**Name: Nermien Elassy**

**ID:900196006**

**Instructor: Dr. Ali Hadi**

**Course: DSCI 141101**

**Data of submission: 12th of December 2020**

**DSCI project Fall 2020**

**Data description:**

**Who:** Observations: Two Hotels (A city hotel and a resort hotel).

**When:** The data was collected in February 2019.

**Where:**

This data was originally written in February 2019 <https://www.sciencedirect.com/science/article/pii/S2352340918315191> by Nuno Antonio, Ana Almeida, and Luis Nunes. Then it was cleaned and sorted in a table by Thomas Mock and Antoine Bichat on 11th of February 2020 <https://www.kaggle.com/jessemostipak/hotel-booking-demand>.

**Why:** This data was collected due to its important role in different fields like data mining, machine learning and revenue management for educational and research purposes. Moreover, the real business data that are available for scientific researchers are rare.

**What?** Here are the 32 variables in my data with their description.

Important note: The whole data set is only analyzed from the perspective of the type of the variables and their unit of measurement. The coding part and the conclusion is only focused on specific variables that matter for my analysis to answer the two specific questions I mentioned in my report. Moreover, there are other variables that are included in my coding part for just explanation of some specific commands (Numerical summaries).

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Type** | **Unit of measurement** | **Description** |
| Hotel | Categorical (binomial) (city, resort) | hotel | Hotel (H1 = Resort Hotel or H2 = City Hotel) |
| is \_canceled | Logical | 0 = false, 1=true |  |
| lead\_time | Numeric(Quantitative) | days | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date |
| arrival\_date\_month | Categorical (multinomial) (January-December) | month | Month of arrival date |
| arrival\_date\_week\_number | Numeric(Quantitative) | Number of the week | Week number of year for arrival date |
| arrival\_date\_day\_of\_month | Numeric(Quantitative) | Number of the dayday | Day of arrival date |
| stays\_in\_weekend\_nights | Numeric(Quantitative) | Number of Weekend nights | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel |
| stays\_in\_week\_nights | Numeric(Quantitative) | Number of Week nights | Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel |
| adults | Numeric(Quantitative) | Number of adults |  |
| distribution\_channel | Multinomial (TA/TO/direct/other) | channel | Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators” |
| is\_repeated\_guest | Logical | (0 = false, 1=true) |  |
| previous\_cancellations | Numeric(Quantitative) | Number of previous cancellations | Number of previous bookings that were cancelled by the customer prior to the current booking |
| reserved\_room\_type | Character string | No unit of measurement | Code of room type reserved. Code is presented instead of designation for anonymity reasons. |
| assigned\_room\_type | Character string | No unit of measurement | Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type. |
| reservation\_status | Categorical(Check-out, Cancelled, other) | Reservation status | Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer. |
| reservation\_status\_date | Character string | No unit of measurement | Date at which the last status was set. This variable can be used in conjunction with the Reservation Status |
| total\_of\_special\_requests | Numeric(Quantitative) | Number of special requests | Number of special requests made by the customer (e.g. twin bed or high floor) |
| required\_car\_parking\_spaces | Numeric(Quantitative) | Number of car parking spaces | Number of car parking spaces required by the customer |
| adr | Numeric(Quantitative) | No unit of measurement | Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights |
| customer\_type | Categorical(multinomial) (transient, transient-party, other) | Customer | Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; |
| days\_in\_waiting\_list | Numeric(Quantitative) | Number of days | Number of days the booking was in the waiting list before it was confirmed to the customer |
| company | Numeric(quantitative) | ID | ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for |
| agent | Numeric(quantitative) | ID | ID of the travel agency that made the booking |
| booking\_changes | Numeric(quantitative) | Number of changes/amendments | Number of changes/amendments made to the booking from the moment the booking was entered on the PMS |
| deposit\_type | Categorical(binomial) | No unit of measurement | Indication on if the customer made a deposit to guarantee the booking. |
| previous\_bookings\_not\_canceled | Numeric(quantitative) | Number of bookings not canceled | Number of previous bookings not cancelled by the customer prior to the current booking |
| Children | Numeric(Quantitative) | Number of children |  |
| babies | Numeric(Quantitative) | Number of babies |  |
| arrival\_date\_year | Numeric(Quantitative) | Year number | Year of arrival date |
| meal | Categorical(multinomial (BB,HB,other) | meal | Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal |
| Countries | Categorical(mutlinomial(PRT,GBR,Other) | country | Country of origin. Categories are represented in the ISO 3155–3:2013 format |
| Market\_segment | Categorical(multinomial)(online, offline, other) | Market segment | Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators |

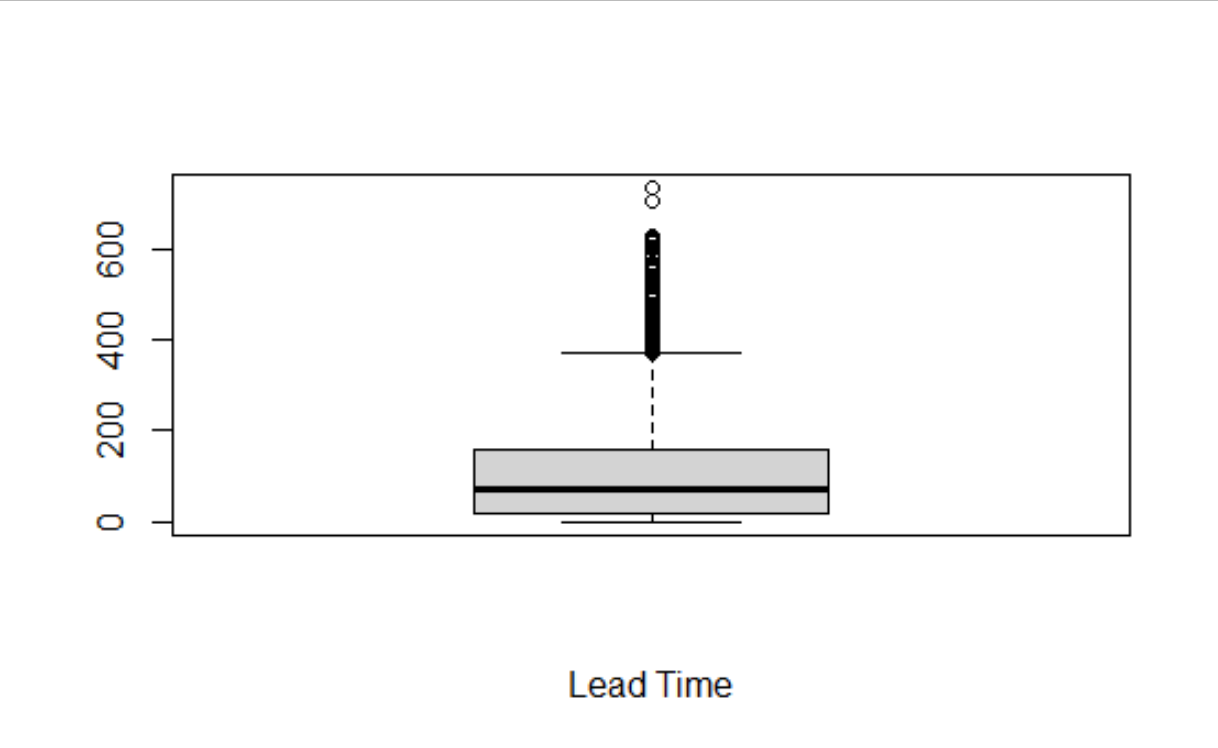
**Research questions:**

**Q1.** when is the best time of year to book a hotel room?

**Q2.** What if you want to predict whether or not a hotel is likely to receive a disproportionately high number of special requests?

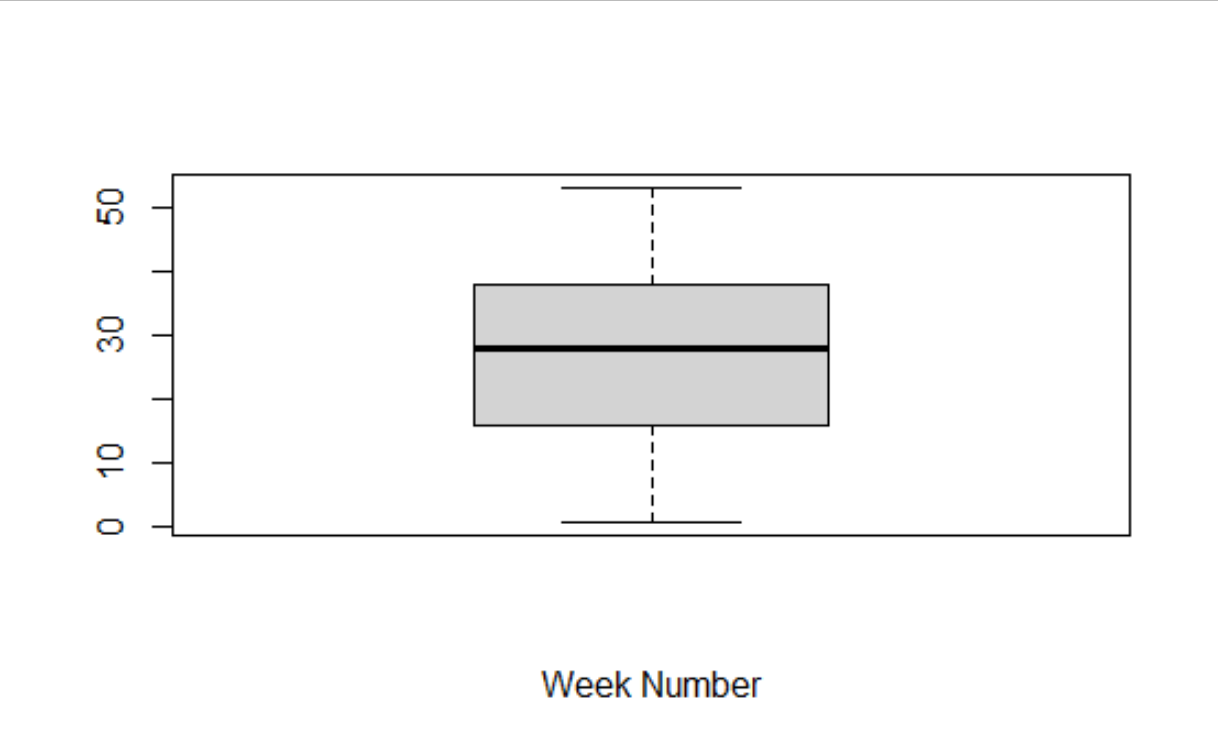
**Analysis:**

1. I start by inserting my whole data (32 variables) inside my Rstudio
2. I omitted all the columns except the ones that are related to the answer of my questions in addition to other columns that I needed to apply the rest of the numerical summaries on them.
3. I cleaned my data from all the NAs values in each column.
4. I started to examine each column by using the boxplot graph to see if it contains if there are any outliers to be removable or not.
5. Starting with the Lead\_time variable which is one of the variables that will not help me in answering none of the two questions, but it will help me applying a special case of the mean which is the trimmed mean because it is obvious from the previous graph that this variable contains many outliers.



Outliers

1. Looking to the third variable which is the arrival day of the week number, it is obvious that this variable doesn’t contain any outliers.



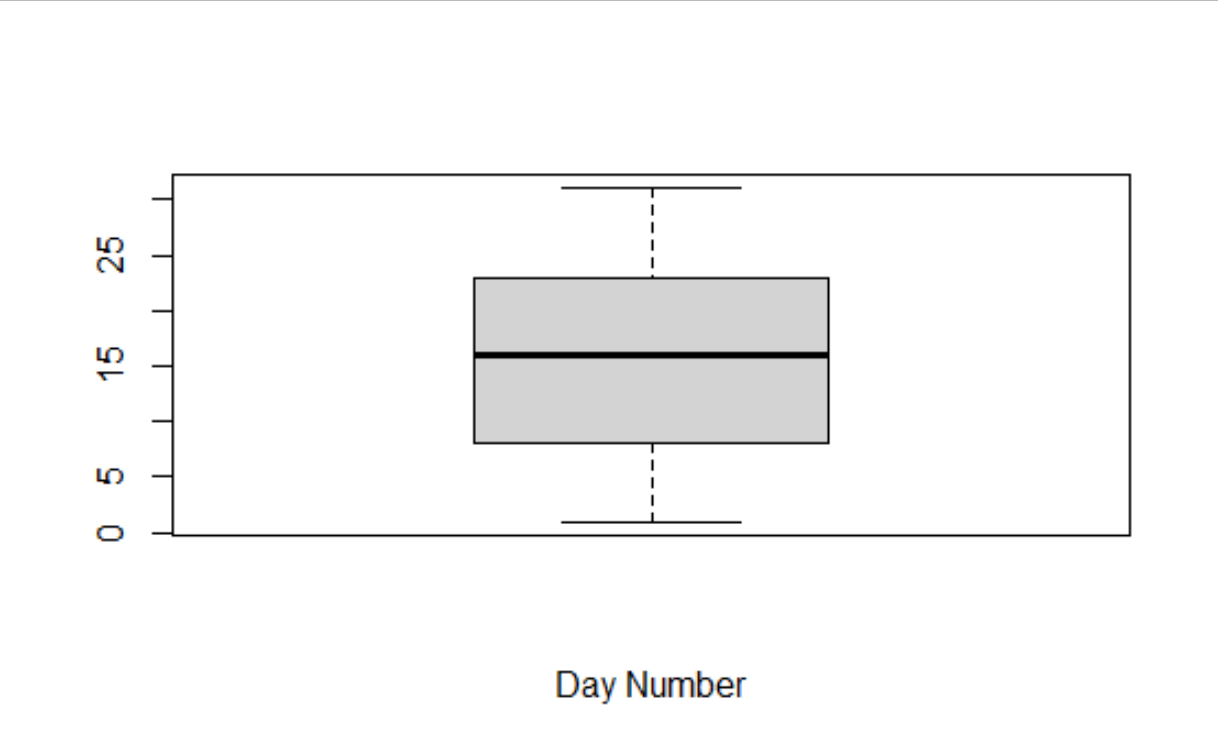
IQR

Median

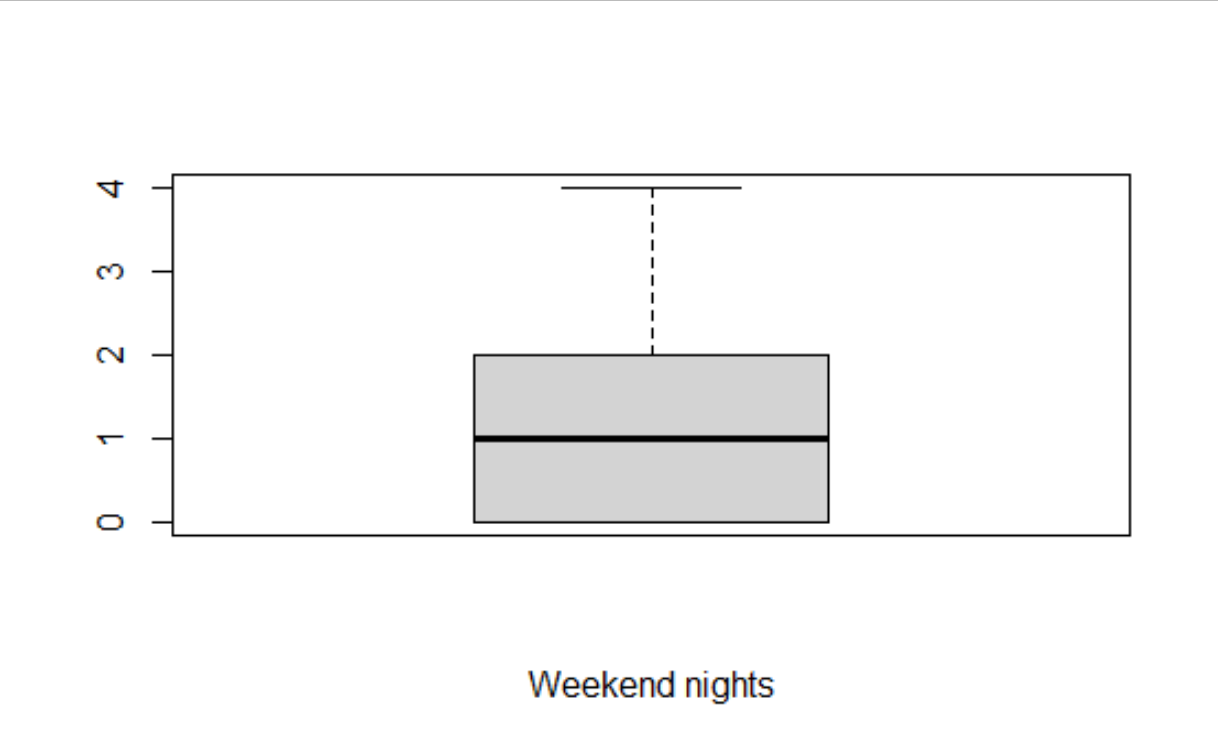
Q1

Q3

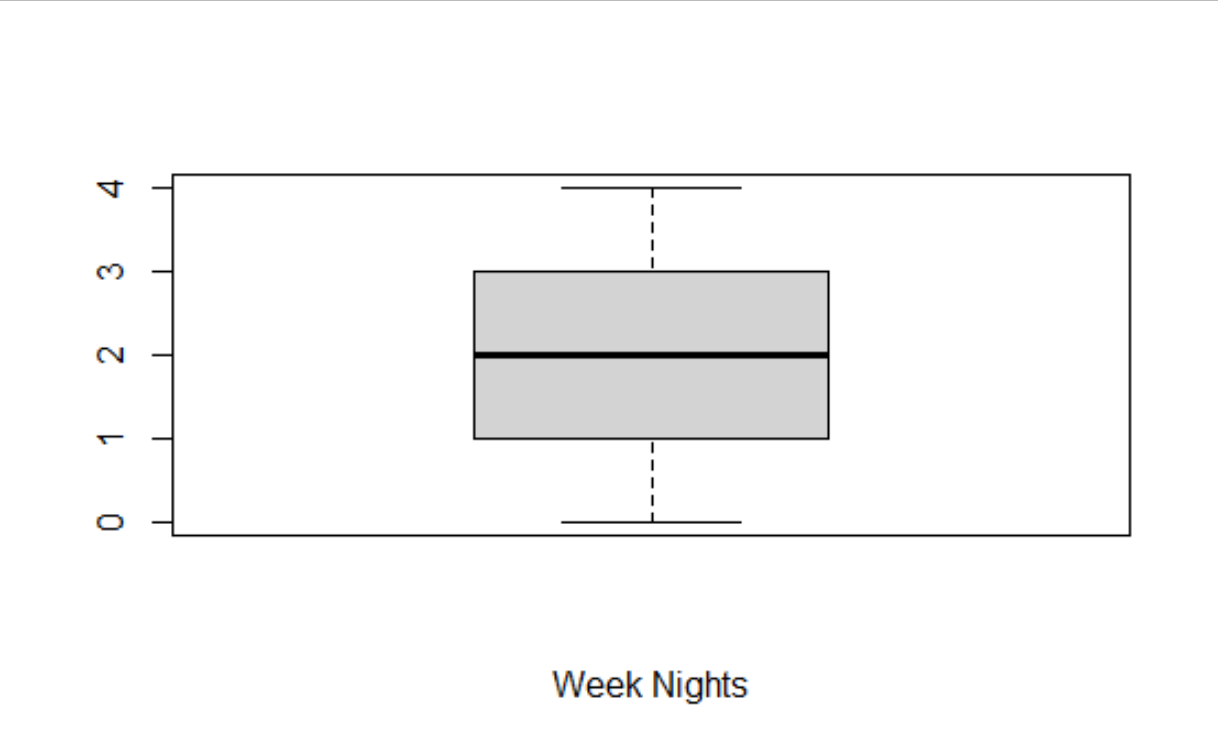
1. Looking to the fourth variable which is the arrival day of the month, it is also obvious that there are no outliers.



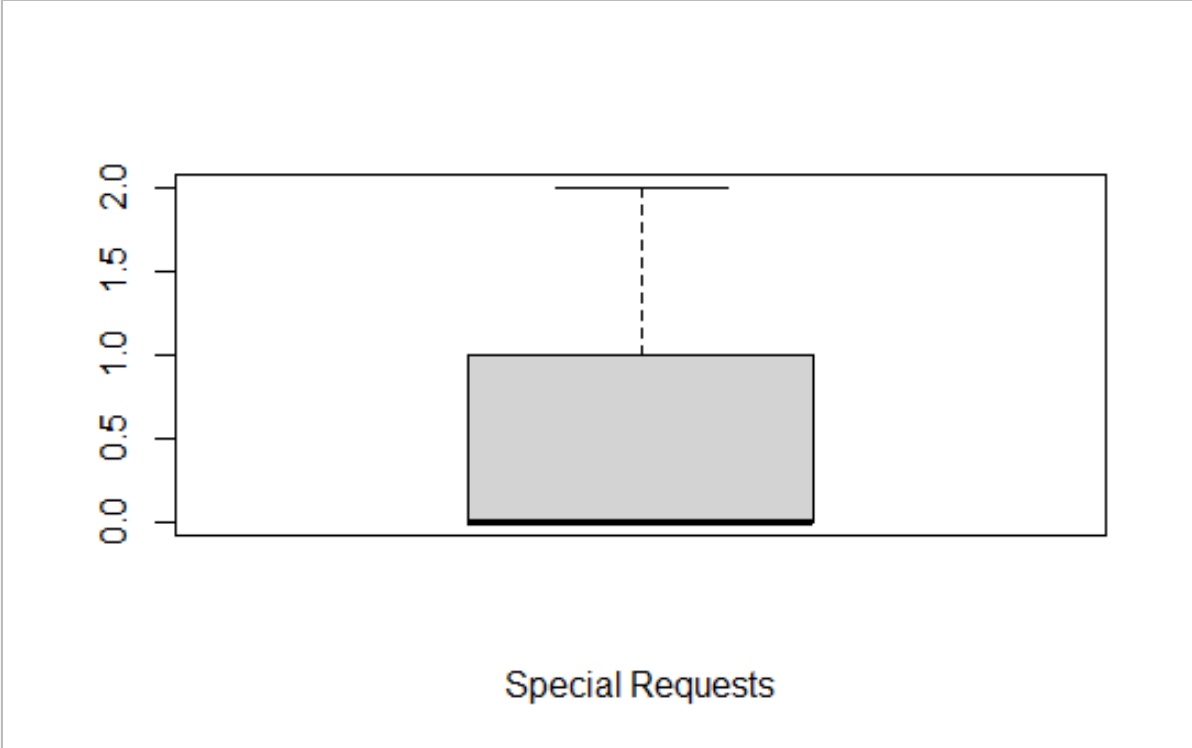
1. Looking to the fifth variable which is the stay in weekend nights, it contains some outliers. So, I removed those outliers and that is the graph without the outliers.



1. Looking to the sixth variable which is the stay in week nights, the case is the same as the previous variable. So, I also removed the outliers and the following graph is the clean version after removing the outliers.



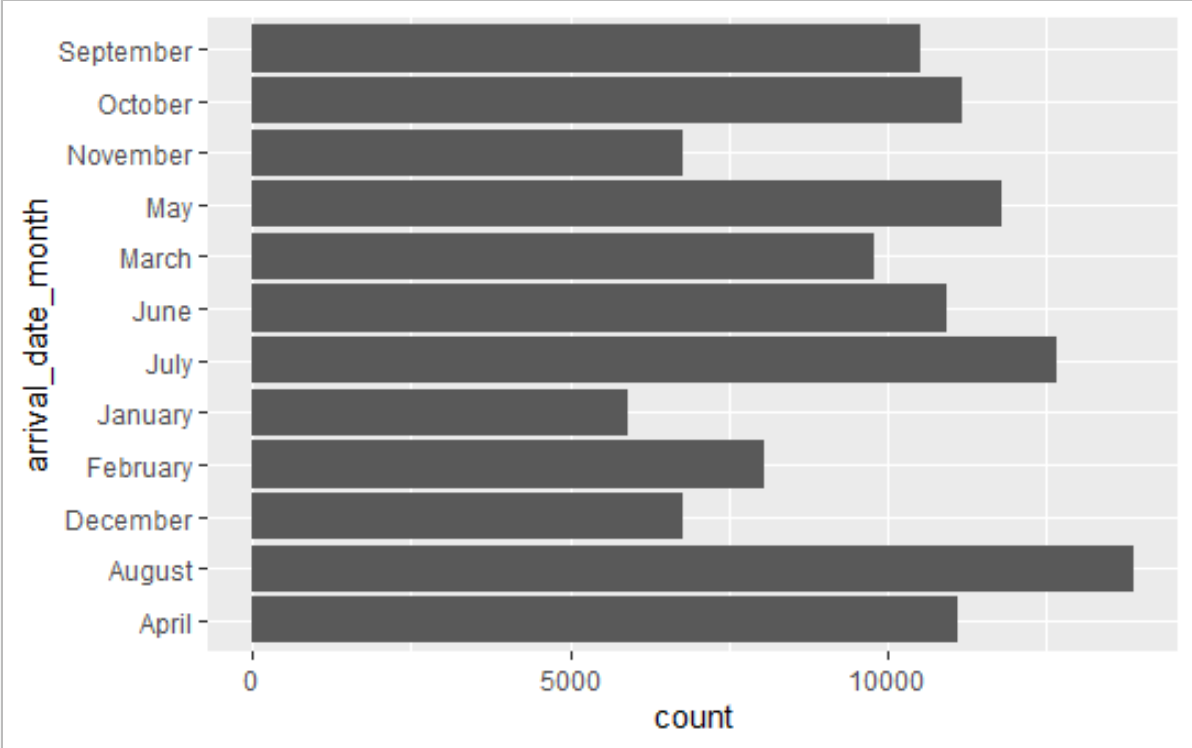
1. Looking the last variable which is the special requests, I observed that it contains a little number of outliers that should be removed and that is what is displayed in the following graph after removing the outliers



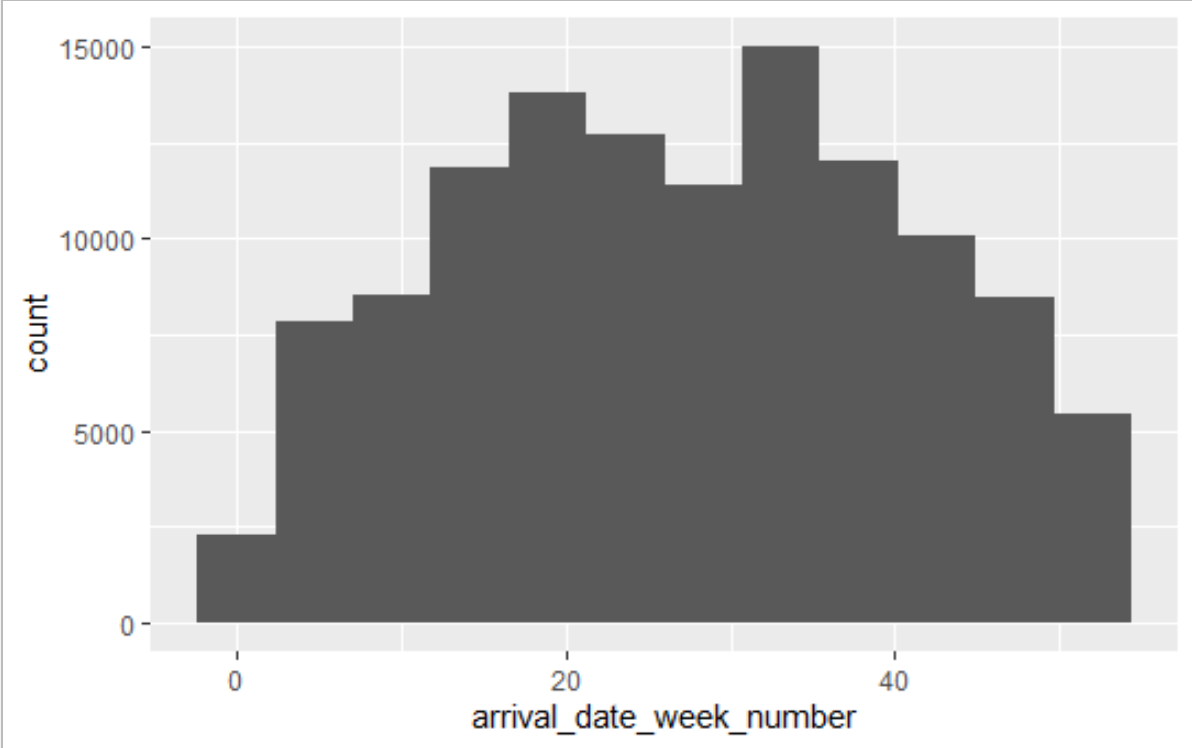
5.Now, my data is clean to start my analysis to answer the two questions I mentioned.

**Part1: What is the best time to book a room in the hotel**

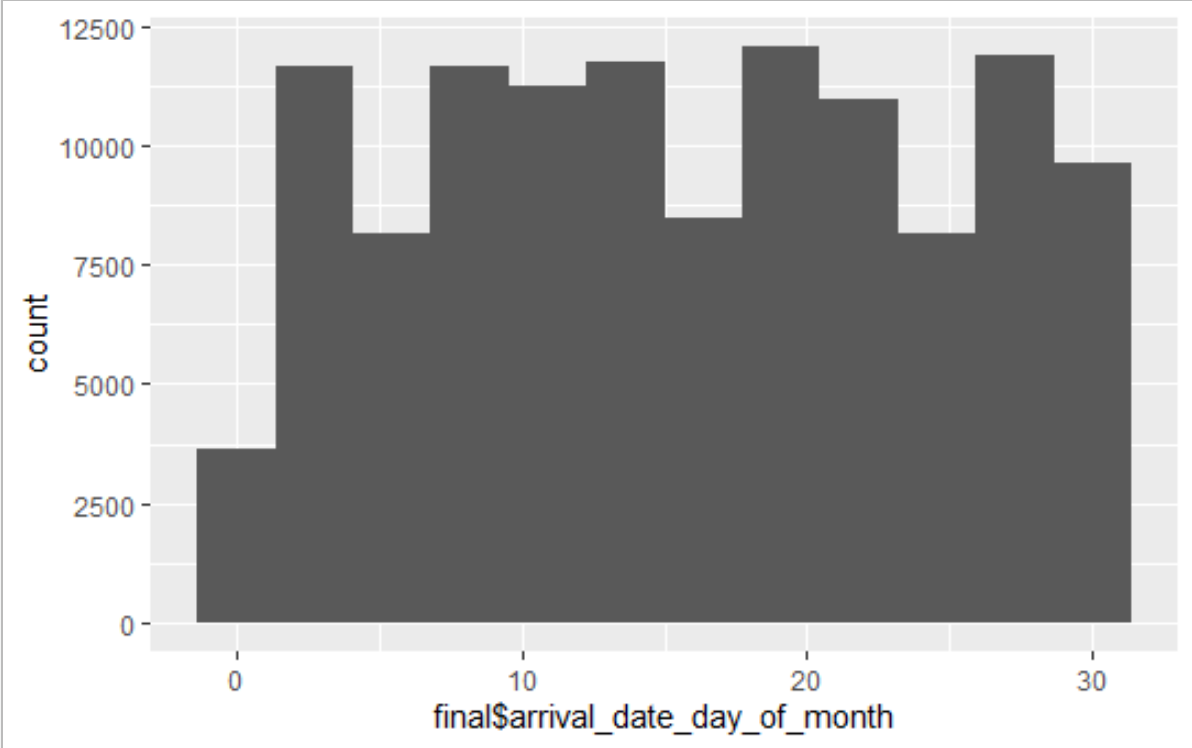
To answer my first question, I had to examine three variables (the month, the week number and the day of the month). From the following graph I can conclude that the most frequent month of arrival is **August.** To make sure that my observation is correct, I used the mode function and I got the same answer.



From the following graph, I observed that the period in terms of the week number in which it contains the most frequent arrivals is the period from 32 to 35. Moreover, to make sure that my estimation is approximated to the reality. I used the summary function which proves that 75% of the data is below week 38.

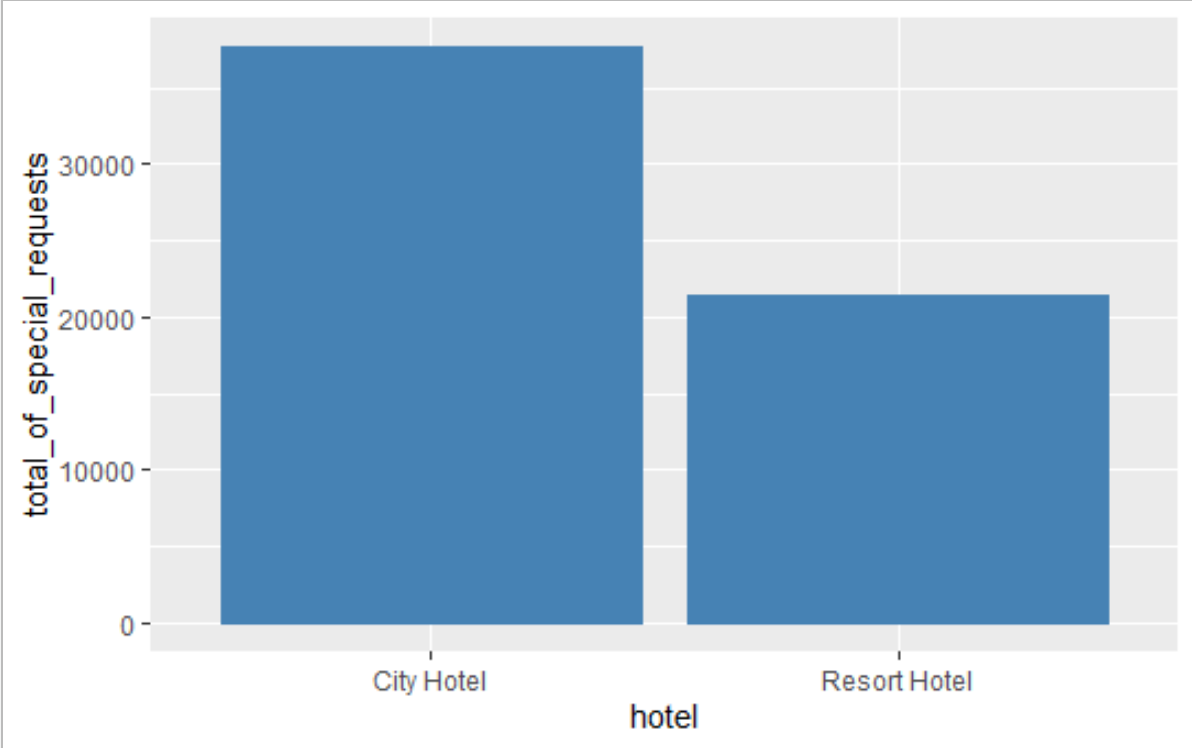


From the following graph, I concluded that the most frequent period in terms of the days of the month is the period from day 17 to day 20. Also to make sure that my estimation is correct, I used the quantile function and I got that 75% of the data is below day 23.



**Part 2: Predicting whether a certain type of the hotel will receive special requests or not.**

In this part I related the variable hotel to the variable special requests and I concluded after observing the graph that a city hotel is more likely to receive special requests than a resort hotel.



**Last part:(applications on the numerical summaries)**

1. I used a function to find the range of the stay in week night variable and I got 5 (the difference between the two extremes of the variable is 5).
2. I used the median () function to find the median of stay in weekends night variable and I got 1
3. I used the trimmed mean function to find the mean of the lead time variable after trimming the outliers and I got 87.24253
4. I used the quantile () function, but with setting the probability to 10% to get the 10 summary numbers of the arrival week number variable
5. I used the mad () function to get the mean value absolute deviations of the days of the months variable and I got 11.8608. (The average distance between each observation in the column and the mean is equal to 11.8608).
6. I used the IQR () function to get the interquartile range of the days of the months and I got 15 (the range of the middle 50% of the data is equal to 15)
7. I used the cov() function to find the variance of the stays in weekends nights and I got 0.7180574 (The average square difference of the mean is 0.7180574)
8. I used the sd () function to find the standard deviation of the stay in week nights variable and I got 1.366076. (square root of the variance) (the observations of the column differ from the mean value of the same column by 1.366076).
9. I used the mean () function to find the mean of the week number variable and I got 27.16517.

**Conclusion:**

1. The best month to book a room is August
2. The best week to book a room is from week 32 to 35
3. The best day to book a room is from day 17 to 20
4. City hotel is more likely to receive special requests than resort hotel

**Appendix:**

**Note: The scales on some graphs are grouped due to the large number of observations (eg.1 is expressed as 10000)**

