Hotel Bookings Demand :

Data Insight

INTRODUCTION

Objective:

- **To gain insights about the demand bookings between two hotels and create visuals to indicate these.**

Some questions to answer in this analysis:

- See which month has the highest number of bookings for families with children and without children;
- See which type of hotel requires more parking spaces.

Dataset:

- For this project, the dataset available has 119390 entries and 32 columns.
- The dataset was taken from Kaggle: https://www.kaggle.com/jessemostipak/hotel-booking-demand

- Columns:
  - 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: day of arrival date
  - Stays_in_weekend_nights: number of weekend nights the guest stayed or booked to stay at the hotel
  - Stays_in_week_nights: number of week nights the guest stayed or booked to stay at the hotel
  - Adults: number of adults
  - Children: number of children
  - Babies: number of babies
  - Meal: type of meal booked
  - Country: country of origin

Dataset

- Market_segment: market segment designation. In categories. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- Distribution_channel: booking distribution channel.
- Is_repeated_guest: value indicating if the booking name was from a repeated guest (1) or not (0)
- Previous_bookings_not_canceled: # of previous bookings not canceled by the customer prior to 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 reasons
- Booking_changes: # of changes made to the booking from the moment the booking was entered on the PMS
- Deposit_type: indication on if the customer made a deposit to guarantee the booking.
- Agent: ID of the travel agency that made the booking
- Company: ID of the company that made the booking
- Days_in_waiting_list: # of days the booking was in the waiting list before it was confirmed to the customer
- Customer_type: type of booking, assuming one of the four type = contract, transient, transient-party, group
- Adr: average daily rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
- Required_car_parking_spaces: # of car parking spaces required by the customer
- Total_of_special_requests: # of special requests made by the customer
- Reservation_status: reservation last status, assuming one of the categories = check-out, canceled, no-show
- Reservation_status_date: date at which the last status was set.

LIBRARIES:

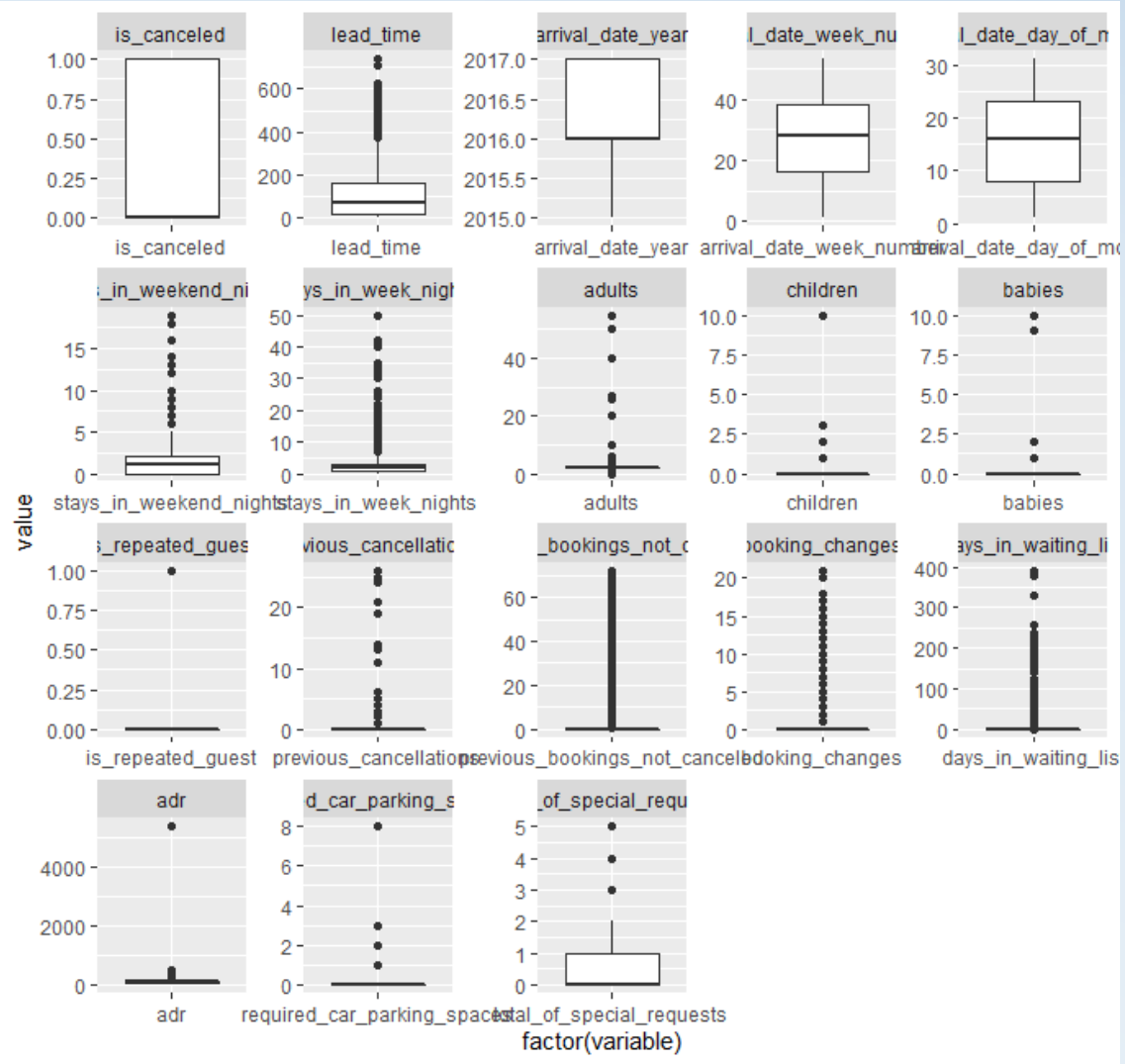
For this project, the following libraries were used:

- dplyr: for data manipulation
- ggplot2: for data visualization
- tidyr: to tidy up the data
- DT: provides an R interface to the JavaScript library
- reshape2: makes it easy to transform data between wide and long formats

DATASET SUMMARY

```
         hotel        is_canceled       lead_time   arrival_date_year arrival_date_month arrival_date_week_number arrival_date_day_of_month
City Hotel  :79330   Min.   :0.0000   Min.   :  0   Min.   :2015      August :13877      Min.   : 1.00            Min.   : 1.0
Resort Hotel:40060   1st Qu.:0.0000   1st Qu.: 18   1st Qu.:2016      July   :12661      1st Qu.:16.00            1st Qu.: 8.0
                     Median :0.0000   Median : 69   Median :2016      May    :11791      Median :28.00            Median :16.0
                     Mean   :0.3704   Mean   :104   Mean   :2016      October:11160      Mean   :27.17            Mean   :15.8
                     3rd Qu.:1.0000   3rd Qu.:160   3rd Qu.:2017      April  :11089      3rd Qu.:38.00            3rd Qu.:23.0
                     Max.   :1.0000   Max.   :737   Max.   :2017      June   :10939      Max.   :53.00            Max.   :31.0
                                                                      (Other):47873
stays_in_weekend_nights stays_in_week_nights     adults          children         babies                 meal          country
Min.   : 0.0000         Min.   : 0.0         Min.   : 0.000   Min.   : 0.0000   Min.   : 0.000000   BB        :92310   PRT    :48590
1st Qu.: 0.0000         1st Qu.: 1.0         1st Qu.: 2.000   1st Qu.: 0.0000   1st Qu.: 0.000000   FB        :  798   GBR    :12129
Median : 1.0000         Median : 2.0         Median : 2.000   Median : 0.0000   Median : 0.000000   HB        :14463   FRA    :10415
Mean   : 0.9276         Mean   : 2.5         Mean   : 1.856   Mean   : 0.1039   Mean   : 0.007949   SC        :10650   ESP    : 8568
3rd Qu.: 2.0000         3rd Qu.: 3.0         3rd Qu.: 2.000   3rd Qu.: 0.0000   3rd Qu.: 0.000000   Undefined: 1169   DEU    : 7287
Max.   :19.0000         Max.   :50.0         Max.   :55.000   Max.   :10.0000   Max.   :10.000000                     ITA    : 3766
                                                              NA's   :4                                               (Other):28635
     market_segment  distribution_channel is_repeated_guest previous_cancellations previous_bookings_not_canceled reserved_room_type
Online TA    :56477   Corporate: 6677      Min.   :0.00000   Min.   : 0.00000       Min.   : 0.0000               A      :85994
Offline TA/TO:24219   Direct   :14645      1st Qu.:0.00000   1st Qu.: 0.00000       1st Qu.: 0.0000               D      :19201
Groups       :19811   GDS      :  193      Median :0.00000   Median : 0.00000       Median : 0.0000               E      : 6535
Direct       :12606   TA/TO    :97870      Mean   :0.03191   Mean   : 0.08712       Mean   : 0.1371               F      : 2897
Corporate    : 5295   Undefined:    5      3rd Qu.:0.00000   3rd Qu.: 0.00000       3rd Qu.: 0.0000               G      : 2094
Complementary:  743                        Max.   :1.00000   Max.   :26.00000       Max.   :72.0000               B      : 1118
(Other)      :  239                                                                                              (Other): 1551
assigned_room_type booking_changes       deposit_type        agent          company       days_in_waiting_list         customer_type
A      :74053      Min.   : 0.0000   No Deposit:104641   9      :31961   NULL   :112593   Min.   :  0.000      Contract       : 4076
D      :25322      1st Qu.: 0.0000   Non Refund: 14587   NULL   :16340   40     :   927   1st Qu.:  0.000      Group          :  577
E      : 7806      Median : 0.0000   Refundable:   162   240    :13922   223    :   784   Median :  0.000      Transient      :89613
F      : 3751      Mean   : 0.2211                       1      : 7191   67     :   267   Mean   :  2.321      Transient-Party:25124
G      : 2553      3rd Qu.: 0.0000                       14     : 3640   45     :   250   3rd Qu.:  0.000
C      : 2375      Max.   :21.0000                       7      : 3539   153    :   215   Max.   :391.000
(Other): 3530                                            (Other):42797   (Other):  4354
     adr          required_car_parking_spaces total_of_special_requests reservation_status reservation_status_date
Min.   :  -6.38   Min.   :0.00000             Min.   :0.0000            Canceled :43017    2015-10-21:  1461
1st Qu.:  69.29   1st Qu.:0.00000             1st Qu.:0.0000            Check-Out:75166    2015-07-06:   805
Median :  94.58   Median :0.00000             Median :0.0000            No-Show  : 1207    2016-11-25:   790
Mean   : 101.83   Mean   :0.06252             Mean   :0.5714                               2015-01-01:   763
3rd Qu.: 126.00   3rd Qu.:0.00000             3rd Qu.:1.0000                               2016-01-18:   625
Max.   :5400.00   Max.   :8.00000             Max.   :5.0000                               2015-07-02:   469
                                                                                           (Other)   :114477
```

Column Boxplots

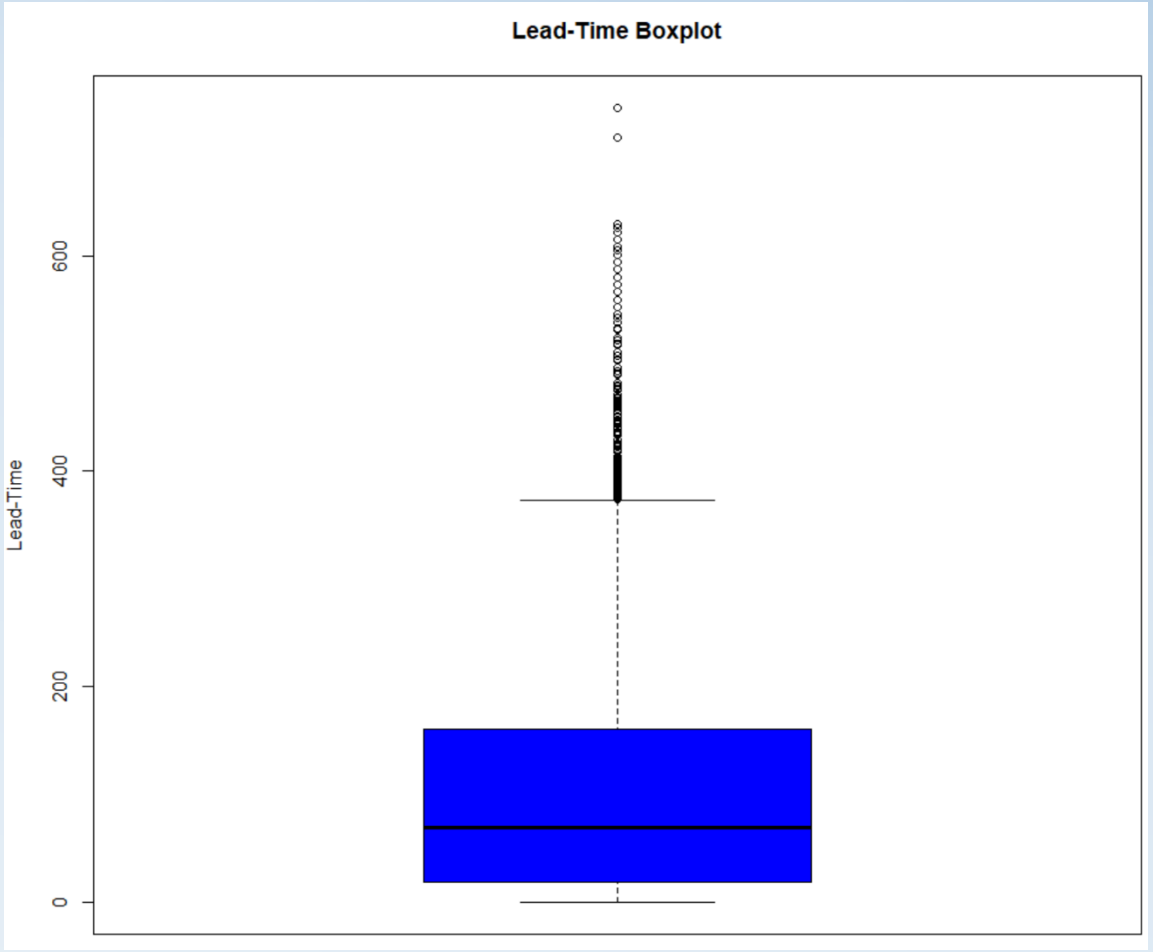


- The graph on the left contains the boxplot of the columns with the exception of some attributes that contains non-finite values like arrival_date_month and deposit_type.
- Most of the columns have outliers with the exception of “is_canceled”, “arrival_date_year”, “arrival_date_week_number”, and “arrival_date_day_of_month”.
- For this analysis, we will mainly focus on “lead_time”, “stays_in_weekend_nights”, and “total_of_special_requests” as these are included on most of the project.

Outlier Detection and Visualization:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	18	69	104	160	737

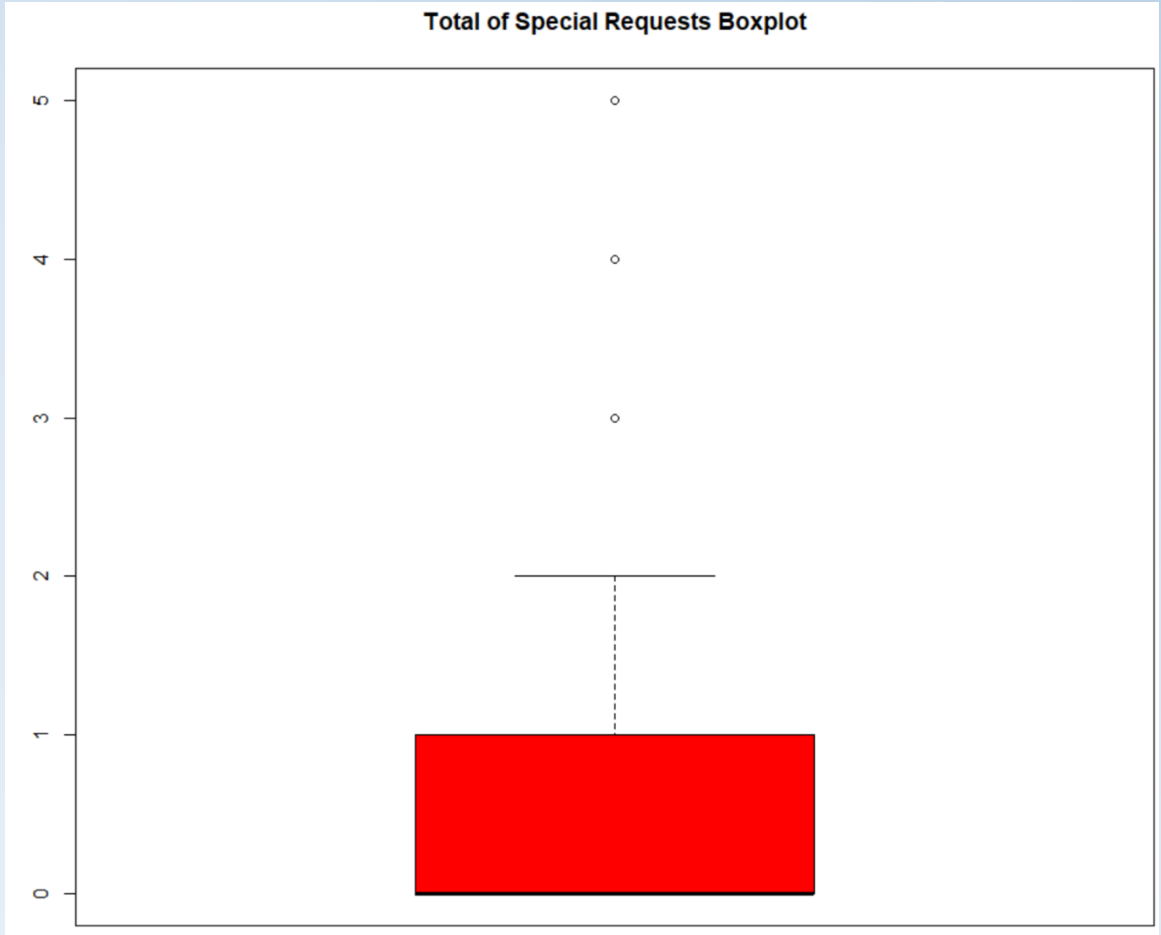
- Before any analysis was done, the whole dataset was checked for any NULL values. After confirming there are no NULL values, attributes were then checked for outliers that could influence results.
- First is the "Lead_Time" attribute. This is the number of days that elapsed between the entering date of the booking into the PMS and the arrival date. Looking at the summary, the maximum value it contains was 737 days.
- There are 3005 outliers in this column. This makes up about 2.5% of the total number of values.



Outlier Detection

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.0000	0.0000	0.5714	1.0000	5.0000

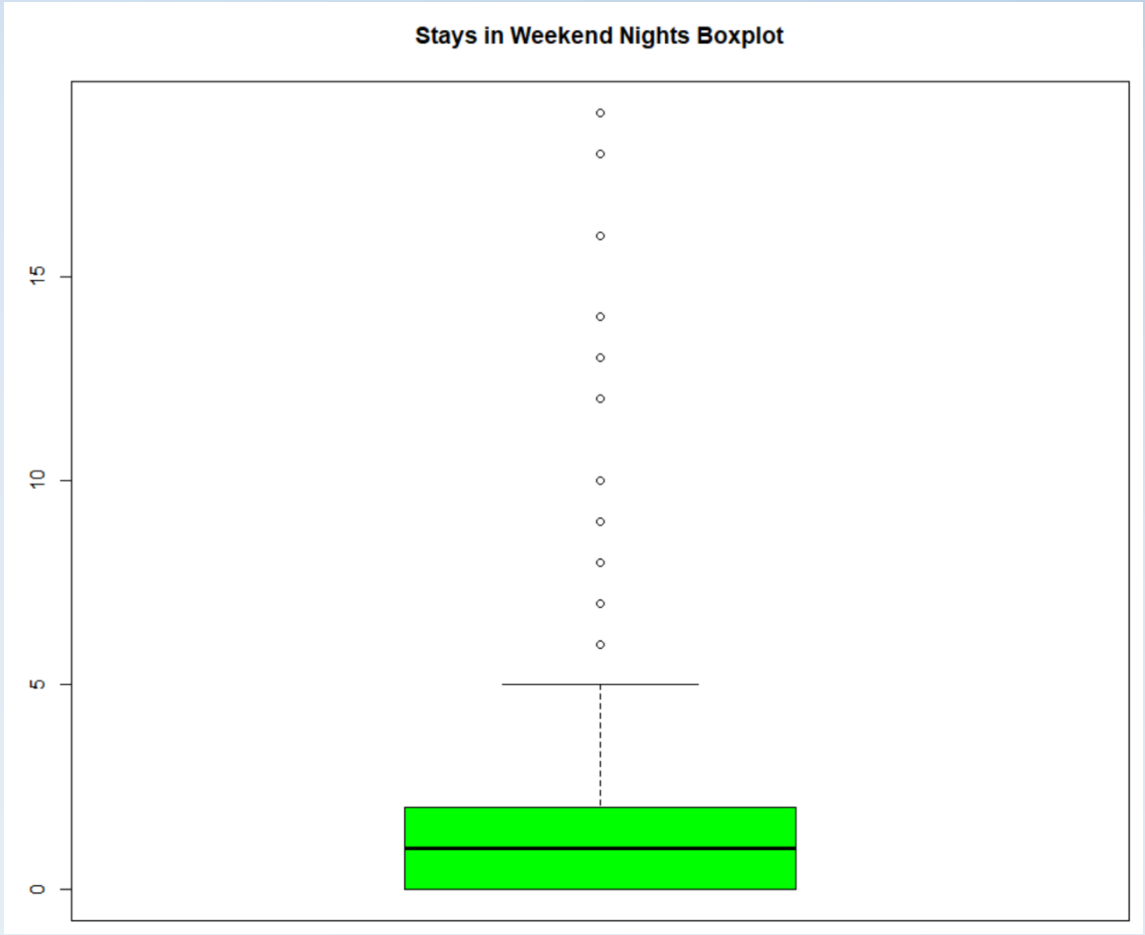
- The "total_of_special_requests" column have a total of 2887 outlier values.
- This is the number of special requests made by the customer (e.g. twin bed or high floor).
- This makes up about 2.4% of the total number of values.



Outlier Detection

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.0000	1.0000	0.9276	2.0000	19.0000

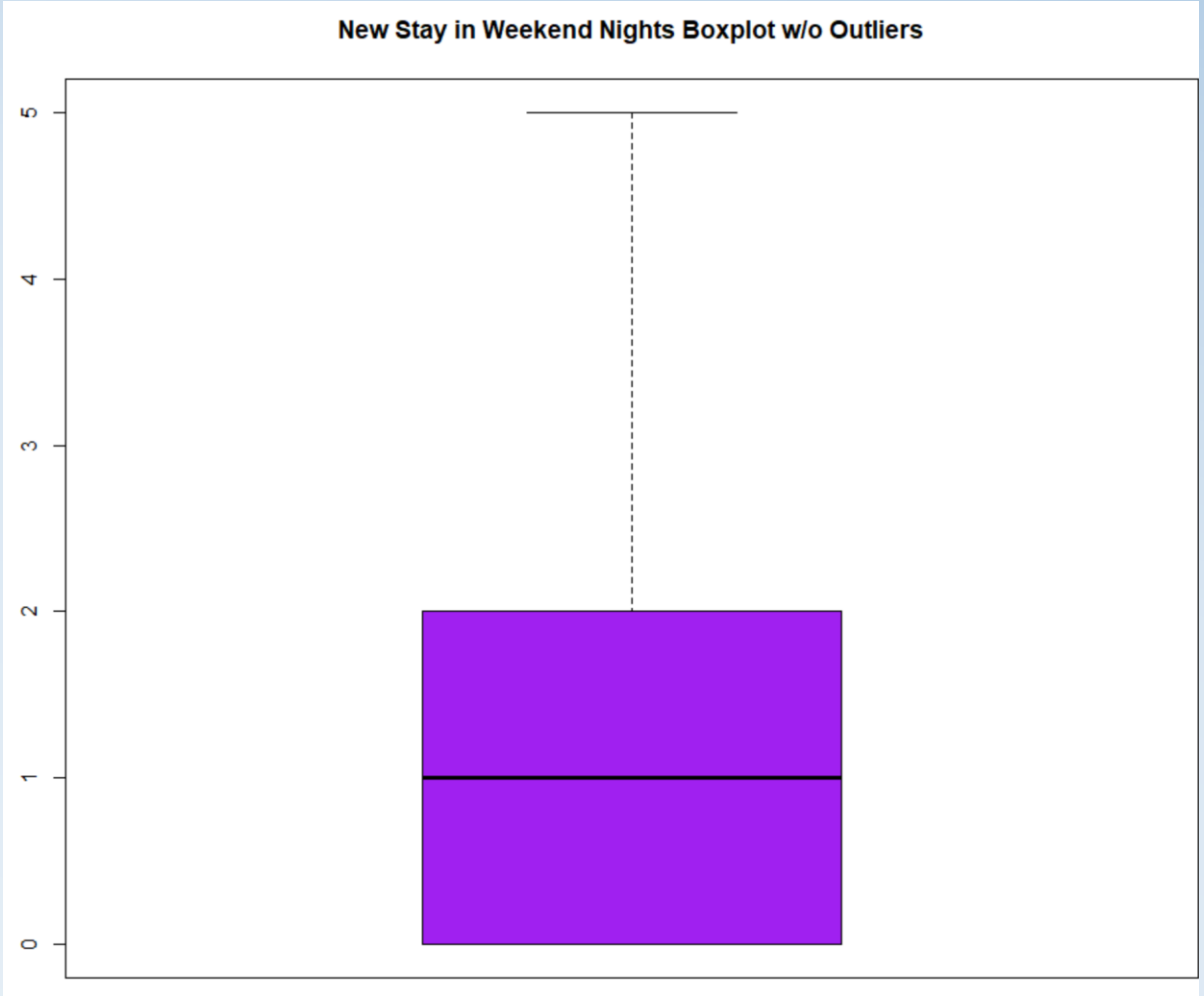
- The "stays_in_weekend_nights" attribute has a total of 265 number of outliers.
- This is the number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
- This is about 0.2% of its total number.
- As this is the only one with low number of outliers, we can therefore drop the rows corresponding to the outliers.



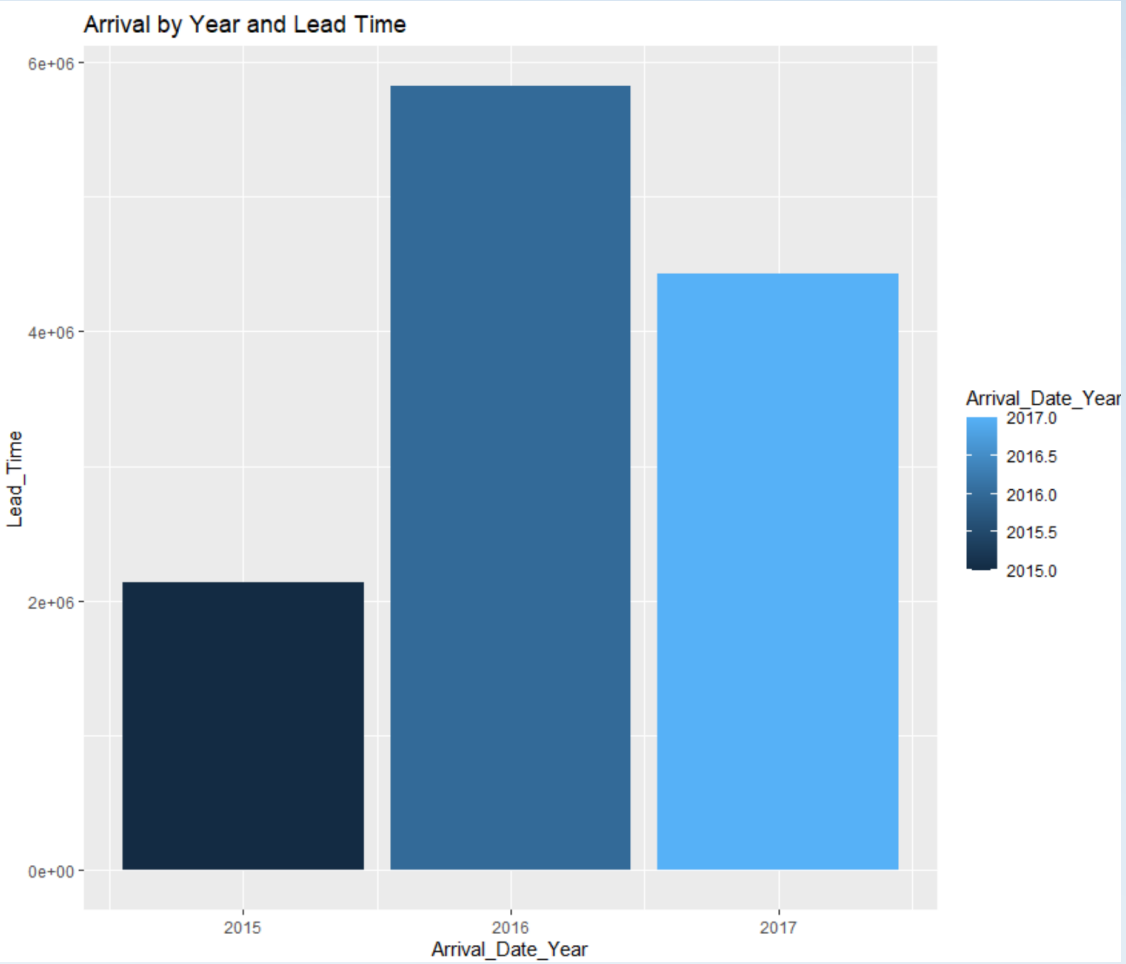
New Dataset

- After deleting the rows corresponding to the outliers of the "stays_in_weekend_nights" column, the new dataset has 119125 entries and still 32 attributes.

- The boxplot on the right is the new boxplot of the column without the outliers.



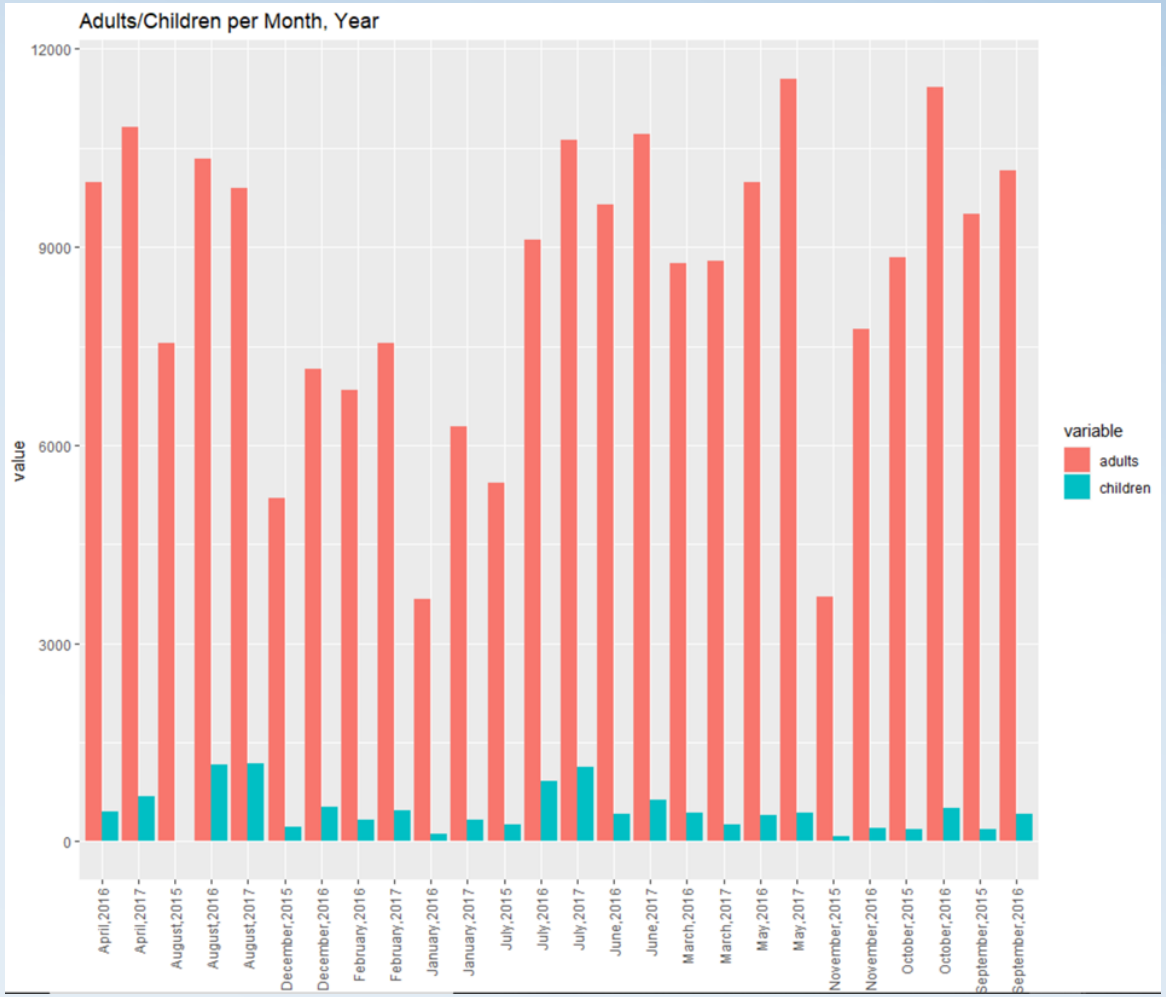
How much in advance are the customers booking before their arrival?



- Looking at the graph, 2015 had the lowest number of days that elapsed between the entering date in the PMS and the arrival date.
- However, there is a big gap between 2015 and 2016. In 2016, guests were booking their reservations more than twice in advance than in 2015.
- Guests are still booking in advance in 2017, though not as early as in 2016, but still more than 2015.

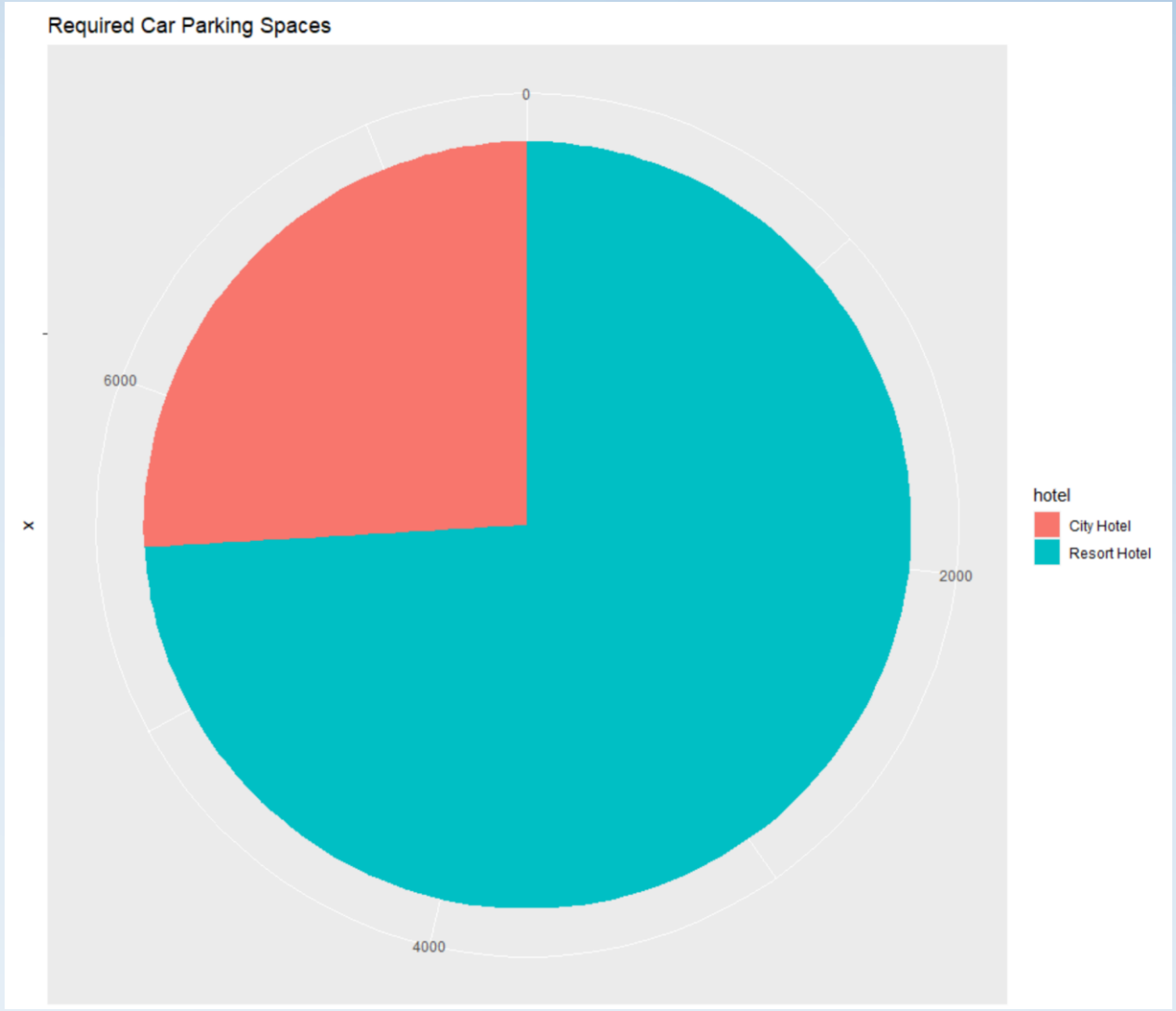
What month is more popular for bookings?

- Looking at the graph, it is obvious that adults are expected to make up most of the guest type any hotel will have.
- May 2017 had the most amount of adult guests from 2015- 2017.
- August 2015 had no children guests present.
- Regardless of the year, there seem to be an increase of the number of guests starting from the month of March. The number of guests start to decrease more at the month of October.

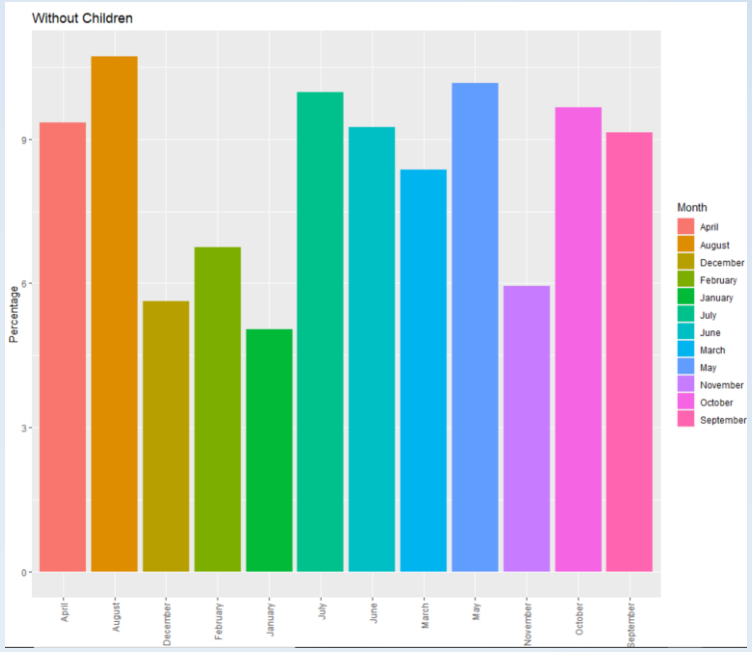


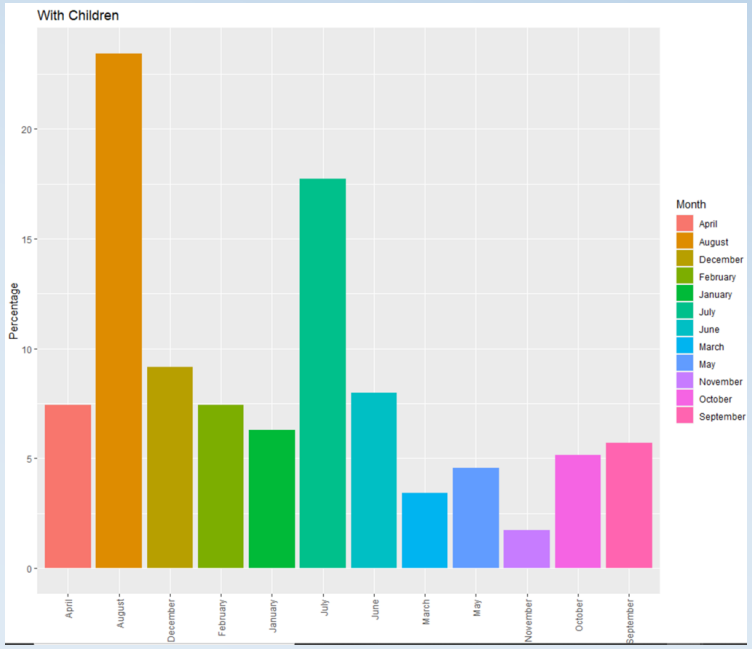
Which hotel requires more parking space?

The graph shows that Resort Hotels require more parking space to accommodate for their guests. City Hotels have only a sum of greater than 2000 parking spaces required by their customers. While Resort Hotels have in total required greater than 5000 parking spaces.



Contrast the arrival timings of adults with children or babies with the that of adults without children or babies:



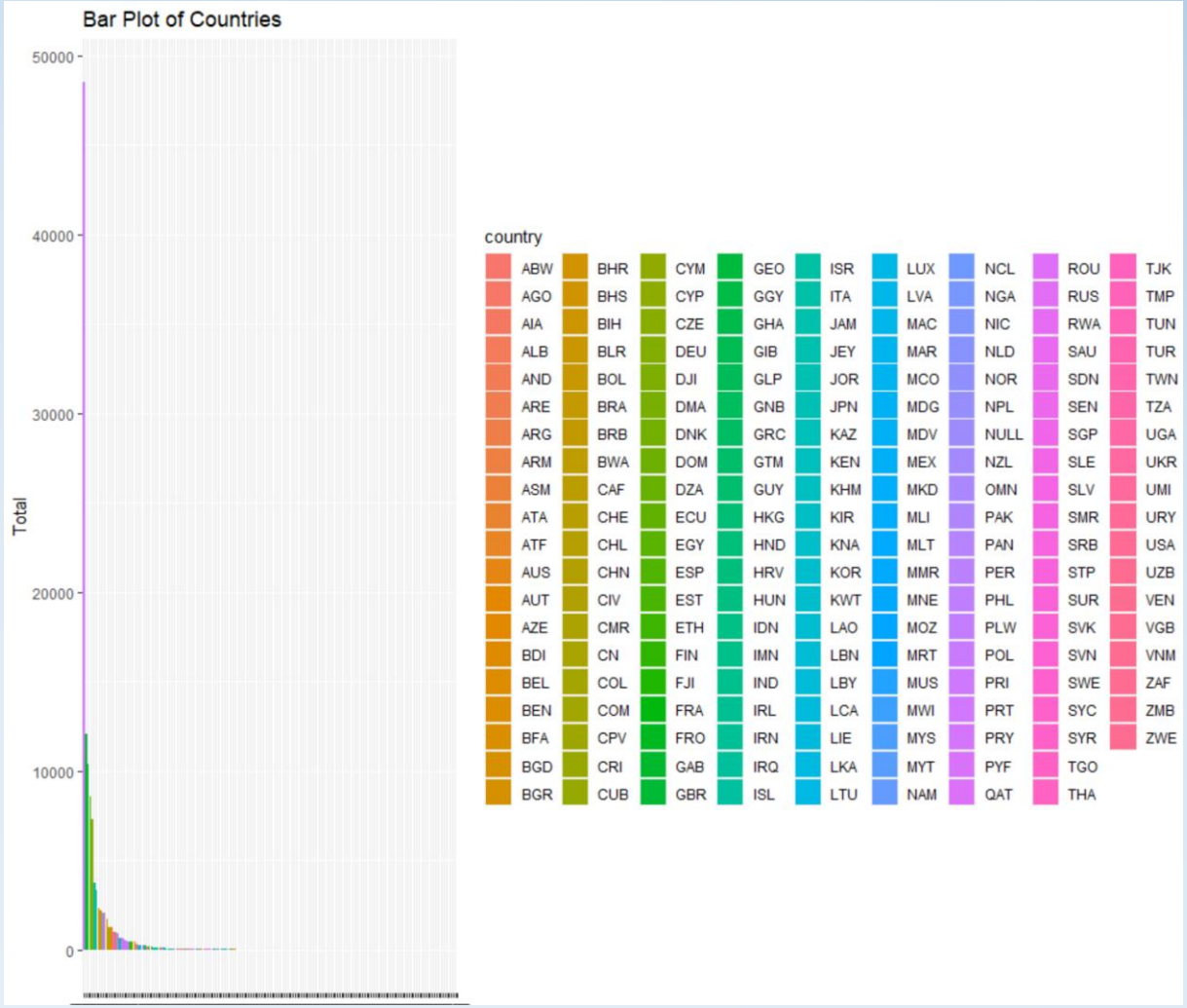


Comparing the two graphs, it is obvious that adults without any children were more likely to make a hotel reservation for any month. Both graphs showed a similarity wherein both have August as the highest number of arrival timings. This make sense as it is during summer when there is no school.

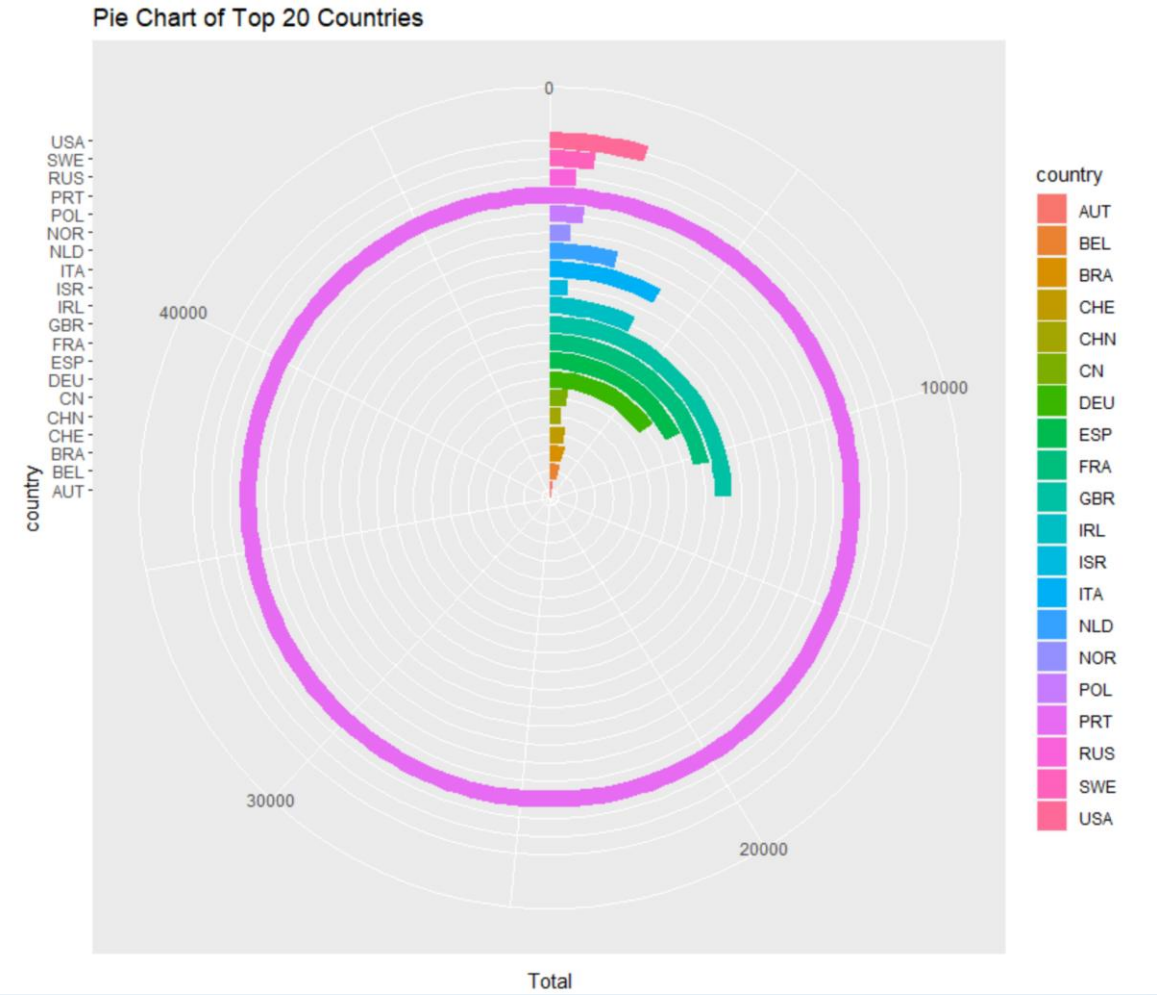
The number of guests arriving from different countries

- This is a bar plot of the countries where the guests arrived from based on the number of reservations. PRT, or Portugal, comprised of the highest number of guests with more than 40,000 hotel reservations made.

- Since this bar plot is hard to look at, a pie chart of the top 20 countries were made.



Top 20 Countries



- Again, Portugal made up most of the reservations made.
- Next is GBR, or the United Kingdom, with more than 10,000 reservations.
- In order of top 10:
  - Portugal
  - United Kingdom
  - France
  - Spain
  - Germany
  - Italy
  - Ireland
  - Belgium
  - Brazil
  - Netherlands

The mean of lead_time bookings that have been canceled