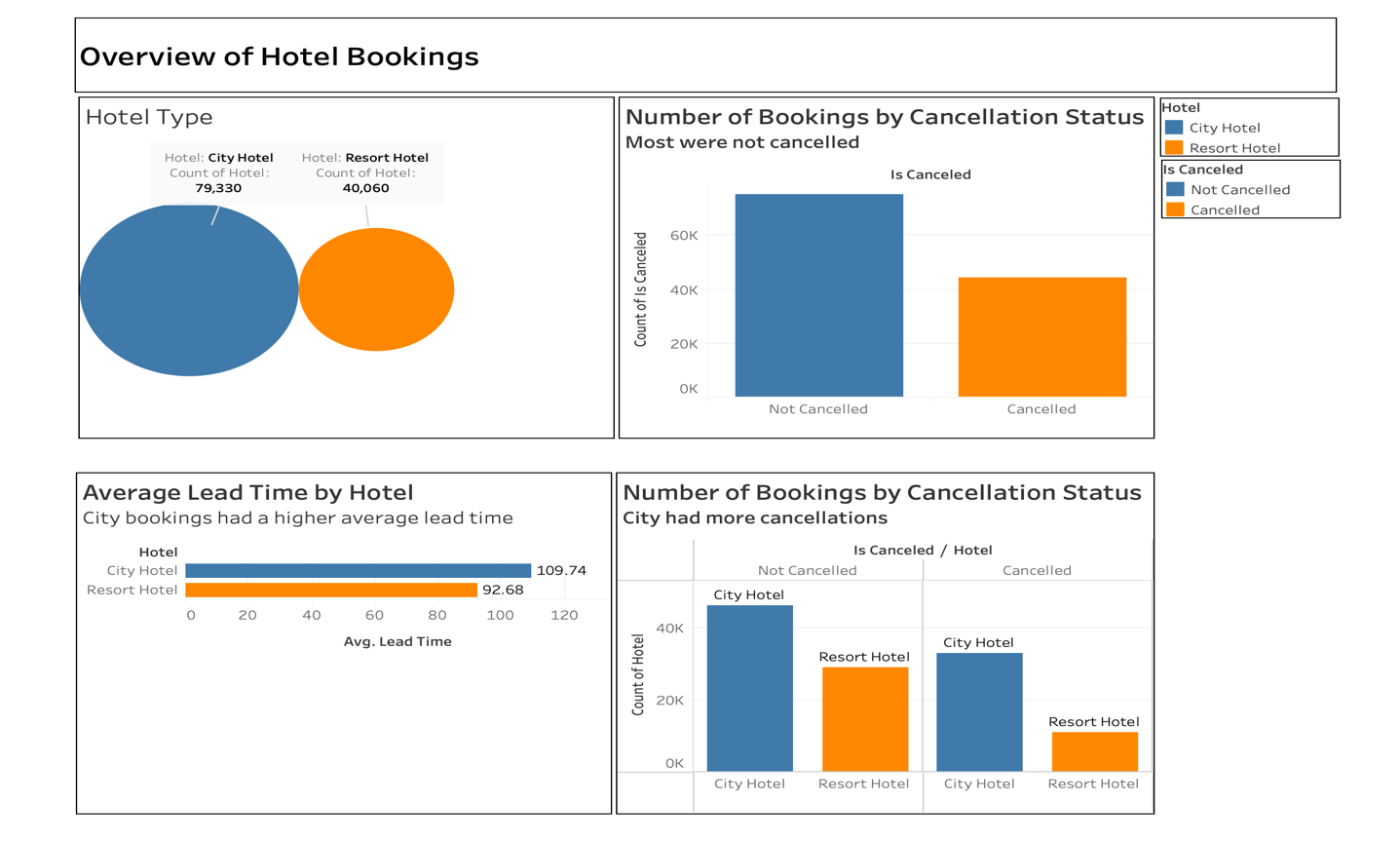
****

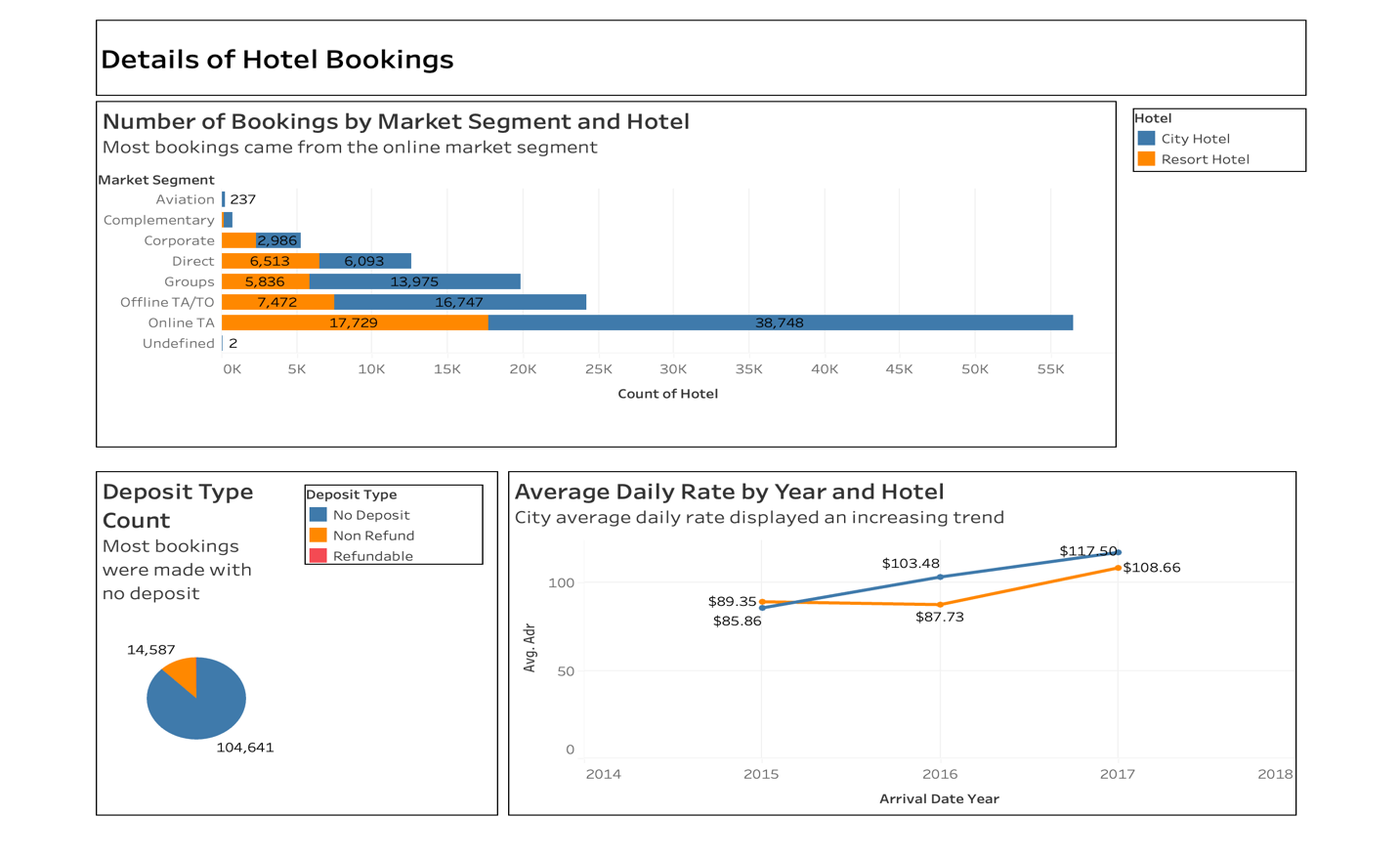
**Average Waiting Days by room type:**

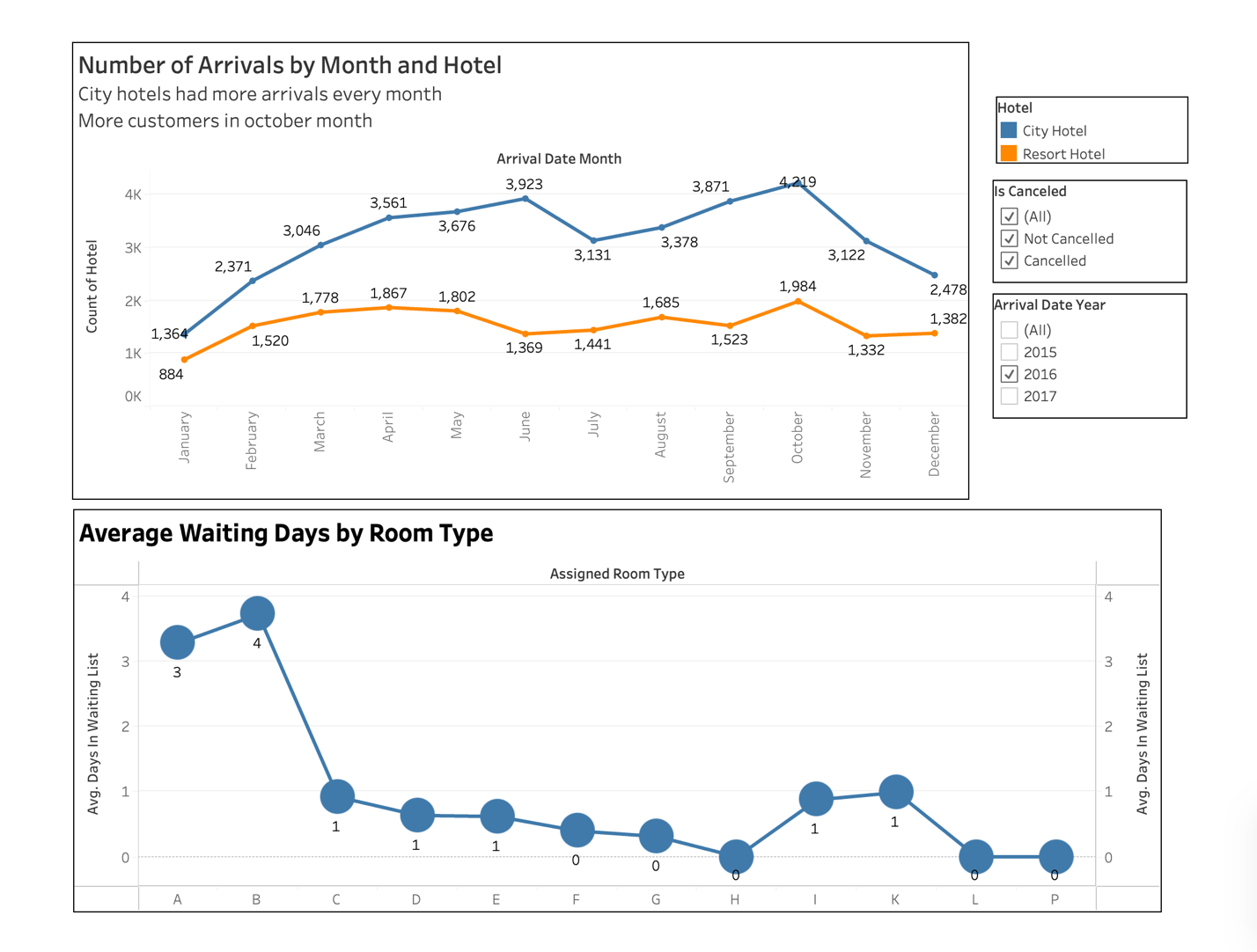
There are different room types like A, B, C, and so on. The average wait days for every type of hotel are displayed in the above graph. Here are the insights:

B type room has the highest average waiting days of 4 days and H, L, and P have the lowest of 0 days.

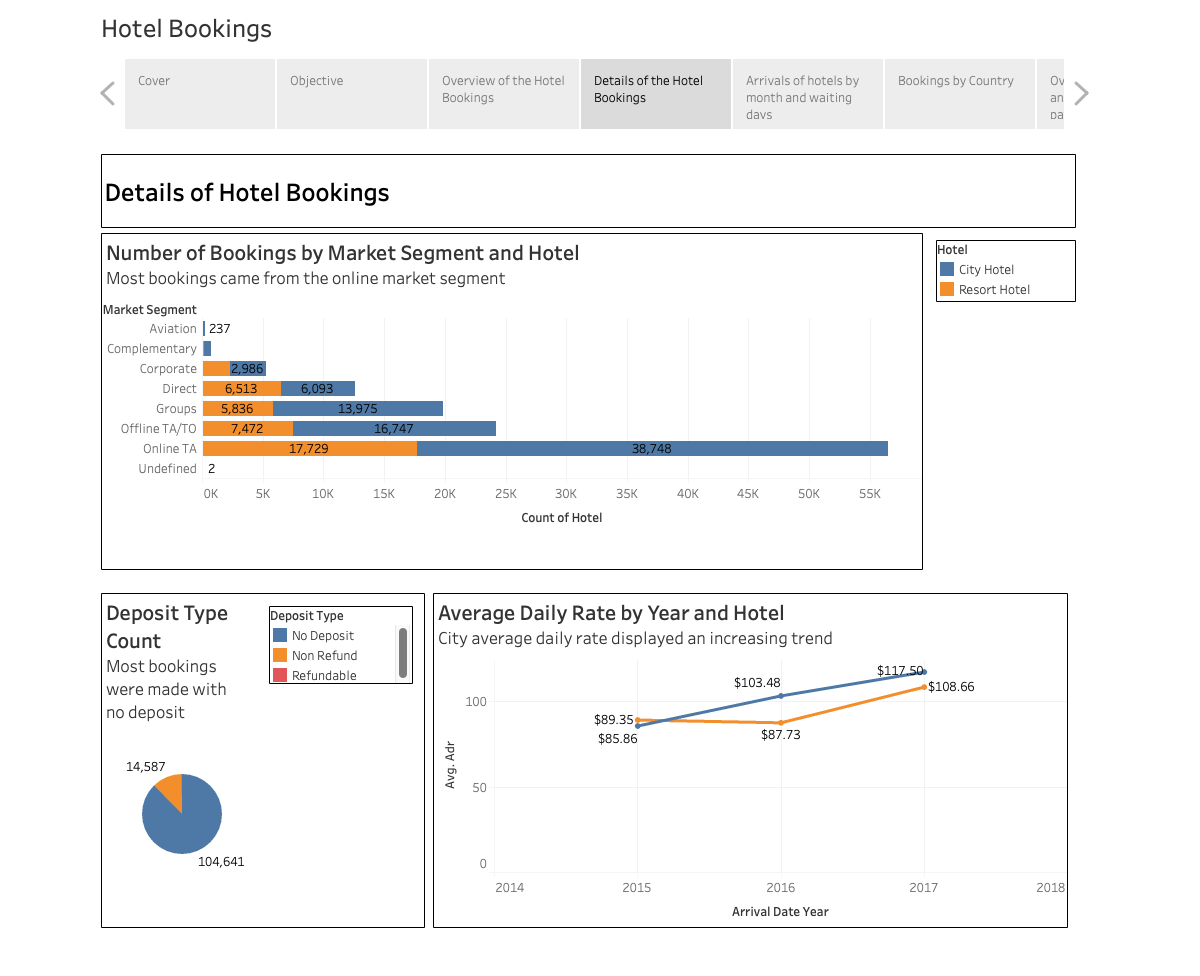
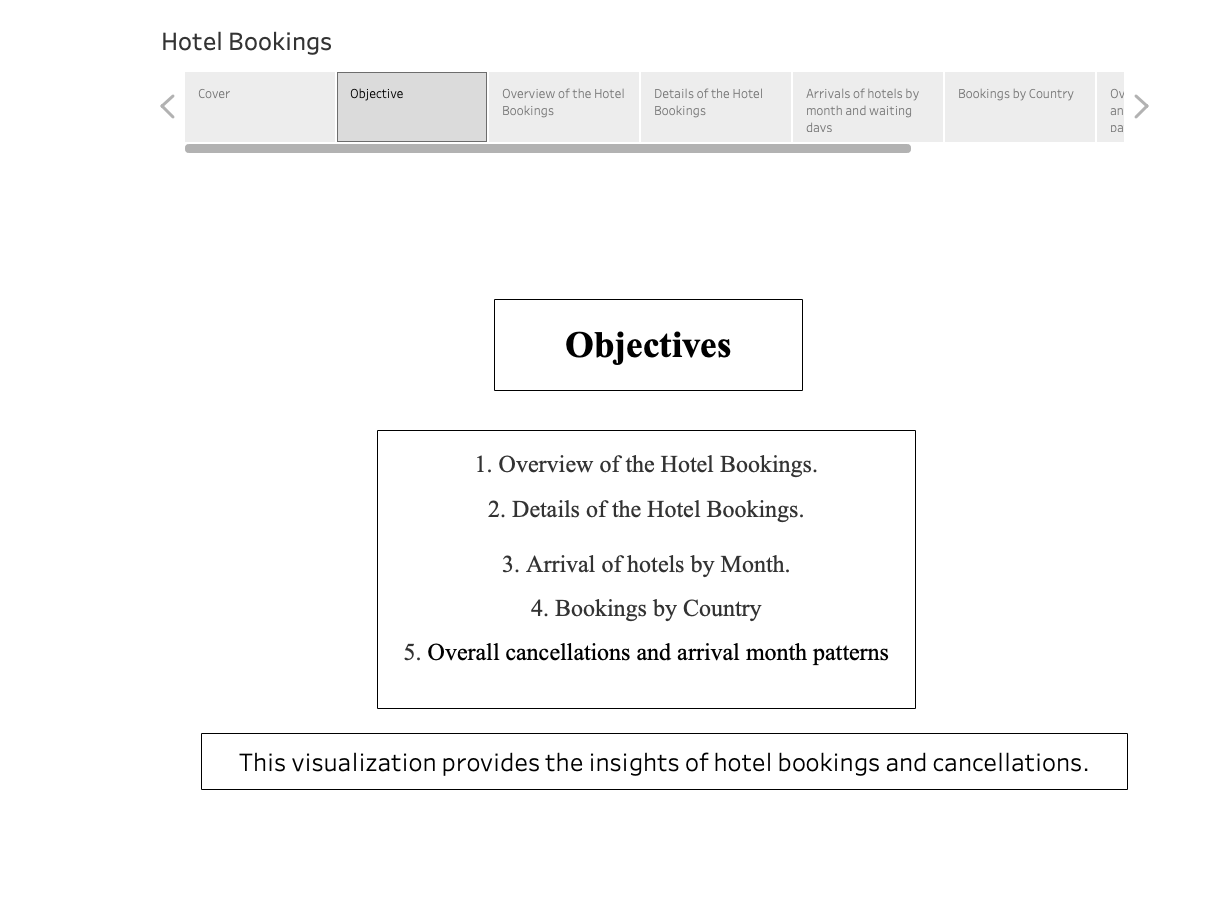
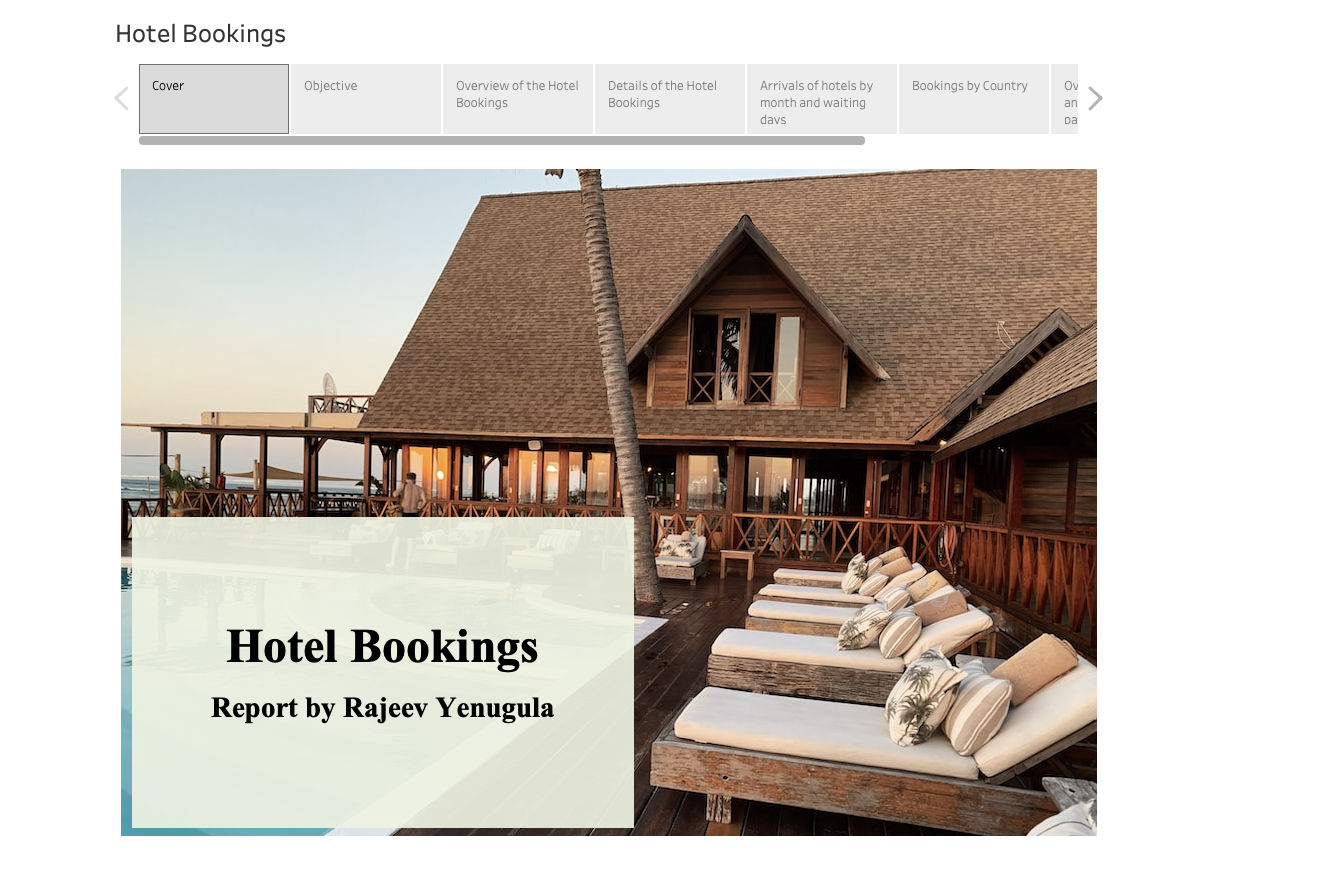
**Dashboard 1: Overview of the dataset**

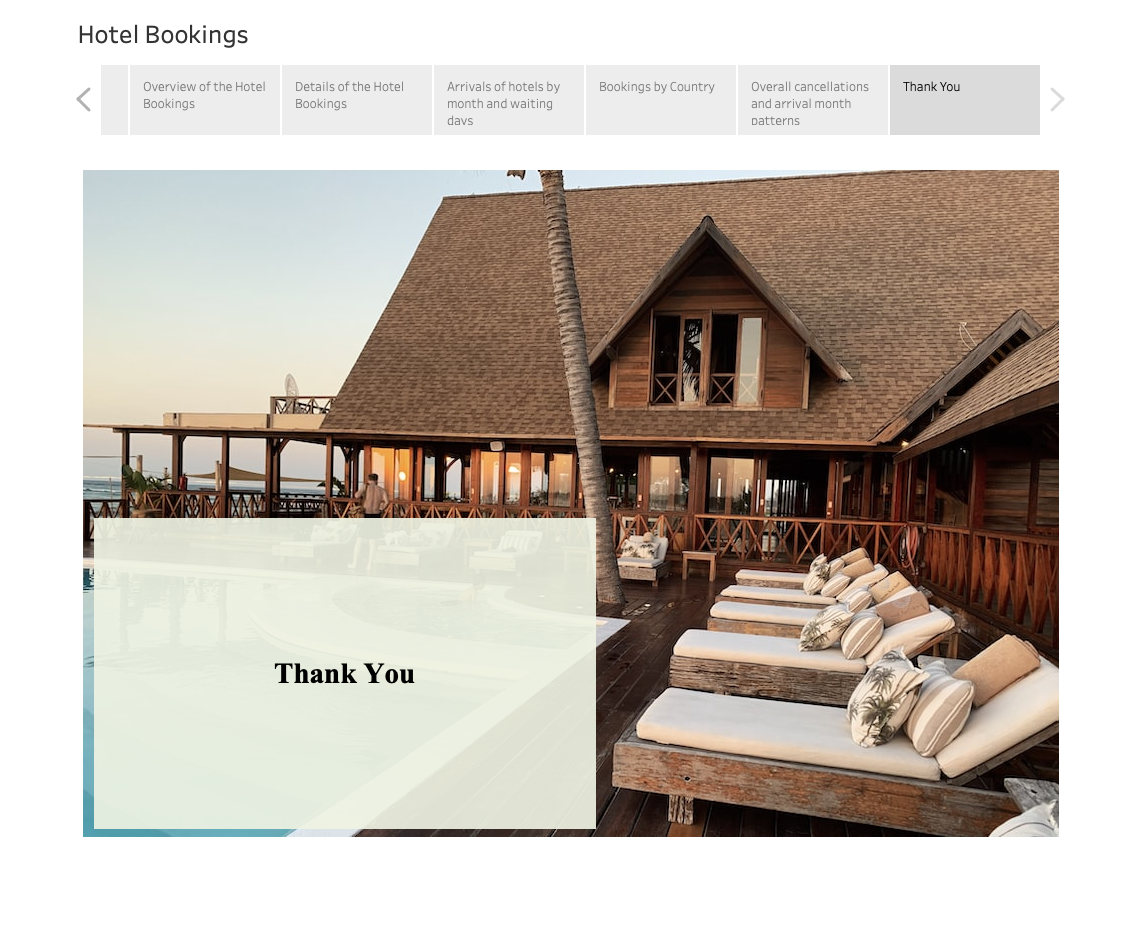
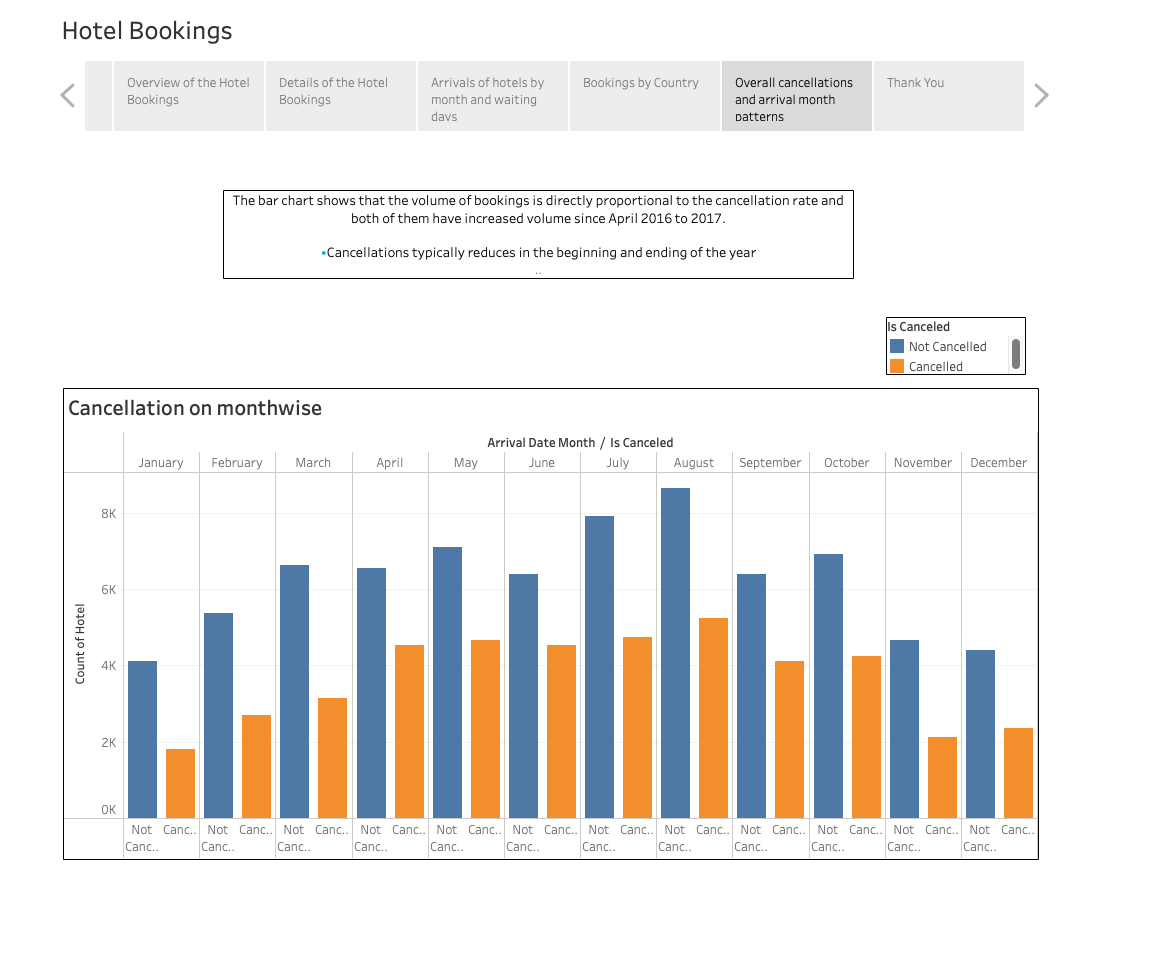
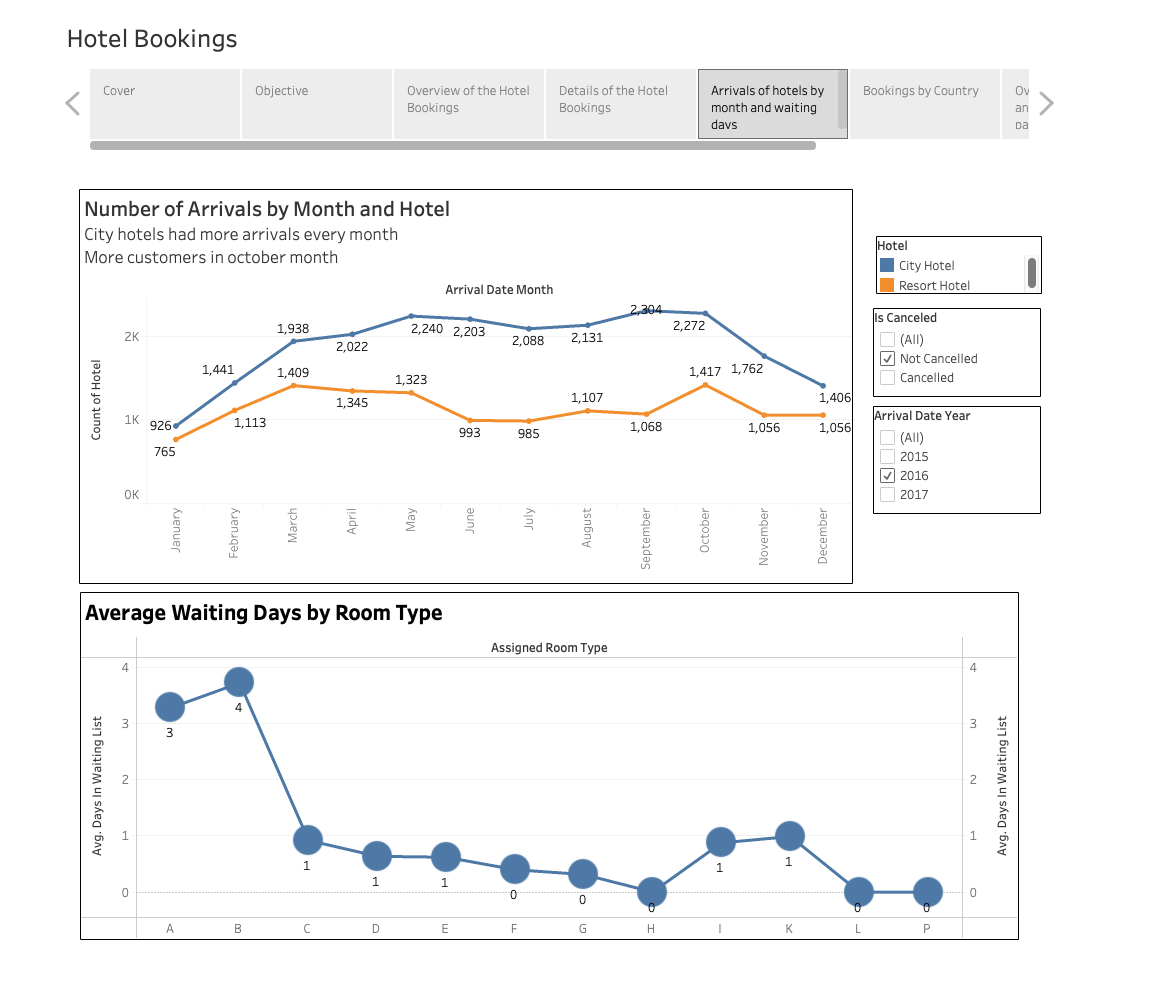
**Dashboard 2: Details of the dataset**

****

**Dashboard 3: Details (CNTD….)**

**Story**

****

****

**Conclusion**

* The month of August and July receives the most no. of booking.
* Booking for city hotels is twice as much for resort hotels.
* Repeated customers cancel their hotel in very rare cases.
* Customers coming from the aviation industry have very less time i.e. they book urgently
* People with no kids prefer to choose city hotels over resort hotels

**Strategies to counter high cancellations at Hotel**

* Since we see, our repetitive customers are the most loyal, to maintain them we can provide them with some bonus points, which can be redeemed in the next booking
* The month of January and December receives less no. of booking, so hotels can offer

discounted packages for these months.

* Families with kids prefer resorts, we can provide holiday family packages.
* Great no. of the bookings are coming from travel agents, so we can provide them some commission.

**References**

Dataset: <https://www.kaggle.com/datasets/mojtaba142/hotel-booking>

The data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.