Lesson 2: GGPlot

Step 1: Import data

The data in this example is originally from the article Hotel Booking Demand Datasets (https://www.sciencedirect.com/science/article/pii/S2352340918315191), written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.

The data was downloaded and cleaned by Thomas Mock and Antoine Bichat for #TidyTuesday during the week of February 11th, 2020 (https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-11/readme.md).

Learn more about the dataset here: https://www.kaggle.com/jessemostipak/hotel-booking-demand

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

Step 2: Look at a sample of data

Using the head() function to preview data:

```
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
```

Use colnames() to get the names of all the columns in the data set.

```
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

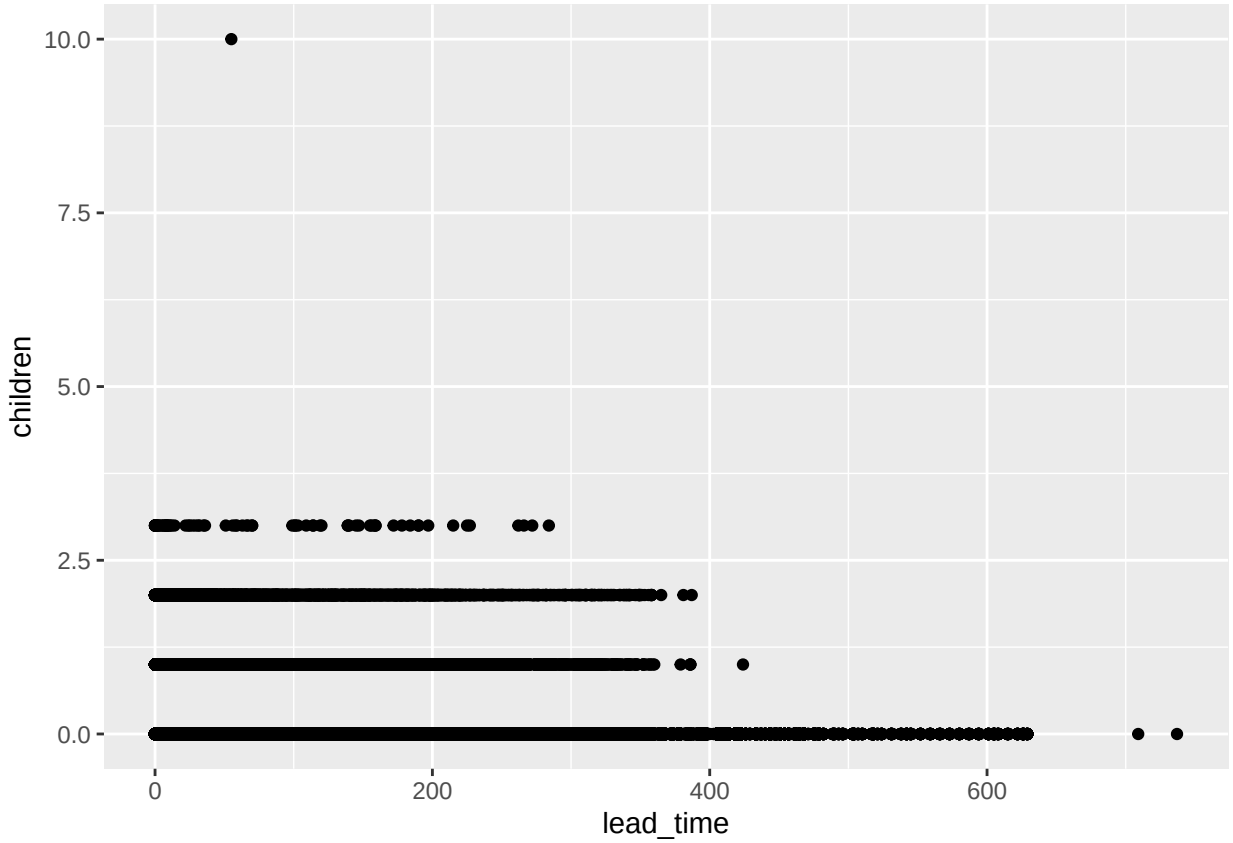
Install and load `ggplot2`.

Step 4: Begin creating a plot

Use `ggplot2` to determine if people with children book hotel rooms in advance. Try running the code below:

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

```
## Warning: Removed 4 rows containing missing values (`geom_point()`).
```



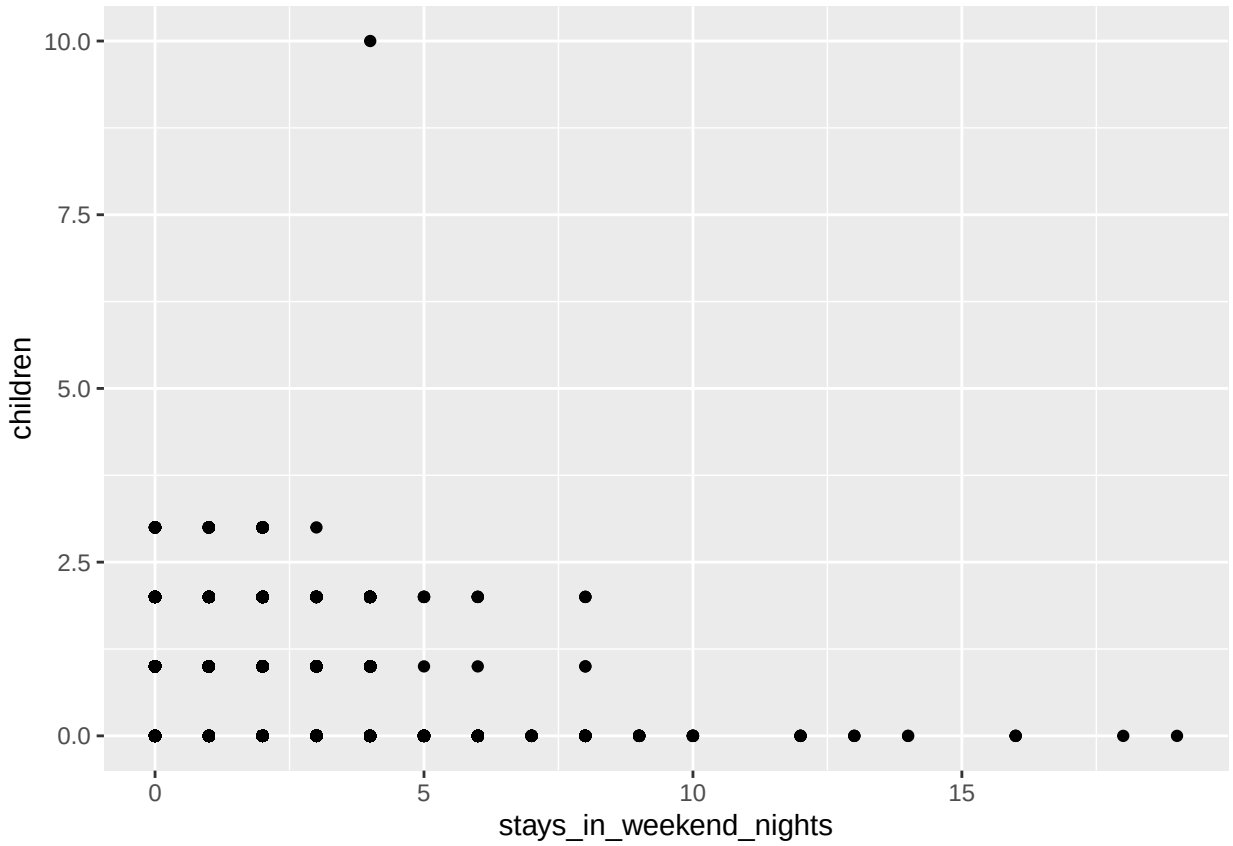
On the x-axis, the plot shows how far in advance a booking is made, with the bookings furthest to the right happening the most in advance. On the y-axis it shows how many children there are in a party.

Step 5: More Plot

Mapping 'stays_in_weekend_nights' on the x-axis and 'children' on the y-axis by filling out the remainder of the code below:

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

```
## Warning: Removed 4 rows containing missing values (`geom_point()`).
```


10.0
7.5
5.0
2.5
0.0
children
0
5
10
15
stays_in_weekend_nights