Hotel booking

# Read the data

library(tidyverse)

## -- Attaching packages ----------------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.3  
## v tibble 3.0.0 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts -------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

theme\_set(theme\_light())  
  
getwd()

## [1] "C:/Users/Lenovo/Documents/Kaggle/Case Study"

hotels <- readr::read\_csv('hotel\_bookings.csv')

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## hotel = col\_character(),  
## `No Response` = col\_logical(),  
## 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\_character()  
## )

## See spec(...) for full column specifications.

glimpse(hotels)

## Rows: 119,390  
## Columns: 33  
## $ hotel <chr> "Resort Hotel", "Resort Hotel", "Res...  
## $ is\_canceled <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, ...  
## $ `No Response` <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, ...  
## $ lead\_time <dbl> 342, 737, 7, 13, 14, 14, 0, 9, 85, 7...  
## $ arrival\_date\_year <dbl> 2015, 2015, 2015, 2015, 2015, 2015, ...  
## $ arrival\_date\_month <chr> "July", "July", "July", "July", "Jul...  
## $ 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, ...  
## $ babies <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ meal <chr> "BB", "BB", "BB", "BB", "BB", "BB", ...  
## $ country <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "...  
## $ market\_segment <chr> "Direct", "Direct", "Direct", "Corpo...  
## $ distribution\_channel <chr> "Direct", "Direct", "Direct", "Corpo...  
## $ is\_repeated\_guest <dbl> 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, ...  
## $ previous\_bookings\_not\_canceled <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ reserved\_room\_type <chr> "C", "C", "A", "A", "A", "A", "C", "...  
## $ assigned\_room\_type <chr> "C", "C", "C", "A", "A", "A", "C", "...  
## $ booking\_changes <dbl> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...  
## $ deposit\_type <chr> "No Deposit", "No Deposit", "No Depo...  
## $ agent <chr> "NULL", "NULL", "NULL", "304", "240"...  
## $ company <chr> "NULL", "NULL", "NULL", "NULL", "NUL...  
## $ days\_in\_waiting\_list <dbl> 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....  
## $ required\_car\_parking\_spaces <dbl> 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, ...  
## $ reservation\_status <chr> "Check-Out", "Check-Out", "Check-Out...  
## $ reservation\_status\_date <chr> "01-07-2015", "01-07-2015", "02-07-2...

hotels %>% count(is\_canceled)

## # A tibble: 3 x 2  
## is\_canceled n  
## <dbl> <int>  
## 1 0 75166  
## 2 1 44046  
## 3 NA 178

hotels %>% select(children,babies)

## # A tibble: 119,390 x 2  
## children babies  
## <dbl> <dbl>  
## 1 0 0  
## 2 0 0  
## 3 0 0  
## 4 0 0  
## 5 0 0  
## 6 0 0  
## 7 0 0  
## 8 0 0  
## 9 0 0  
## 10 0 0  
## # ... with 119,380 more rows

#Explore the data

The hotel stay had more reservation without children than with children

hotels %>%   
 filter(is\_canceled == 0) %>%   
 mutate(children = case\_when(children+babies > 0 ~ "children", TRUE ~ "none")) %>% count(children)

## # A tibble: 2 x 2  
## children n  
## <chr> <int>  
## 1 children 6073  
## 2 none 69093

Explore the hotel stay with respect to car parking Cleaning the hotel column

hotel\_stays<-hotels %>%   
 filter(is\_canceled == 0) %>%   
 mutate(children = case\_when(children+babies > 0 ~ "children", TRUE ~ "none"),  
 hotel = recode(hotel,'Citi Hotel'='City Hotel','Reosrt Hotel'='Resort Hotel'),  
 required\_car\_parking\_spaces = case\_when(required\_car\_parking\_spaces>0~'parking', TRUE~ "none")) %>% select(-is\_canceled, -reservation\_status,-babies)  
  
hotel\_stays

## # A tibble: 75,166 x 30  
## hotel `No Response` lead\_time arrival\_date\_ye~ arrival\_date\_mo~  
## <chr> <lgl> <dbl> <dbl> <chr>   
## 1 Reso~ NA 342 2015 July   
## 2 Reso~ NA 737 2015 July   
## 3 Reso~ NA 7 2015 July   
## 4 Reso~ NA 13 2015 July   
## 5 Reso~ NA 14 2015 July   
## 6 Reso~ NA 14 2015 July   
## 7 Reso~ NA 0 2015 July   
## 8 Reso~ NA 9 2015 July   
## 9 Reso~ NA 35 2015 July   
## 10 Reso~ NA 68 2015 July   
## # ... with 75,156 more rows, and 25 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 <chr>, 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 <chr>,  
## # total\_of\_special\_requests <dbl>, reservation\_status\_date <chr>

hotel\_stays %>% count(children)

## # A tibble: 2 x 2  
## children n  
## <chr> <int>  
## 1 children 6073  
## 2 none 69093

hotel\_stays %>% count(hotel)

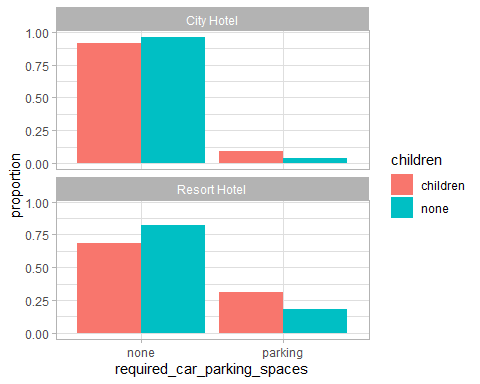
## # A tibble: 2 x 2  
## hotel n  
## <chr> <int>  
## 1 City Hotel 46228  
## 2 Resort Hotel 28938

hotel\_stays %>%   
 mutate(arrival\_date\_month = factor(arrival\_date\_month,levels = month.name)) %>%   
 count(hotel,arrival\_date\_month,children) %>%   
 group\_by(hotel,children) %>%   
 mutate(proportion = n/sum(n)) %>%   
 ggplot(aes(arrival\_date\_month,proportion,fill=children))+  
 geom\_col(position="dodge")+  
 facet\_wrap(~hotel,nrow=2)



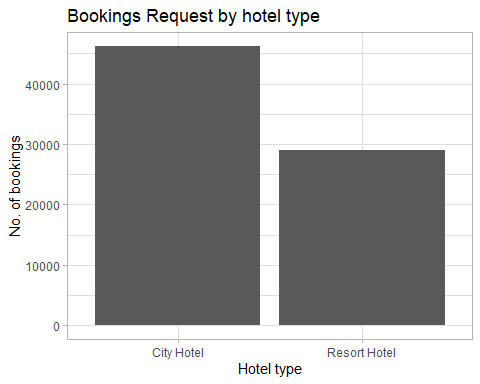
Lets compare the car parking with with children and without children

hotel\_stays %>%   
 count(hotel,required\_car\_parking\_spaces,children) %>%   
 group\_by(hotel,children) %>%   
 mutate(proportion = n/sum(n)) %>%   
 ggplot(aes(required\_car\_parking\_spaces,proportion,fill=children))+  
 geom\_col(position="dodge")+  
 facet\_wrap(~hotel,nrow=2)



#Exploring booking reuest by hotel type

ggplot(data = hotel\_stays,aes(x=hotel))+  
 geom\_bar(stat = "count")+  
 labs(title = "Bookings Request by hotel type",  
 x = "Hotel type",  
 y = "No. of bookings")+  
 scale\_color\_brewer(palette = "Set2")



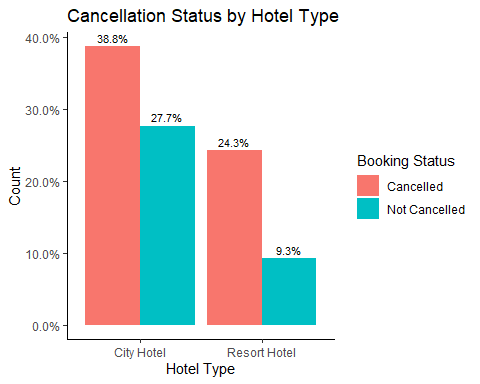
# City hotel has more booking requests than Resort hotel

# Check the dstribution of hotel type for cancellation

hotel\_stays\_overall<-hotels %>%   
 filter(is\_canceled %in% c(1,0)) %>%   
 mutate(children = case\_when(children+babies > 0 ~ "children", TRUE ~ "none"),  
 hotel = recode(hotel,'Citi Hotel'='City Hotel','Reosrt Hotel'='Resort Hotel'),  
 required\_car\_parking\_spaces = case\_when(required\_car\_parking\_spaces>0~'parking', TRUE~ "none")) %>% select(-reservation\_status,-babies)  
  
  
hotel\_stays\_overall %>%   
 count(is\_canceled)

## # A tibble: 2 x 2  
## is\_canceled n  
## <dbl> <int>  
## 1 0 75166  
## 2 1 44046

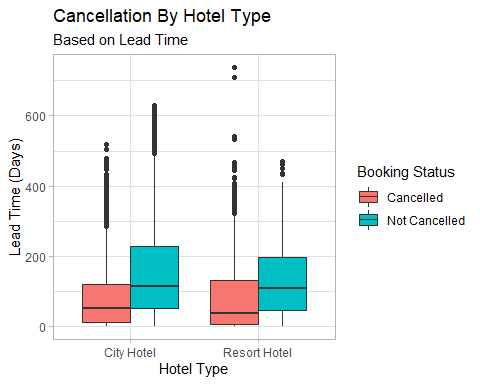
ggplot(data = hotel\_stays\_overall,  
 aes(  
 x = hotel,  
 y = prop.table(stat(count)),  
 fill = factor(is\_canceled),  
 label = scales::percent(prop.table(stat(count)))  
 )) +  
   
 geom\_bar(position = position\_dodge()) +  
 geom\_text(  
 stat = "count",  
 position = position\_dodge(.9),  
 vjust = -0.5,  
 size = 3  
 ) +  
 scale\_y\_continuous(labels = scales::percent) +  
 labs(title = "Cancellation Status by Hotel Type",  
 x = "Hotel Type",  
 y = "Count") +  
 theme\_classic() +  
 scale\_fill\_discrete(  
 name = "Booking Status",  
 breaks = c("0", "1"),  
 labels = c("Cancelled", "Not Cancelled")  
 )



Obervation : Out of the bookings that were made more than 66% were done for City Hotel, and around 34% were for Resort Hotel. However both the hotel types that proportion of cancellation more than the confirmed status.

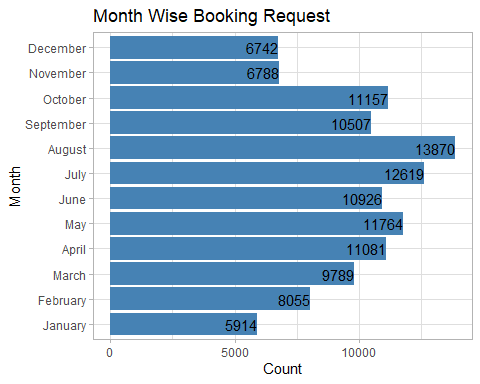
#Cancellation ratio by hotel type based on the lead time. Lead time is the gap between Booking made and actual date check in

ggplot(data = hotel\_stays\_overall, aes(  
 x = hotel,  
 y = lead\_time,  
 fill = factor(is\_canceled)  
)) +  
 geom\_boxplot(position = position\_dodge()) +  
 labs(  
 title = "Cancellation By Hotel Type",  
 subtitle = "Based on Lead Time",  
 x = "Hotel Type",  
 y = "Lead Time (Days)"  
 ) +  
 scale\_fill\_discrete(  
 name = "Booking Status",  
 breaks = c("0", "1"),  
 labels = c("Cancelled", "Not Cancelled")  
 )



#Explore month favorable for hotels when they cane expect maximum demand

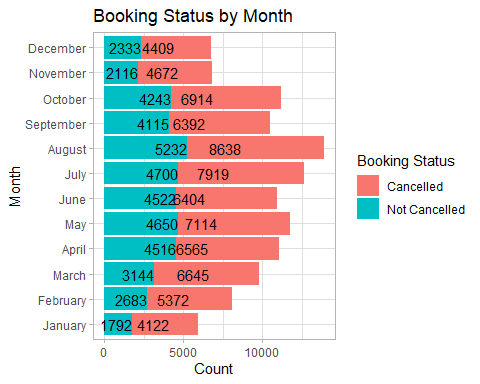
hotel\_stays\_overall$arrival\_date\_month <-  
 factor(hotel\_stays\_overall$arrival\_date\_month, levels = month.name)  
# Visualize Hotel traffic on Monthly basis  
ggplot(data = hotel\_stays\_overall, aes(x = arrival\_date\_month)) +  
 geom\_bar(fill = "steelblue") +  
 geom\_text(stat = "count", aes(label = ..count..), hjust = 1) +  
 coord\_flip() + labs(title = "Month Wise Booking Request",  
 x = "Month",  
 y = "Count")



Obervation: Month wise booking analysis depicts that more number of booking request is in July ,August respectively.

# Explore the booking made in not confirmed statuses month wise.

ggplot(hotel\_stays\_overall, aes(arrival\_date\_month, fill = factor(is\_canceled))) +  
 geom\_bar() + geom\_text(stat = "count", aes(label = ..count..), hjust = 1) +  
 coord\_flip() + scale\_fill\_discrete(  
 name = "Booking Status",  
 breaks = c("0", "1"),  
 label = c("Cancelled", "Not Cancelled")  
 ) +  
 labs(title = "Booking Status by Month",  
 x = "Month",  
 y = "Count")



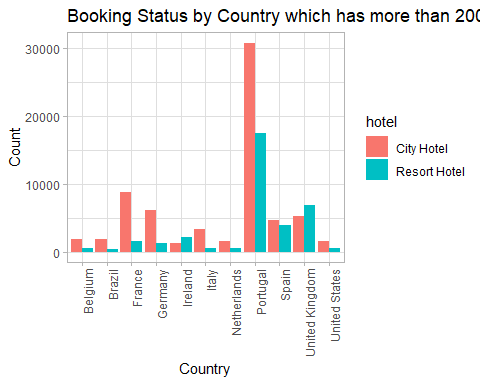
library(countrycode)  
hotel\_stays\_overall$country\_name <- countrycode(hotel\_stays\_overall$country,   
 origin = "iso3c",  
 destination = "country.name")

## Warning in countrycode(hotel\_stays\_overall$country, origin = "iso3c", destination = "country.name"): Some values were not matched unambiguously: CN, NULL, TMP

hotel\_stays\_overall %>% count(country\_name)

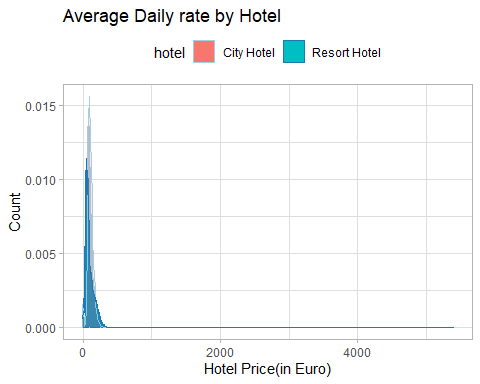
## # A tibble: 176 x 2  
## country\_name n  
## <chr> <int>  
## 1 Albania 12  
## 2 Algeria 102  
## 3 American Samoa 1  
## 4 Andorra 7  
## 5 Angola 362  
## 6 Anguilla 1  
## 7 Antarctica 2  
## 8 Argentina 214  
## 9 Armenia 8  
## 10 Aruba 2  
## # ... with 166 more rows

# Traveller by Country per hotel wise  
hotel\_stays\_overall %>%   
 group\_by(country) %>%   
 filter(n()>2000) %>%   
ggplot(aes(country\_name, fill = hotel)) +   
 geom\_bar(stat = "count", position = position\_dodge()) +   
 labs(title = "Booking Status by Country which has more than 2000 reservation",  
 x = "Country",  
 y = "Count") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 panel.background = element\_blank())



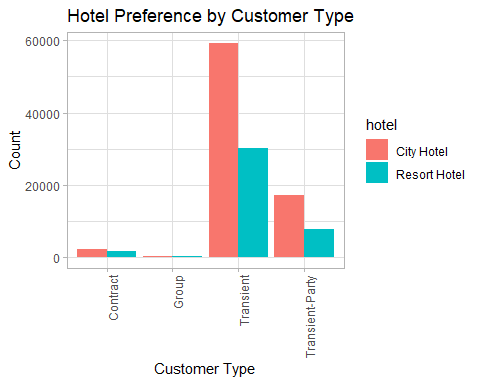
#Average daily rate by hotel type

hotel\_stays\_overall %>%   
ggplot(aes(x = adr, fill = hotel, color = hotel)) +   
 geom\_histogram(aes(y = ..density..), position = position\_dodge(), binwidth = 10 ) +  
 geom\_density(alpha = 0.2) +   
 labs(title = "Average Daily rate by Hotel",  
 x = "Hotel Price(in Euro)",  
 y = "Count") + scale\_color\_brewer(palette = "Paired") +   
 theme(legend.position = "top")



# Hotel prefernce ny customer type

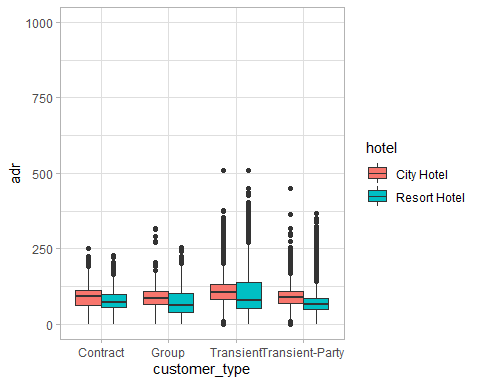
hotel\_stays\_overall %>%   
ggplot(aes(customer\_type, fill = hotel)) +   
 geom\_bar(stat = "count", position = position\_dodge()) +   
 labs(title = "Hotel Preference by Customer Type",  
 x = "Customer Type",  
 y = "Count") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 panel.background = element\_blank())



#Does the hotel charge differently for different customer type

hotel\_stays\_overall %>%   
ggplot(aes(x = customer\_type, y = adr, fill = hotel)) +   
 geom\_boxplot(position = position\_dodge()) +   
 ylim(0,1000)

## Warning: Removed 2 rows containing non-finite values (stat\_boxplot).



labs(title = "Price Charged by Hotel Type",  
 subtitle = "for Customer Type",  
 x = "Customer Type",  
 y = "Price per night(in Euro)")

## $x  
## [1] "Customer Type"  
##   
## $y  
## [1] "Price per night(in Euro)"  
##   
## $title  
## [1] "Price Charged by Hotel Type"  
##   
## $subtitle  
## [1] "for Customer Type"  
##   
## attr(,"class")  
## [1] "labels"