STAT 495
Spring 2021
Dr. Xiyue Liao

# Hotel Booking Demands
# Data Analysis

---

**By Noah Gallagher, Alex Gonzalez, Chris Ibarra**

## Introduction

This data set compares various booking information between two hotels, a city hotel and a resort hotel. The data was collected between July of 2015 and August of 2017 and was posted on Kaggle. Each entry in this dataset represents a single hotel booking made by a guest. The data includes information such as when the booking was made, the guest's length of stay, the number of special requests made, etc.

## Questions of Interest

For this Presentation, we will be focusing on a few topics that relate to our dataset. These topics include:

- When is the best time to book a Hotel Room?
- How far in advance do people make Hotel Bookings?
- How are Hotel Bookings related to Cancellations?
- Can we make any predictions from this dataset?

## Analysis

### Exploratory Data Analysis

The data contains 119,390 rows with 32 attributes.

```
## Rows: 119,390
## Columns: 32
## $ hotel                          <chr> "Resort Hotel", "Resort Hotel",
## $ is_canceled                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0,
## $ lead_time                      <dbl> 342, 737, 7, 13, 14, 14, 0, 9, 85,
## $ arrival_date_year              <dbl> 2015, 2015, 2015, 2015, 2015, 2015,
## $ arrival_date_month             <chr> "July", "July", "July", "July",
## $ arrival_date_week_number       <dbl> 27, 27, 27, 27, 27, 27, 27, 27, 27,
## $ arrival_date_day_of_month      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
## $ stays_in_weekend_nights        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ stays_in_week_nights           <dbl> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4,
## $ adults                         <dbl> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2,
## $ children                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

## Summary Statistics

```r
library(skimr)
skim(hotel.data)
```

*Data summary*

| | |
|---|---|
| Name | hotel.data |
| Number of rows | 119390 |
| Number of columns | 32 |
| ________________________ | |
| Column type frequency: | |
| character | 13 |
| Date | 1 |
| 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 |

**Variable type: Date**

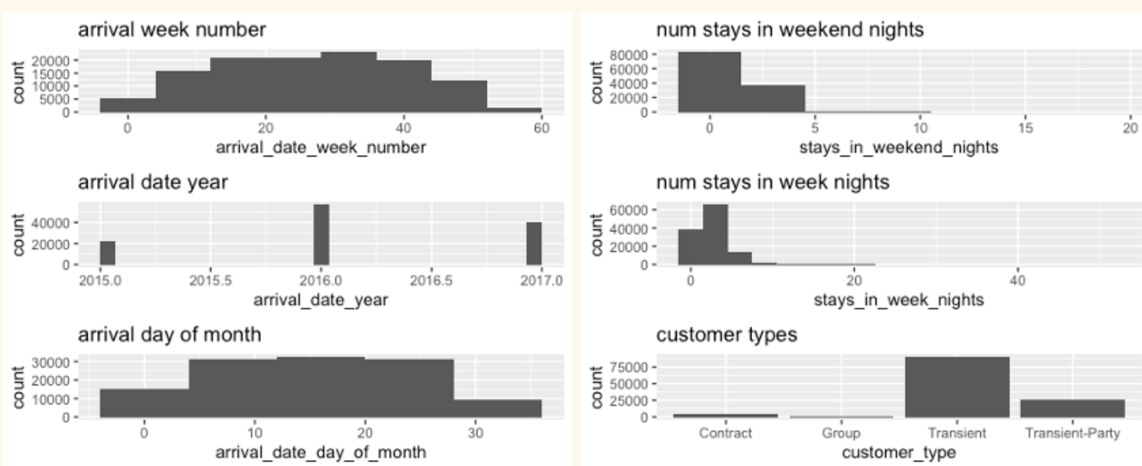
| skim_variable | n_missing | complete_rate | min | max | median | n_unique |
|---|---|---|---|---|---|---|
| reservation_status_date | 0 | 1 | 2014-10-17 | 2017-09-14 | 2016-08-07 | 926 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| is_canceled | 0 | 1 | 0.37 | 0.48 | 0.00 | 0.00 | 0.00 | 1 | 1 | |
| lead_time | 0 | 1 | 104.01 | 106.86 | 0.00 | 18.00 | 69.00 | 160 | 737 | |
| arrival_date_year | 0 | 1 | 2016.16 | 0.71 | 2015.00 | 2016.00 | 2016.00 | 2017 | 2017 | |
| arrival_date_week_number | 0 | 1 | 27.17 | 13.61 | 1.00 | 16.00 | 28.00 | 38 | 53 | |
| arrival_date_day_of_month | 0 | 1 | 15.80 | 8.78 | 1.00 | 8.00 | 16.00 | 23 | 31 | |
| stays_in_weekend_nights | 0 | 1 | 0.93 | 1.00 | 0.00 | 0.00 | 1.00 | 2 | 19 | |
| stays_in_week_nights | 0 | 1 | 2.50 | 1.91 | 0.00 | 1.00 | 2.00 | 3 | 50 | |
| adults | 0 | 1 | 1.86 | 0.58 | 0.00 | 2.00 | 2.00 | 2 | 55 | |
| children | 4 | 1 | 0.10 | 0.40 | 0.00 | 0.00 | 0.00 | 0 | 10 | |
| babies | 0 | 1 | 0.01 | 0.10 | 0.00 | 0.00 | 0.00 | 0 | 10 | |
| is_repeated_guest | 0 | 1 | 0.03 | 0.18 | 0.00 | 0.00 | 0.00 | 0 | 1 | |
| previous_cancellations | 0 | 1 | 0.09 | 0.84 | 0.00 | 0.00 | 0.00 | 0 | 26 | |
| previous_bookings_not_canceled | 0 | 1 | 0.14 | 1.50 | 0.00 | 0.00 | 0.00 | 0 | 72 | |
| booking_changes | 0 | 1 | 0.22 | 0.65 | 0.00 | 0.00 | 0.00 | 0 | 21 | |
| days_in_waiting_list | 0 | 1 | 2.32 | 17.59 | 0.00 | 0.00 | 0.00 | 0 | 391 | |
| adr | 0 | 1 | 101.83 | 50.54 | -6.38 | 69.29 | 94.58 | 126 | 5400 | |
| required_car_parking_spaces | 0 | 1 | 0.06 | 0.25 | 0.00 | 0.00 | 0.00 | 0 | 8 | |
| total_of_special_requests | 0 | 1 | 0.57 | 0.79 | 0.00 | 0.00 | 0.00 | 1 | 5 | |

## Data Visualization

Visualizing numeric variables to better understand data and distribution.



**Topic 1: When is the Best Time of the Year to Book a Hotel Room?**

For topic 1, our goal was to find when is the best time of the year to book a hotel room.

### 1. Average Booking Dates

The average week of arrival is the 27th week of the year (June).

The average day of arrival is the 16th of the month.

| week_mean | day_month_mean |
|---|---|
| <dbl> | <dbl> |
| 27.16517 | 15.79824 |

1 row

### 2. Busiest and Slowest Times of the Year

The busiest time of the year to book is August with 13877 bookings.

| arrival_date_month | count |
|---|---|
| <chr> | <int> |
| August | 13877 |
| July | 12661 |
| May | 11791 |
| October | 11160 |
| April | 11089 |

The slowest time of the year to book is January with 5929 bookings.

| arrival_date_month | count |
|---|---|
| <chr> | <int> |
| January | 5929 |
| December | 6780 |
| November | 6794 |
| February | 8068 |
| March | 9794 |

### 3. Hotel Fees

The average price of stay per night for the busiest time of the year was about $140.11.

| arrival_date_month | count | price_mean |
|---|---|---|
| <chr> | <int> | <dbl> |
| August | 13877 | 140.11152 |
| July | 12661 | 126.78801 |
| May | 11791 | 108.69552 |

Price for the slowest times of the year per night was an average price of $70.36.

| arrival_date_month | count | price_mean |
|---|---|---|
| <chr> | <int> | <dbl> |
| January | 5929 | 70.36124 |
| December | 6780 | 81.07678 |
| November | 6794 | 73.79496 |

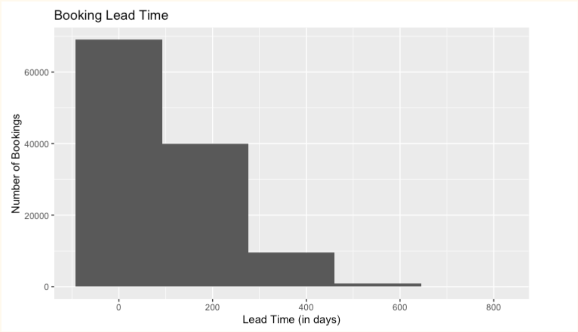
***In Conclusion,*** the busier it is in a hotel the higher the price for a booking and vice versa when the time of the year is slow.

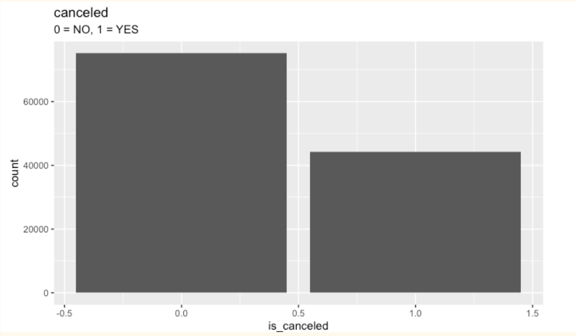
**Topic 2: How Far in Advance do People Make Bookings? Is this related to**

## Cancellations?

This topic focuses on the Lead Time of Hotel Bookings as well as determining if there is a relationship between Lead Time and Cancellation Rates. We chose to study this topic because it can help the Hotel staff better prepare and anticipate Hotel bookings.

### 1. Visualizing Lead Time & Cancellation




The distribution of Lead Time is Right Skewed, meaning that most guests typically make reservations closer to their expected arrival date.

Guests typically book hotel rooms 104 days prior to their expected arrival date.

We can see that about half of the Bookings are Cancelled, related to not being cancelled.

Summary statistics tell us that 37% of all bookings are cancelled.

### 2. Is Lead Time Related to Cancellations?

*Population Correlation:*

- The correlation between Lead Time and Cancellations is 0.29.
- The two variables are 29% correlated.

*Now, let's split the Population data into two groups, City Hotel and Resort Hotel.*

- For the City Hotel:
  - The Correlation between Lead Time and Cancellation is 0.31
- For the Resort Hotel:
  - The Correlation between Lead Time and Cancellation is 0.23

Thus, we can say that on average, guests staying at the City Hotel are more inclined to cancel.

| hotel <chr> | entries <int> | corr <dbl> |
|---|---|---|
| City Hotel | 79330 | 0.3092419 |
| Resort Hotel | 40060 | 0.2294438 |

What is the Percent of Cancellation for both groups?

| hotel <chr> | entries <int> | avg_lead <dbl> | avg_can <dbl> | percent_cancell <dbl> |
|---|---|---|---|---|
| City Hotel | 79330 | 109.73572 | 0.4172696 | 41.73 |
| Resort Hotel | 40060 | 92.67569 | 0.2776335 | 27.76 |

ANSWER: City Hotels have a higher cancellation rate and booking Lead Time than Resort Hotels.

## 3. Bootstrapping

We can use Bootstrapping to:

- Create sample datasets and generate random replicates.
- Get a more precise estimate for Lead Time and Cancellation Rates

How can we do this?

1. Split the data into two datasets: City Hotel and Resort Hotel.
2. Bootstrap 50 samples of size 1000.
3. Get sample statistics.

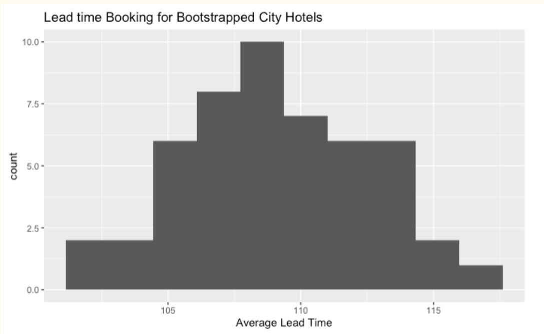
| replicate <int> | week_mean <dbl> | day_month_mean <dbl> | avg_lead <dbl> | corr <dbl> | avg_can <dbl> | percent_cancell <dbl> |
|---|---|---|---|---|---|---|
| 1 | 26.741 | 15.645 | 99.287 | 0.2081555 | 0.264 | 26.4 |
| 2 | 27.813 | 15.926 | 91.536 | 0.2440234 | 0.283 | 28.3 |
| 3 | 26.862 | 15.816 | 94.321 | 0.2133312 | 0.298 | 29.8 |
| 4 | 27.849 | 15.829 | 90.315 | 0.2385710 | 0.275 | 27.5 |
| 5 | 26.641 | 15.702 | 93.093 | 0.2255115 | 0.277 | 27.7 |
| 6 | 27.440 | 15.677 | 93.205 | 0.1747793 | 0.286 | 28.6 |
| 7 | 26.437 | 15.498 | 92.245 | 0.1927994 | 0.273 | 27.3 |
| 8 | 26.759 | 15.864 | 90.719 | 0.2175132 | 0.271 | 27.1 |
| 9 | 27.096 | 15.506 | 90.257 | 0.2512578 | 0.310 | 31.0 |
| 10 | 26.706 | 15.924 | 94.193 | 0.2186166 | 0.279 | 27.9 |

1–10 of 50 rows Previous 1 2 3 4 5 Next

| replicate <int> | week_mean <dbl> | day_month_mean <dbl> | avg_lead <dbl> | corr <dbl> | avg_can <dbl> | percent_cancell <dbl> |
|---|---|---|---|---|---|---|
| 1 | 26.991 | 15.767 | 109.188 | 0.3356216 | 0.401 | 40.1 |
| 2 | 27.230 | 15.977 | 105.065 | 0.3780940 | 0.397 | 39.7 |
| 3 | 26.513 | 15.889 | 112.151 | 0.2640341 | 0.400 | 40.0 |
| 4 | 27.600 | 15.181 | 113.696 | 0.3012203 | 0.418 | 41.8 |
| 5 | 27.486 | 15.710 | 106.785 | 0.3605139 | 0.383 | 38.3 |
| 6 | 26.367 | 15.809 | 109.979 | 0.3326268 | 0.431 | 43.1 |
| 7 | 26.809 | 15.844 | 107.085 | 0.3768247 | 0.409 | 40.9 |
| 8 | 27.009 | 15.665 | 105.765 | 0.2501399 | 0.445 | 44.5 |
| 9 | 26.761 | 15.573 | 116.444 | 0.2818956 | 0.447 | 44.7 |
| 10 | 27.157 | 15.991 | 114.986 | 0.2762142 | 0.447 | 44.7 |

1–10 of 50 rows Previous 1 2 3 4 5 Next

4. Visualize sample statistics.




5. Compare and Interpret our findings.

From the original data, we found:
- **Lead Time:**
  - City Hotel = 110 days
  - Resort Hotel = 93 days
- **Correlation:**
  - City Hotel = 0.31
  - Resort Hotel = 0.23
- **Cancellation Rate:**
  - City Hotel = 42%
  - Resort Hotel = 28%

From the Bootstrapped data, we found:
- **Lead Time:**
  - City Hotel = 110 days
  - Resort Hotel = 93 days
- **Correlation:**
  - City Hotel = 0.31
  - Resort Hotel = 0.23
- **Cancellation Rate:**
  - City Hotel = 41.4%
  - Resort Hotel = 28.1%

*In Conclusion*, We can confidently say that guests typically make bookings about 100 days before their arrival and about 35% of these bookings result in a cancellation.

## <u>Topic 3: Can we Make any Predictions?</u>

This topic uses Regression Analysis to make predictions about certain scenarios within the Hotel Booking Data. We do this because:
- Cancellations can lead to loss of revenue.
- It's common for room changes.

### 1. Predicting Cancellations

*<u>Regression Table & Equation:</u>*

| term <chr> | estimate <dbl> | std_error <dbl> | statistic <dbl> | p_value <dbl> | lower_ci <dbl> | upper_ci <dbl> |
|---|---|---|---|---|---|---|
| intercept | 0.235 | 0.002 | 125.511 | 0 | 0.232 | 0.239 |
| lead_time | 0.001 | 0.000 | 104.767 | 0 | 0.001 | 0.001 |
| previous_bookings_not_canceled | -0.012 | 0.001 | -12.982 | 0 | -0.013 | -0.010 |

$$\hat{Cancel} = 0.235 + leadTime * (0.001) + previousBooking * (-0.012)$$

*<u>Significant Coefficients:</u>*

Based on our model we see the both lead time and previous cancellations coefficients are significant at a 5% significance level.

*<u>Application:</u> Can we Predict a Cancellation for a Booking Lead Time of 100 days with no previous cancellation?*

$$\hat{Cancel} = 0.235 + (100) * (0.001) + (0) * (-0.012) = 0.335$$

There is a predicted cancellation rate of 33.5%

### 2. Predicting Arrival Dates

*<u>Regression Table & Equation:</u>*

| term | estimate | std_error | statistic | p_value | lower_ci | upper_ci |
|---|---|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| intercept | 15.821 | 0.085 | 185.237 | 0.000 | 15.654 | 15.988 |
| adults | −0.030 | 0.044 | −0.682 | 0.495 | −0.116 | 0.056 |
| children | 0.322 | 0.064 | 5.047 | 0.000 | 0.197 | 0.447 |
| babies | −0.049 | 0.261 | −0.190 | 0.850 | −0.561 | 0.462 |

*Significant Coefficients:*

We can see that only children is a significant coefficient at the 5% level.

$$\hat{ArrivalDay} = 15.821 + Adults * (-0.030) + Children * (0.322) + Babies * (-0.049)$$

*Application:* *Can we Predict what day of the month a booking with 2 adults, 2 children and 1 baby will arrive on?*

$$\hat{ArrivalDay} = 15.821 + 2 * (-0.030) + 2 * (0.322) + 1 * (-0.049) = 16.356$$

This booking is predicted to arrive on the 16th day of the month.

### 3. Predicting the Number of Special Requests

*Regression Table & Equation:*

| term | estimate | std_error | statistic | p_value | lower_ci | upper_ci |
|---|---|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| intercept | 0.248 | 0.008 | 32.603 | 0 | 0.233 | 0.263 |
| adults | 0.163 | 0.004 | 41.666 | 0 | 0.155 | 0.170 |
| children | 0.151 | 0.006 | 26.597 | 0 | 0.140 | 0.162 |
| babies | 0.764 | 0.023 | 32.932 | 0 | 0.719 | 0.810 |

$$Special\hat{Request} = .248 + 2 * Adults * (.163) + Kids * (.151) + Babies * (.764)$$

*Significant Coefficients:*

Using the outputted coefficients we see that adults, children, and babies are all significant at a 5% level.

*Application:* *Can we predict the number of special requests with 2 adults, 1 child and 2 two babies?*

$$2.25 \approx .248 + 2 * (.163) + 1 * (.151) + 2 * (.764)$$

2.25 Special Requests are predicted with families with one child and 2 babies.

## Conclusion

From our analysis we were able to find that, the busier time of the year the more expensive the night will be for each hotel and vice versa when it comes to the slowest time of the year. Guests typically make bookings about 100 days before their arrival and about 35% of these bookings result in a cancellation. We were able to accurately predict the number of cancellations, arrival dates, and number of special requests using a regression model. From this analysis, hotels may have a better understanding of their guests which will allow hotels to make better accommodations and expectations for the future.

# APPENDIX

## Read in Data

```
library(tidyverse)

## ── Attaching packages
─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.3     ✓ purrr   0.3.4
## ✓ tibble  3.0.3     ✓ dplyr   1.0.2
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.0

## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──

## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

#Reading in data
hotel.data <- read_csv('hotel_bookings.csv')

##
## ── Column specification
────────────────────────────────────────────────────────
──

## cols(
##   .default = col_double(),
##   hotel = col_character(),
##   arrival_date_month = col_character(),
```

```
##   meal = col_character(),
##   country = col_character(),
##   market_segment = col_character(),
##   distribution_channel = col_character(),
##   reserved_room_type = col_character(),
##   assigned_room_type = col_character(),
##   deposit_type = col_character(),
##   agent = col_character(),
##   company = col_character(),
##   customer_type = col_character(),
##   reservation_status = col_character(),
##   reservation_status_date = col_date(format = "")
## )
## ℹ Use `spec()` for the full column specifications.
```

## Exploratory Data Analysis

First we will look at the raw data values.

```
glimpse(hotel.data)
## Rows: 119,390
## Columns: 32
## $ hotel                          <chr> "Resort Hotel", "Resort Hotel", "Resor…
## $ is_canceled                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,…
## $ lead_time                      <dbl> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75,…
```

```
## $ arrival_date_year              <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 20...
## $ arrival_date_month             <chr> "July", "July", "July", "July", "July"...
## $ arrival_date_week_number       <dbl> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27...
## $ arrival_date_day_of_month      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ stays_in_weekend_nights        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ stays_in_week_nights           <dbl> 0, 0, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4, 4, 4,...
## $ adults                         <dbl> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2,...
## $ children                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ babies                         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ meal                           <chr> "BB", "BB", "BB", "BB", "BB", "BB", "B...
## $ country                        <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "GB...
## $ market_segment                 <chr> "Direct", "Direct", "Direct", "Corpora...
## $ distribution_channel           <chr> "Direct", "Direct", "Direct", "Corpora...
## $ is_repeated_guest              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ previous_cancellations         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ previous_bookings_not_canceled <dbl> 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                <dbl> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ deposit_type                   <chr> "No Deposit", "No Deposit", "No Deposi...
## $ agent                          <chr> "NULL", "NULL", "NULL", "304", "240", ...
## $ company                        <chr> "NULL", "NULL", "NULL", "NULL", "NULL"...
## $ days_in_waiting_list           <dbl> 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      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ total_of_special_requests        <dbl> 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          <date> 2015-07-01, 2015-07-01, 2015-07-02, 2...
```

```
head(hotel.data)
```

```
## # A tibble: 6 x 32
##   hotel is_canceled lead_time arrival_date_ye... arrival_date_mo... arrival_date_we...
##   <chr>       <dbl>     <dbl>             <dbl> <chr>                        <dbl>
## 1 Reso...            0        342              2015 July                        27
## 2 Reso...            0        737              2015 July                      27
## 3 Reso...            0        7         2015 July                  27
## 4 Reso...            0        13        2015 July                      27
## 5 Reso...            0        14        2015 July                  27
## 6 Reso...            0        14        2015 July                  27
## # ... with 26 more variables: arrival_date_day_of_month <dbl>,
## #   stays_in_weekend_nights <dbl>, stays_in_week_nights <dbl>, adults <dbl>,
## #   children <dbl>, babies <dbl>, meal <chr>, country <chr>,
## #   market_segment <chr>, distribution_channel <chr>, is_repeated_guest <dbl>,
## #   previous_cancellations <dbl>, previous_bookings_not_canceled <dbl>,
## #   reserved_room_type <chr>, assigned_room_type <chr>, booking_changes <dbl>,
## #   deposit_type <chr>, agent <chr>, company <chr>, days_in_waiting_list <dbl>,
```

```
## #   customer_type <chr>, adr <dbl>, required_car_parking_spaces <dbl>,
## #   total_of_special_requests <dbl>, reservation_status <chr>,
## #   reservation_status_date <date>
```

Next we can compute the summary statistics

```
library(skimr)
skim(hotel.data)
```

*Data summary*

| | |
|---|---|
| Name | hotel.data |
| Number of rows | 119390 |
| Number of columns | 32 |
| _______________________ | |
| Column type frequency: | |
| character | 13 |
| Date | 1 |
| 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 |

**Variable type: Date**

| skim_variable | n_missing | complete_rate | min | max | median | n_unique |
|---|---|---|---|---|---|---|
| reservation_status_date | 0 | 1 | 2014-10-17 | 2017-09-14 | 2016-08-07 | 926 |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| is_canceled | 0 | 1 | 0.37 | 0.48 | 0.00 | 0.00 | 0.00 | 1 | 1 | ▇▁▁▁▅ |
| lead_time | 0 | 1 | 104.01 | 106.86 | 0.00 | 18.00 | 69.00 | 160 | 737 | ▇▃▁▁▁ |
| arrival_date_year | 0 | 1 | 2016.16 | 0.71 | 2015.00 | 2016.00 | 2016.00 | 2017 | 2017 | ▃▁▇▁▆ |
| arrival_date_week_number | 0 | 1 | 27.17 | 13.61 | 1.00 | 16.00 | 28.00 | 38 | 53 | ▅▇▇▇▅ |
| arrival_date_day_of_month | 0 | 1 | 15.80 | 8.78 | 1.00 | 8.00 | 16.00 | 23 | 31 | ▇▇▇▇▆ |
| stays_in_weekend_nights | 0 | 1 | 0.93 | 1.00 | 0.00 | 0.00 | 1.00 | 2 | 19 | ▇▁▁▁▁ |
| stays_in_week_nights | 0 | 1 | 2.50 | 1.91 | 0.00 | 1.00 | 2.00 | 3 | 50 | ▇▁▁▁▁ |
| adults | 0 | 1 | 1.86 | 0.58 | 0.00 | 2.00 | 2.00 | 2 | 55 | ▇▁▁▁▁ |
| children | 4 | 1 | 0.10 | 0.40 | 0.00 | 0.00 | 0.00 | 0 | 10 | ▇▁▁▁▁ |
| babies | 0 | 1 | 0.01 | 0.10 | 0.00 | 0.00 | 0.00 | 0 | 10 | ▇▁▁▁▁ |
| is_repeated_guest | 0 | 1 | 0.03 | 0.18 | 0.00 | 0.00 | 0.00 | 0 | 1 | ▇▁▁▁▁ |

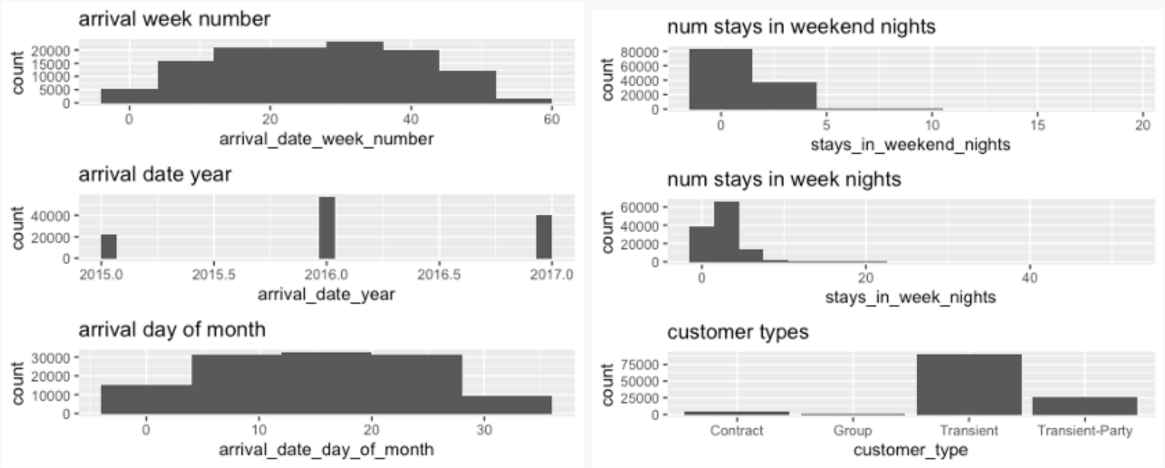
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| previous_cancellations | 0 | 1 | 0.09 | 0.84 | 0.00 | 0.00 | 0.00 | 0 | 26 | ▇▁▁▁▁ |
| previous_bookings_not_canceled | 0 | 1 | 0.14 | 1.50 | 0.00 | 0.00 | 0.00 | 0 | 72 | ▇▁▁▁▁ |
| booking_changes | 0 | 1 | 0.22 | 0.65 | 0.00 | 0.00 | 0.00 | 0 | 21 | ▇▁▁▁▁ |
| days_in_waiting_list | 0 | 1 | 2.32 | 17.59 | 0.00 | 0.00 | 0.00 | 0 | 391 | ▇▁▁▁▁ |
| adr | 0 | 1 | 101.83 | 50.54 | -6.38 | 69.29 | 94.58 | 126 | 5400 | ▇▁▁▁▁ |
| required_car_parking_spaces | 0 | 1 | 0.06 | 0.25 | 0.00 | 0.00 | 0.00 | 0 | 8 | ▇▁▁▁▁ |
| total_of_special_requests | 0 | 1 | 0.57 | 0.79 | 0.00 | 0.00 | 0.00 | 1 | 5 | ▇▁▁▁▁ |

Now we can start Data Visualization.

```
library(gridExtra)

##

## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':

##

##      combine

# visualizing week number

a = ggplot(hotel.data, aes(x= arrival_date_week_number))+

 geom_histogram(binwidth = 8)+
```

```
 labs(title = "arrival week number")


# arrival date year
b = ggplot(hotel.data, aes(x= arrival_date_year))+
 geom_histogram(binwidth=)+
 labs(title = "arrival date year")


# arrival day of month
c = ggplot(hotel.data, aes(x= arrival_date_day_of_month))+
 geom_histogram(binwidth = 8)+
 labs(title = "arrival day of month")


# stay in weekend nights
d = ggplot(hotel.data, aes(x= stays_in_weekend_nights))+
 geom_histogram(binwidth = 3)+
 labs(title = "num stays in weekend nights")


# stay in week nights
e = ggplot(hotel.data, aes(x= stays_in_week_nights))+
 geom_histogram(binwidth = 3)+
 labs(title = "num stays in week nights")


# customer type
```

```
f = ggplot(hotel.data, aes(x= customer_type))+
  geom_bar()+
  labs(title = "customer types")


grid.arrange(a, b, c, ncol=1, nrow=3)
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
grid.arrange(d, e, f, ncol=1, nrow=3)
```




## Question: When is the best time of year to book a hotel room?

```
# years 2015-2017
# months july 2015-august 2017
stats<- hotel.data%>%
  summarize(
       week_mean = mean(arrival_date_week_number),
       day_month_mean = mean(arrival_date_day_of_month)
  )
```

```
stats
## # A tibble: 1 x 2
##   week_mean day_month_mean
##       <dbl>          <dbl>
## 1      27.2           15.8
most_freq<-hotel.data%>%
 group_by(arrival_date_month)%>%
 summarize(
         count=n()
 )%>%
 arrange(desc(count))%>%
 head(1)
## `summarise()` ungrouping output (override with `.groups` argument)
most_freq
## # A tibble: 1 x 2
##   arrival_date_month count
##   <chr>              <int>
## 1 August             13877
least_freq<-hotel.data%>%
 group_by(arrival_date_month)%>%
 summarize(
         count=n()
 )%>%
```

```
arrange(count)
## `summarise()` ungrouping output (override with `.groups` argument)
least_freq
## # A tibble: 12 x 2
##    arrival_date_month count
##    <chr>              <int>
##  1 January             5929
##  2 December            6780
##  3 November            6794
##  4 February            8068
##  5 March               9794
##  6 September          10508
##  7 June               10939
##  8 April              11089
##  9 October            11160
## 10 May                11791
## 11 July               12661
## 12 August             13877
```

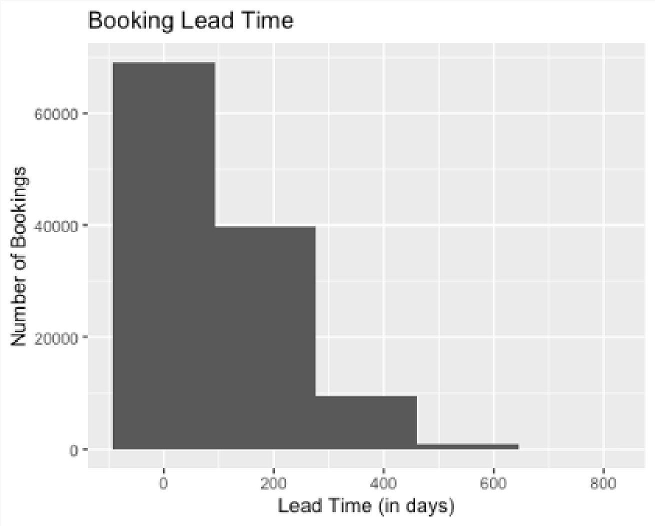
Guests typically arrive on the 15th on the month. Guests typically arrive the 27th week of the year. Guests typically arrive in August.

Least busy: January Most busy: August

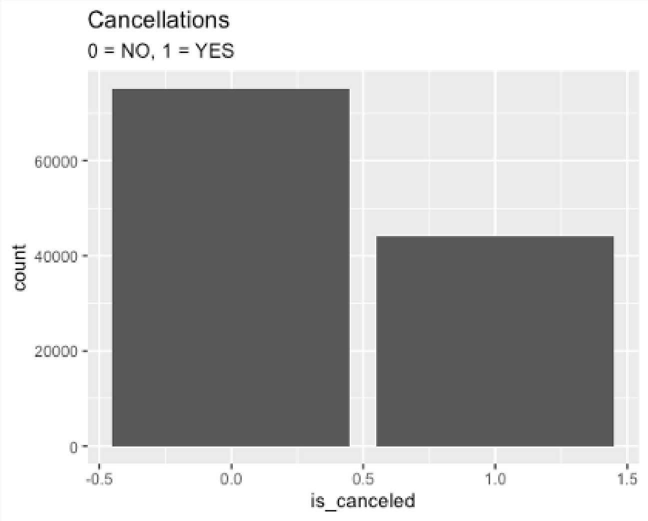
## Question: How far in advance do people make bookings? Are they more inclined to cancel?

Lead time represents the number of days that elapsed between the entering date of the booking into the PMS and the arrival date.

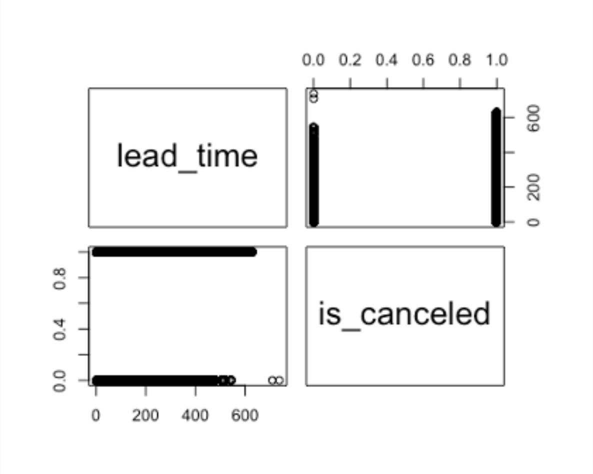
```
attach(hotel.data)


# visualizing lead time FOR ALL
ggplot(hotel.data, aes(x= lead_time))+
  geom_histogram(bins=5)+
  labs(title = "Booking Lead Time", x = "Lead Time (in days)", y = "Number of Bookings")
```




```
# visualizing cancellation rate
ggplot(hotel.data, aes(x= is_canceled))+
  geom_bar(bins=5)+
  labs(title = "Cancellations", subtitle = '0 = NO, 1 = YES')
## Warning: Ignoring unknown parameters: bins
```




```
#average lead time

cat("the average lead time is: ", mean(lead_time))

## the average lead time is:  104.0114

# Is the higher lead time associated with higher cancellation?

cat("\nthe correlation btw lead time and cancellation is: ",cor(lead_time, is_canceled))

##

## the correlation btw lead time and cancellation is:  0.2931234

cancellation<-hotel.data%>%

  select(lead_time, is_canceled)

pairs(cancellation)
```




```
mean(is_canceled)

## [1] 0.3704163
```

Guests typically book hotels 104 days prior to their arrival.

Lead time and Cancellations have a positive correlation of 0.29

## Seperating the data into City hotels and Resport hotels

```
seperate_hotels <- hotel.data%>%
 group_by(hotel)%>%
 summarise(
        entries = n(),
        #week_mean = mean(arrival_date_week_number),
        #day_month_mean = mean(arrival_date_day_of_month),
        avg_lead = mean(lead_time),
        #corr = cor(lead_time, is_canceled),
        avg_can = mean(is_canceled)
 )%>%
```

```
 mutate(
        percent_cancell = round(avg_can*100, digits = 2)
 )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
seperate_hotels
```

```
## # A tibble: 2 x 5
##   hotel        entries avg_lead avg_can percent_cancell
##   <chr>          <int>    <dbl>   <dbl>           <dbl>
## 1 City Hotel     79330    110.    0.417            41.7
## 2 Resort Hotel   40060     92.7   0.278            27.8
```

Cancellation Rate: - City Hotels have 41.73% cancellation rate. - Resort Hotels have 27.76% cancellation rate.

Correlation: - City Hotels - cancellation and lead time is 31% correlated (positive) - Resort Hotels - cancellation and lead time is 23% correlated (positive)

## Bootstrapping

```
library(infer)
set.seed(99999)


#Size 1,000 City
city_sample <- hotel.data%>%
 filter(hotel == 'City Hotel')%>%
 rep_sample_n(size = 1000, reps = 50)
city_sample
```

```
## # A tibble: 50,000 x 33
## # Groups:   replicate [50]
##       replicate hotel is_canceled lead_time arrival_date_ye… arrival_date_mo…
##       <int> <chr>     <dbl>   <dbl>            <dbl> <chr>
##  1             1 City…          0         1        2017 July
##  2             1 City…          0         1        2017 June
##  3             1 City…          0         66       2017 March
##  4             1 City…          1         54       2016 March
##  5             1 City…          1         156              2017 April
##  6             1 City…          0         7        2016 June
##  7             1 City…          1         139              2016 July
##  8             1 City…          0      83          2017 May
##  9             1 City…          0         130              2016 September
## 10             1 City…          0         87       2016 October
## # … with 49,990 more rows, and 27 more variables:
## #   arrival_date_week_number <dbl>, arrival_date_day_of_month <dbl>,
## #   stays_in_weekend_nights <dbl>, stays_in_week_nights <dbl>, adults <dbl>,
## #   children <dbl>, babies <dbl>, meal <chr>, country <chr>,
## #   market_segment <chr>, distribution_channel <chr>, is_repeated_guest <dbl>,
## #   previous_cancellations <dbl>, previous_bookings_not_canceled <dbl>,
## #   reserved_room_type <chr>, assigned_room_type <chr>, booking_changes <dbl>,
## #   deposit_type <chr>, agent <chr>, company <chr>, days_in_waiting_list <dbl>,
## #   customer_type <chr>, adr <dbl>, required_car_parking_spaces <dbl>,
```

```
## #   total_of_special_requests <dbl>, reservation_status <chr>,
## #   reservation_status_date <date>
#Size 1,000 Resort
resort_sample <- hotel.data%>%
  filter(hotel == 'Resort Hotel')%>%
  rep_sample_n(size = 1000, reps = 50)
resort_sample
## # A tibble: 50,000 x 33
## # Groups:   replicate [50]
##       replicate hotel is_canceled lead_time arrival_date_ye… arrival_date_mo…
##           <int> <chr>       <dbl>     <dbl>            <dbl> <chr>
##  1            1 Reso…           0         0             2017 August
##  2            1 Reso…           1        86             2016 February
##  3            1 Reso…           1       232             2017 May
##  4            1 Reso…           0        22             2015 November
##  5            1 Reso…           0       294             2016 June
##  6            1 Reso…           1         0             2016 March
##  7            1 Reso…           0       154             2016 December
##  8            1 Reso…           0       147             2016 December
##  9            1 Reso…           1        38             2015 September
## 10            1 Reso…           0        37             2017 June
## # … with 49,990 more rows, and 27 more variables:
## #   arrival_date_week_number <dbl>, arrival_date_day_of_month <dbl>,
```

```
## #   stays_in_weekend_nights <dbl>, stays_in_week_nights <dbl>, adults <dbl>,
## #   children <dbl>, babies <dbl>, meal <chr>, country <chr>,
## #   market_segment <chr>, distribution_channel <chr>, is_repeated_guest <dbl>,
## #   previous_cancellations <dbl>, previous_bookings_not_canceled <dbl>,
## #   reserved_room_type <chr>, assigned_room_type <chr>, booking_changes <dbl>,
## #   deposit_type <chr>, agent <chr>, company <chr>, days_in_waiting_list <dbl>,
## #   customer_type <chr>, adr <dbl>, required_car_parking_spaces <dbl>,
## #   total_of_special_requests <dbl>, reservation_status <chr>,
## #   reservation_status_date <date>
# getting stats from city sample
city_sample_stats <- city_sample%>%
 group_by(replicate)%>%
 summarise(
      week_mean = mean(arrival_date_week_number),
      day_month_mean = mean(arrival_date_day_of_month),
      avg_lead = mean(lead_time),
      corr = cor(lead_time, is_canceled),
      avg_can = mean(is_canceled)
 )%>%
 mutate(
      percent_cancell = round(avg_can*100, digits = 2)
 )
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
city_sample_stats

## # A tibble: 50 x 7
##    replicate week_mean day_month_mean avg_lead  corr avg_can percent_cancell
##        <int>     <dbl>          <dbl>    <dbl> <dbl>   <dbl>           <dbl>
##  1         1      27.0           15.8     109. 0.336   0.401            40.1
##  2         2      27.2           16.0     105. 0.378   0.397            39.7
##  3         3      26.5           15.9     112. 0.264   0.4              40
##  4         4      27.6           15.2     114. 0.301   0.418            41.8
##  5         5      27.5           15.7     107. 0.361   0.383            38.3
##  6         6      26.4           15.8     110. 0.333   0.431            43.1
##  7         7      26.8           15.8     107. 0.377   0.409            40.9
##  8         8      27.0           15.7     106. 0.250   0.445            44.5
##  9         9      26.8           15.6     116. 0.282   0.447            44.7
## 10        10      27.2           16.0     115. 0.276   0.447            44.7
## # ... with 40 more rows

mean(city_sample_stats$avg_lead)

## [1] 109.0995

mean(city_sample_stats$corr)

## [1] 0.3098421

mean(city_sample_stats$percent_cancell)

## [1] 41.43

# getting stats from resort sample

resort_sample_stats <- resort_sample%>%
```

```
group_by(replicate)%>%
summarise(
    week_mean = mean(arrival_date_week_number),
    day_month_mean = mean(arrival_date_day_of_month),
    avg_lead = mean(lead_time),
    corr = cor(lead_time, is_canceled),
    avg_can = mean(is_canceled)
)%>%
mutate(
    percent_cancell = round(avg_can*100, digits = 2)
)
## `summarise()` ungrouping output (override with `.groups` argument)
resort_sample_stats
## # A tibble: 50 x 7
##   replicate week_mean day_month_mean avg_lead  corr avg_can percent_cancell
##       <int>     <dbl>          <dbl>    <dbl> <dbl>   <dbl>           <dbl>
## 1         1      26.7           15.6     99.3 0.208   0.264            26.4
## 2         2      27.8           15.9     91.5 0.244   0.283            28.3
## 3         3      26.9           15.8     94.3 0.213   0.298            29.8
## 4         4      27.8           15.8     90.3 0.239   0.275            27.5
## 5         5      26.6           15.7     93.1 0.226   0.277            27.7
## 6         6      27.4           15.7     93.2 0.175   0.286            28.6
## 7         7      26.4           15.5     92.2 0.193   0.273            27.3
```

```
## 8     8    26.8    15.9    90.7 0.218   0.271    27.1
## 9     9    27.1    15.5    90.3 0.251   0.31     31
## 10    10   26.7    15.9    94.2 0.219   0.279    27.9
## # ... with 40 more rows
mean(resort_sample_stats$avg_lead)
## [1] 93.17952
mean(resort_sample_stats$corr)
## [1] 0.2286904
mean(resort_sample_stats$percent_cancell)
## [1] 28.078
# plotting avg lead time for city sample
ggplot(city_sample_stats, aes(x= avg_lead))+
  geom_histogram(bins=10)+
  labs(title = "Lead time Booking for Bootstrapped City Hotels", x = "Average Lead Time")
```


```
# plotting avg lead time for resort sample
ggplot(resort_sample_stats, aes(x= avg_lead))+
```

```
geom_histogram(bins=10)+
labs(title = "Lead time Booking for Bootstrapped Resort Hotels", x = "Average Lead Time")
```


## Prediction from Samples

Q: Can we predict the arrival date for 2 adults with 2 children.

```
library(moderndive)

# Prediction using City Hotel data
city_predict <- lm(arrival_date_day_of_month ~ adults + children, data=city_sample)
get_regression_table(city_predict)
## # A tibble: 3 x 7
##   term       estimate std_error statistic p_value lower_ci upper_ci
##   <chr>      <dbl>    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>
## 1 intercept   15.9     0.146    109.      0        15.6     16.2
## 2 adults     -0.062    0.076    -0.824    0.41    -0.211    0.086
## 3 children    0.354    0.106    3.35      0.001    0.147    0.561
```

```
# Prediction using Resort Hotel data
resort_predict <- lm(arrival_date_day_of_month ~ adults + children, data=resort_sample)
get_regression_table(resort_predict)
## # A tibble: 3 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>     <dbl>    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>
## 1 intercept  15.8     0.128    123.       0       15.5     16.0
## 2 adults     0.018    0.065    0.271    0.786   -0.11     0.146
## 3 children   0.387    0.092    4.22       0       0.207    0.566
#Prediction using ALL hotels - this is on teh presentation
hotel_arrival <- lm(arrival_date_day_of_month ~ adults + children + babies, data=hotel.data)
get_regression_table(hotel_arrival)
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>     <dbl>    <dbl>     <dbl>     <dbl>   <dbl>    <dbl>
## 1 intercept  15.8     0.085    185.       0       15.7     16.0
## 2 adults     -0.03    0.044    -0.682   0.495   -0.116 0.056
## 3 children   0.322    0.064    5.05       0       0.197    0.447
## 4 babies     -0.049 0.261     -0.19     0.85    -0.561  0.462
# Prediction using Resort Hotel data
resort_predict_ad2_ch2_resort <- lm(is_canceled~lead_time+previous_bookings_not_canceled,
data=resort_sample)
get_regression_table(resort_predict_ad2_ch2_resort)
```

```
## # A tibble: 3 x 7
##   term                   estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                     <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 0.188     0.003      68.5       0    0.183    0.194
## 2 lead_time                 0.001     0          50.9       0    0.001    0.001
## 3 previous_bookings_not_…  -0.026     0.002     -12.1       0   -0.03    -0.022
# number of special requests
attach(hotel.data)
## The following objects are masked from hotel.data (pos = 5):
##
##     adr, adults, agent, arrival_date_day_of_month, arrival_date_month,
##     arrival_date_week_number, arrival_date_year, assigned_room_type,
##     babies, booking_changes, children, company, country, customer_type,
##     days_in_waiting_list, deposit_type, distribution_channel, hotel,
##     is_canceled, is_repeated_guest, lead_time, market_segment, meal,
##     previous_bookings_not_canceled, previous_cancellations,
##     required_car_parking_spaces, reservation_status,
##     reservation_status_date, reserved_room_type, stays_in_week_nights,
##     stays_in_weekend_nights, total_of_special_requests
spec_req <- lm(total_of_special_requests ~ adults + children + babies)
get_regression_table(spec_req)
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
```

```
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 intercept 0.248 0.008 32.6 0 0.233 0.263
## 2 adults 0.163 0.004 41.7 0 0.155 0.17
## 3 children 0.151 0.006 26.6 0 0.14 0.162
## 4 babies 0.764 0.023 32.9 0 0.719 0.81
```

```
# number of cancellation
hotel_cancelled<- lm(is_canceled ~ lead_time + previous_bookings_not_canceled,
data=hotel.data)
get_regression_table(hotel_cancelled)
```

```
## # A tibble: 3 x 7
## term estimate std_error statistic p_value lower_ci upper_ci
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 intercept 0.235 0.002 126. 0 0.232 0.239
## 2 lead_time 0.001 0 105. 0 0.001 0.001
## 3 previous_bookings_not_… -0.012 0.001 -13.0 0 -0.013 -0.01
```

City Hotel: Prediction date of 15.917 + 2(-0.062) + 2(0.354) = 16.5 - Arrival will be on the 17th

Resort Hotel: Prediction date of 15.756 + 2(0.018) + 2(0.387) = 16.566 = 17 - Arrival will be on the 17th

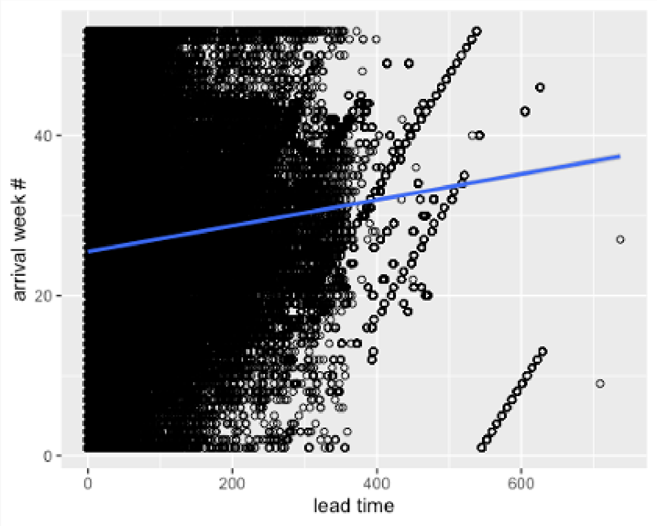
## Regression : lead time vs. arrival week number

Using regression to form a line that tells us the relationship between how far in advance guests book their stays and their arrival week number.

```
ggplot(hotel.data, aes(y = arrival_date_week_number, x = lead_time))+
 geom_point(shape = 1) +
```

```
geom_smooth(method = lm) +
#geom_jitter(shape = 1)+
labs(y='arrival week #',x='lead time')
```

## `geom_smooth()` using formula 'y ~ x'




Findings: People that arrive later in the year, often book further in advance.