**Universidad Peruana de Ciencias Aplicadas (UPC)**



**Proyecto:**

Anàlisis exploratorio de un conjunto de datos en R/Rstudio

**Curso:**

Administraciuòn de la Informaciòn

**Integrantes:**

* Carhuancho Bazan, Alessandro David u201913933
* Nuñez Lazo, Sergio Antonio u201910357
* Cruz Mamani, Jack Yefri u201912400

**Ciclo:**

2021-01

**Indice**

1. Caso de análisis
   1. Origen de los datos
   2. Casos de uso aplicables
2. Conjunto de datos

2.1. Descripción de la estructura de datos

2.2. Resúmen de la estructura de datos

2.3. Deteccion de valores NA

2.4. Detección de valores vacios

2.5. Deteccion de valores outliers

1. Análsis de datos exploratorios

3.1. Cantitdad de clientes hospedados en cada hotel por año

3.2. Cantidad de reservas canceladas y no canceladas por año

3.3. Distribución de clientes que vienen solo entre adultos y con menores de edad por hotel

1. Conclusiones preliminares
2. Referencias
3. **Caso de análisis** 
   1. **Origen de los datos**

El dataset analizado en el presente trabajo fue obtenido desde los repositorios de Kaggle, sin embargo, este dataset ya ha pasado por el proceso de limpieza de datos, el cual fue realizado Thomas Mock y Antoine Bichat para los repositorios de Tidy Tuesday en GitHub, un proyecto semanal enfocado en el análisis de datos usando el lenguaje de programación R. El commit fue realizado el 11 de Febrero del 2020.

El dataset original proviene del artículo Hotel Booking Demand Datasets. Según el artículo, el dataset fue extraído de las bases de datos Property Management System SQL de los dos hoteles examinados. Dichos hoteles se ubican en Portugal, precisamente en el Resort de Algarve, y en la Ciudad de Lisboa.

El artículo está disponible en los repositorios de Science Direct desde el 29 de noviembre del 2018. En el mes de Febrero del año 2019, el artículo fue publicado en el volumen 22 de la revista Data in Brief. Los autores son Nuno Antonio a, b, Ana Almeida a, c, d, y Luis Nunes a, b, d, quienes realizaron estudios en las siguientes instituciones de Portugal:

a. Instituto Universitário de Lisboa, Lisboa

b. Instituto de Telecomunicações, Lisboa

c. CISUC, Coimbra

d. ISTAR-IUL, Lisboa

* 1. **Casos de uso aplicables**

La utilidad del dato es innegable, sin embargo, la importancia de este dataset en particular se refleja en los siguientes campos:

* Análisis descriptivo para el entendimiento de patrones, tendencias y anomalías en la data.
* Investigación para la predicción de cancelación de una reserva y satisfacción del cliente, dicha información beneficia a los hoteles.
* Benchmarking de algoritmos de clasificación y segmentación, el cual beneficia a los investigadores y educadores de machine learning.
* Entrenamiento de modelos de data mining y estadística para beneficio de los educadores.

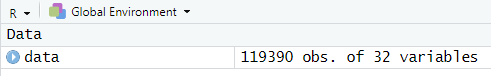
1. **Conjunto de datos (DataSet)**
   1. **Descripción de la estructura de datos**

Para la descripciòn de la estructura de datos se utilizaron los siguientes comandos:

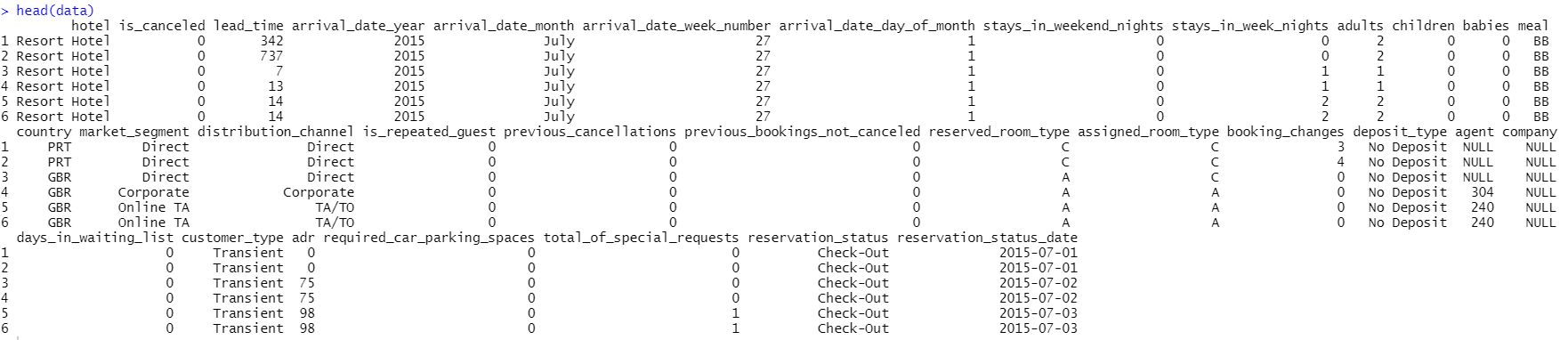
1. Para leer el conjunto de datos del archivo .csv usamos la instrucción **read.csv()**.



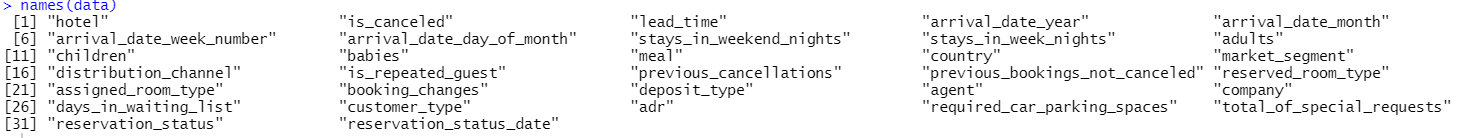
Al lado izquierdo de Rstudio, se puede observar que presenta 119390 observaciones y 32 variables.



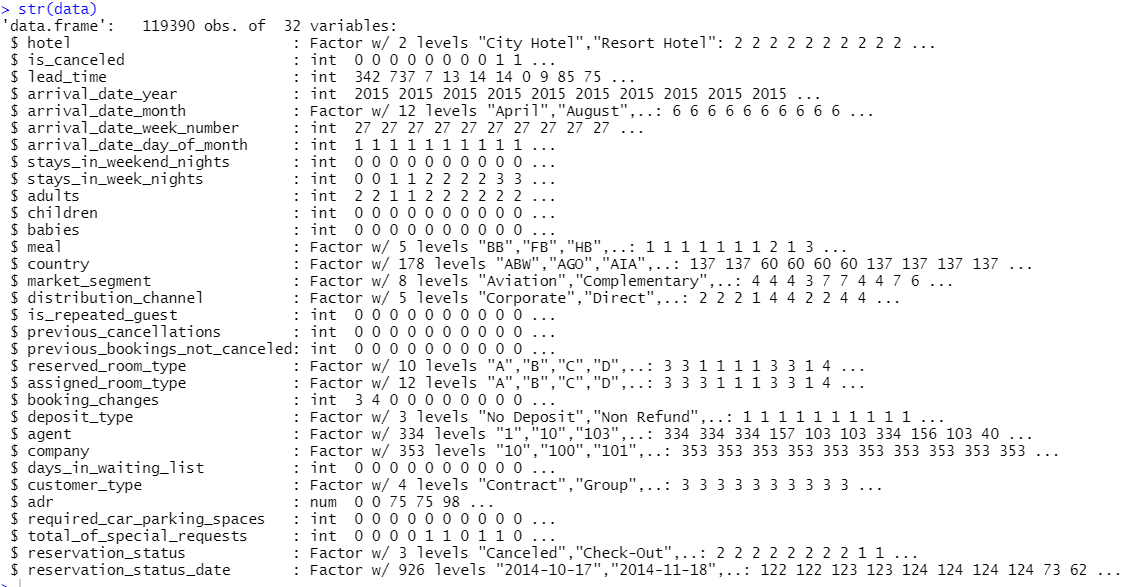
1. Usamos la instrucción **head()** para obtener las cinco primeras observaciones.



1. Obtenemos los nombres de las columnas con la instrucción **names().**



1. Obtenemos el tipo de dato de cada variable y los valores que contiene con la instrucción **str().**

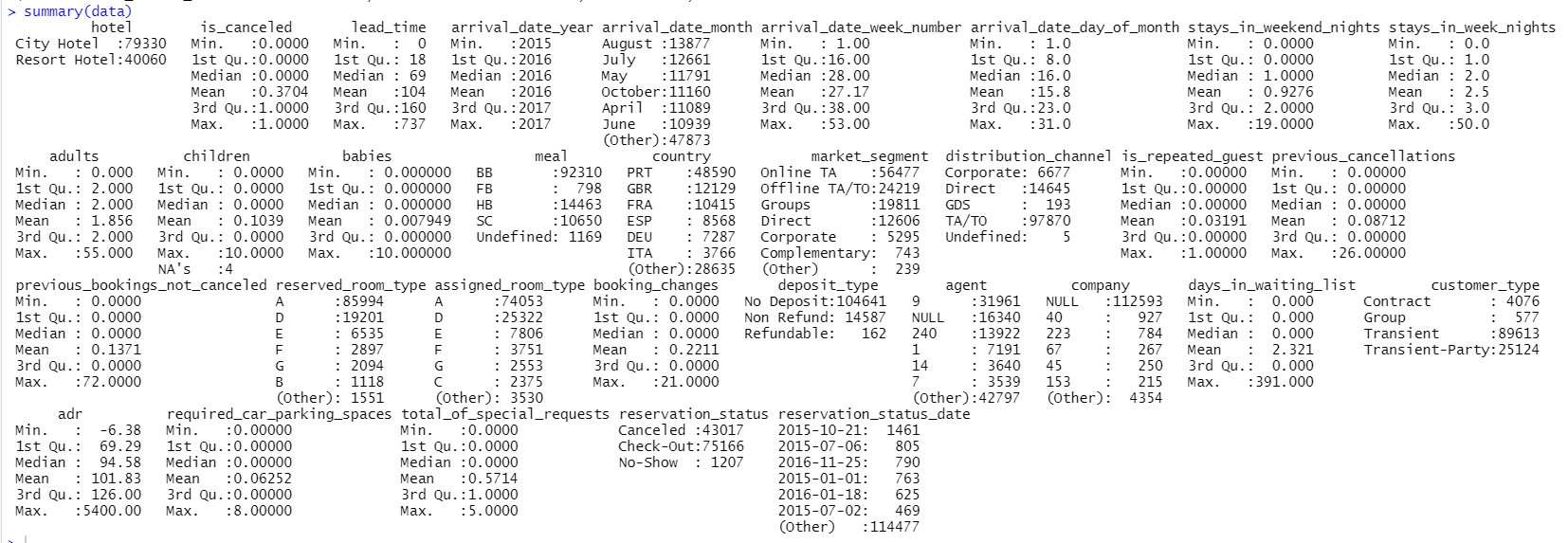


1. Obtenemos un resumen de cada variable con la instrucción **summary()**. Esto nos devuelve, en caso de variables numéricas, los siguientes valores:

* el valor mínimo (Min)
* el primer cuartil (1st Qu)
* la mediana (Median)
* la media (Mean)
* el tercer cuartil (3rd Qu)
* el valor máximo (Max)

Por otro lado, para las variables de tipo String o categórica:

* tipo 1 : cantidad 1
* tipo 2 : cantidad 2
* tipo 3 : cantidad 3
* tipo n : cantidad n



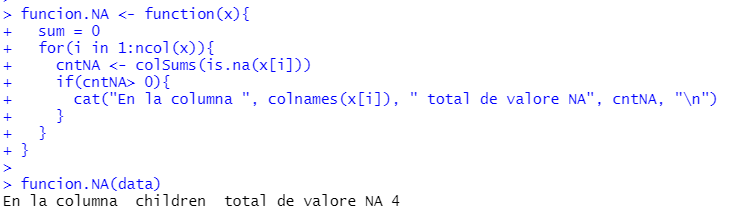
* 1. **Resúmen de la estructura de datos**

El dataset presenta 119390 observaciones (40060 resort hotel + 793 30 city hotel) y 32 variables. Sus variables se muestran en la siguiente tabla:

|  |  |  |
| --- | --- | --- |
| Nombre | Tipo de dato | Descripción |
| hotel | Factor | El tipo de hotel (city, resort) |
| is\_canceled | Int | Si se ha cancelado (1), sino (0) |
| lead\_time | Int | El tiempo entre la reserva y la llegada al hotel (días) |
| Arrival\_date\_year | Int | Año de llegada al hotel |
| Arrival\_date\_month | Int | Mes de llegada(12 categorias de “Enero” a “Diciembre” ) |
| Arrival\_data\_week\_number | Int | Número de semana de llegada al hotel |
| Arrival\_date\_day\_of\_month | Int | Día del mes de la llegada al hotel |
| Stays\_in\_weedend\_nights | Int | Nùmero de noches de fines de semanas reservados o prescenciados (“Sábado o Domingo”) |
| Stays\_in\_week\_nights | Int | Número de noches de semana reservados o prescenciados (“Lunes” a “Viernes”) |
| Adults | Int | Número de adultos |
| Children | Int | Número de niños |
| Babies | Int | Número de bebes |
| Meal | Factor | Tipo de comida reservada:  - NA/“SC”: Sin paquete de comida  - “BB”: Bed & Breakfast  - “HB”: Half board, solo desayuno, no cena |
| Country | Factor | País de origen (las categorías estan representadas en ISO 3155-3:2013 format) |
| Market\_segment | Factor | Segmento de mercado destinado  - “TA”: agencia de viajes  - “TO”: operaciones de tour |
| Distribution\_channel | Factor | Medio por el cual se realizó la reseva:  - “TA”: Travel agents  - “TO”: Tour operations |
| Is\_repeated\_guest | Int | Si es usuario repetido (1), sino (0) |
| Previous\_cancellations | Int | Número de reservaciones canceladas anterior a la reserva |
| Previous\_bookings\_not\_canceled | Int | Número de reservaciones no canceladas anterior a la reseva |
| Reserved\_room\_type | Factor | Código de tipo de habitación reservado |
| Assigned\_room\_type | Factor | Código de tipo de habitación asignado (algunas veces no se le asigna la habitación reservada, ya sea por doblereserva o petición del usuario) |
| Booking\_changes | Int | Número de cambios hechos a la reserva desde la fecha de reserva hasta el check-in. |
| Deposit\_type | Factor | Si realizó deposito para garantizar la reserva:  - “ No Deposit”: no se deposito  - “Not refund”: se deposito el valor total o mayor de la reserva  - “Refundable”: se deposito un valor menor al costo total |
| Agent | Factor | El Id de la agencia de viajes el cual realizo la reserva |
| Company | Factor | El Id de la compañía que realizó la reserva o responsable a quien se pago la reserva. |
| Days\_in\_waiting\_list | Int | Número de días que la reserva estaba en lista antes de que fuera confirmada al cliente |
| Customer\_type | Factor | Tipo de reserva:  - Contract: La reserva es asociado a un contrato o persona quien funciona como intermediario  - Group: La reserva es asociada a un grupo  - Transient: No es parte de grup or contract, y no asociado a una reserva transitorio  - Transient-party: Cuando ha sido asociado con al menos un transitorio. |
| Adr | Num | Promedio de ventas diarias |
| Requiered\_car\_parking\_spaces | Int | Número de espacios de para estacionar el vehículo por el cliente |
| Total\_of\_special\_rerquest | Int | Número de pedidos espciales por el cliente (dos camas, piso mayor, etc.) |
| Reservation\_status | Factor | El último estado de la reservación:  Cancelled: Fue cancellado por el cliente  Check-out: Realizó el check-in, pero ya se fue.  No Show: El cliente no realizó check-in pero informó al hotel los motivos |
| Reservation\_status\_date | Factor | Fecha en el cual el último Reservation\_status fue establecido. |

* 1. **Detección de valores NA**

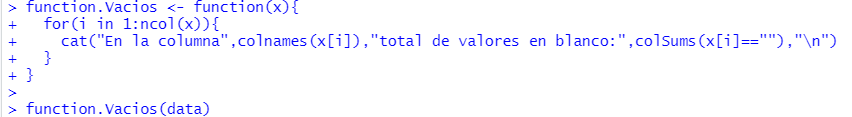
Para la detección de valores NA, creamos una funcion que muestre solo solo valores NA de un dataframe.



Obtenemos que la columna children presenta valores NA.

* 1. **Detección de valores vacios**

Para la detección de valores vacios creamos una funcion que nos muestre de todas las columnas, sus valores vacios.

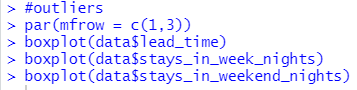


Obtenemos el siguiente resultado:

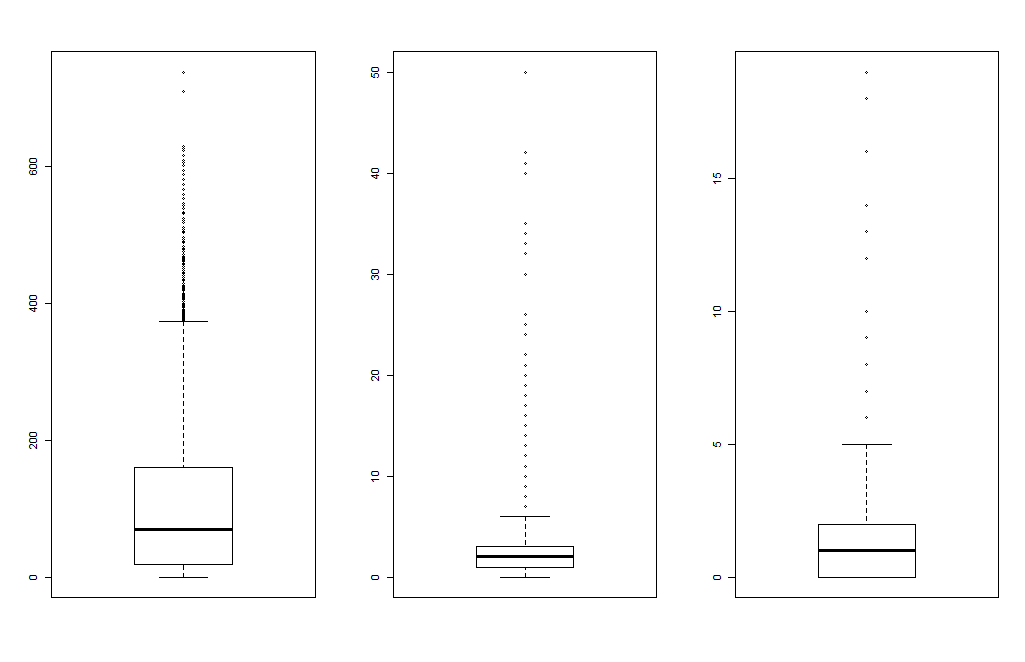


* 1. **Detección de valores outliers**

Para la detección de outliers utilizamos la función **boxplot(),** este modelo usa la detección de outliers por el método de las cajas.



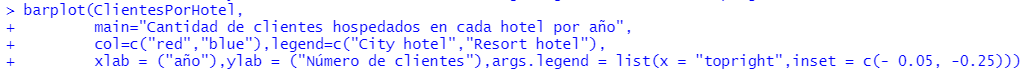
Obteniendo el siguiente resultado, donde los puntos representan los outliers:



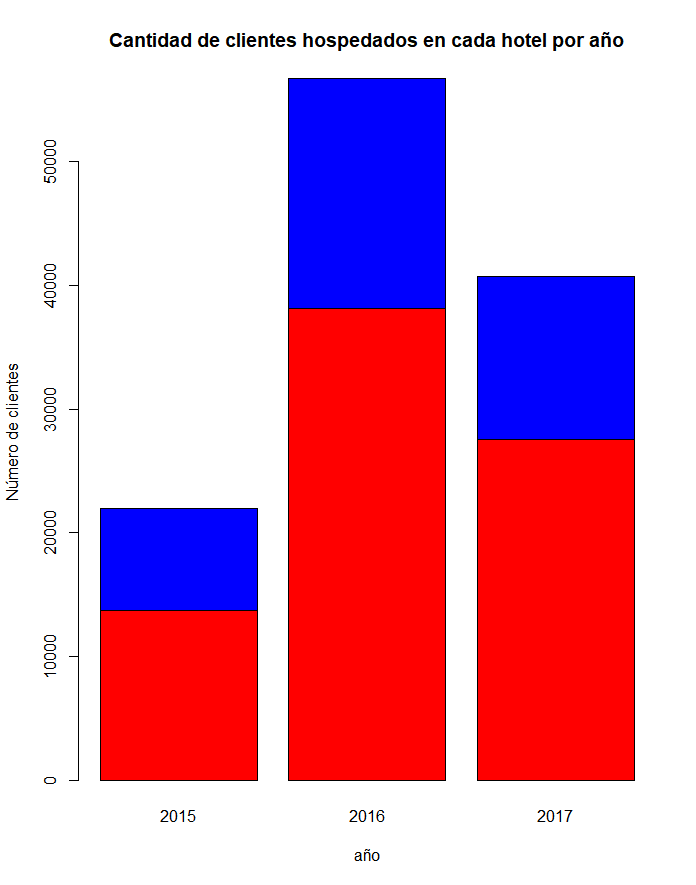
1. **Análisis de datos exploratorios**
   1. **Cantitdad de clientes hospedados en cada hotel por año**

Para hallar la cantidad de clientes hospedados en cada hotel por año, utilizamos la instrucción **table()**, luego usamos **barplot()** para imprimirlo.





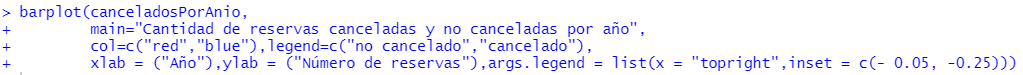
Obteniendo el siguiente resultado:



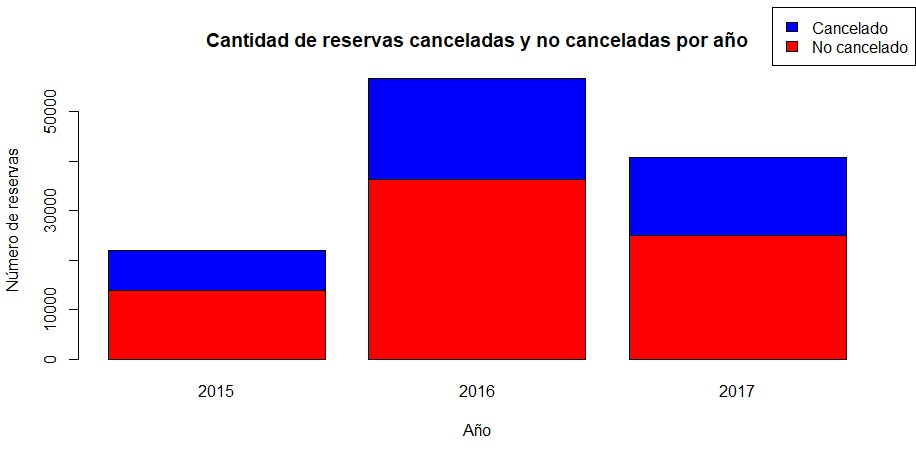
* 1. **Cantitdad de reservas canceladas y no canceladas por años**

Para hallar la cantidad de reservas canceladas y no canceladas por años, utilizamos la instrucción **table()**, luego usamos **barplot()** para imprimirlo.



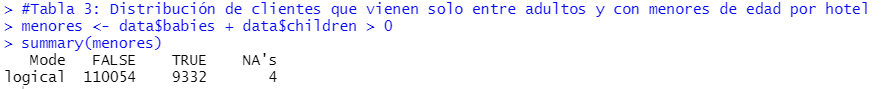


Obteniendo el siguiente resultado:

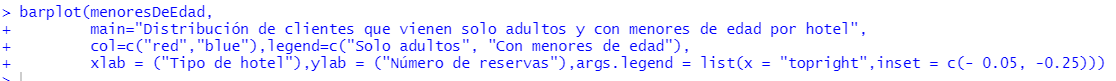


* 1. **Distribución de clientes que vienen solo entre adultos y con menores de edad por hotel**

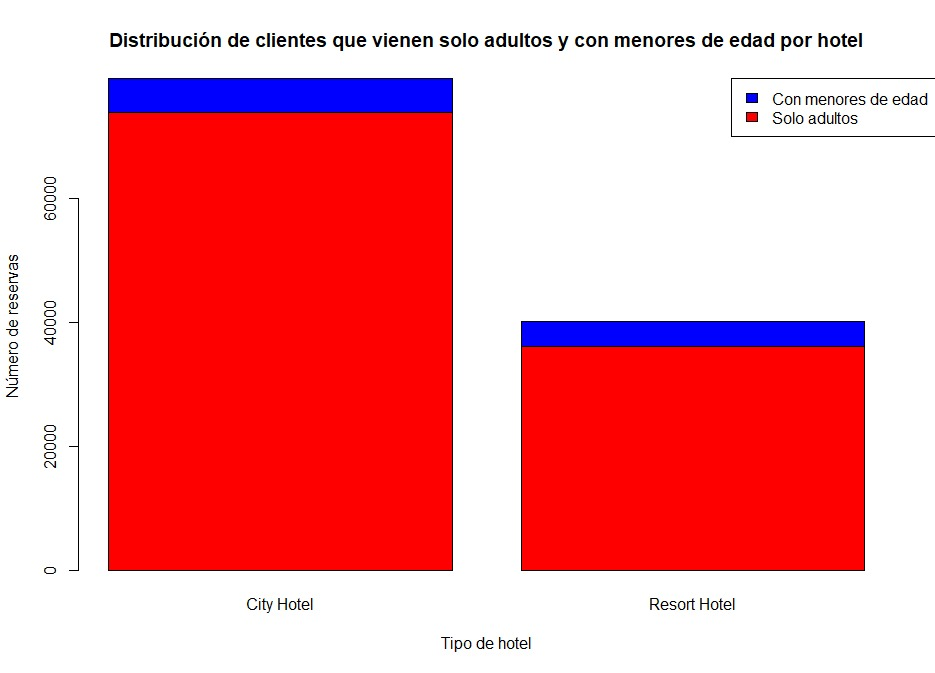
Para hallar la cantidad de clientes que vienen solo entre adultos y con menores de edad por hotel, sumamos los cantidad de reservas con bebes y con niños, luego, si el resultado es mayor a cero, entonces la reserva se realizó con menores, sino solo entre adultos. Finalmente, utilizamos la instrucción **table()** y **barplot()** para imprimirlo.







Obteniendo el siguiente resultado:



1. **Conclusiones preliminares**

Dados el análisis de las tres tablas anteriores podemos deducir las siguientes conclusiones:

1. El hotel de tipo ciudad es más concurrido que el de tipo resort durante los 3 años.

1. El año 2016 fue el año con más turistas, pero a la vez fue el año con más cancelaciones, además, si el número de reservas crece, tambien lo hace las cancelaciones.
2. La mayor cantidad de personas viajan solos o en grupo, pero que entren en la categoría de adultos.
3. **Referencias**

Hotel booking demand datasets Antonio, de Almeida, & Nunes - Data in Brief – 2019. Recuperado de <https://doi.org/10.1016/j.dib.2018.11.126>