Exploratory Data Analysis on Hotel Dataset to Draw Insights

Kar

5/30/2021

Summary

The Scenario

In this scenario, I am working for a hotel booking company as a data analyst, a cleaned dataset has been given to me for exploratory data analysis to draw insights to share to the team.

data and functions import

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
hotel <- read.csv("hotel_bookings.csv")
hotel <- as_tibble(hotel)
```

Data exploration

First, explore the basic structure of the dataset before looking into further details such as the summary of individual levels within each column. The basic structures that we should first to understand could be the number of rows, columns (to be perceived as variables within the dataset), the existence of categorical and numerical variables and who are them.

```
skim_without_charts(hotel)
```

Table 1: Data summary

Name	hotel
Number of rows	119390
Number of columns	32
Column type frequency:	
character	14
numeric	18
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
hotel	0	1	10	12	0	2	0
arrival_date_month	0	1	3	9	0	12	0
meal	0	1	2	9	0	5	0
country	0	1	2	4	0	178	0
market_segment	0	1	6	13	0	8	0
distribution_channel	0	1	3	9	0	5	0
reserved_room_type	0	1	1	1	0	10	0
assigned_room_type	0	1	1	1	0	12	0
deposit_type	0	1	10	10	0	3	0
agent	0	1	1	4	0	334	0
company	0	1	1	4	0	353	0
customer_type	0	1	5	15	0	4	0
reservation_status	0	1	7	9	0	3	0
reservation_status_date	0	1	10	10	0	926	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
is_canceled	0	1	0.37	0.48	0.00	0.00	0.00	1
lead_time	0	1	104.01	106.86	0.00	18.00	69.00	160
arrival_date_year	0	1	2016.16	0.71	2015.00	2016.00	2016.00	2017
arrival_date_week_number	0	1	27.17	13.61	1.00	16.00	28.00	38
arrival_date_day_of_month	0	1	15.80	8.78	1.00	8.00	16.00	23
stays_in_weekend_nights	0	1	0.93	1.00	0.00	0.00	1.00	2
stays_in_week_nights	0	1	2.50	1.91	0.00	1.00	2.00	3
adults	0	1	1.86	0.58	0.00	2.00	2.00	2
children	4	1	0.10	0.40	0.00	0.00	0.00	0
babies	0	1	0.01	0.10	0.00	0.00	0.00	0
is_repeated_guest	0	1	0.03	0.18	0.00	0.00	0.00	0
previous_cancellations	0	1	0.09	0.84	0.00	0.00	0.00	0
previous_bookings_not_canceled	0	1	0.14	1.50	0.00	0.00	0.00	0
booking_changes	0	1	0.22	0.65	0.00	0.00	0.00	0

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
days_in_waiting_list	0	1	2.32	17.59	0.00	0.00	0.00	0
adr	0	1	101.83	50.54	-6.38	69.29	94.58	126
required_car_parking_spaces	0	1	0.06	0.25	0.00	0.00	0.00	0
total_of_special_requests	0	1	0.57	0.79	0.00	0.00	0.00	1

We can use other code quickly visualise the names of all the columns in the dataset.

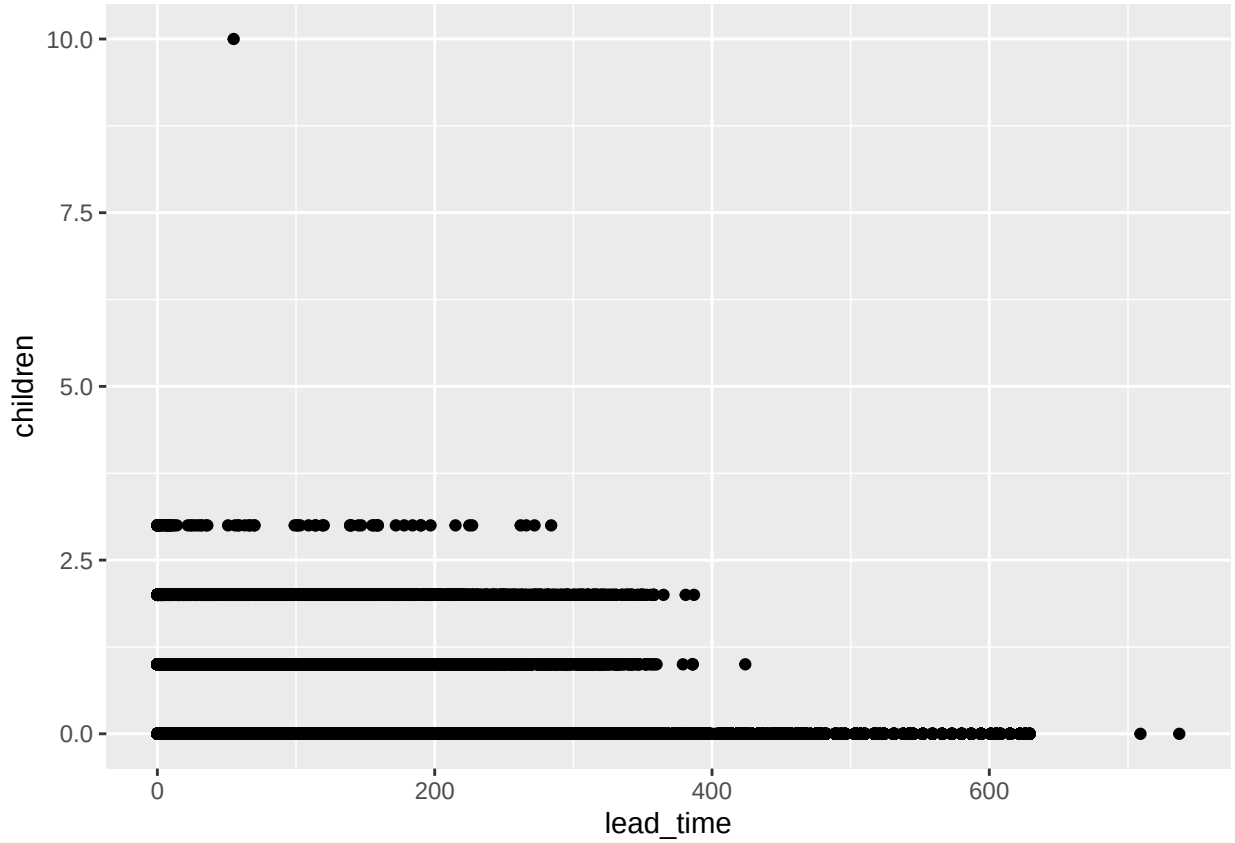
```
glimpse(hotel)
```

```
## Rows: 119,390
## Columns: 32
## $ hotel                           <chr> "Resort Hotel", "Resort Hotel", "Resort~
## $ is_canceled                     <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, ~
## $ lead_time                       <int> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, ~
## $ arrival_date_year               <int> 2015, 2015, 2015, 2015, 2015, 2015, 201~
## $ arrival_date_month              <chr> "July", "July", "July", "July", "July",~
## $ arrival_date_week_number        <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,~
## $ arrival_date_day_of_month       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ stays_in_weekend_nights         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ stays_in_week_nights            <int> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, ~
## $ adults                          <int> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ children                        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ babies                          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ meal                            <chr> "BB", "BB", "BB", "BB", "BB", "BB", "BB~
## $ country                         <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "GBR~
## $ market_segment                  <chr> "Direct", "Direct", "Direct", "Corporat~
## $ distribution_channel            <chr> "Direct", "Direct", "Direct", "Corporat~
## $ is_repeated_guest               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_cancellations          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_bookings_not_canceled <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ reserved_room_type              <chr> "C", "C", "A", "A", "A", "A", "C", "C",~
## $ assigned_room_type              <chr> "C", "C", "C", "A", "A", "A", "C", "C",~
## $ booking_changes                 <int> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ deposit_type                    <chr> "No Deposit", "No Deposit", "No Deposit~
## $ agent                           <chr> "NULL", "NULL", "NULL", "304", "240", "~
## $ company                         <chr> "NULL", "NULL", "NULL", "NULL", "NULL",~
## $ days_in_waiting_list            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ customer_type                   <chr> "Transient", "Transient", "Transient", ~
## $ adr                             <dbl> 0.00, 0.00, 75.00, 75.00, 98.00, 98.00,~
## $ required_car_parking_spaces     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ total_of_special_requests       <int> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 3, ~
## $ reservation_status              <chr> "Check-Out", "Check-Out", "Check-Out", ~
## $ reservation_status_date         <chr> "2015-07-01", "2015-07-01", "2015-07-02~
```

Data visualisation

Objective 1: A stakeholder tells you, "I want to target people who book early, and I have a hypothesis that people with children have to book in advance." Now, I will start to explore the data to create a visualization to see how true that statement is.

```
ggplot(hotel, aes(x = lead_time, y = children)) +
  geom_point()
```



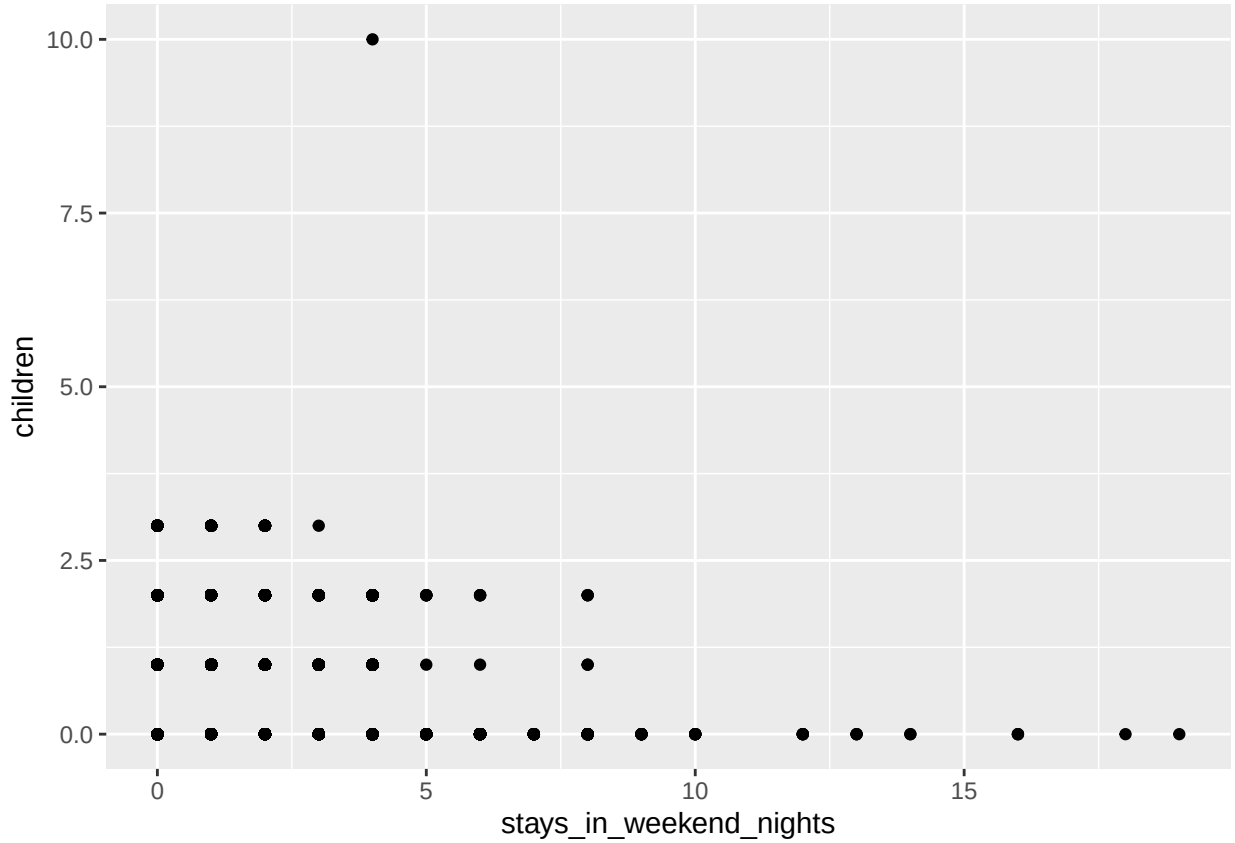
On the x-axis, the plot shows how far in advance a booking is made, with the booking furthest to the right happening the most in advance. On the y-axis, it shows how many children there are in the party. The plot reveals that your stakeholder's hypothesis is incorrect. You report back to your stakeholder that many of the advanced bookings are being made by people with 0 children.

Objective 2: Next, your stakeholder says that she wants to increase weekend bookings, an important source of revenue for the hotel. Your stakeholder wants to know what group of guests book the most weekend nights in order to target that group in a new marketing campaign.

She suggests that guests without children book the most weekend nights. Is this true?

```
ggplot(hotel, aes(x = stays_in_weekend_nights, y = children)) +
  geom_point()
```

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

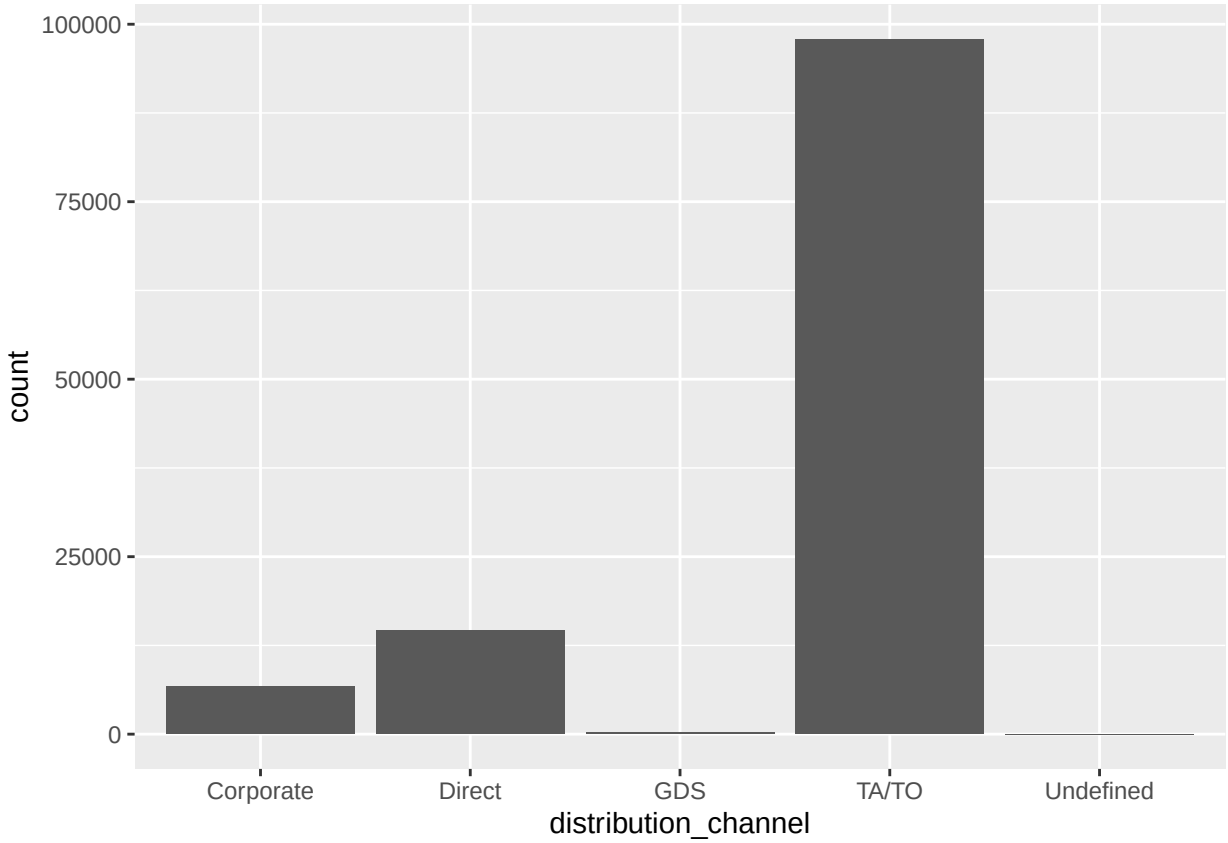


Objective 3: Your stakeholder is interested in developing promotions based on different booking distributions, but first they need to know how many of the transactions are occurring for each different distribution type.

```
hotel$distribution_channel <- as.factor(hotel$distribution_channel)
levels(hotel$distribution_channel)
```

```
## [1] "Corporate" "Direct"    "GDS"       "TA/TO"     "Undefined"
```

```
ggplot(hotel, aes(x = distribution_channel)) +
  geom_bar()
```



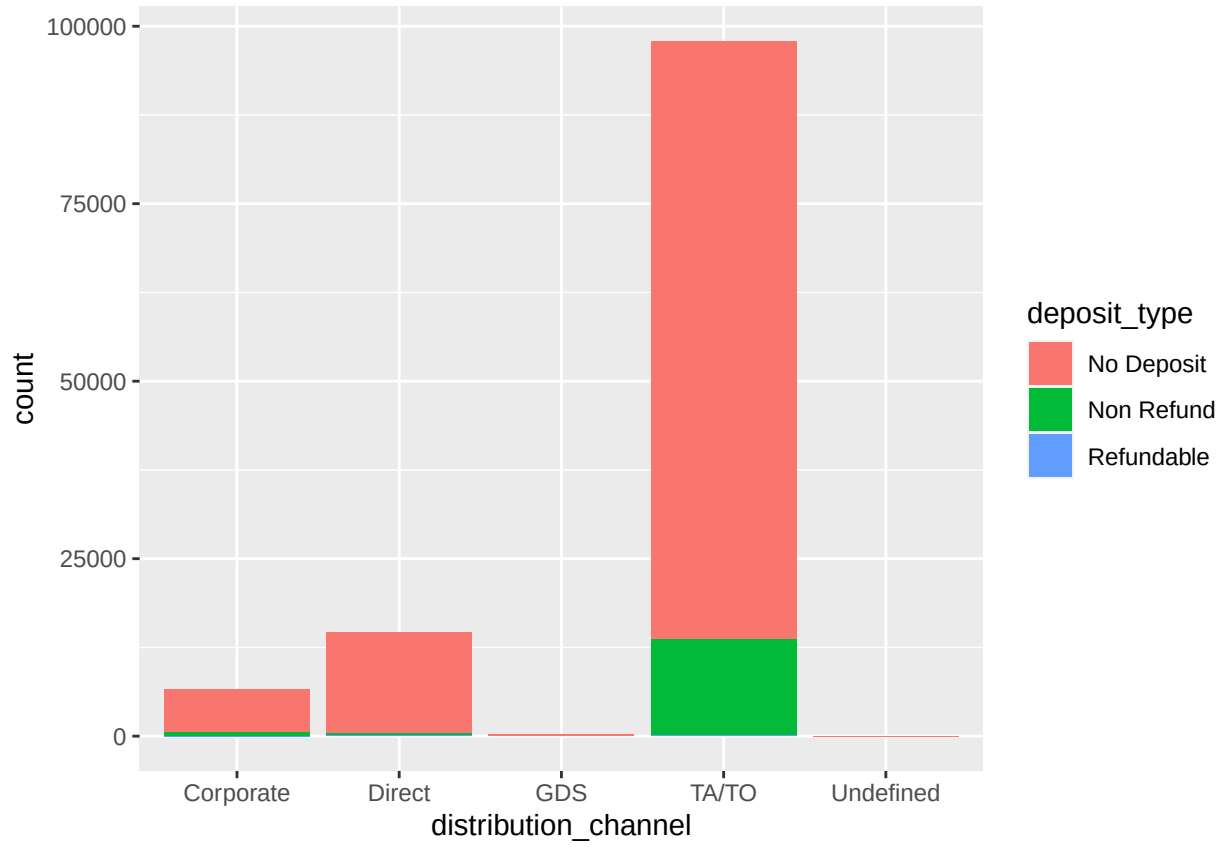
We found that TA/TO is the most used booking channel.

Objective 5: After observing the TA/TO being the best booking channel, now a question arisen regrading if the number of bookings for each distribution type is different depending on whether or not there was a deposit or what market segment they represent.

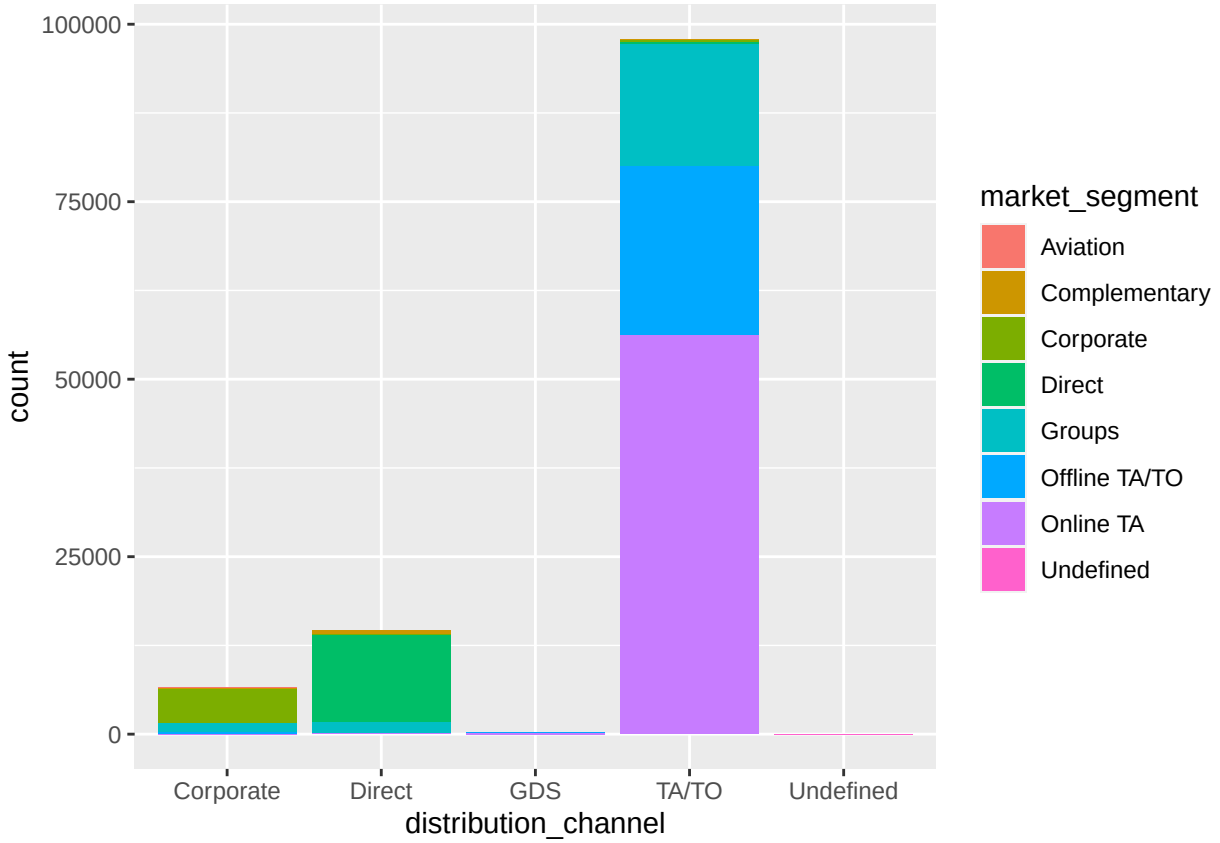
```
hotel$deposit_type <- as.factor(hotel$deposit_type)
levels(hotel$deposit_type)
```

```
## [1] "No Deposit" "Non Refund" "Refundable"
```

```
ggplot(hotel, aes(x = distribution_channel, fill = deposit_type)) +
  geom_bar()
```

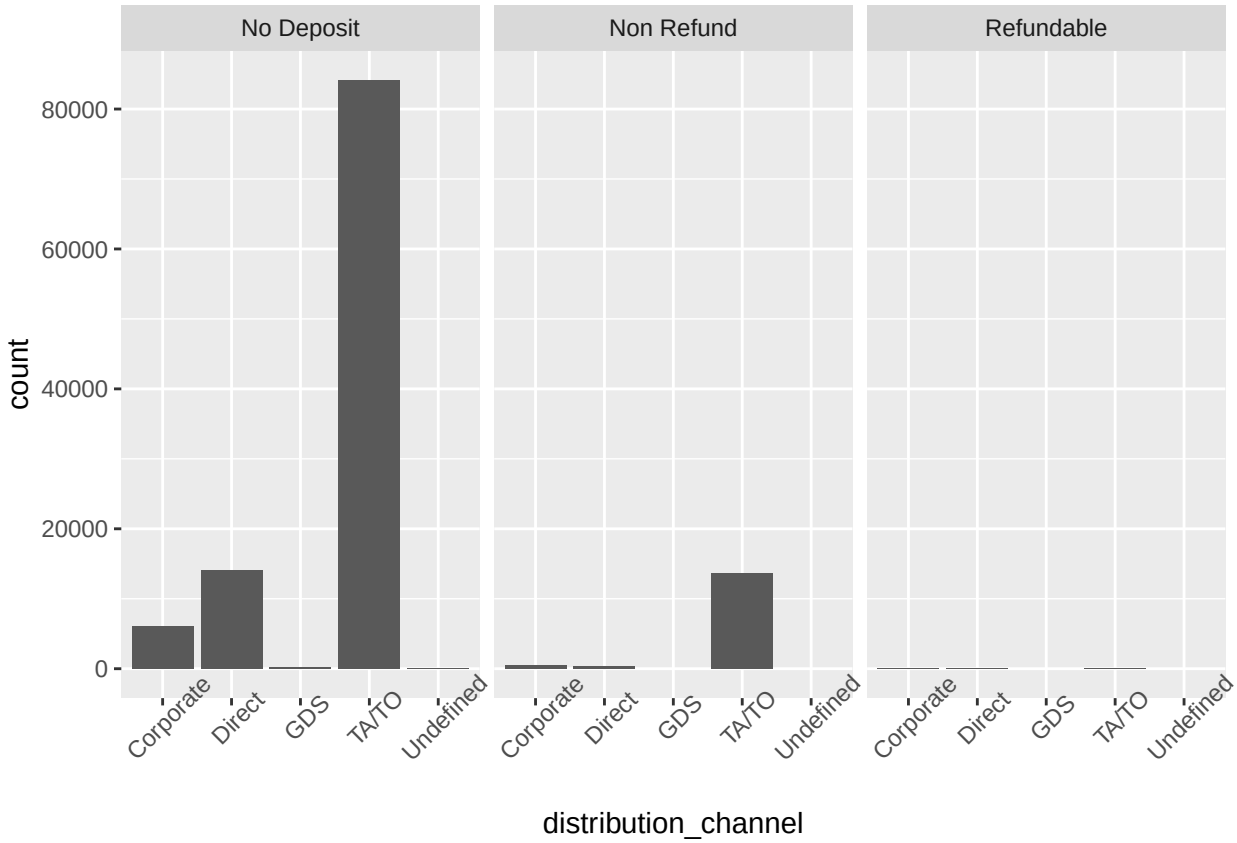



```
ggplot(hotel, aes(x = distribution_channel, fill = market_segment)) +
  geom_bar()
```

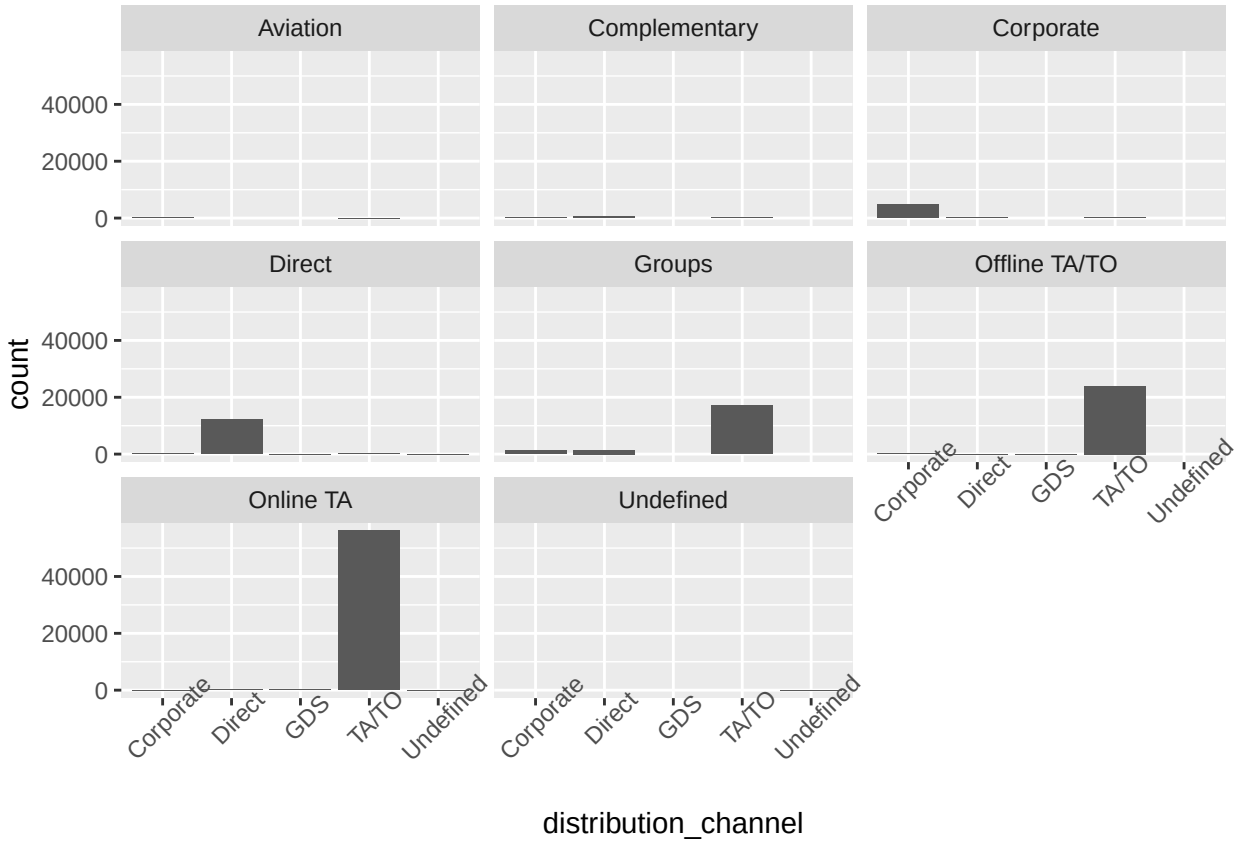


Objective 6: After reviewing the new charts, your stakeholder asks you to create separate charts for each deposit type and market segment to help them understand the differences more clearly.

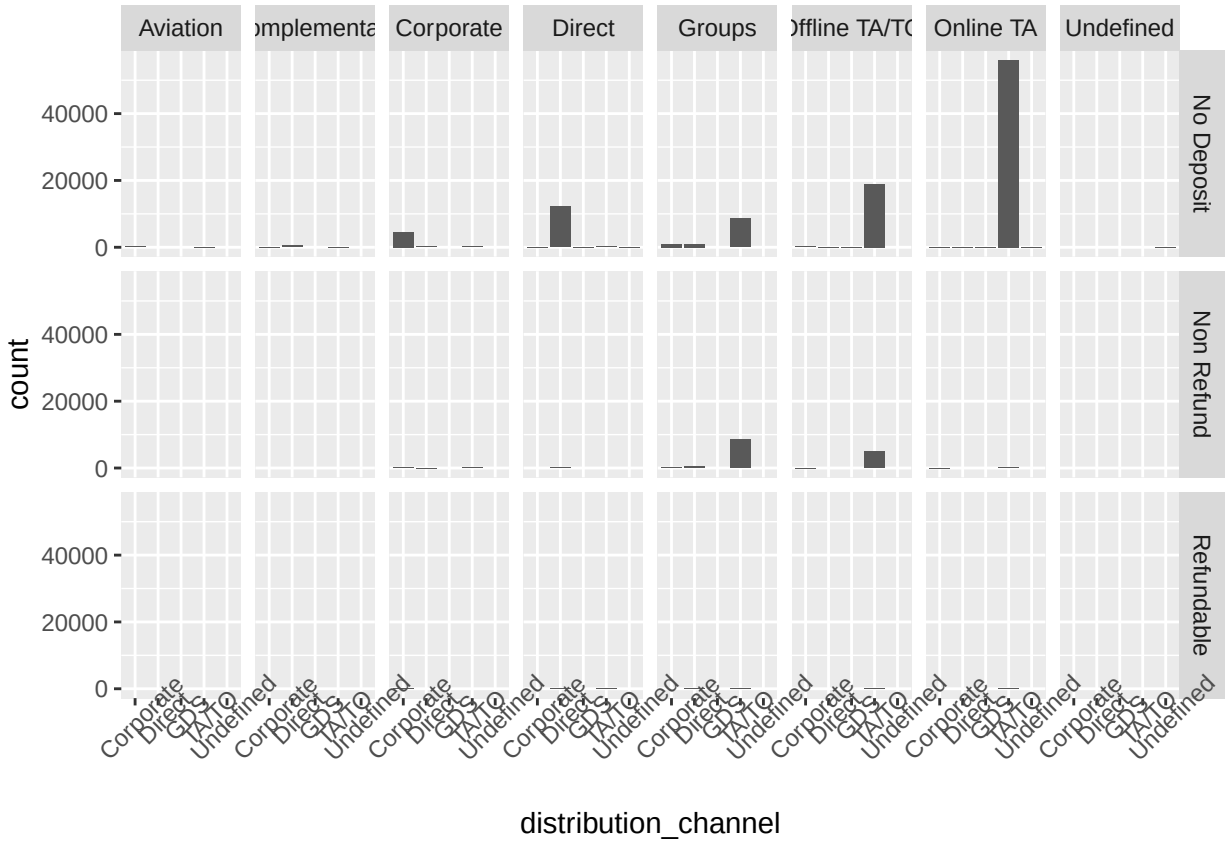
```
ggplot(hotel, aes(x = distribution_channel)) +
  geom_bar() +
  facet_wrap(~deposit_type) +
  theme(axis.text.x = element_text(angle = 45))
```




```
ggplot(hotel, aes(x = distribution_channel)) +
  geom_bar() +
  facet_wrap(~market_segment) +
  theme(axis.text.x = element_text(angle = 45))
```




```
ggplot(hotel, aes(x = distribution_channel)) +
  geom_bar() +
  facet_grid(deposit_type~market_segment) +
  theme(axis.text.x = element_text(angle = 45))
```



These charts are probably overwhelming and too hard to read, but it can be useful if you are exploring your data through visualizations.

Objective 7: After considering all the data, your stakeholder decides to send the promotion to families that make TA online bookings for city hotels. The online segment is the fastest growing segment, and families tend to spend more at city hotels than other types of guests.

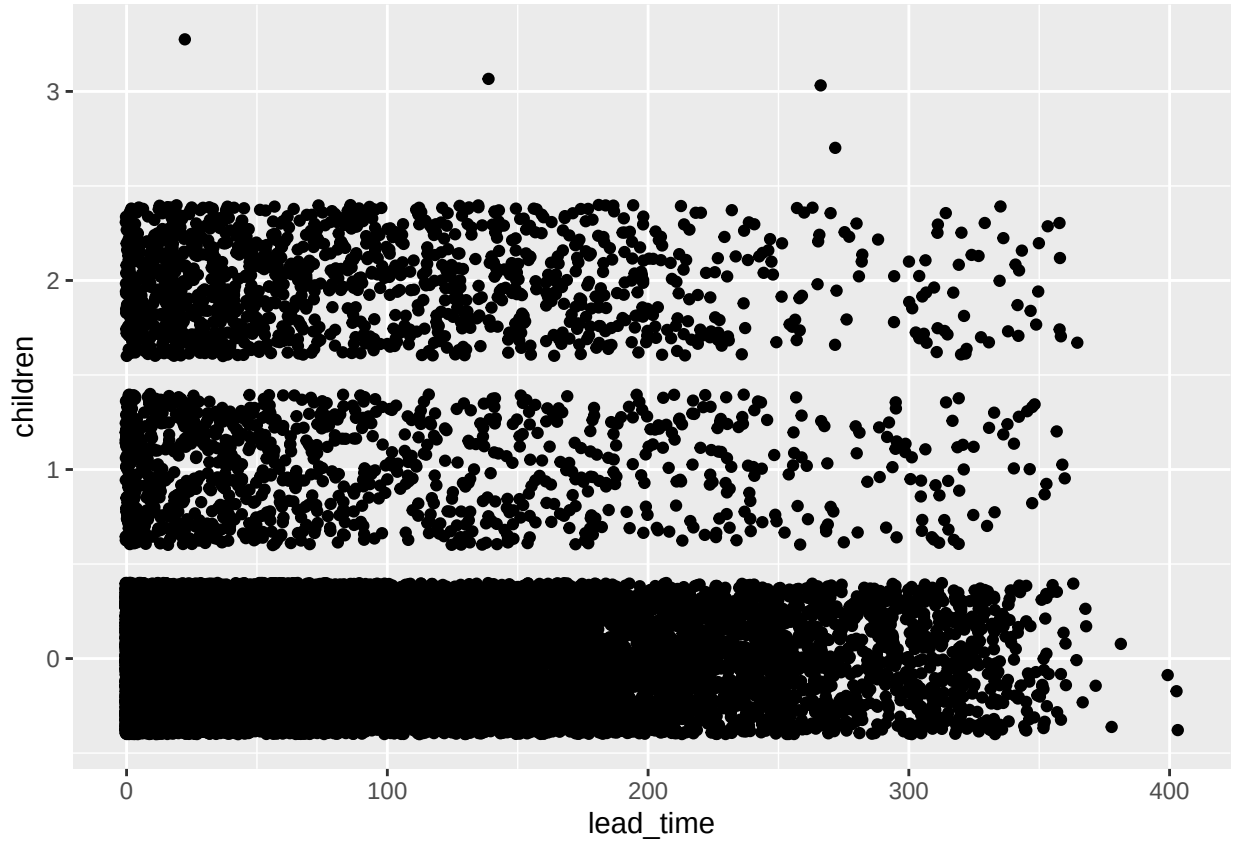
Your stakeholder asks if you can create a plot that shows the relationship between lead time and guests traveling with children for online bookings at city hotels. This will give her a better idea of the specific timing for the promotion.

```
hotel$hotel <- as.factor(hotel$hotel)
levels(hotel$hotel)
```

```
## [1] "City Hotel"   "Resort Hotel"
```

```
onlineta_city_hotels <- hotel %>%
  filter(hotel == "Resort Hotel",
         market_segment == "Online TA")

ggplot(onlineta_city_hotels, aes(x = lead_time, y = children)) +
  geom_jitter()
```

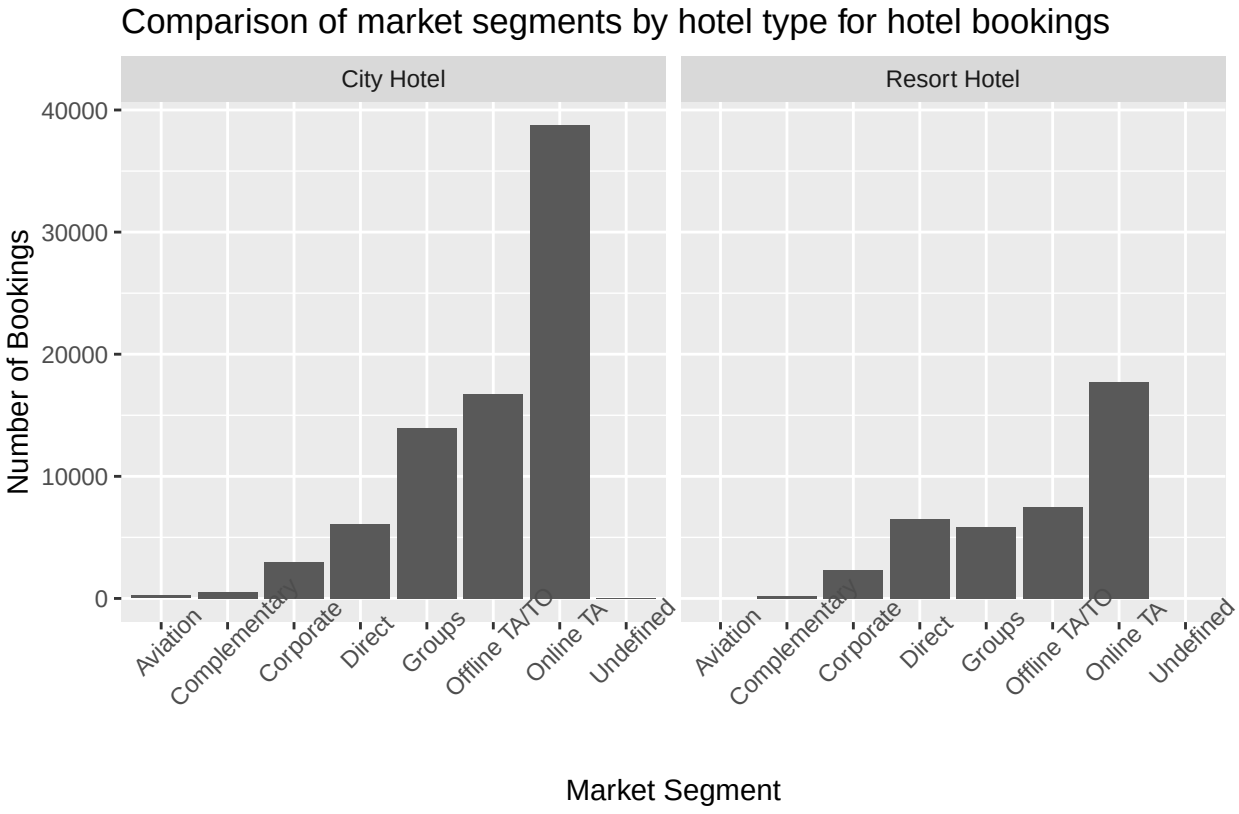


The plot reveals that bookings with children tend to have a shorter lead time, and bookings with 3 children have a significantly shorter lead time (<200 days). So, promotions targeting families can be made closer to the valid booking dates.

Objective 8 Your stakeholder tells you that they would like you to share they visualization breaking down payment type by city because it will help inform how the company targets promotions in the future. They ask you to create a cleaned and labeled version and save it as a .png file for them to include in a presentation.

```
mindate <- min(hotel$arrival_date_year)
maxdate <- max(hotel$arrival_date_year)


ggplot(hotel, aes(x = market_segment)) +
  geom_bar() +
  facet_wrap(~hotel) +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(title = "Comparison of market segments by hotel type for hotel bookings",
       caption = paste0("Data from: ",
                        mindate,
                        " to ",
                        maxdate),
       x = "Market Segment",
       y = "Number of Bookings")
```


Comparison of market segments by hotel type for hotel bookings
City Hotel
Resort Hotel
Number of Bookings
40000
30000
20000
10000
0
Aviation
Complementary
Corporate
Direct
Groups
Offline TA/TO
Online TA
Undefined
Market Segment
Data from: 2015 to 2017