Lesson 3: Filters and Plots

Background for this activity

En esta actividad, revisará un escenario y practicará la creación de una visualización de datos con ggplot2. Aprenderá a utilizar las funciones de filtros y facetas de ggplot2 para crear visualizaciones personalizadas basadas en diferentes criterios.

A lo largo de esta actividad, también tendrá la oportunidad de practicar la escritura de su propio código realizando usted mismo cambios en los fragmentos de código. Si encuentra un error o se queda atascado, siempre puede consultar el archivo Lesson3_Filters_Solutions .rmd en la carpeta Soluciones en la Semana 4 para obtener el código completo y correcto.

The Scenario

Como analista de datos junior para una empresa de reservas de hoteles, se le ha pedido que limpie los datos de reservas de hoteles, cree visualizaciones con “ggplot2” para obtener información sobre los datos y presente diferentes facetas de los datos a través de la visualización. Ahora, continuará con el trabajo que realizó anteriormente para aplicar filtros a sus visualizaciones de datos en `ggplot2` .

Step 1: Import your data

Si no ha salido de RStudio desde la última vez que importó estos datos, puede omitir estos pasos. Sin embargo, volver a ejecutar estos fragmentos de código no afectará a tu consola si quieres ejecutarlos por si acaso.

Si esta línea causa un error, cópiela en la línea setwd(“/cloud/project/Course 7/Week 4”) anterior.

Ejecute el siguiente código para leer el archivo ‘hotel_bookings.csv’ en un marco de datos:

```
hotel_bookings <- read.csv("hotel_bookings.csv")
```

Step 2: Refresh Your Memory

A estas alturas ya estás bastante familiarizado con este conjunto de datos. Pero puedes refrescar tu memoria con las funciones `head()` y `colnames()` . Ejecute dos fragmentos de código a continuación para obtener una muestra de los datos y también obtener una vista previa de todos los nombres de las columnas:

```
head(hotel_bookings)
```

```
##         hotel is_canceled lead_time arrival_date_year arrival_date_month
## 1 Resort Hotel           0       342              2015               July
## 2 Resort Hotel           0       737              2015               July
## 3 Resort Hotel           0         7              2015               July
## 4 Resort Hotel           0        13              2015               July
## 5 Resort Hotel           0        14              2015               July
## 6 Resort Hotel           0        14              2015               July
##   arrival_date_week_number arrival_date_day_of_month stays_in_weekend_nights
## 1                       27                         1                     0
## 2                       27                         1                     0
## 3                       27                         1                     0
## 4                       27                         1                     0
## 5                       27                         1                     0
## 6                       27                         1                     0
##   stays_in_week_nights adults children babies meal country market_segment
## 1                    0      2        0      0   BB     PRT         Direct
## 2                    0      2        0      0   BB     PRT         Direct
## 3                    1      1        0      0   BB     GBR         Direct
## 4                    1      1        0      0   BB     GBR      Corporate
## 5                    2      2        0      0   BB     GBR      Online TA
## 6                    2      2        0      0   BB     GBR      Online TA
##   distribution_channel is_repeated_guest previous_cancellations
## 1               Direct                 0                      0
## 2               Direct                 0                      0
## 3               Direct                 0                      0
## 4            Corporate                 0                      0
## 5                TA/TO                 0                      0
## 6                TA/TO                 0                      0
##   previous_bookings_not_canceled reserved_room_type assigned_room_type
## 1                              0                 C                  C
## 2                              0                 C                  C
## 3                              0                 A                  C
## 4                              0                 A                  A
## 5                              0                 A                  A
## 6                              0                 A                  A
##   booking_changes deposit_type agent company days_in_waiting_list customer_type
## 1               3   No Deposit  NULL    NULL                    0     Transient
## 2               4   No Deposit  NULL    NULL                    0     Transient
## 3               0   No Deposit  NULL    NULL                    0     Transient
## 4               0   No Deposit   304    NULL                    0     Transient
## 5               0   No Deposit   240    NULL                    0     Transient
## 6               0   No Deposit   240    NULL                    0     Transient
##   adr required_car_parking_spaces total_of_special_requests reservation_status
## 1   0                           0                         0          Check-Out
## 2   0                           0                         0          Check-Out
## 3  75                           0                         0          Check-Out
## 4  75                           0                         0          Check-Out
## 5  98                           0                         1          Check-Out
## 6  98                           0                         1          Check-Out
##   reservation_status_date
## 1              2015-07-01
## 2              2015-07-01
```

```
## 3              2015-07-02
## 4              2015-07-02
## 5              2015-07-03
## 6              2015-07-03
```

```
colnames(hotel_bookings)
```

```
##  [1] "hotel"                          "is_canceled"
##  [3] "lead_time"                      "arrival_date_year"
##  [5] "arrival_date_month"             "arrival_date_week_number"
##  [7] "arrival_date_day_of_month"      "stays_in_weekend_nights"
##  [9] "stays_in_week_nights"           "adults"
## [11] "children"                       "babies"
## [13] "meal"                           "country"
## [15] "market_segment"                 "distribution_channel"
## [17] "is_repeated_guest"              "previous_cancellations"
## [19] "previous_bookings_not_canceled" "reserved_room_type"
## [21] "assigned_room_type"             "booking_changes"
## [23] "deposit_type"                   "agent"
## [25] "company"                        "days_in_waiting_list"
## [27] "customer_type"                  "adr"
## [29] "required_car_parking_spaces"    "total_of_special_requests"
## [31] "reservation_status"             "reservation_status_date"
```

Step 3: Install and load the 'ggplot2' package (optional)

Si aún no ha instalado y cargado el paquete `ggplot2`, deberá hacerlo antes de poder usar la función `ggplot()`. Sólo tienes que hacer esto una vez, no cada vez que llames a `ggplot()`.

También puede omitir este paso si no ha cerrado su cuenta de RStudio desde que realizó la última actividad. Si no está seguro, puede ejecutar el fragmento de código y presionar "cancelar" si aparece un mensaje de advertencia que le indica que ya descargó el paquete "ggplot2".

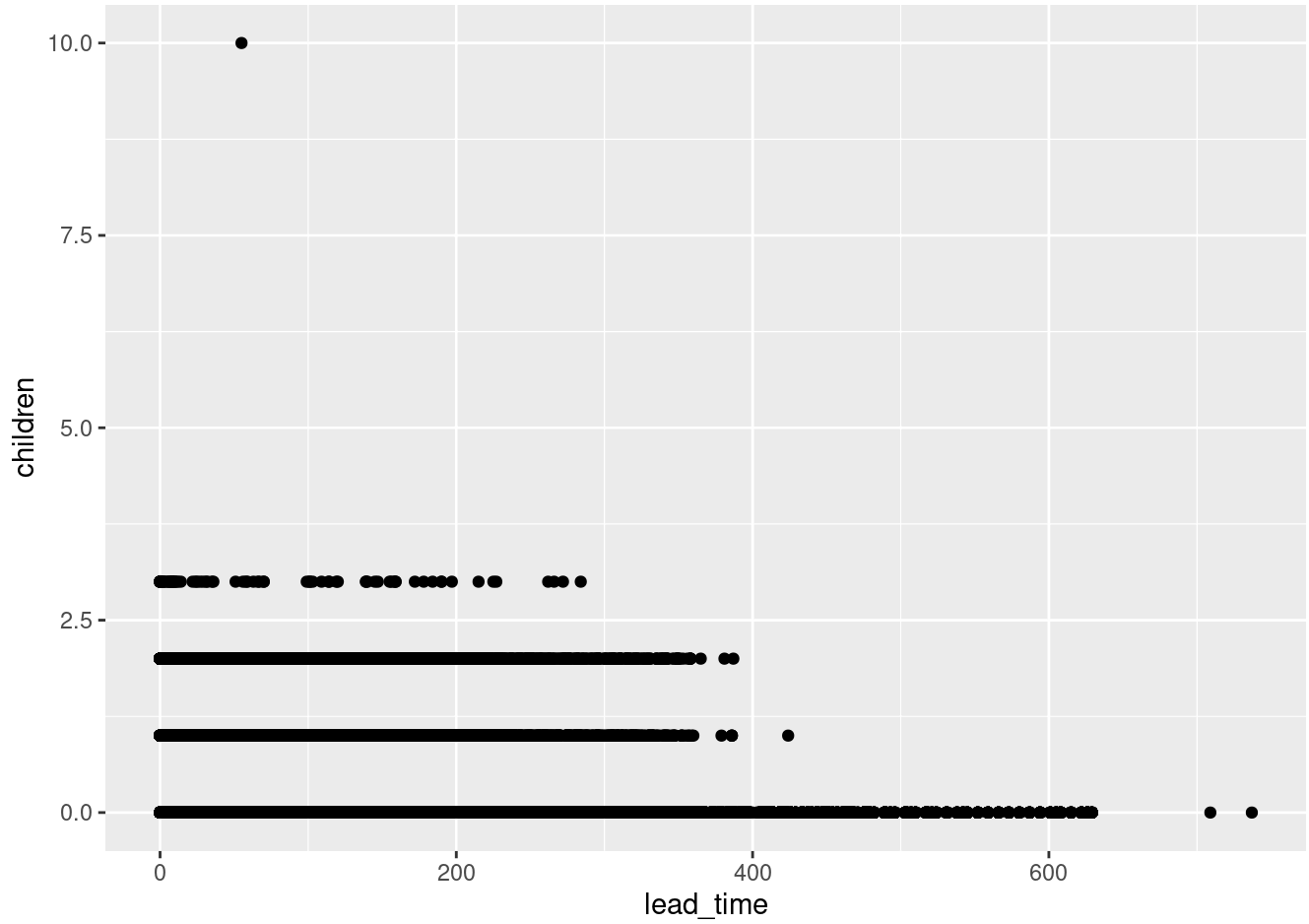
Ejecute el siguiente fragmento de código para instalar y cargar `ggplot2`. ¡Esto puede tomar unos pocos minutos!

Step 4: Making many different charts

Anteriormente, creó un diagrama de dispersión para explorar la relación entre el tiempo de espera de la reserva y los huéspedes que viajan con niños. Como repaso, aquí está el código:

```
ggplot(data = hotel_bookings) +
  geom_point(mapping = aes(x = lead_time, y = children))
```

```
## Warning: Removed 4 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

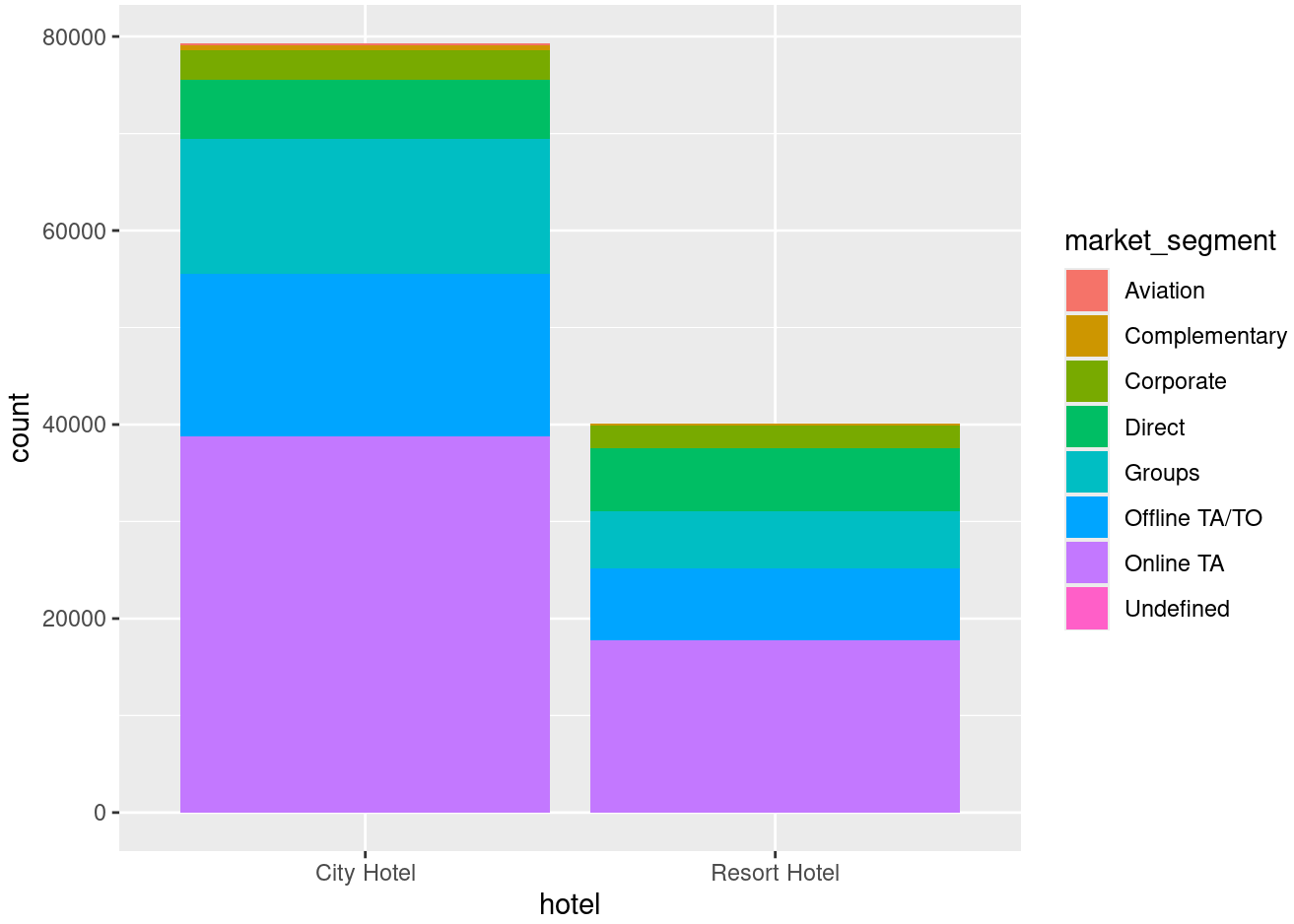


Su parte interesada preguntó sobre el grupo de huéspedes que suelen hacer reservas anticipadas y este gráfico mostró que muchos de estos huéspedes no tienen hijos.

Ahora, su parte interesada quiere realizar una promoción familiar dirigida a segmentos clave del mercado. Quiere saber qué segmentos del mercado generan el mayor número de reservas y dónde se realizan estas reservas (hoteles urbanos u hoteles turísticos).

Primero, decide crear un gráfico de barras que muestre cada tipo de hotel y segmento de mercado. Utiliza diferentes colores para representar cada segmento de mercado:

```
ggplot(data = hotel_bookings) +
  geom_bar(mapping = aes(x = hotel, fill = market_segment))
```

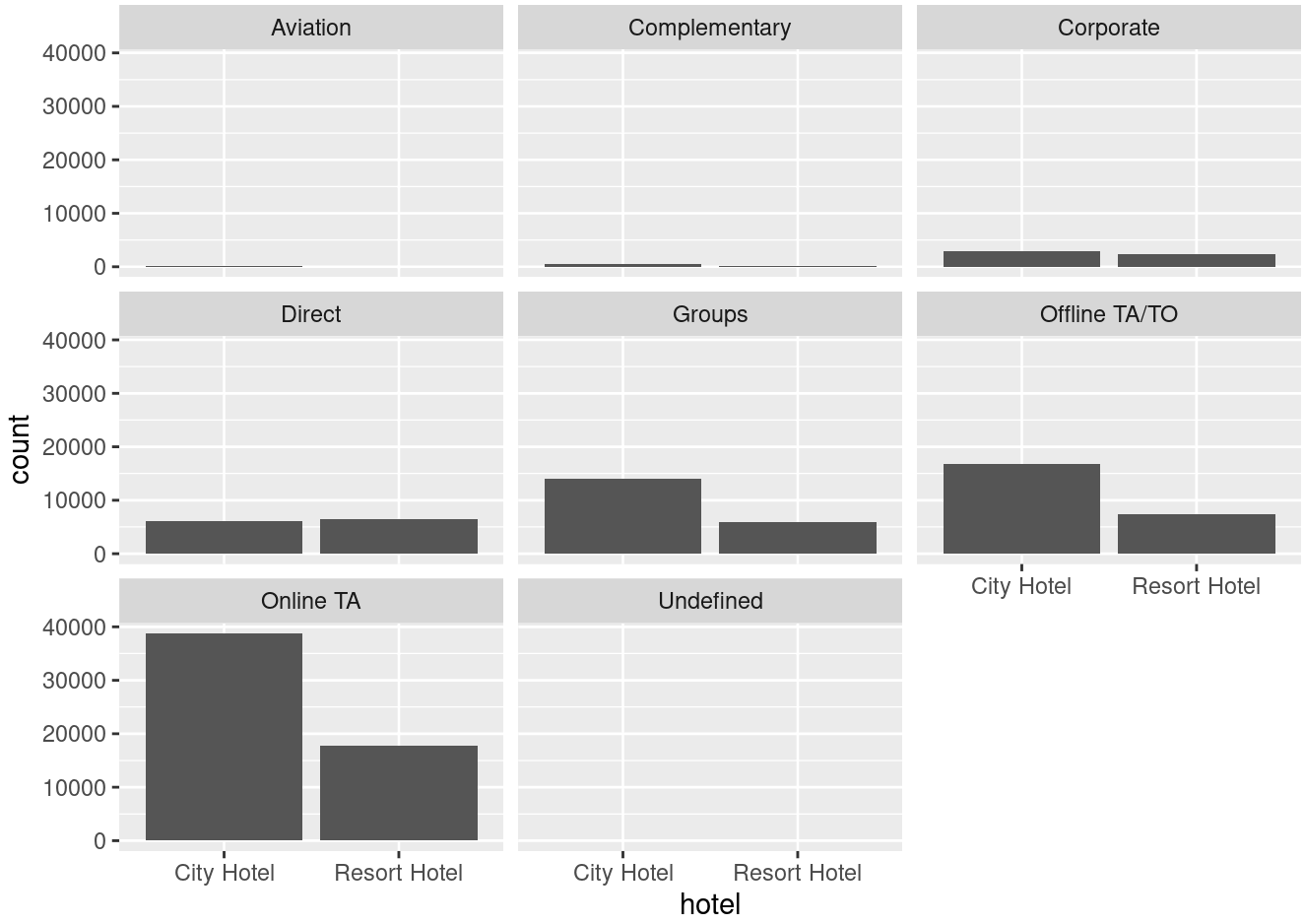


La función geom_bar() utiliza barras para crear un gráfico de barras. El gráfico tiene “hotel” en el eje x y “recuento” en el eje y (nota: si no especifica una variable para el eje y, el código predeterminado es “recuento”). El código asigna la estética de ‘relleno’ a la variable ‘market_segment’ para generar secciones codificadas por colores dentro de cada barra.

Después de crear este gráfico de barras, te das cuenta de que es difícil comparar el tamaño de los segmentos de mercado en la parte superior de las barras. Quiere que su parte interesada pueda comparar claramente cada segmento.

Decide utilizar la función facet_wrap() para crear un gráfico separado para cada segmento de mercado. Entre paréntesis de la función facet_wrap(), agregue la variable ‘market_segment’ después del símbolo de tilde (~):

```
ggplot(data = hotel_bookings) +
  geom_bar(mapping = aes(x = hotel)) +
  facet_wrap(~market_segment)
```



Ahora tienes un gráfico de barras independiente para cada segmento de mercado. Su parte interesada tiene una idea más clara del tamaño de cada segmento de mercado, así como de los datos correspondientes a cada tipo de hotel.

Step 5: Filtering

Para el siguiente paso, necesitarás tener instalado y cargado el paquete `tidyverse` . Es posible que vea una ventana emergente que le preguntará si desea instalar; Si ese es el caso, haga clic en “Instalar”. ¡Esto puede tomar unos pocos minutos!

Si ya has hecho esto porque estás usando el paquete `tidyverse` por tu cuenta, puedes omitir este fragmento de código.

```
#install.packages('tidyverse')
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ lubridate 1.9.2     ✓ tibble    3.2.1
## ✓ purrr     1.0.1     ✓ tidyr     1.3.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

Después de considerar todos los datos, su interesado decide enviar la promoción a las familias que realizan reservas online en hoteles de ciudad. El segmento online es el de más rápido crecimiento y las familias tienden a gastar más en hoteles urbanos que otro tipo de huéspedes.

Su parte interesada le pregunta si puede crear un gráfico que muestre la relación entre el tiempo de espera y los huéspedes que viajan con niños para reservas online en hoteles urbanos. Esto le dará una mejor idea del momento específico de la promoción.

Lo piensas y te das cuenta de que tienes todas las herramientas que necesitas para cumplir con la solicitud. Lo divides en los siguientes dos pasos: 1) filtrar tus datos; 2) trazar sus datos filtrados.

Para el primer paso, puede utilizar la función `filter()` para crear un conjunto de datos que solo incluya los datos que desea. Ingrese 'City Hotel' en el primer conjunto de comillas y 'Online TA' en el segundo conjunto de comillas para especificar sus criterios:

```
onlineta_city_hotels <- filter(hotel_bookings,
                           (hotel=="City Hotel" &
                             hotel_bookings$market_segment== "Online TA"))
```

Tenga en cuenta que puede utilizar el carácter '&' para demostrar que desea que dos condiciones diferentes sean verdaderas. Además, puede utilizar el carácter '$' para especificar a qué columna del marco de datos 'hotel_bookings' hace referencia (por ejemplo, 'market_segment').

Puede utilizar la función `View` () para comprobar su nuevo marco de datos:

```
#View(onlineta_city_hotels)
```

También hay otra forma de hacer esto. ¡Puedes usar el operador de tubería (%>%) para hacer esto en pasos!

Nombra este marco de datos `onlineta_city_hotels_v2` :

```
onlineta_city_hotels_v2 <- hotel_bookings %>%
  filter(hotel=="City Hotel") %>%
  filter(market_segment=="Online TA")
```

Observe cómo en el fragmento de código anterior, el símbolo %>% se usa para anotar los pasos lógicos de este código. Primero, comienza con el nombre del marco de datos, `onlineta_city_hotels_v2` , Y LUEGO le dice a `R` que comience con el marco de datos original `hotel_bookings` . Luego le dice que filtre por la columna 'hotel'; finalmente, le indica que filtre por la columna 'market_segment'.

Este fragmento de código genera el mismo marco de datos usando la función `View()` :

```
#View(onlineta_city_hotels_v2)
```

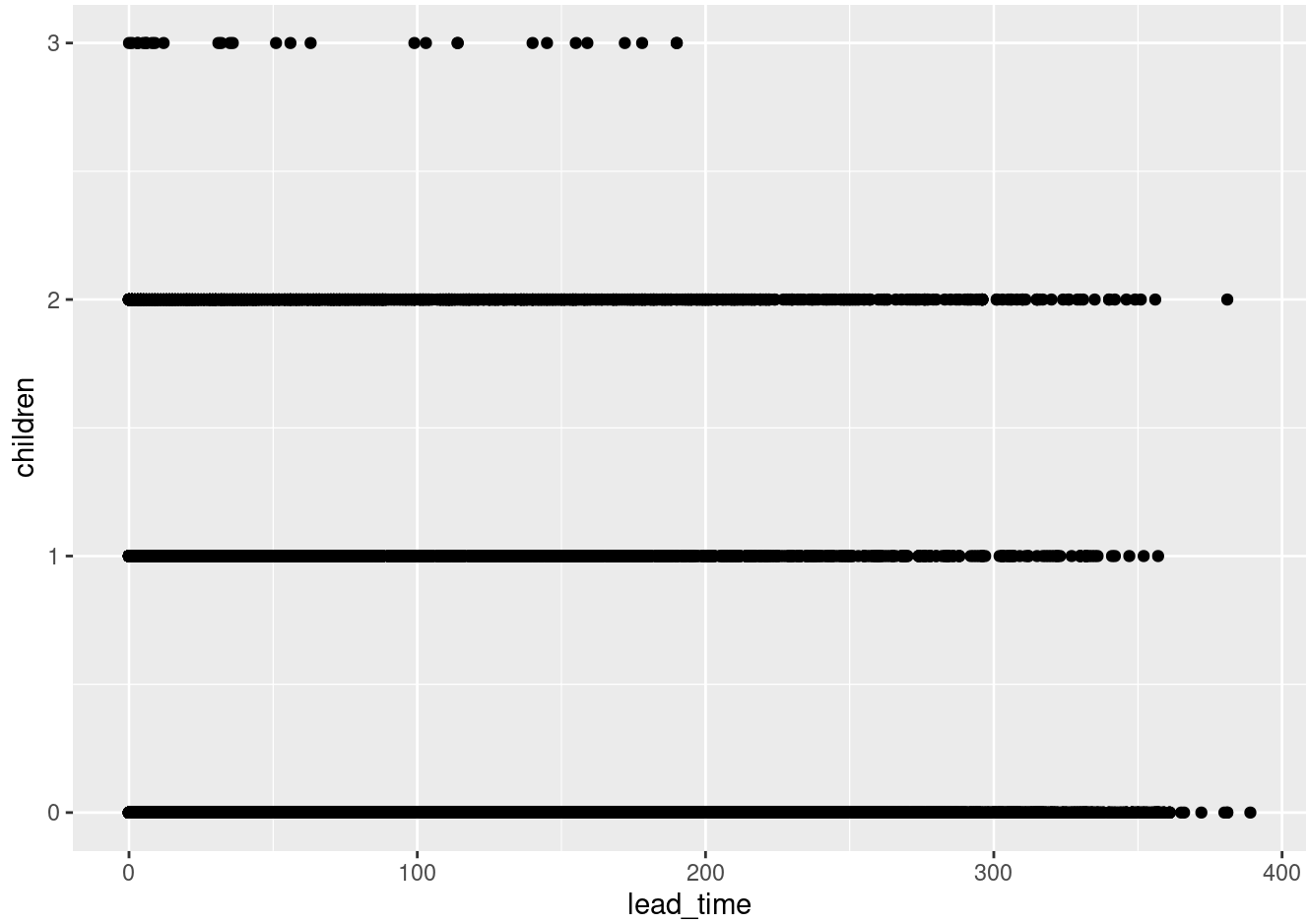
Step 6: Use your new dataframe

Puede utilizar cualquiera de los marcos de datos que creó anteriormente para sus nuevos gráficos porque son iguales.

Utilizando el código de su diagrama de dispersión anterior, reemplace `variable_name` en el siguiente fragmento de código con `onlineta_city_hotels` o `onlineta_city_hotels_v2` para trazar los datos que solicitó su parte interesada:

```
ggplot(data = onlineta_city_hotels) +
  geom_point(mapping = aes(x = lead_time, y = children))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



Según su filtro anterior, este diagrama de dispersión muestra datos de reservas en línea para hoteles urbanos. El gráfico revela que las reservas con niños tienden a tener un plazo de entrega más corto, y las reservas con 3 niños tienen un plazo de entrega significativamente más corto (<200 días). De esta forma, las promociones dirigidas a familias se podrán realizar más cerca de las fechas de reserva válidas.

Activity Wrap Up

Los filtros le permiten crear diferentes vistas de sus datos y le permiten investigar relaciones más específicas dentro de sus datos. Puede practicar estas habilidades modificando los fragmentos de código en el archivo rmd o utilizar este código como punto de partida en su propia consola de proyecto. Ahora que su parte interesada ha tenido la oportunidad de revisar estos gráficos, está interesada en agregar anotaciones que puedan usar para explicar los datos en una presentación. Afortunadamente, `ggplot2` tiene una función que te permitirá hacer precisamente eso. ¡Aprenderás más sobre `ggplot2` en la siguiente actividad!