IST687 Project Master Code

Megha Banerjee, Kevin Harmer

12/9/2021

1. Collective Libraries Used for Analysis

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

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

```
library(ggplot2)
library(stats)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':
##
##     some
```

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(kernlab)
```

```
##
## Attaching package: 'kernlab'

## The following object is masked from 'package:purrr':
##
##     cross

## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```
library(e1071)
library(rworldmap)
```

```
## Warning: package 'rworldmap' was built under R version 4.1.2

## Loading required package: sp

## ### Welcome to rworldmap ###

## For a short introduction type :   vignette('rworldmap')
```

```
library(arules)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

##
## Attaching package: 'arules'

## The following object is masked from 'package:kernlab':
##
##     size

## The following object is masked from 'package:car':
##
##     recode
```

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

## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(arulesViz)
library(performance)
library(DHARMa)
```

```
## This is DHARMa 0.4.4. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')
```

```
library(stats)
library(dplyr)
library(purrr)
library(tidyr)
```

2. Importing Data

```
hotels <- read_csv('https://intro-datascience.s3.us-east-2.amazonaws.com/Resort01.csv') #importing the
```

```
## Rows: 40060 Columns: 20

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): Meal, Country, MarketSegment, ReservedRoomType, AssignedRoomType, ...
## dbl (13): IsCanceled, LeadTime, StaysInWeekendNights, StaysInWeekNights, Adu...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

3. Checking for Any NA values and Cleaning

```
#View(hotels) #viewing data set in other window
str(hotels) #view data set structure
```

```
## spec_tbl_df [40,060 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ IsCanceled                : num [1:40060] 0 0 0 0 0 0 0 0 1 1 ...
##  $ LeadTime                  : num [1:40060] 342 737 7 13 14 14 0 9 85 75 ...
##  $ StaysInWeekendNights      : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
##  $ StaysInWeekNights         : num [1:40060] 0 0 1 1 2 2 2 2 3 3 ...
##  $ Adults                    : num [1:40060] 2 2 1 1 2 2 2 2 2 2 ...
##  $ Children                  : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Babies                    : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Meal                      : chr [1:40060] "BB" "BB" "BB" "BB" ...
##  $ Country                   : chr [1:40060] "PRT" "PRT" "GBR" "GBR" ...
##  $ MarketSegment             : chr [1:40060] "Direct" "Direct" "Direct" "Corporate" ...
```

```
##  $ IsRepeatedGuest            : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
##  $ PreviousCancellations      : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
##  $ PreviousBookingsNotCanceled: num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ReservedRoomType           : chr [1:40060] "C" "C" "A" "A" ...
##  $ AssignedRoomType           : chr [1:40060] "C" "C" "C" "A" ...
##  $ BookingChanges             : num [1:40060] 3 4 0 0 0 0 0 0 0 0 ...
##  $ DepositType                : chr [1:40060] "No Deposit" "No Deposit" "No Deposit" "No Deposit" ..
##  $ CustomerType               : chr [1:40060] "Transient" "Transient" "Transient" "Transient" ...
##  $ RequiredCarParkingSpaces   : num [1:40060] 0 0 0 0 0 0 0 0 0 0 ...
##  $ TotalOfSpecialRequests     : num [1:40060] 0 0 0 0 1 1 0 1 1 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   IsCanceled = col_double(),
##   ..   LeadTime = col_double(),
##   ..   StaysInWeekendNights = col_double(),
##   ..   StaysInWeekNights = col_double(),
##   ..   Adults = col_double(),
##   ..   Children = col_double(),
##   ..   Babies = col_double(),
##   ..   Meal = col_character(),
##   ..   Country = col_character(),
##   ..   MarketSegment = col_character(),
##   ..   IsRepeatedGuest = col_double(),
##   ..   PreviousCancellations = col_double(),
##   ..   PreviousBookingsNotCanceled = col_double(),
##   ..   ReservedRoomType = col_character(),
##   ..   AssignedRoomType = col_character(),
##   ..   BookingChanges = col_double(),
##   ..   DepositType = col_character(),
##   ..   CustomerType = col_character(),
##   ..   RequiredCarParkingSpaces = col_double(),
##   ..   TotalOfSpecialRequests = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(hotels) #view the first 5 inputs
```

```
## # A tibble: 6 x 20
##   IsCanceled LeadTime StaysInWeekendNig~ StaysInWeekNigh~ Adults Children Babies
##        <dbl>    <dbl>              <dbl>            <dbl>  <dbl>    <dbl>  <dbl>
## 1          0      342                  0                0      2        0      0
## 2          0      737                  0                0      2        0      0
## 3          0        7                  0                1      1        0      0
## 4          0       13                  0                1      1        0      0
## 5          0       14                  0                2      2        0      0
## 6          0       14                  0                2      2        0      0
## # ... with 13 more variables: Meal <chr>, Country <chr>, MarketSegment <chr>,
## #   IsRepeatedGuest <dbl>, PreviousCancellations <dbl>,
## #   PreviousBookingsNotCanceled <dbl>, ReservedRoomType <chr>,
## #   AssignedRoomType <chr>, BookingChanges <dbl>, DepositType <chr>,
## #   CustomerType <chr>, RequiredCarParkingSpaces <dbl>,
## #   TotalOfSpecialRequests <dbl>
```

```r
for (i in c(1:20)) {
  print(sum(hotels[,i] == 'NULL')) #checking for NULL inputs; also checked for NA values but there were
} #amount of null values in data set
```

```
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 464
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
```

```r
hotels <- hotels[!hotels$Country == 'NULL',] #removing inputs without a country input
```

```r
cancel <- hotels[hotels$IsCanceled==1,] #creating a new data set with only canceled observations
nrow(cancel) #looking at amount of data observations
```

```
## [1] 11077
```

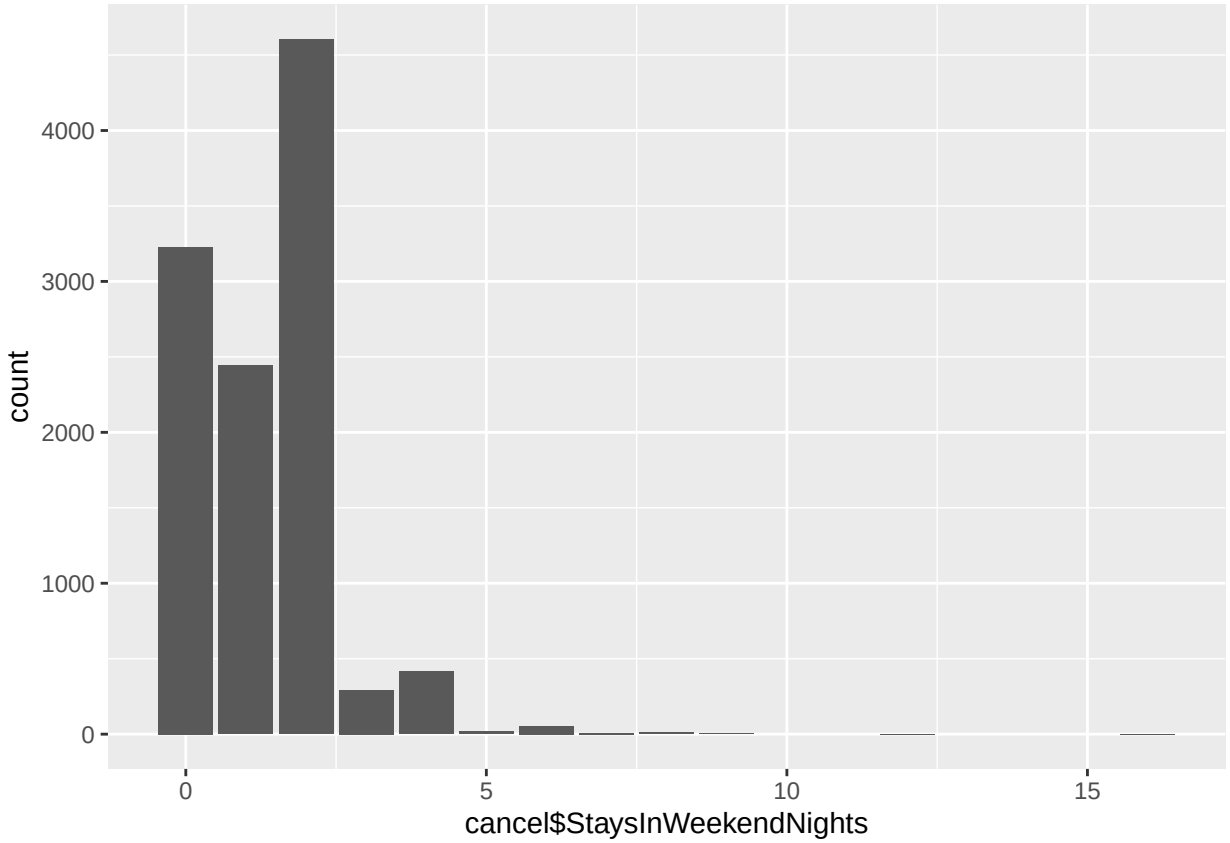
```r
stay <- hotels[hotels$IsCanceled==0,] #creating a new data set with only non-canceled observations
nrow(stay) #looking at amount of dat observations
```

```
## [1] 28519
```

4. Looking at Some Bar Plots

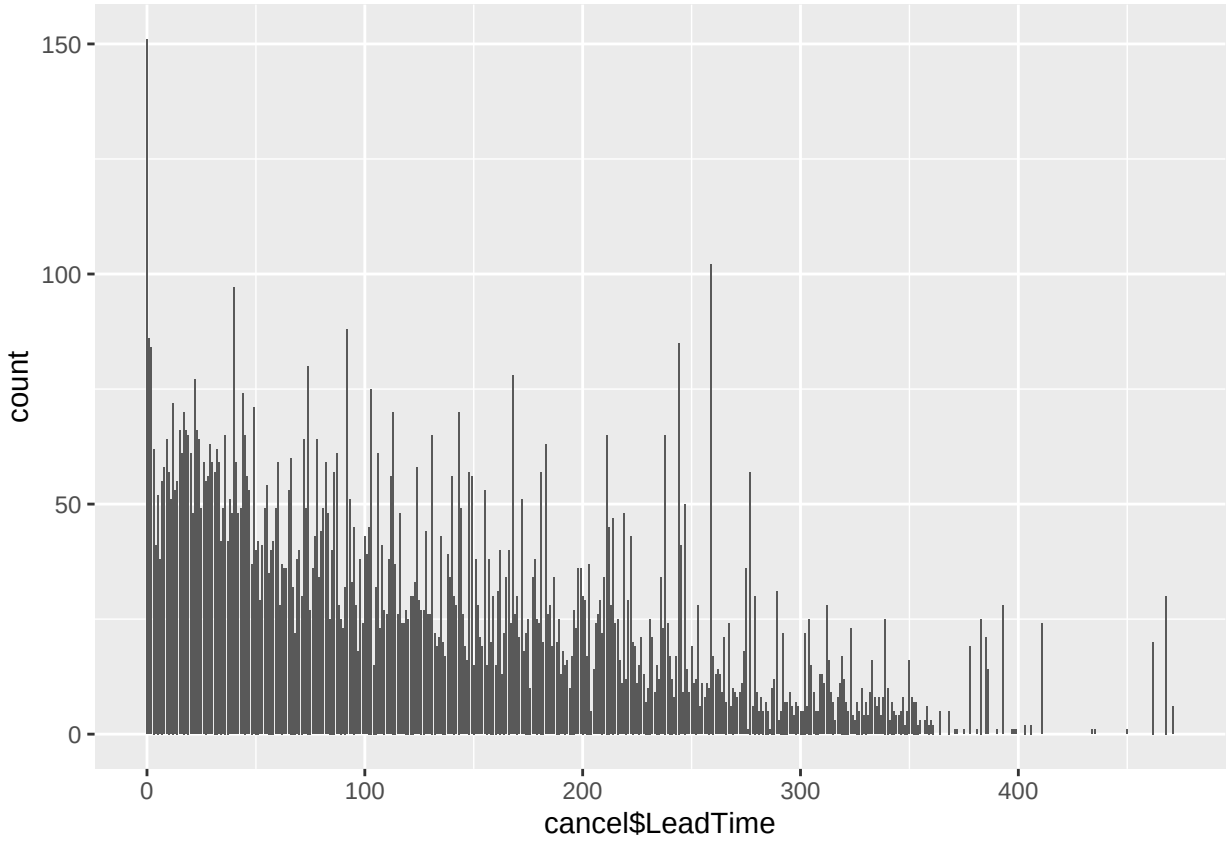
```r
# COLUMN NOT REQUIRED for analysis of cancellation, no distinct measure shows up
plotyplot <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$StaysInWeekendNights)
plotyplot
```

```
## Warning: Use of ‘cancel$StaysInWeekendNights‘ is discouraged. Use
## ‘StaysInWeekendNights‘ instead.
```



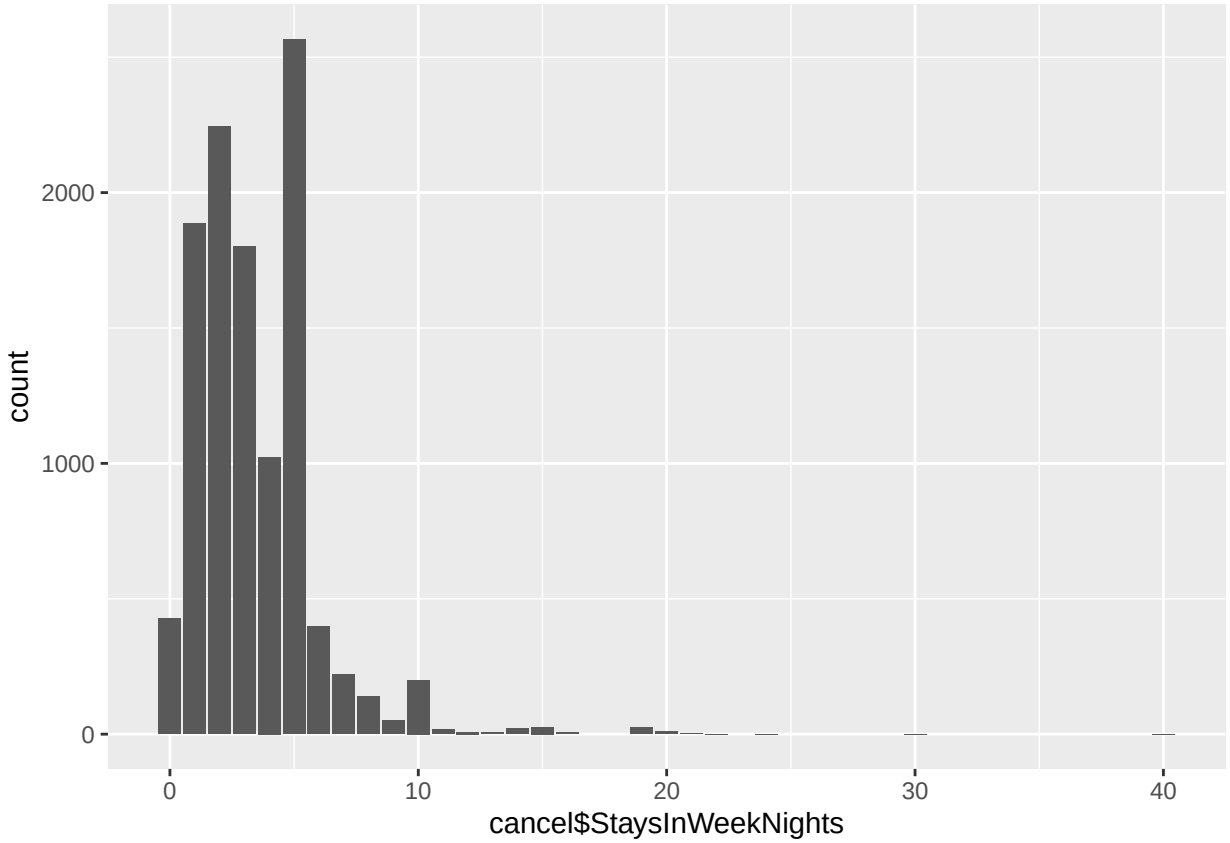

```
# COLUMN NOT REQUIRED for analysis of cancellation, no distinct measure shows up
plotyplot_2 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$LeadTime)
plotyplot_2
```

```
## Warning: Use of ‘cancel$LeadTime‘ is discouraged. Use ‘LeadTime‘ instead.
```



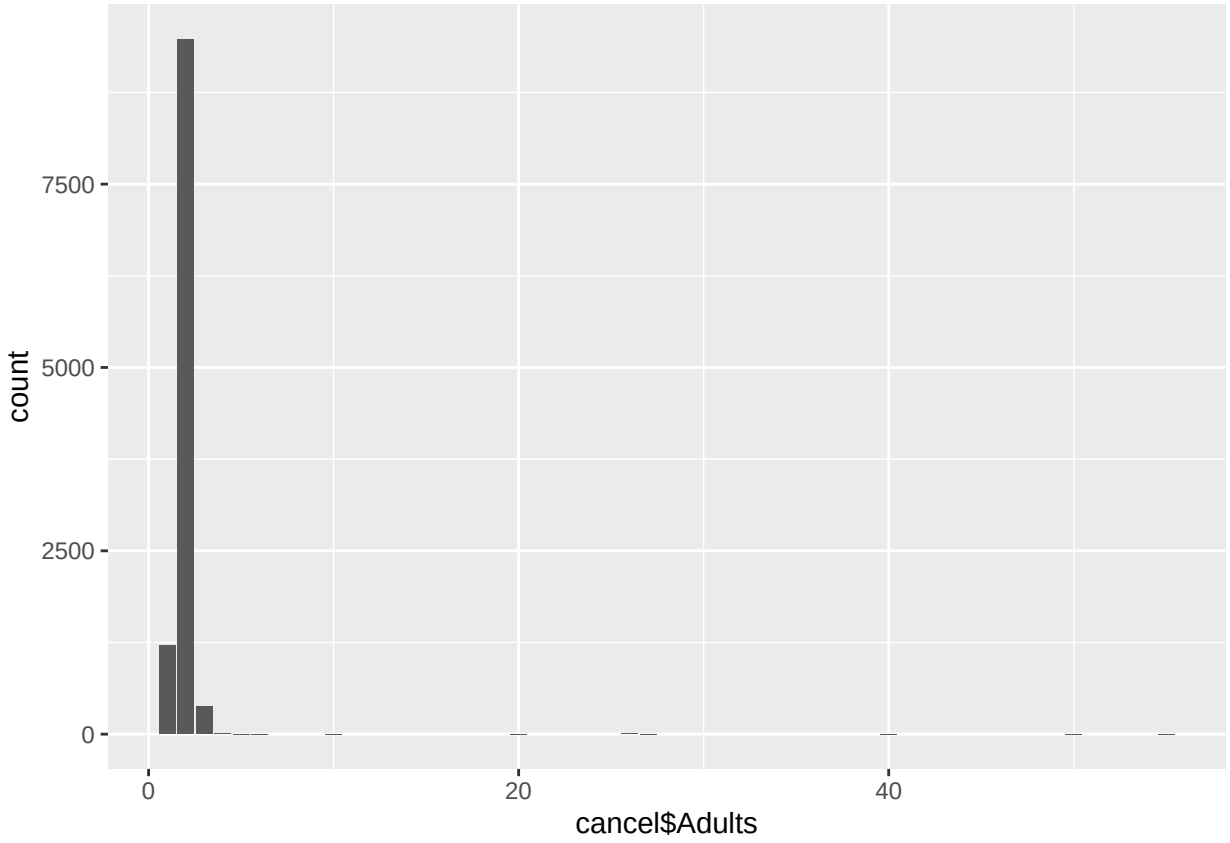

```
# COLUMN NOT REQUIRED for analysis of cancellation, no distinct measure shows up
plotyplot_3 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$StaysInWeekNights)
plotyplot_3
```

```
## Warning: Use of ‘cancel$StaysInWeekNights‘ is discouraged. Use
## ‘StaysInWeekNights‘ instead.
```



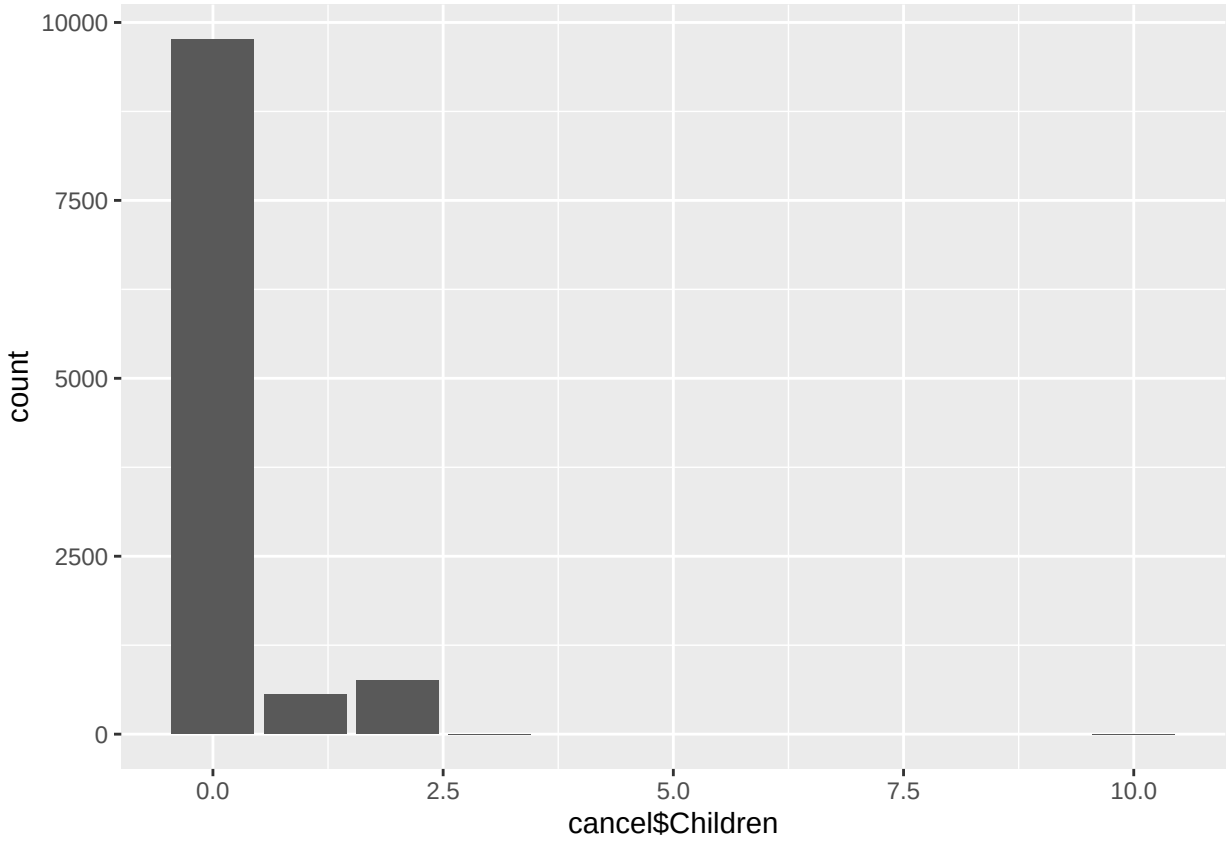

```
# bookings with 2 adults often tend to get cancelled more frequently , around 75 percent of the times i
plotyplot_4 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$Adults)
plotyplot_4
```

```
## Warning: Use of ‘cancel$Adults‘ is discouraged. Use ‘Adults‘ instead.
```



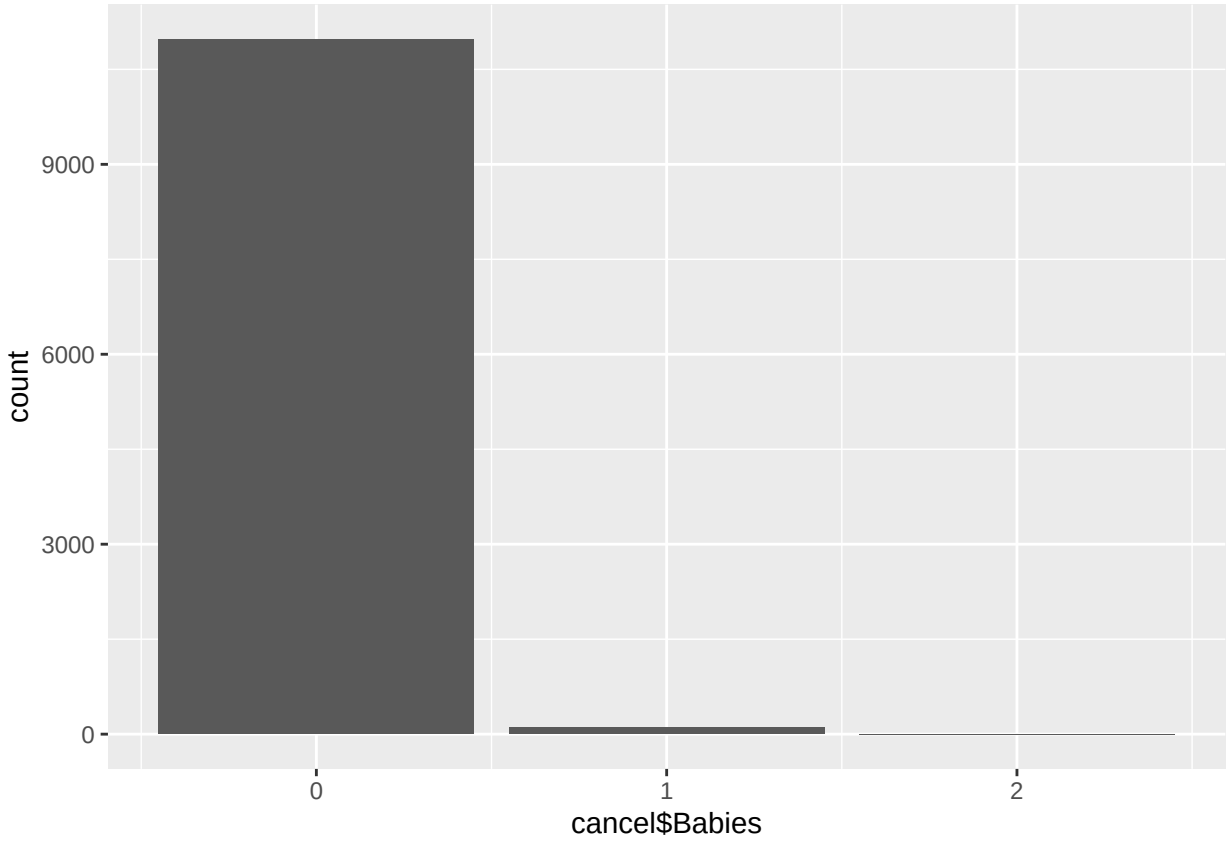

```
# Guests who have children hardly cancel but there cancel rate is higher than with babies
plotyplot_5 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$Children)
plotyplot_5
```

```
## Warning: Use of ‘cancel$Children‘ is discouraged. Use ‘Children‘ instead.
```



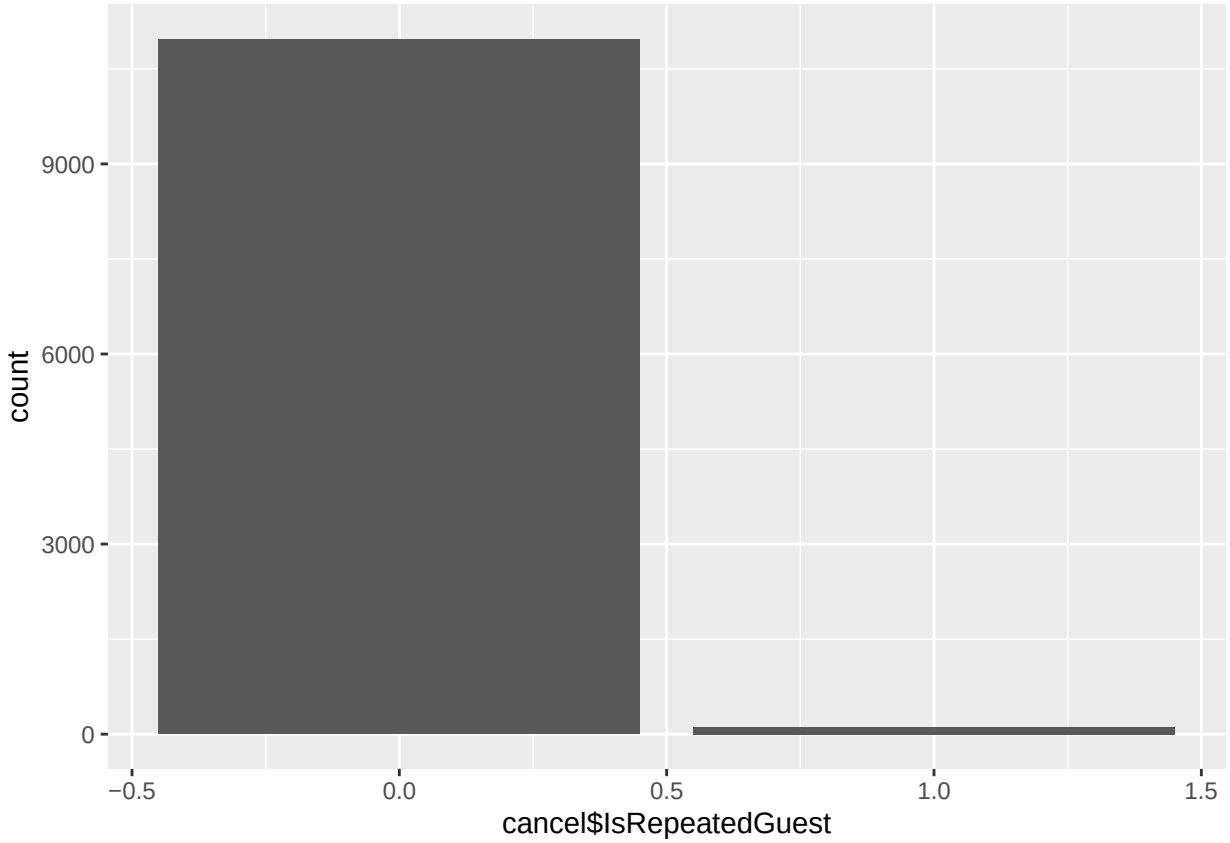

```
# Guests who have babies hardly cancel.
plotyplot_6 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$Babies)
plotyplot_6
```

```
## Warning: Use of ‘cancel$Babies‘ is discouraged. Use ‘Babies‘ instead.
```



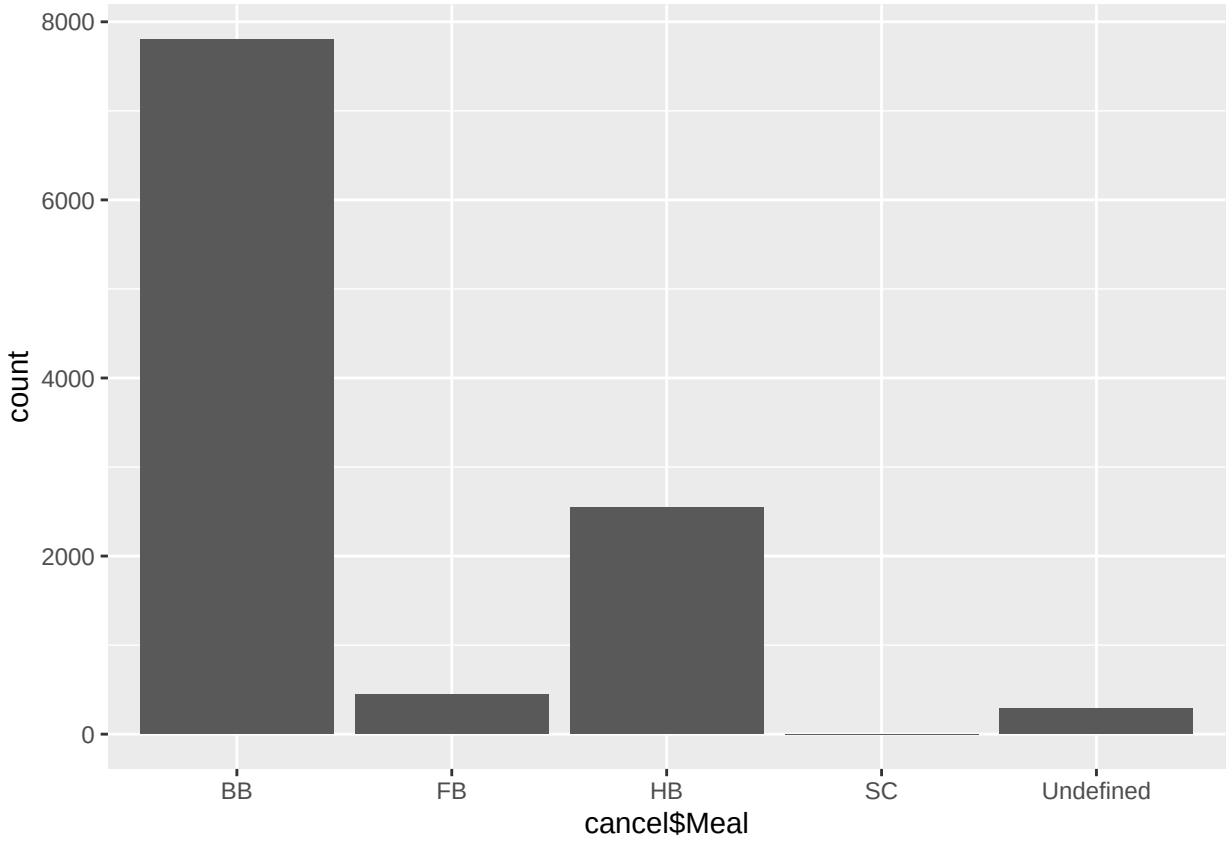

```
# only 1 percent of the repeated guests cancel on the hotels.
plotyplot_7 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$IsRepeatedGuest)
plotyplot_7
```

```
## Warning: Use of ‘cancel$IsRepeatedGuest‘ is discouraged. Use ‘IsRepeatedGuest‘
## instead.
```



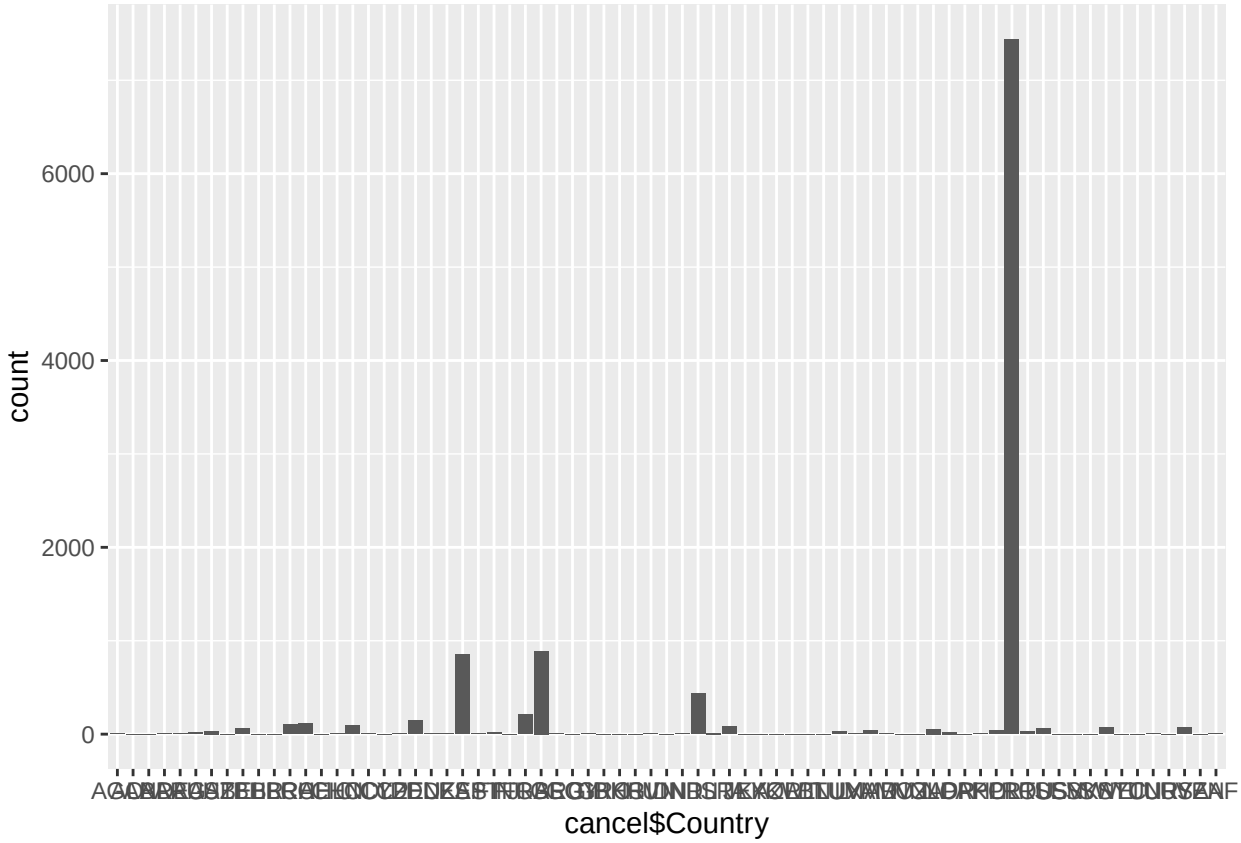

```
# People with Meal SC and FB hardly cancel whereas people taking BB meals cancel 80 percent of the time
plotyplot_8 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$Meal)
plotyplot_8
```

```
## Warning: Use of ‘cancel$Meal‘ is discouraged. Use ‘Meal‘ instead.
```



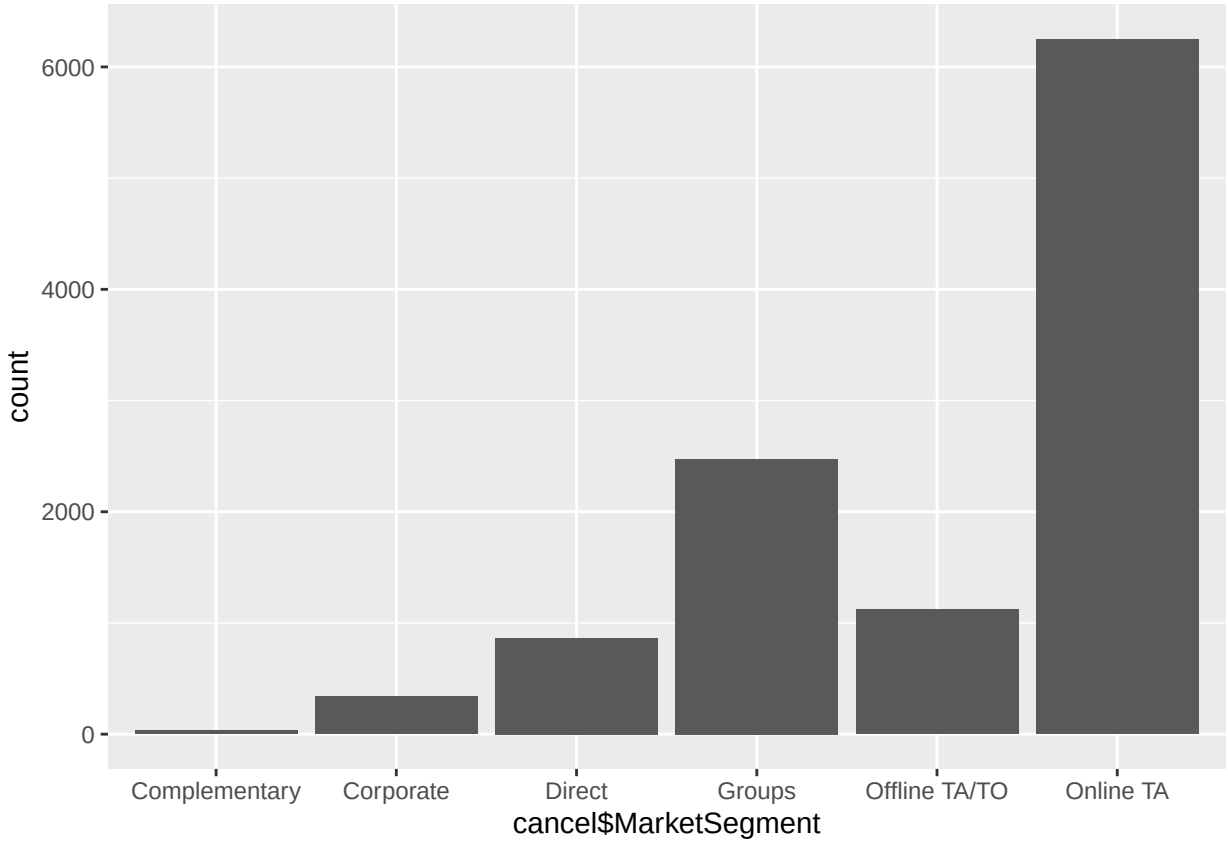

```
# COLUMN NOT REQUIRED for analysis

# there is only one country from where bookings are cancelled the most. While we can also infer that mo
plotyplot_9 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x=cancel$Country)
plotyplot_9
```

```
## Warning: Use of ‘cancel$Country‘ is discouraged. Use ‘Country‘ instead.
```



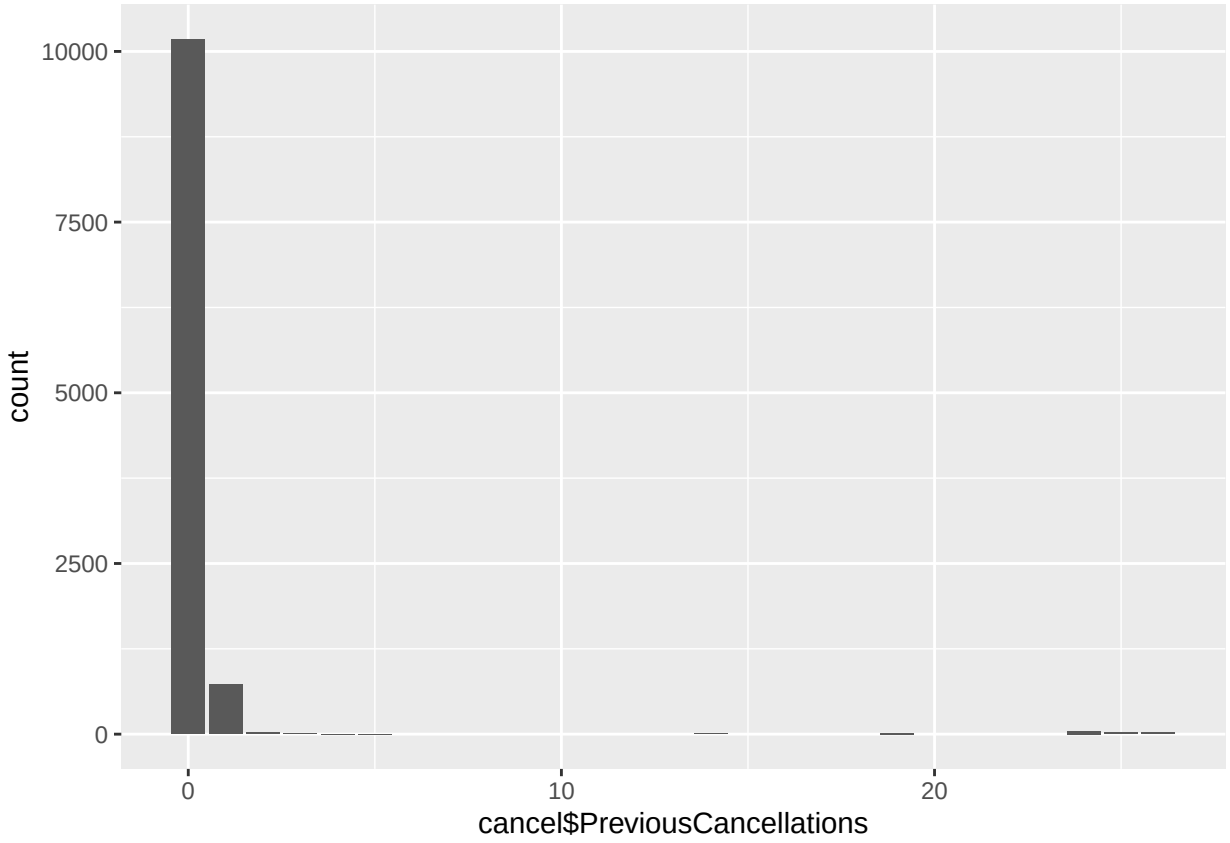

```
# the market segment which comprises of online TA is responsible for most of the cancellations. Booking
plotyplot_10 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$MarketSegment)
plotyplot_10
```

```
## Warning: Use of ‘cancel$MarketSegment‘ is discouraged. Use ‘MarketSegment‘
## instead.
```



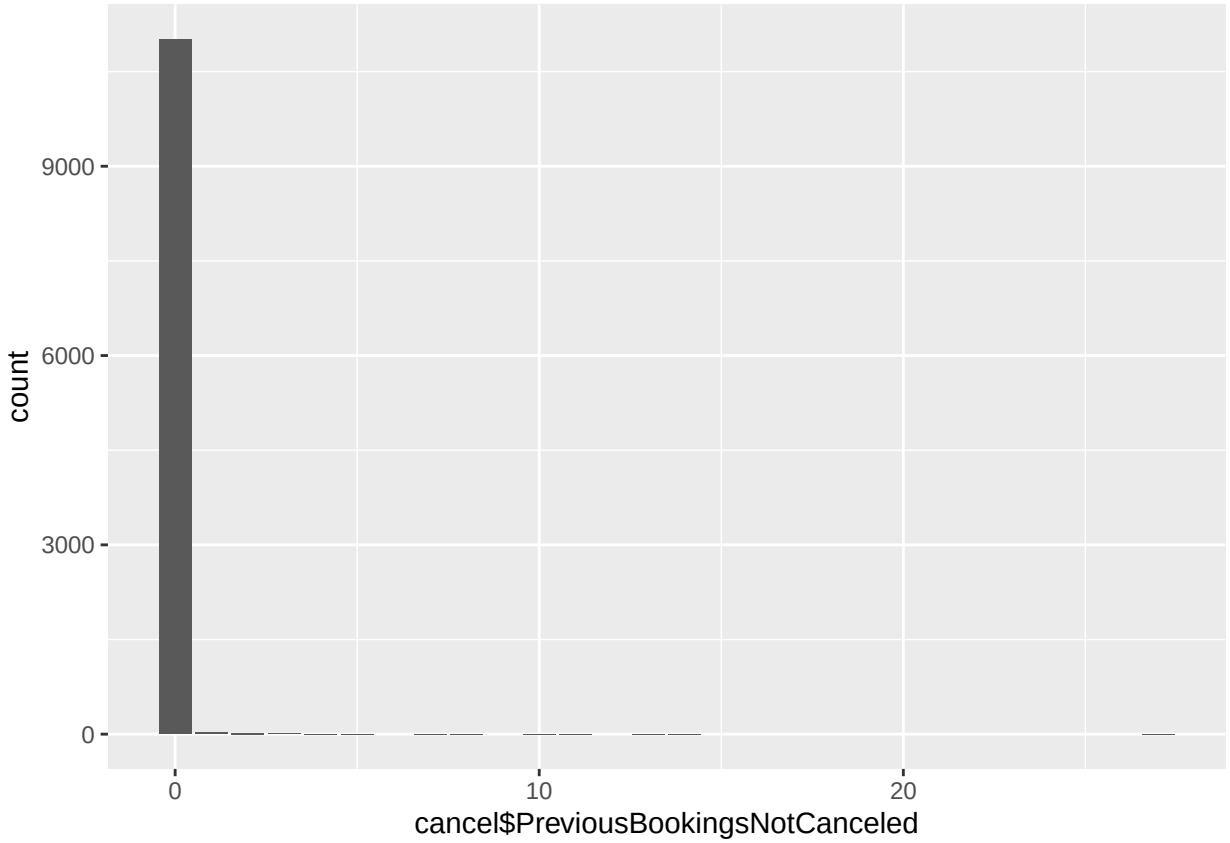

```
# COLUMN NOT REQUIRED for analysis

# We cannot use this as a measure to identify potential risk of someone cancelling their booking. Since
plotyplot_11 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$PreviousCancellations)
plotyplot_11
```

```
## Warning: Use of ‘cancel$PreviousCancellations‘ is discouraged. Use
## ‘PreviousCancellations‘ instead.
```



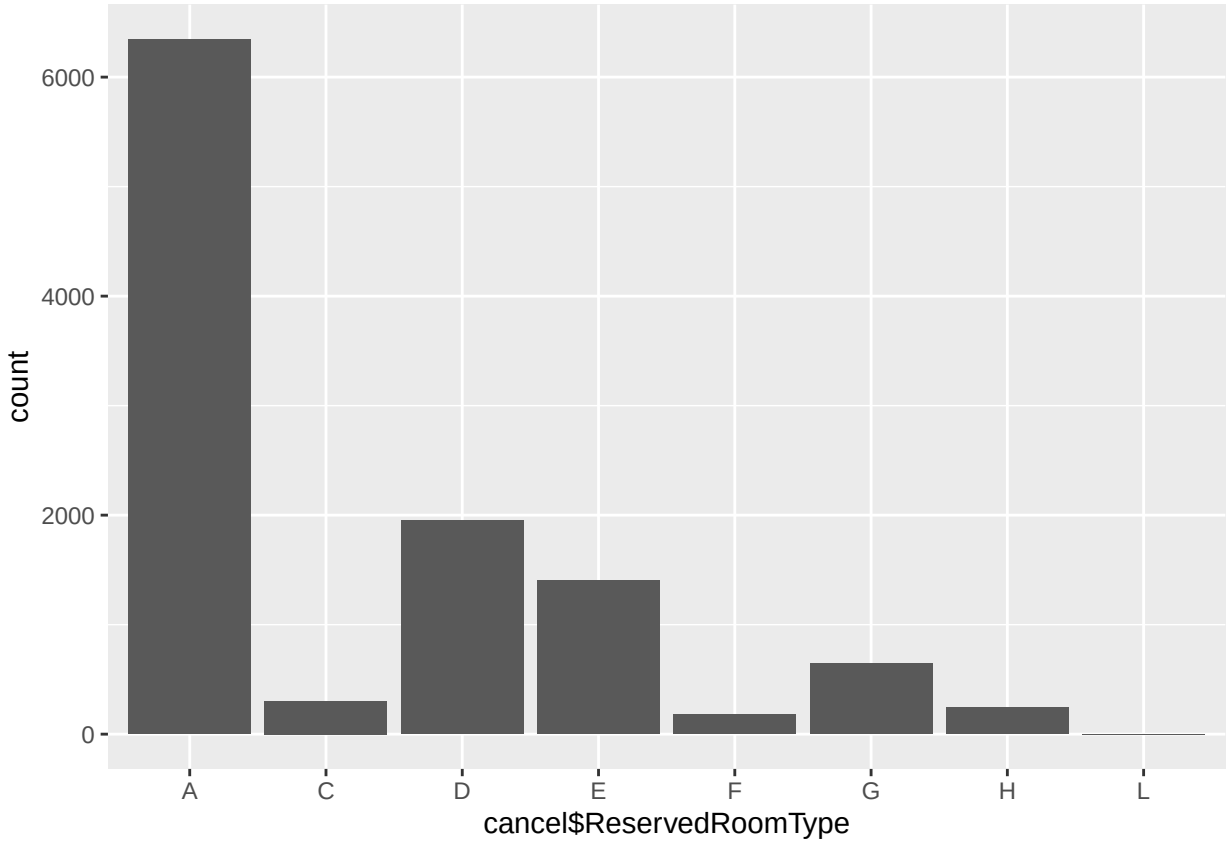

```
# COLUMN NOT REQUIRED for analysis

# We cannot use this as a measure to identify potential risk of someone cancelling their booking. Since
plotyplot_12 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$PreviousBookingsNotCanceled)
plotyplot_12
```

```
## Warning: Use of ‘cancel$PreviousBookingsNotCanceled‘ is discouraged. Use
## ‘PreviousBookingsNotCanceled‘ instead.
```



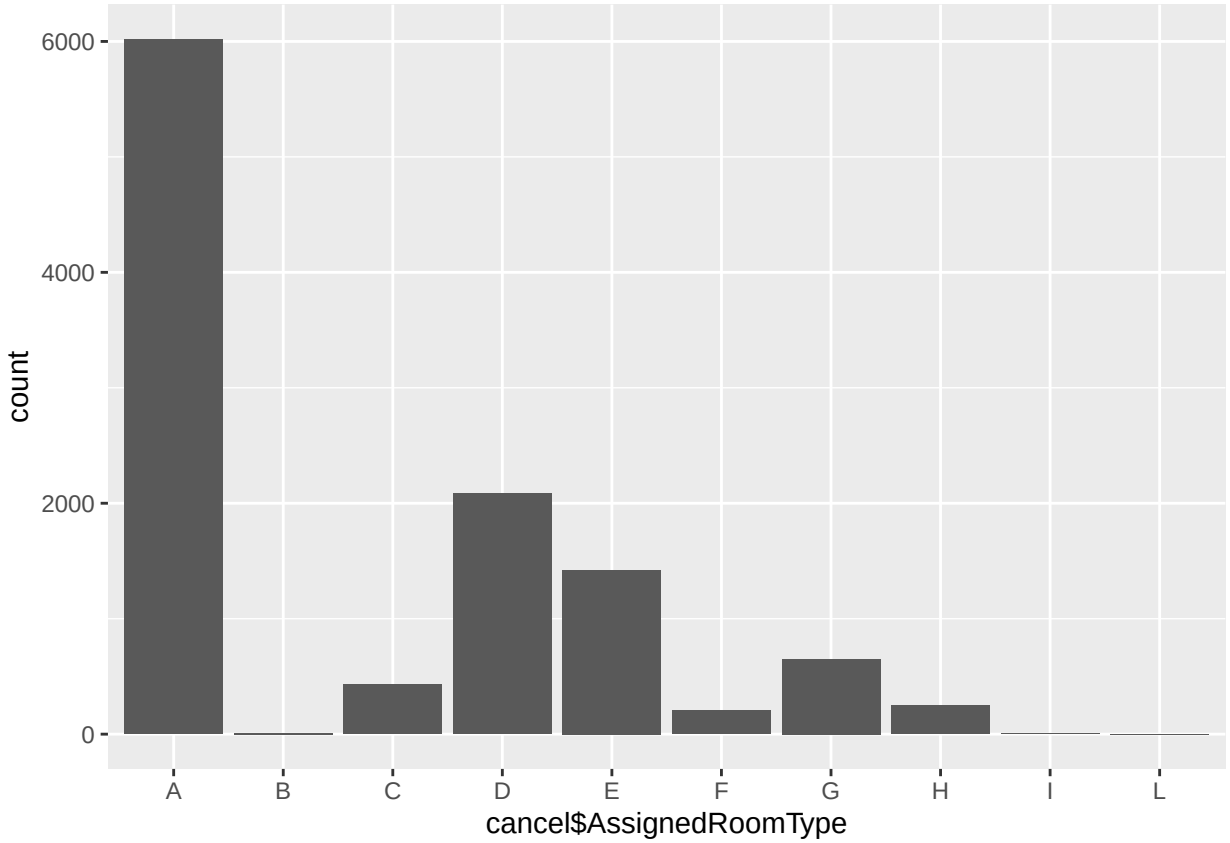

```
# people assigned with room type A has seen most of the cancellations. To avoid cancellations , people
plotyplot_13 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$ReservedRoomType)
plotyplot_13
```

```
## Warning: Use of ‘cancel$ReservedRoomType‘ is discouraged. Use ‘ReservedRoomType‘
## instead.
```



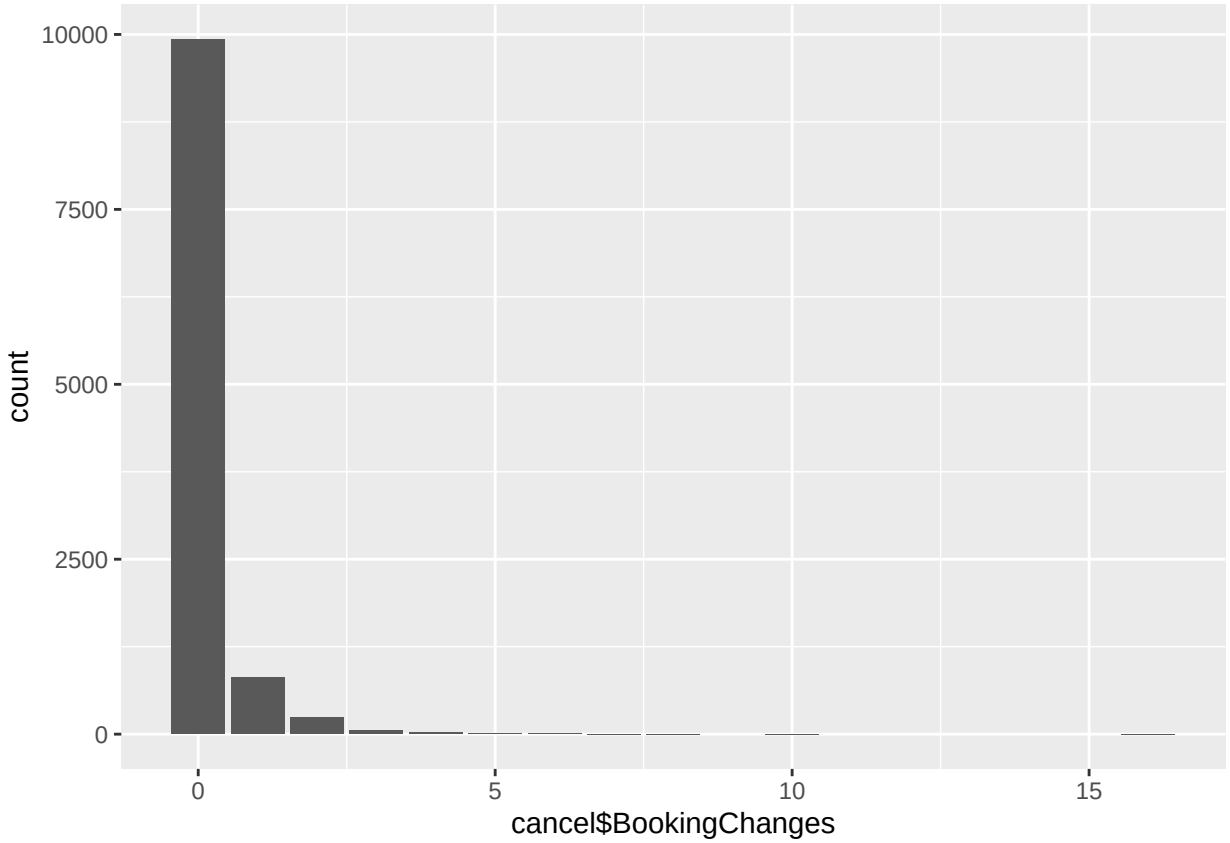

```
# people assigned with room type A has seen most of the cancellations. To avoid cancellations , people
plotyplot_14 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$AssignedRoomType)
plotyplot_14
```

```
## Warning: Use of ‘cancel$AssignedRoomType‘ is discouraged. Use ‘AssignedRoomType‘
## instead.
```



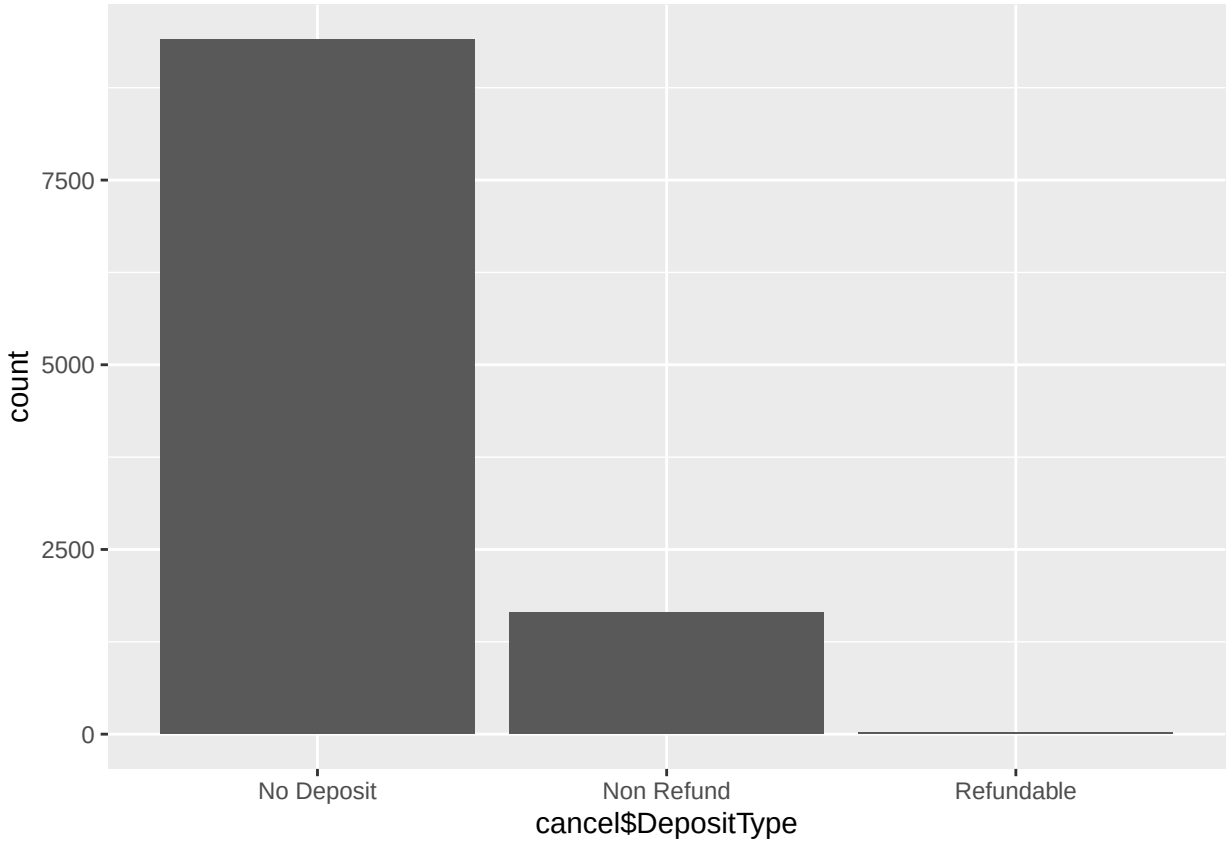

```
# most people directly cancel there plans. But some have to change their bookings to accomodate their p
plotyplot_15 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$BookingChanges)
plotyplot_15
```

```
## Warning: Use of ‘cancel$BookingChanges‘ is discouraged. Use ‘BookingChanges‘
## instead.
```



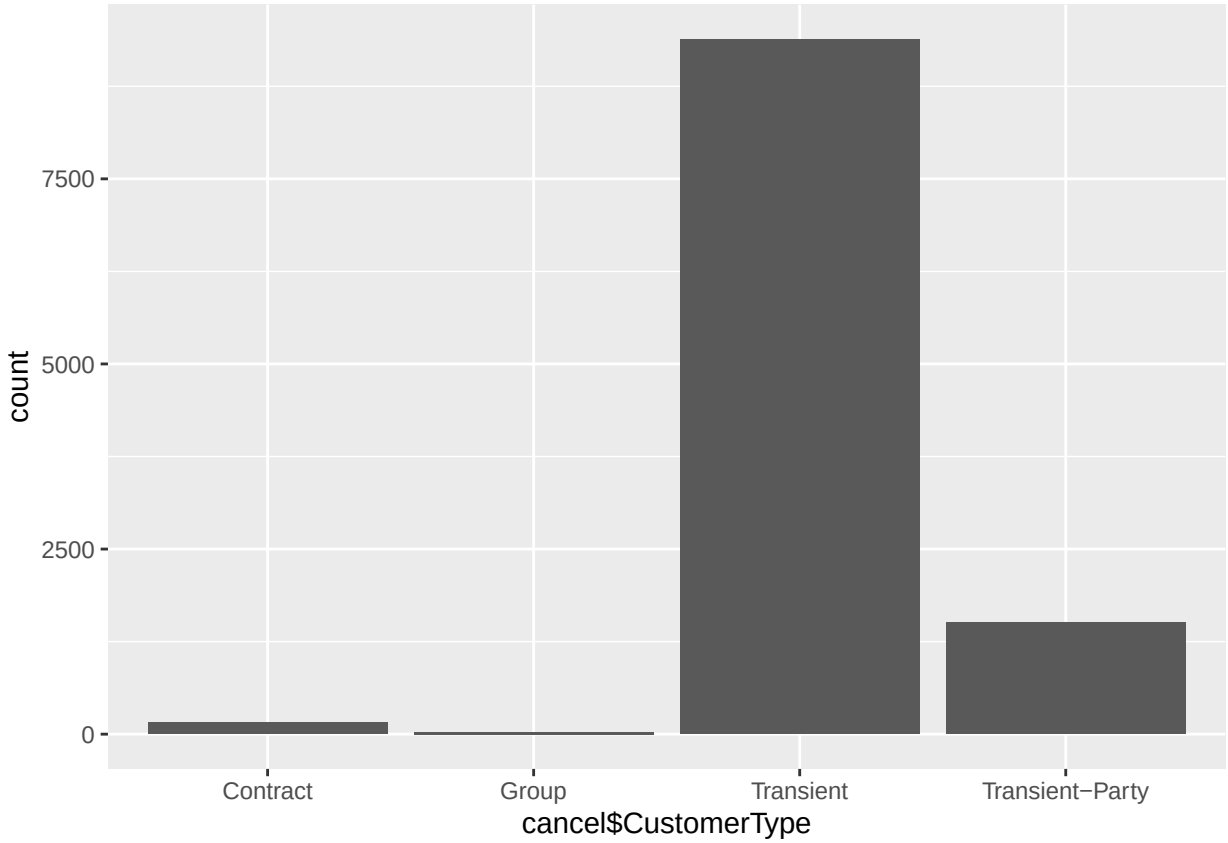

```
# when booking made with no deposit
plotyplot_16 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$DepositType)
plotyplot_16
```

```
## Warning: Use of ‘cancel$DepositType‘ is discouraged. Use ‘DepositType‘ instead.
```



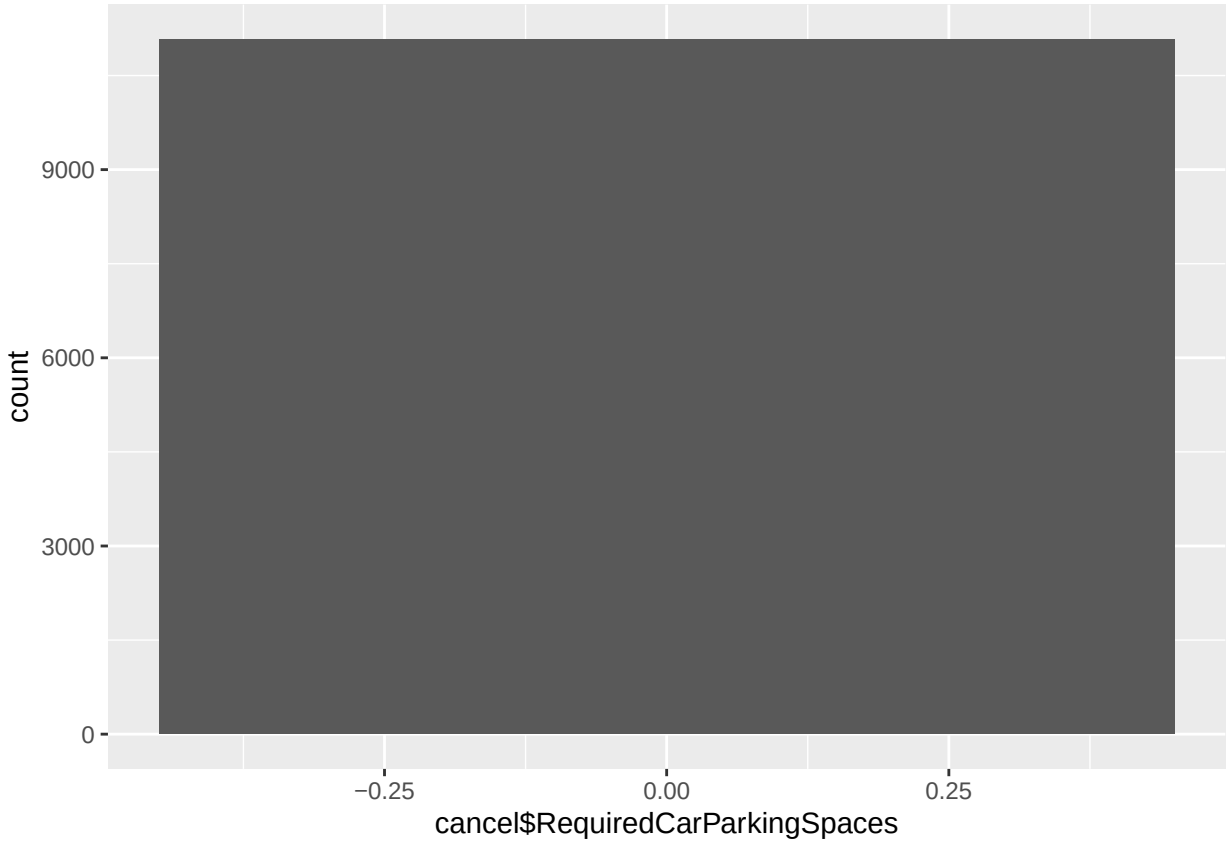

```
# transient customers are the types who are cancelling the booking the most
plotyplot_17 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$CustomerType)
plotyplot_17
```

```
## Warning: Use of ‘cancel$CustomerType‘ is discouraged. Use ‘CustomerType‘
## instead.
```



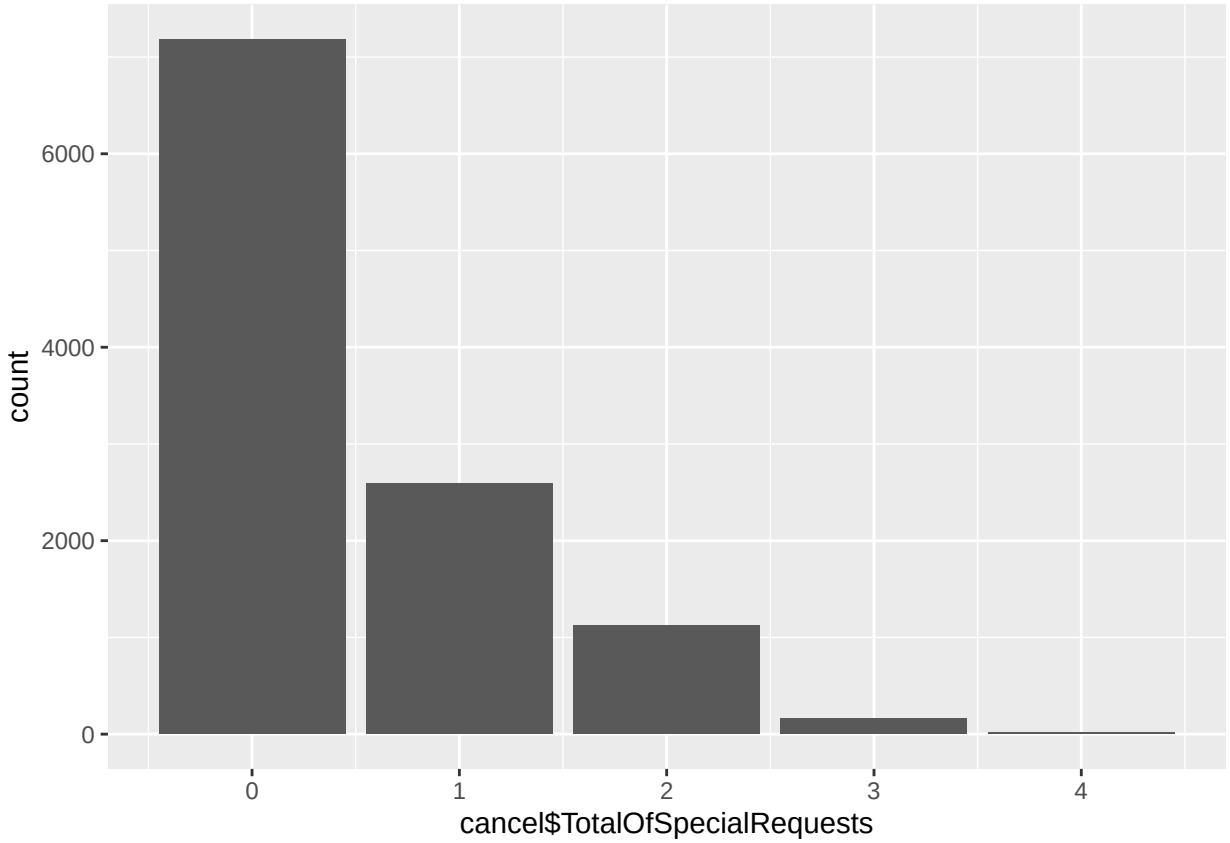

```
# people who cancelled their bookings do not ask for a parking space in advance.
plotyplot_18 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$RequiredCarParkingSpaces)
plotyplot_18
```

```
## Warning: Use of ‘cancel$RequiredCarParkingSpaces‘ is discouraged. Use
## ‘RequiredCarParkingSpaces‘ instead.
```




```
# people with 0 special requests tend to cancel their bookings. We should be sure the booking won't be
plotyplot_19 <-  ggplot(cancel) + geom_bar(data=cancel)+ aes(x= cancel$TotalOfSpecialRequests)
plotyplot_19
```

```
## Warning: Use of ‘cancel$TotalOfSpecialRequests‘ is discouraged. Use
## ‘TotalOfSpecialRequests‘ instead.
```



Actionable Insights: 1) While selecting Room type either reserved or assigned room type , avoid giving room type A to people show patterns of high cancellations risk. 2) Try incorporating a structure where deposit amount is compulsory because most of the people cancelled had no deposit 3) When people make an effort to change their bookings greet them with more importance because it shows inclination towards coming to stay in the hotels. Which means there is less chance of cancellation 4) Make an effort to get special requests from the customers, to be very safe , at least 4 or more. Customers making special requests hardly cancel their bookings 5) Always ask for parking space from the customer prior to the booking. Reserving parking space for the customers will lead to reduction in cancellation. 6) The above 5 points should be focused especially more on the market segment "Online TA" which is responsible for most of the cancellations 7) Special care of the first 5 points should be given to transient customers as they cancel more than other type of customers 8) Customers who are families especially who have children and babies would hardly cancel bookings in your hotels, focus on customers in groups of 2.

5. Geogrphic Analysis

```
x<- hotels %>% #filtering hotel data set
  group_by(Country) %>% #by countries
  summarise(Frequency = sum(as.numeric(IsCanceled))) #and summing
```

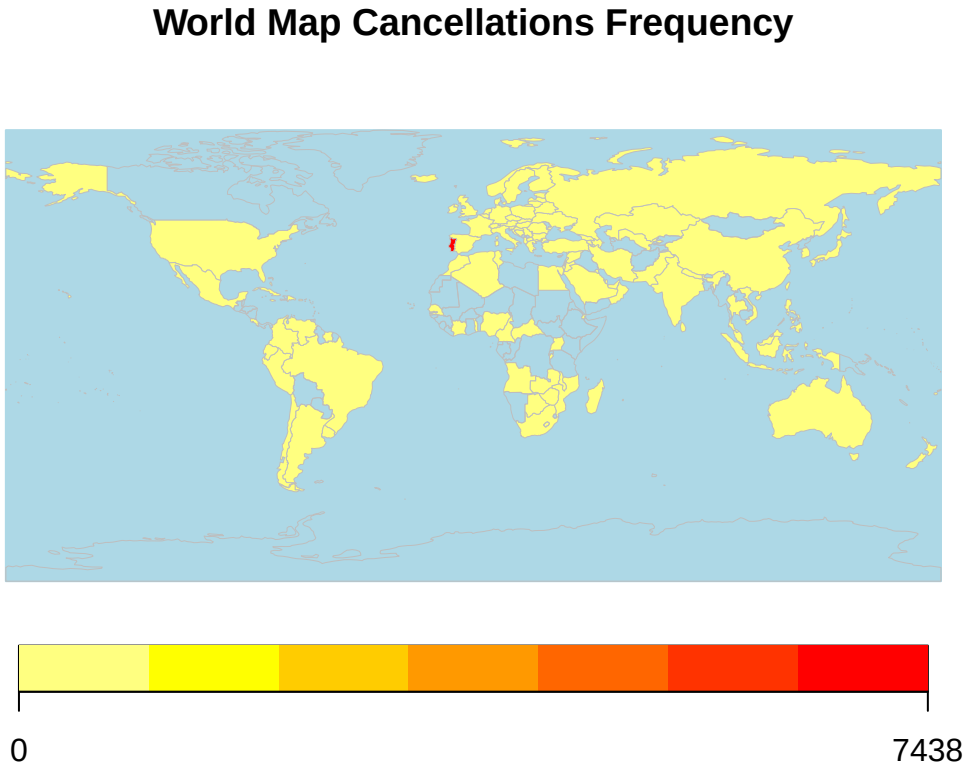
```
joinData <- joinCountryData2Map( x, #joining our country subset
                                 joinCode = "ISO3", #with code
                                 nameJoinColumn = "Country") #to countries
```

```
## 123 codes from your data successfully matched countries in the map
## 2 codes from your data failed to match with a country code in the map
## 120 codes from the map weren't represented in your data
```

```
#WorldMap
theMap <- mapCountryData( joinData, nameColumnToPlot="Frequency", mapTitle= 'World Map Cancellations Fr
```

```
## Warning in rwmGetClassBreaks(dataCategorised, catMethod = catMethod, numCats = numCats, : classifica
## setting to fixedWidth as default
```

```
## Warning in rwmGetColours(colourPalette, numColours): colourPalette should be set to either a vector
## setting to heat colours as default
```

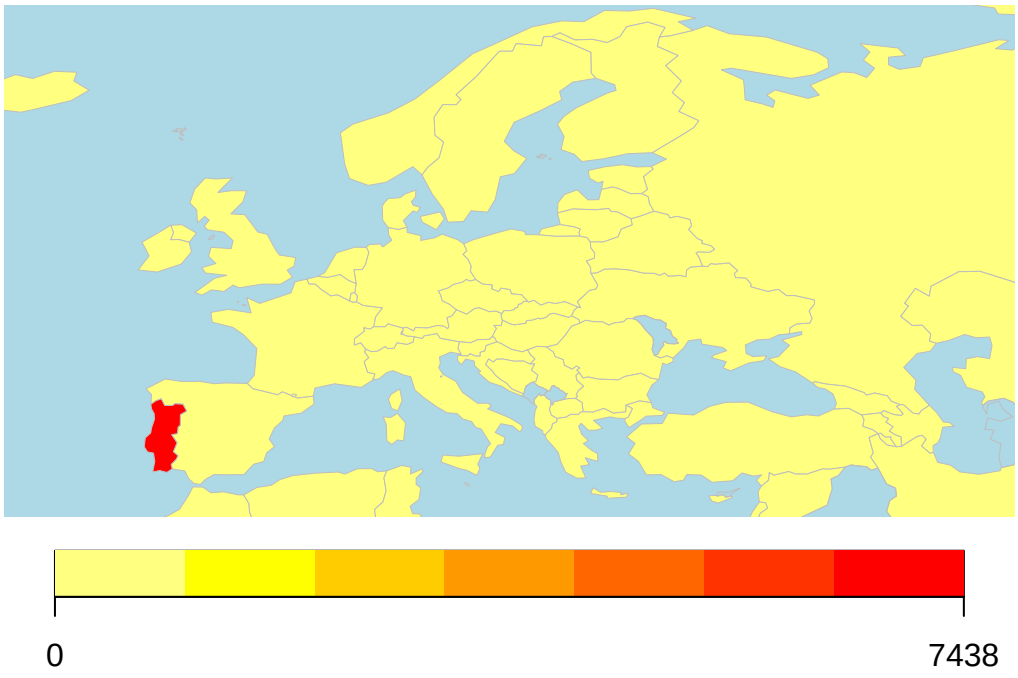



```
#do.call( addMapLegend, c(theMap, legendWidth=1, legendMar = 2))
```

```
#EUROPE
theMap <- mapCountryData( joinData, nameColumnToPlot="Frequency", mapTitle= 'Europe Cancellations Frequ
```

```
## Warning in rwmGetClassBreaks(dataCategorised, catMethod = catMethod, numCats = numCats, : classifica
## setting to fixedWidth as default
```

```
## Warning in rwmGetColours(colourPalette, numColours): colourPalette should be set to either a vector
## setting to heat colours as default
```

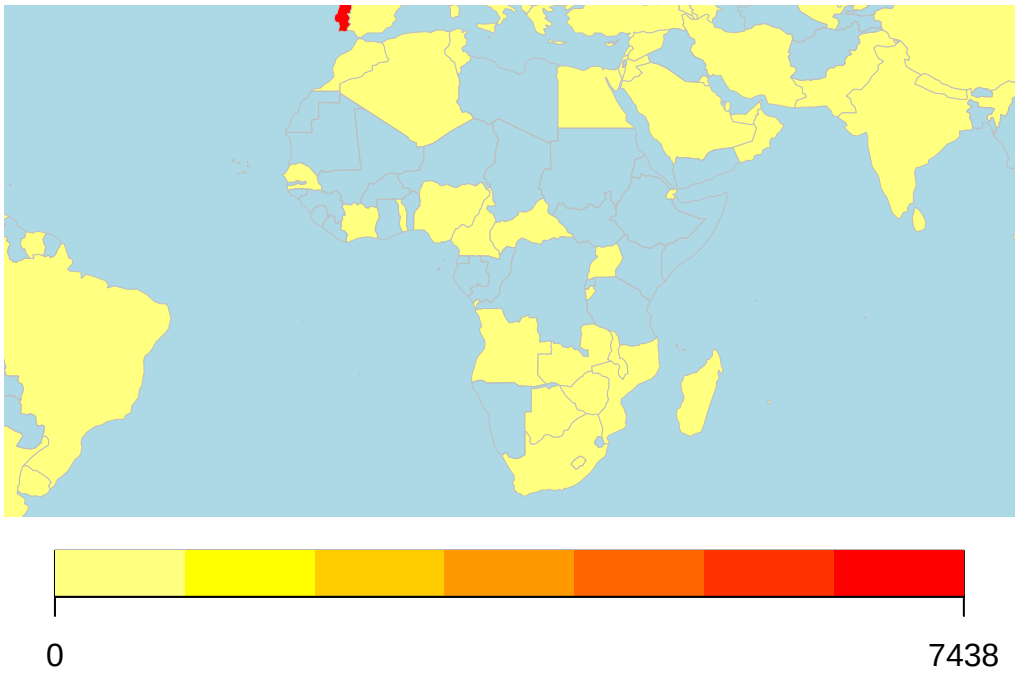



```
#do.call( addMapLegend, c(theMap, legendWidth=1, legendMar = 2))

#country_coord<-data.frame(coordinates(joinData),stringsAsFactors=F)
# label the countries
#text(x=country_coord$X1,y=country_coord$X2,labels=row.names(country_coord))
```

```
#Africa
theMap <- mapCountryData( joinData, nameColumnToPlot="Frequency", mapTitle= 'Africa Cancellations Frequ
```

```
## Warning in rwmGetClassBreaks(dataCategorised, catMethod = catMethod, numCats = numCats, : classifica
## setting to fixedWidth as default
```

```
## Warning in rwmGetColours(colourPalette, numColours): colourPalette should be set to either a vector
## setting to heat colours as default
```

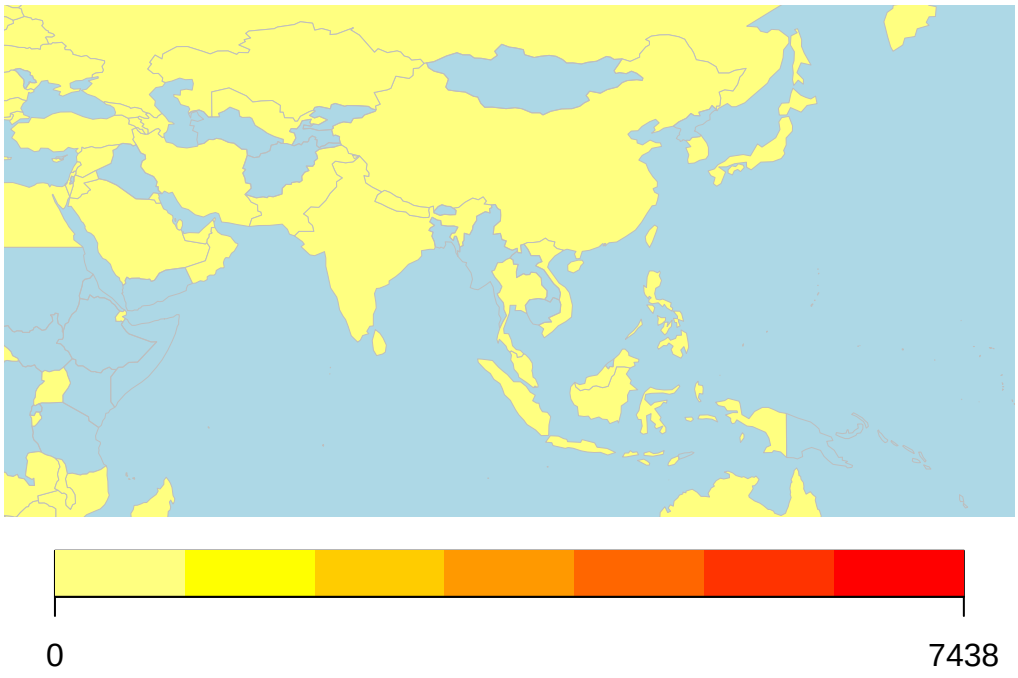



```
#do.call( addMapLegend, c(theMap, legendWidth=1, legendMar = 2))

#country_coord<-data.frame(coordinates(joinData),stringsAsFactors=F)
# label the countries
#text(x=country_coord$X1,y=country_coord$X2,labels=row.names(country_coord))
```

```
#ASIA

theMap <- mapCountryData( joinData, nameColumnToPlot="Frequency", mapTitle= 'Asia Cancellations Frequen
```

```
## Warning in rwmGetClassBreaks(dataCategorised, catMethod = catMethod, numCats = numCats, : classifica
## setting to fixedWidth as default
```

```
## Warning in rwmGetColours(colourPalette, numColours): colourPalette should be set to either a vector
## setting to heat colours as default
```



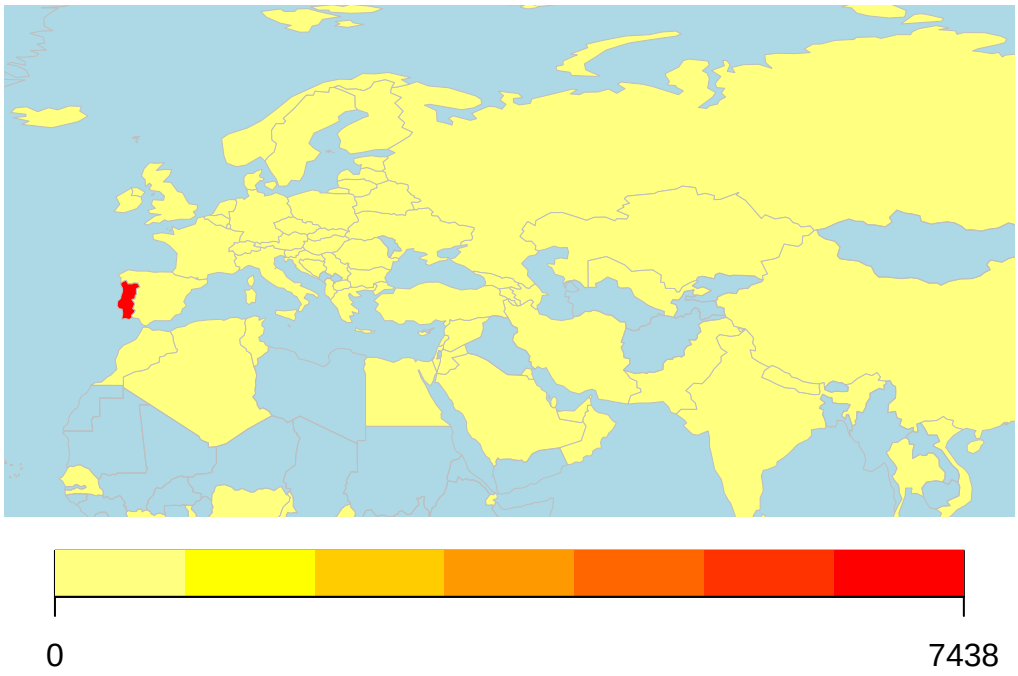

```
#do.call( addMapLegend, c(theMap, legendWidth=1, legendMar = 2))

#country_coord<-data.frame(coordinates(joinData),stringsAsFactors=F)
# label the countries
#text(x=country_coord$X1,y=country_coord$X2,labels=row.names(country_coord))
```

```
#Europe-Asia

theMap <- mapCountryData( joinData, nameColumnToPlot="Frequency", mapTitle= 'Eurasia Cancellations Frequ
```

```
## Warning in rwmGetClassBreaks(dataCategorised, catMethod = catMethod, numCats = numCats, : classificat
## setting to fixedWidth as default
```

```
## Warning in rwmGetColours(colourPalette, numColours): colourPalette should be set to either a vector o
## setting to heat colours as default
```

Eurasia Cancellations Frequency

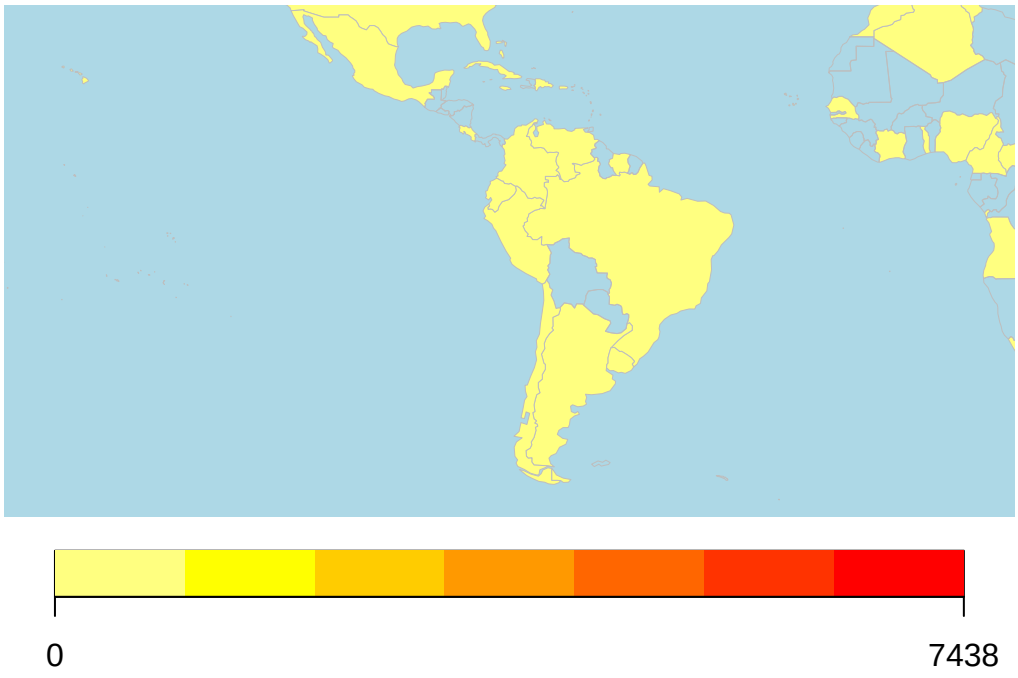



```
#do.call( addMapLegend, c(theMap, legendWidth=1, legendMar = 2))

#country_coord<-data.frame(coordinates(joinData),stringsAsFactors=F)
# label the countries
#text(x=country_coord$X1,y=country_coord$X2,labels=row.names(country_coord))
```

```
#Latin america

theMap <- mapCountryData( joinData, nameColumnToPlot="Frequency", mapTitle= 'Latin America Cancellations
```

```
## Warning in rwmGetClassBreaks(dataCategorised, catMethod = catMethod, numCats = numCats, : classifica
## setting to fixedWidth as default
```

```
## Warning in rwmGetColours(colourPalette, numColours): colourPalette should be set to either a vector
## setting to heat colours as default
```

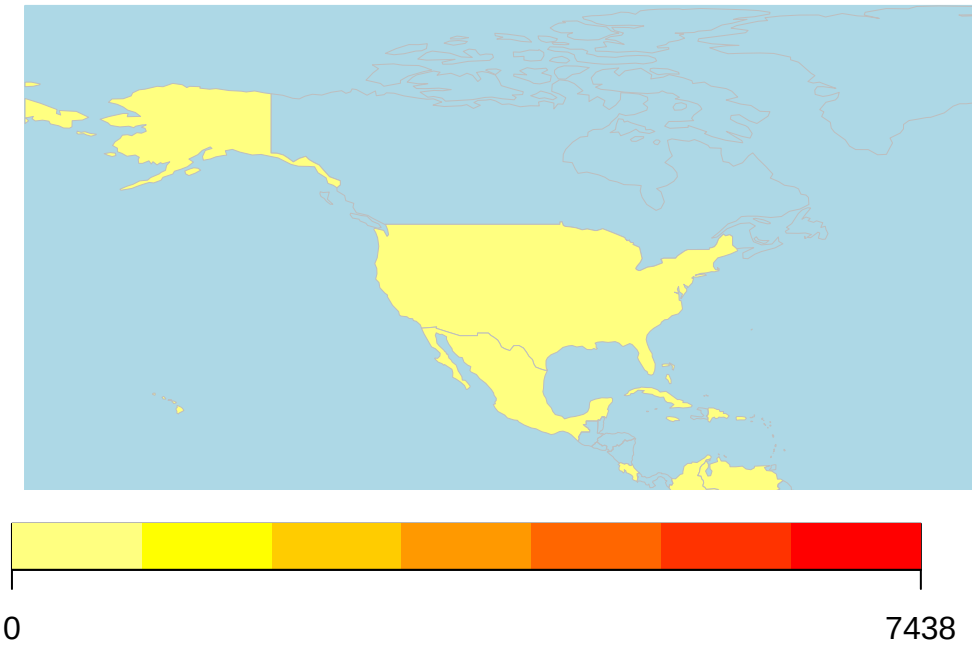



```
#do.call( addMapLegend, c(theMap, legendWidth=1, legendMar = 2))

#country_coord<-data.frame(coordinates(joinData),stringsAsFactors=F)
# label the countries
#text(x=country_coord$X1,y=country_coord$X2,labels=row.names(country_coord))
```

```
#North america

theMap <- mapCountryData( joinData, nameColumnToPlot="Frequency", mapTitle= 'North America Cancellations
```

```
## Warning in rwmGetClassBreaks(dataCategorised, catMethod = catMethod, numCats = numCats, : classifica
## setting to fixedWidth as default
```

```
## Warning in rwmGetColours(colourPalette, numColours): colourPalette should be set to either a vector
## setting to heat colours as default
```

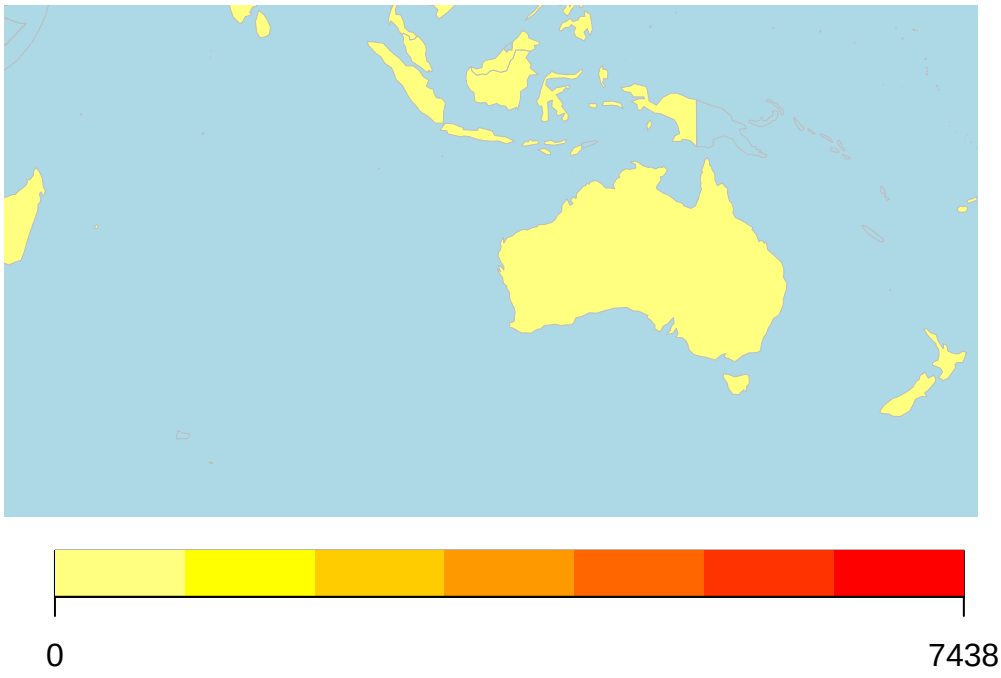



```
#do.call( addMapLegend, c(theMap, legendWidth=1, legendMar = 2))

#country_coord<-data.frame(coordinates(joinData),stringsAsFactors=F)
# label the countries
#text(x=country_coord$X1,y=country_coord$X2,labels=row.names(country_coord))
```

```
#Oceania

theMap <- mapCountryData( joinData, nameColumnToPlot="Frequency", mapTitle= 'Oceania Cancellations Frequ
```

```
## Warning in rwmGetClassBreaks(dataCategorised, catMethod = catMethod, numCats = numCats, : classificat
## setting to fixedWidth as default
```

```
## Warning in rwmGetColours(colourPalette, numColours): colourPalette should be set to either a vector o
## setting to heat colours as default
```

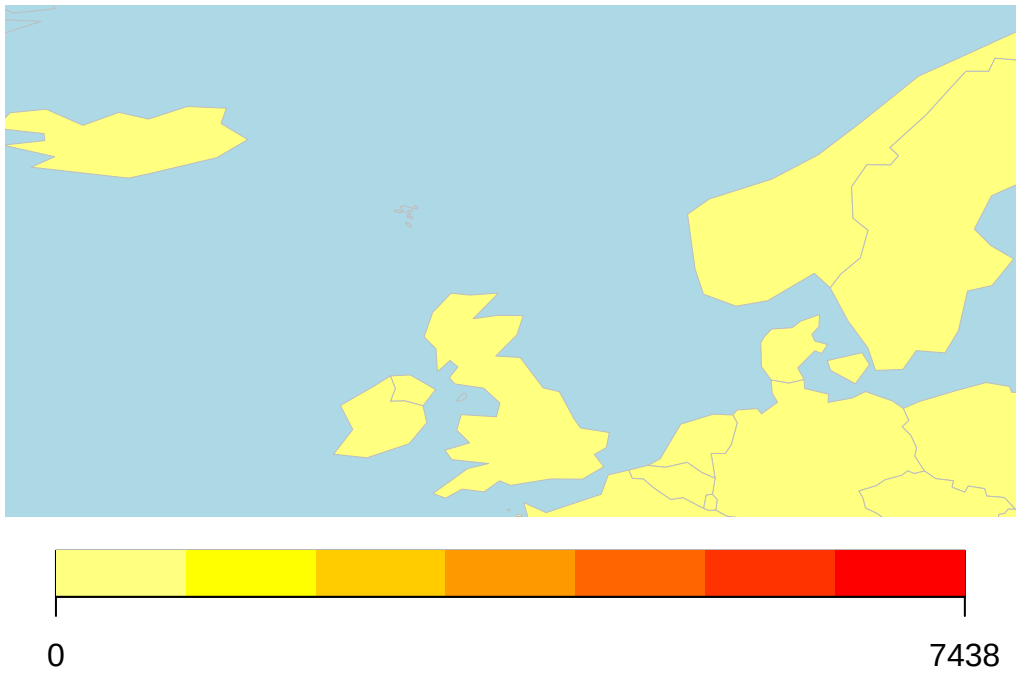



```
#do.call( addMapLegend, c(theMap, legendWidth=1, legendMar = 2))

#country_coord<-data.frame(coordinates(joinData),stringsAsFactors=F)
# label the countries
#text(x=country_coord$X1,y=country_coord$X2,labels=row.names(country_coord))
```

```
#United Kingdom

theMap <- mapCountryData( joinData, nameColumnToPlot="Frequency", mapTitle= 'United Kingdom Cancellatio
```

```
## Warning in rwmGetClassBreaks(dataCategorised, catMethod = catMethod, numCats = numCats, : classifica
## setting to fixedWidth as default
```

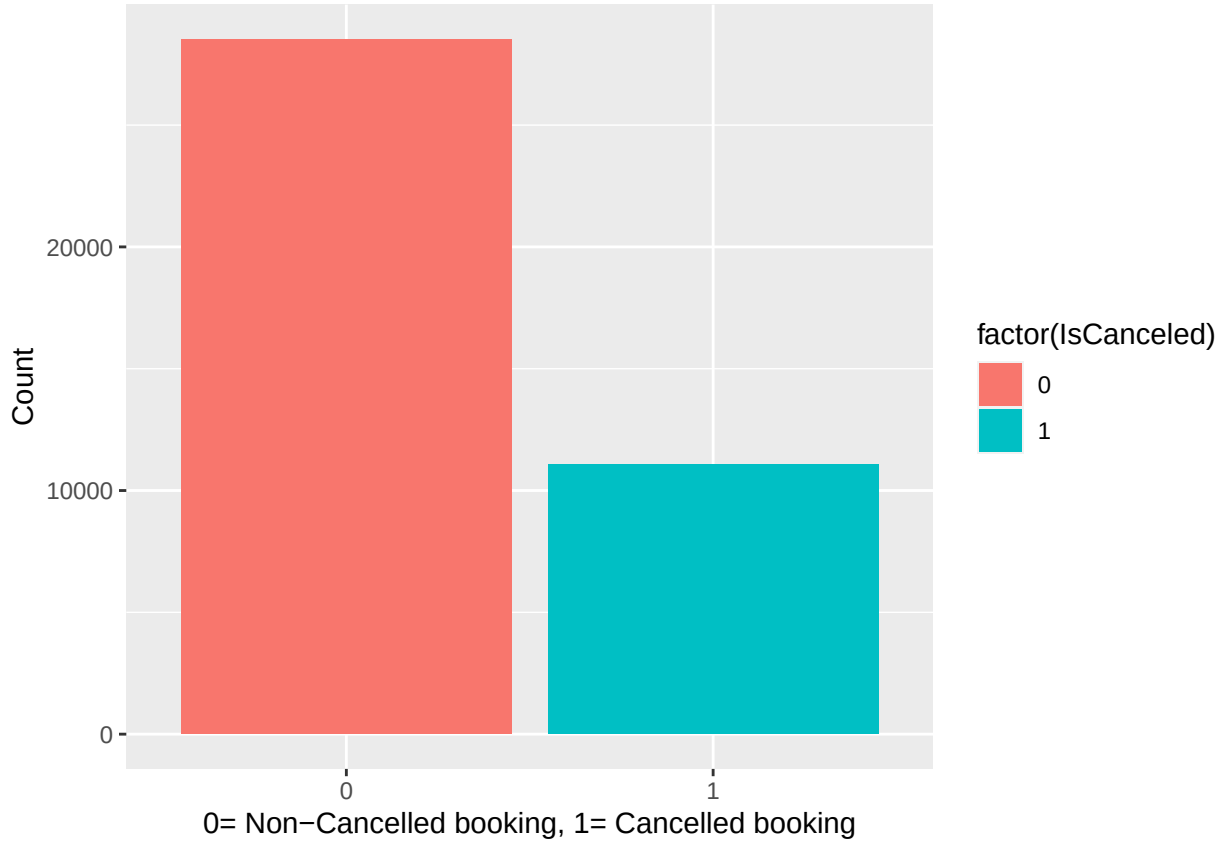
```
## Warning in rwmGetColours(colourPalette, numColours): colourPalette should be set to either a vector
## setting to heat colours as default
```



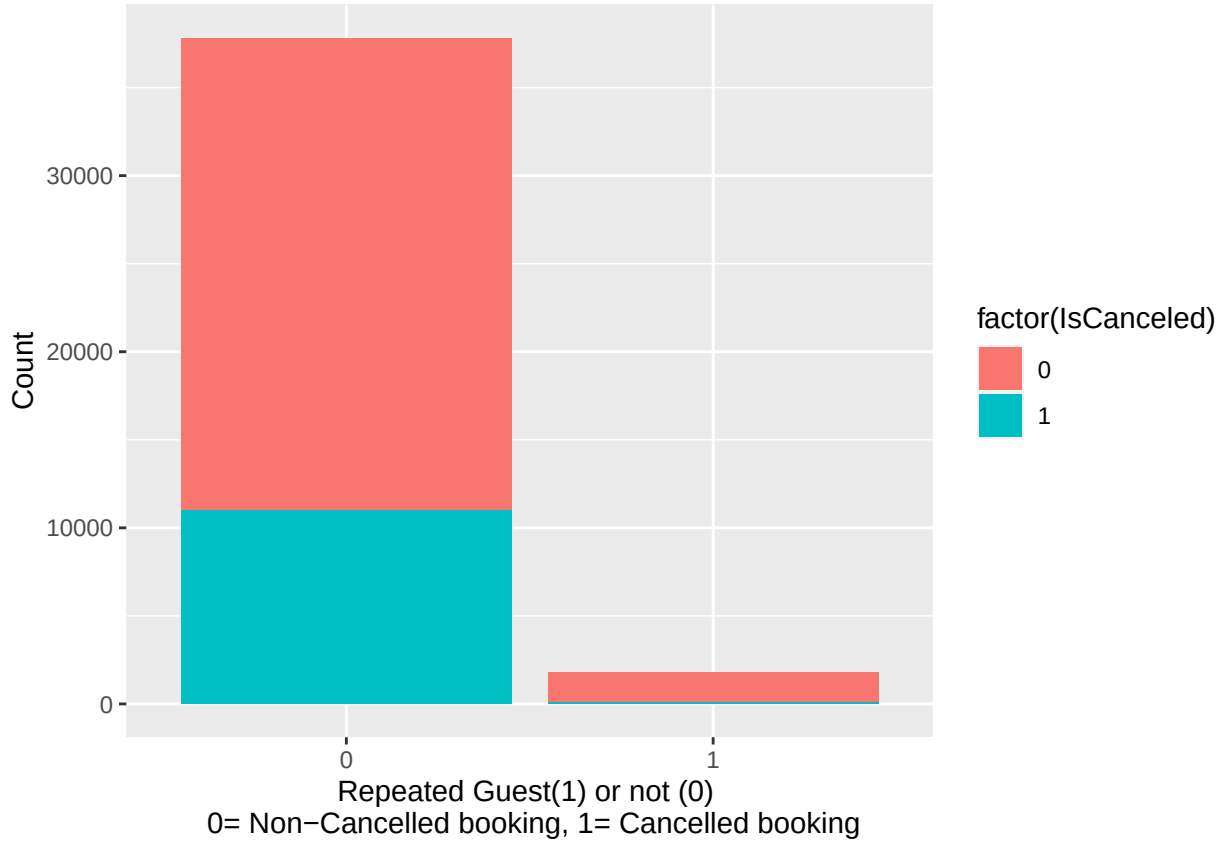

```
#do.call( addMapLegend, c(theMap, legendWidth=1, legendMar = 2))

#country_coord<-data.frame(coordinates(joinData),stringsAsFactors=F)
# label the countries
#text(x=country_coord$X1,y=country_coord$X2,labels=row.names(country_coord))
```

6. Miscellaneous Graphing

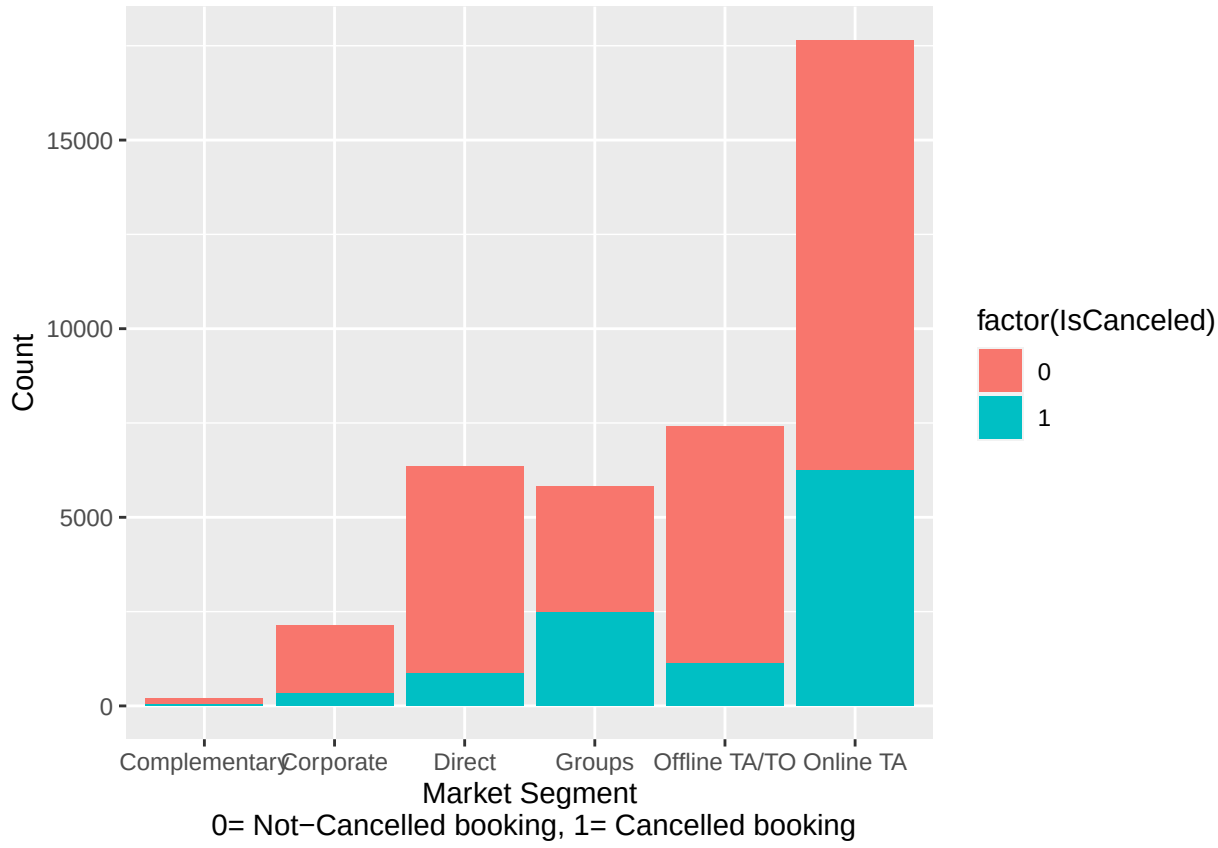
```
ggplot(hotels, aes(factor(IsCanceled), #bar plot of only the IsCanceled variable
       fill = factor(IsCanceled))) +geom_bar()+ xlab("0= Non-Cancelled booking, 1= Cancelled booking")+
```



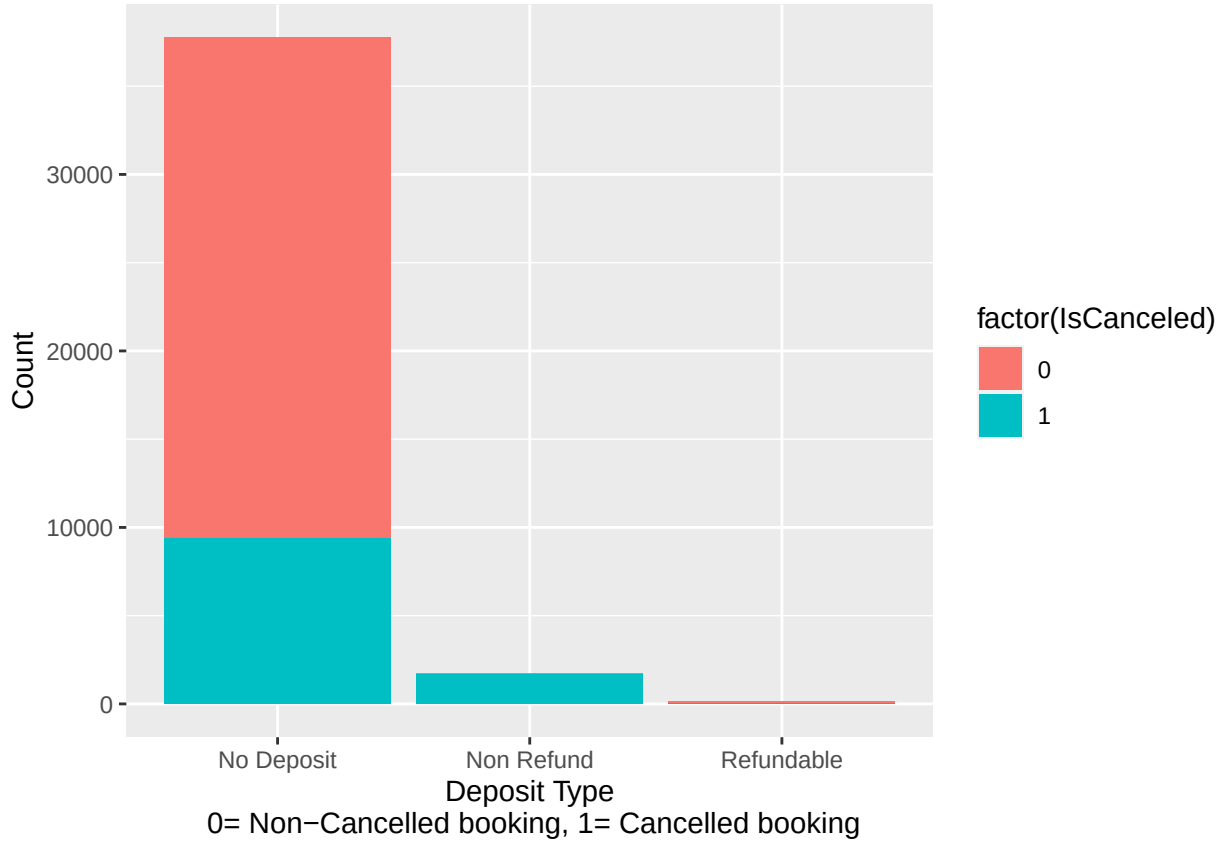

```
ggplot(hotels, aes(factor(IsRepeatedGuest), #bar plot describing the IsRepeatedGuest variable with IsCa
        fill = factor(IsCanceled))) +geom_bar() + xlab("Repeated Guest(1) or not (0) \n 0= Non-Cancelle
```



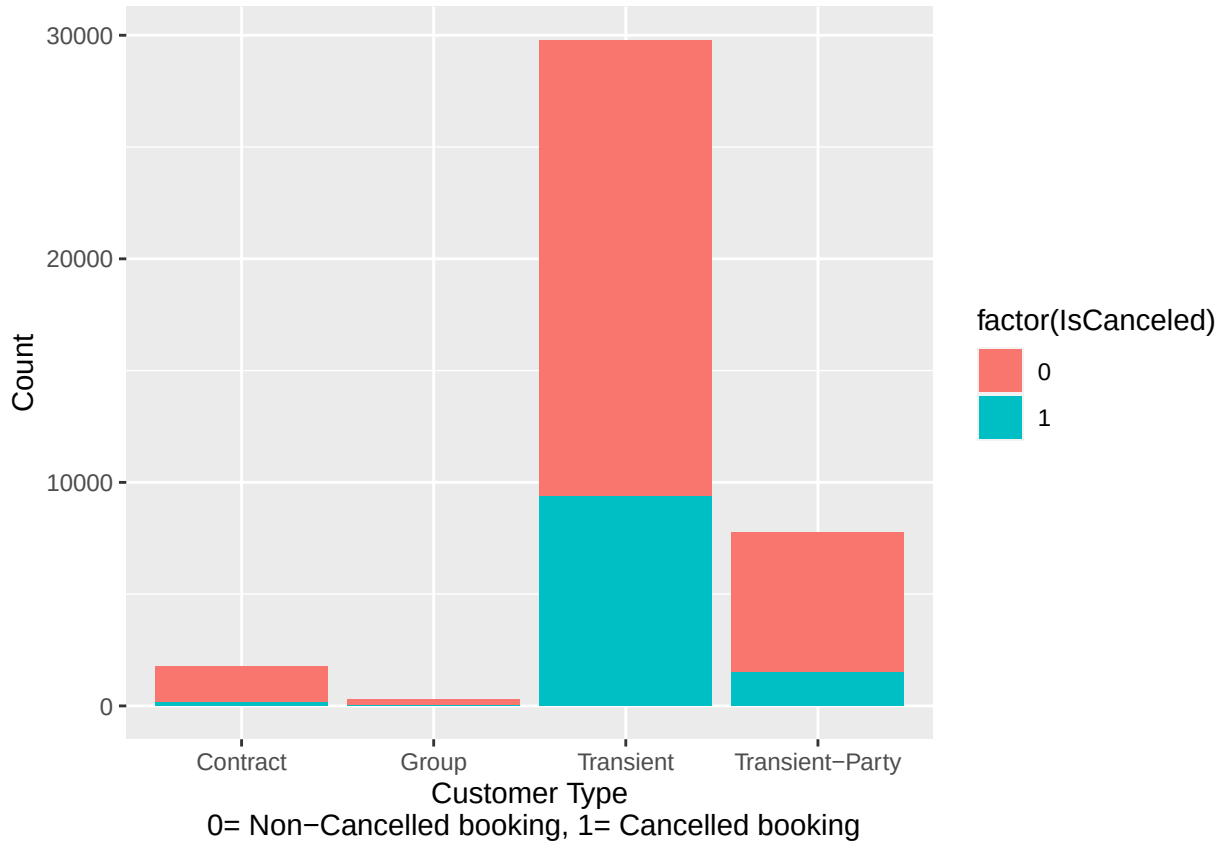

```
ggplot(hotels, aes(factor(MarketSegment), #bar plot describing the MarketSegment variable with IsCancel
        fill = factor(IsCanceled))) +geom_bar()+ xlab("Market Segment\n 0= Not-Cancelled booking, 1= Ca
```



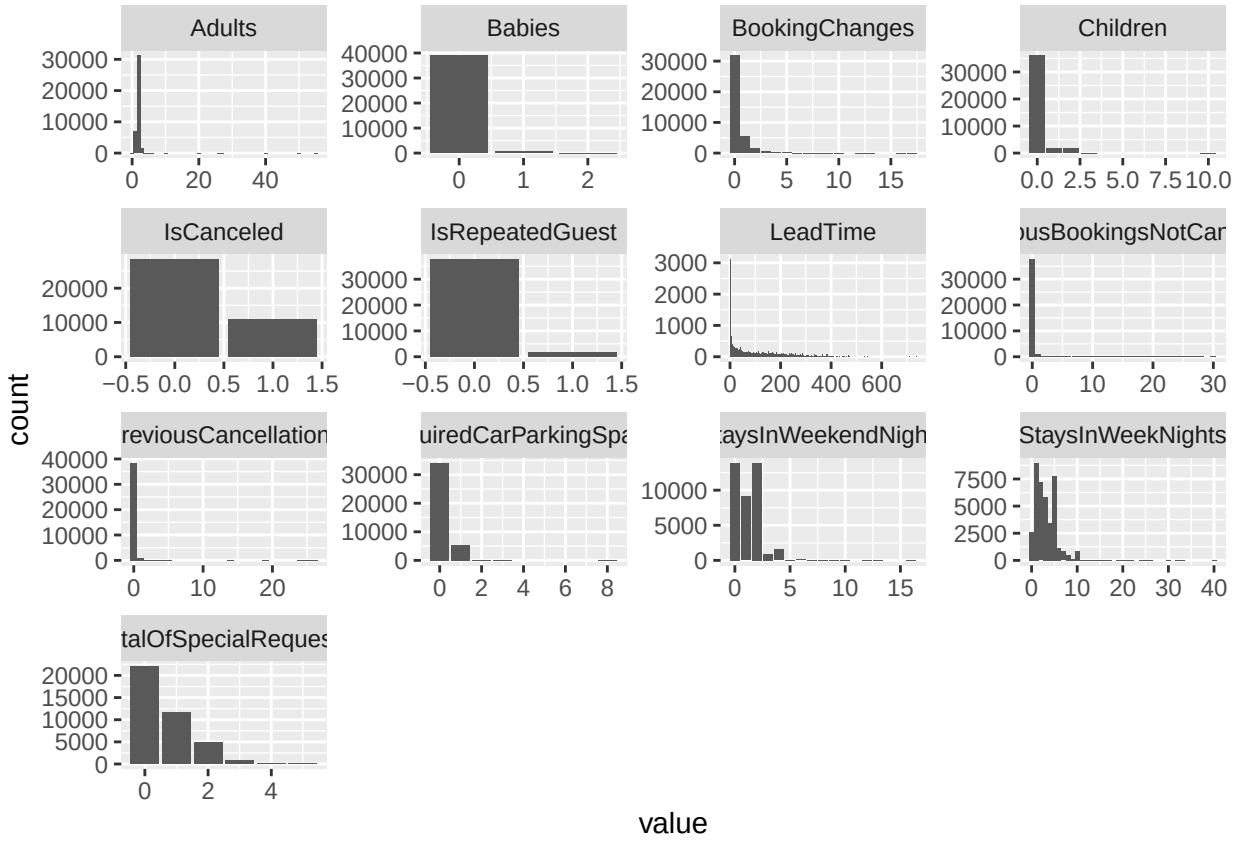

```
ggplot(hotels, aes(factor(DepositType), #bar plot describing the DepositType variable with IsCanceled f
        fill = factor(IsCanceled))) +geom_bar()+ xlab("Deposit Type\n 0= Non-Cancelled booking, 1= Canc
```




```
ggplot(hotels, aes(factor(CustomerType), #bar plot describing the CustomerType variable with IsCanceled
        fill = factor(IsCanceled))) +geom_bar()+ xlab("Customer Type\n 0= Non-Cancelled booking, 1= Can
```




```
hotels %>% #generalized bar plots with each of the variables
  keep(is.numeric) %>% #changing factors to numeric where possible
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_bar()
```



7. Descriptive Statistics and Associated Plots

```
numdata <- hotels[,c('IsCanceled','LeadTime','StaysInWeekendNights','StaysInWeekNights',
                     'PreviousCancellations','PreviousBookingsNotCanceled','BookingChanges',
                     'RequiredCarParkingSpaces')] #subsetting only the numerical columns
numdata$IsCanceled <- factor(numdata$IsCanceled, levels = c(0,1), labels = c('Not Canceled', 'Canceled')
newstay <- numdata[numdata$IsCanceled == 'Not Canceled',][,-1] #data set with all of the non-cancellati
newcancel <- numdata[numdata$IsCanceled == 'Canceled',][,-1] #data set with all of the cancellations
summary(newstay) #summary statistics of stay observations
```

```
##     LeadTime       StaysInWeekendNights StaysInWeekNights PreviousCancellations
##  Min.   :  0.00   Min.   : 0.000       Min.   : 0.000    Min.   :0.000000
##  1st Qu.:  6.00   1st Qu.: 0.000       1st Qu.: 1.000    1st Qu.:0.000000
##  Median : 39.00   Median : 1.000       Median : 3.000    Median :0.000000
##  Mean   : 79.69   Mean   : 1.142       Mean   : 3.025    Mean   :0.006802
##  3rd Qu.:132.00   3rd Qu.: 2.000       3rd Qu.: 5.000    3rd Qu.:0.000000
##  Max.   :737.00   Max.   :16.000       Max.   :40.000    Max.   :5.000000
##  PreviousBookingsNotCanceled BookingChanges    RequiredCarParkingSpaces
##  Min.   : 0.000              Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.000              1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 0.000              Median : 0.0000   Median :0.0000
##  Mean   : 0.173              Mean   : 0.3414   Mean   :0.1902
##  3rd Qu.: 0.000              3rd Qu.: 0.0000   3rd Qu.:0.0000
##  Max.   :30.000              Max.   :17.0000   Max.   :8.0000
```

```
summary(newcancel) #summary statistics of cancel observations
```

```
##     LeadTime     StaysInWeekendNights StaysInWeekNights PreviousCancellations
##  Min.   :  0.0   Min.   : 0.000       Min.   : 0.000    Min.   : 0.0000
##  1st Qu.: 45.0   1st Qu.: 0.000       1st Qu.: 2.000    1st Qu.: 0.0000
##  Median :109.0   Median : 1.000       Median : 3.000    Median : 0.0000
##  Mean   :128.8   Mean   : 1.336       Mean   : 3.443    Mean   : 0.3468
##  3rd Qu.:198.0   3rd Qu.: 2.000       3rd Qu.: 5.000    3rd Qu.: 0.0000
##  Max.   :471.0   Max.   :16.000       Max.   :40.000    Max.   :26.0000
##  PreviousBookingsNotCanceled BookingChanges    RequiredCarParkingSpaces
##  Min.   : 0.00000            Min.   : 0.0000   Min.   :0
##  1st Qu.: 0.00000            1st Qu.: 0.0000   1st Qu.:0
##  Median : 0.00000            Median : 0.0000   Median :0
##  Mean   : 0.01959            Mean   : 0.1533   Mean   :0
##  3rd Qu.: 0.00000            3rd Qu.: 0.0000   3rd Qu.:0
##  Max.   :27.00000            Max.   :16.0000   Max.   :0
```

```
numdata$IsCanceled <- as.numeric(numdata$IsCanceled)-1 #bringing back to numeric only for the next func
cor(numdata) #correlation matrix between all of the numeric variables and IsCanceled as numeric
```

```
##                               IsCanceled    LeadTime StaysInWeekendNights
## IsCanceled                    1.00000000  0.22645266          0.076344636
## LeadTime                      0.22645266  1.00000000          0.322805327
## StaysInWeekendNights          0.07634464  0.32280533          1.000000000
## StaysInWeekNights             0.07676576  0.38657921          0.712534947
## PreviousCancellations         0.11366616  0.09383440         -0.006760373
## PreviousBookingsNotCanceled -0.07366120 -0.10193216         -0.090932352
## BookingChanges               -0.11590875  0.07478166          0.054844031
## RequiredCarParkingSpaces     -0.24398288 -0.15128853         -0.091773289
##                               StaysInWeekNights PreviousCancellations
## IsCanceled                        0.076765756          0.113666164
## LeadTime                          0.386579211          0.093834404
## StaysInWeekendNights              0.712534947         -0.006760373
## StaysInWeekNights                 1.000000000         -0.006614897
## PreviousCancellations            -0.006614897          1.000000000
## PreviousBookingsNotCanceled      -0.086396111          0.023189819
## BookingChanges                    0.083981249         -0.026604530
## RequiredCarParkingSpaces         -0.103550224         -0.026974868
##                               PreviousBookingsNotCanceled BookingChanges
## IsCanceled                                 -7.366120e-02  -1.159087e-01
## LeadTime                                   -1.019322e-01   7.478166e-02
## StaysInWeekendNights                       -9.093235e-02   5.484403e-02
## StaysInWeekNights                          -8.639611e-02   8.398125e-02
## PreviousCancellations                       2.318982e-02  -2.660453e-02
## PreviousBookingsNotCanceled                 1.000000e+00  -7.966498e-05
## BookingChanges                             -7.966498e-05   1.000000e+00
## RequiredCarParkingSpaces                    4.704985e-02   6.245549e-02
##                               RequiredCarParkingSpaces
## IsCanceled                                -0.24398288
## LeadTime                                  -0.15128853
## StaysInWeekendNights                      -0.09177329
## StaysInWeekNights                         -0.10355022
```

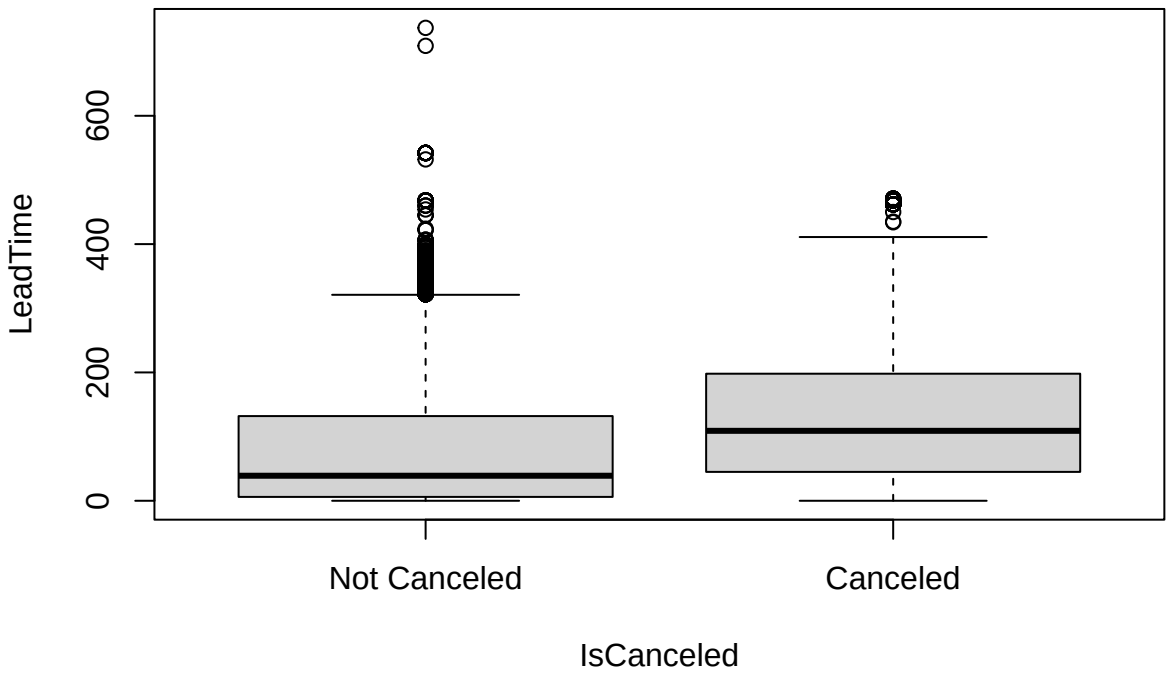
```
## PreviousCancellations                    -0.02697487
## PreviousBookingsNotCanceled               0.04704985
## BookingChanges                            0.06245549
## RequiredCarParkingSpaces                  1.00000000
```

```
numdata$IsCanceled <- factor(numdata$IsCanceled, levels = c(0,1), labels = c('Not Canceled', 'Canceled'
```

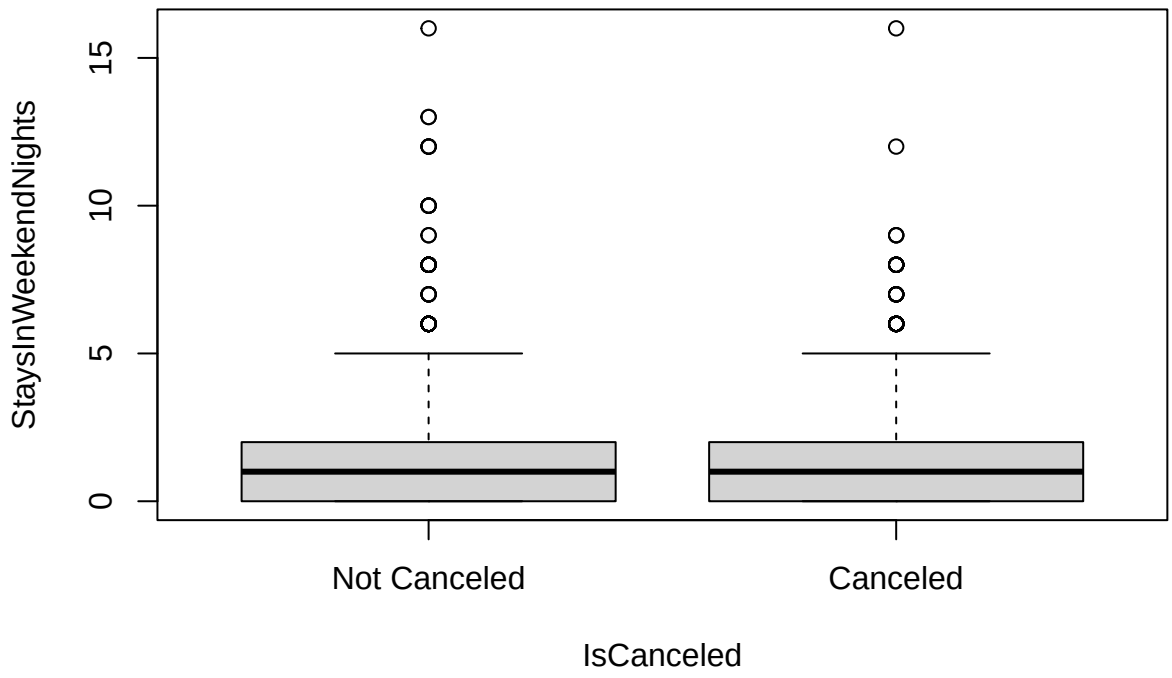
Descriptive Statistics to note: - Lead time appears much larger in cancellation data set

- IsRepeatedGuest has a significantly higher mean in stays
- PreviousCancellations has a significantly higher mean in canceled
- PreviousBookingsNotCanceled has a significantly higher mean in stays
- BookingChanges has a higher mean in stays

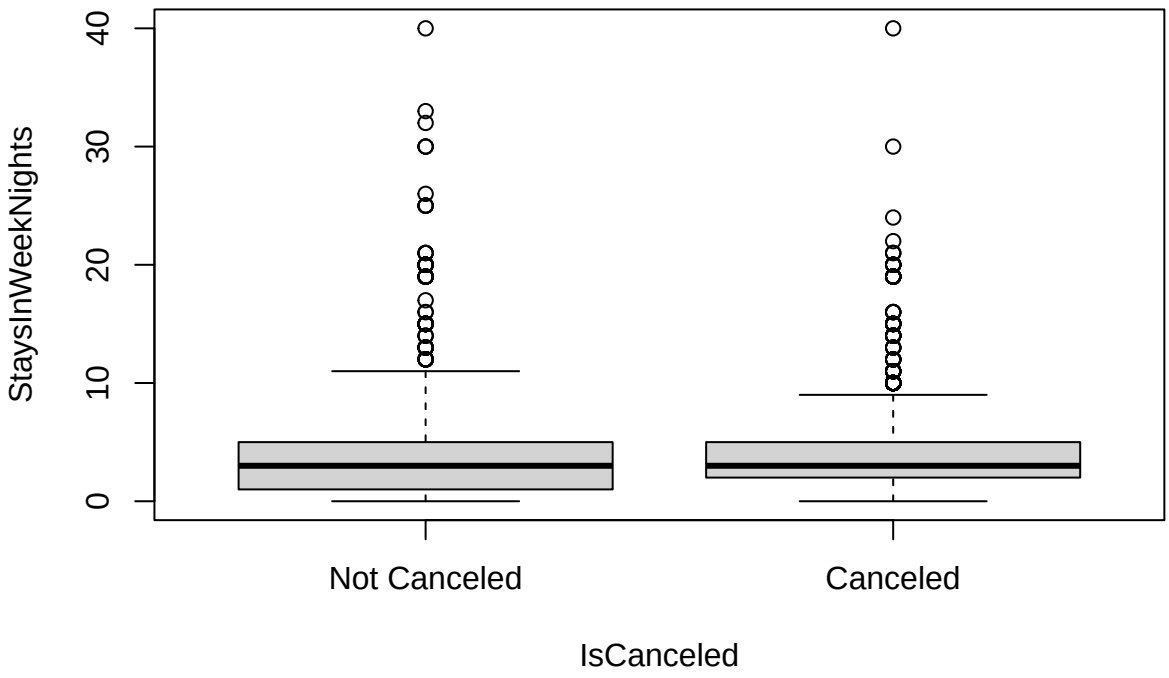
```
boxplot(LeadTime ~ IsCanceled, data = numdata) #LeadTime
```



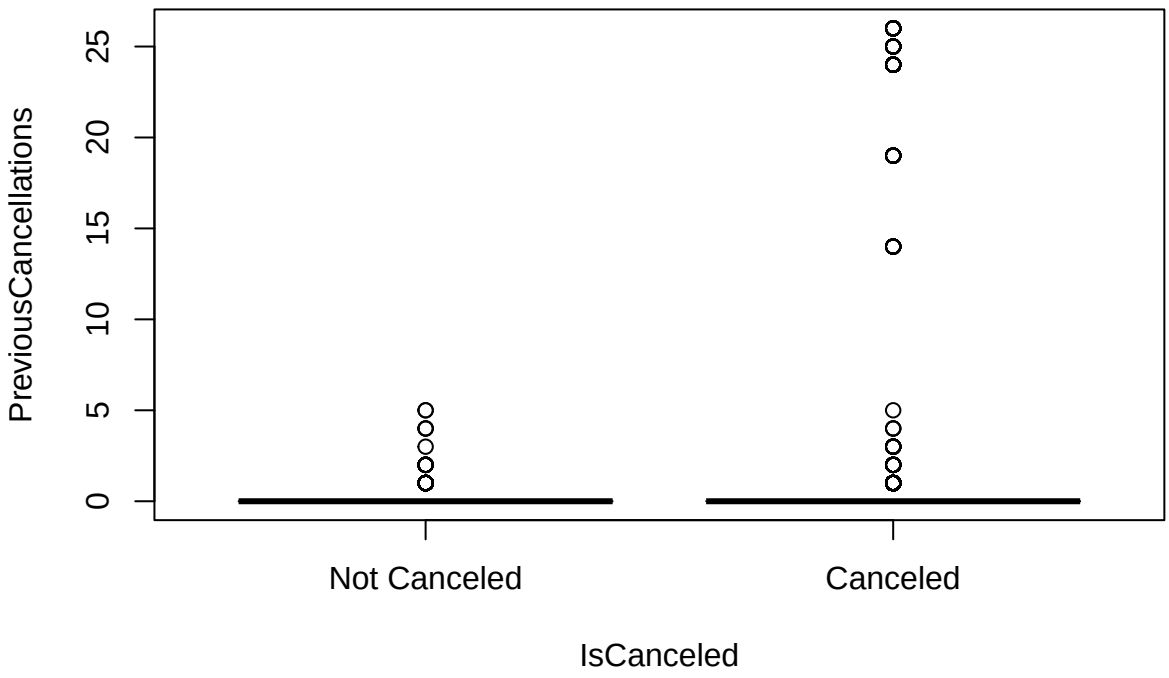

```
boxplot(StaysInWeekendNights ~ IsCanceled, data = numdata) #StaysInWeekendNights
```



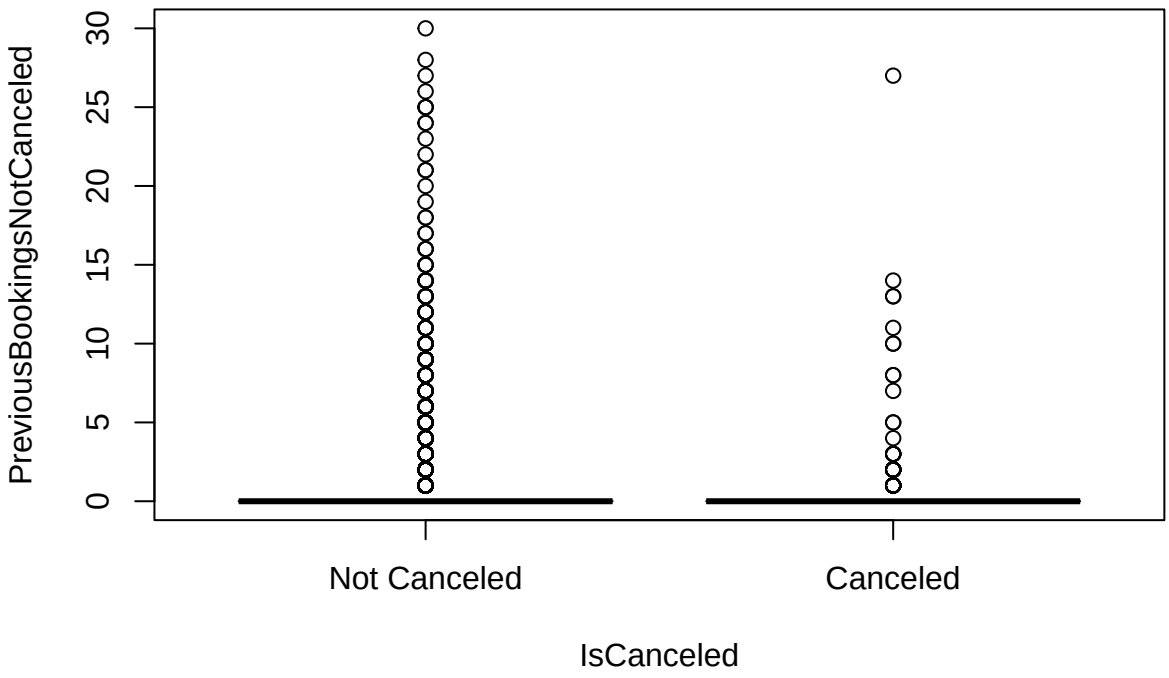

```
boxplot(StaysInWeekNights ~ IsCanceled, data = numdata) #StaysInWeekNights
```



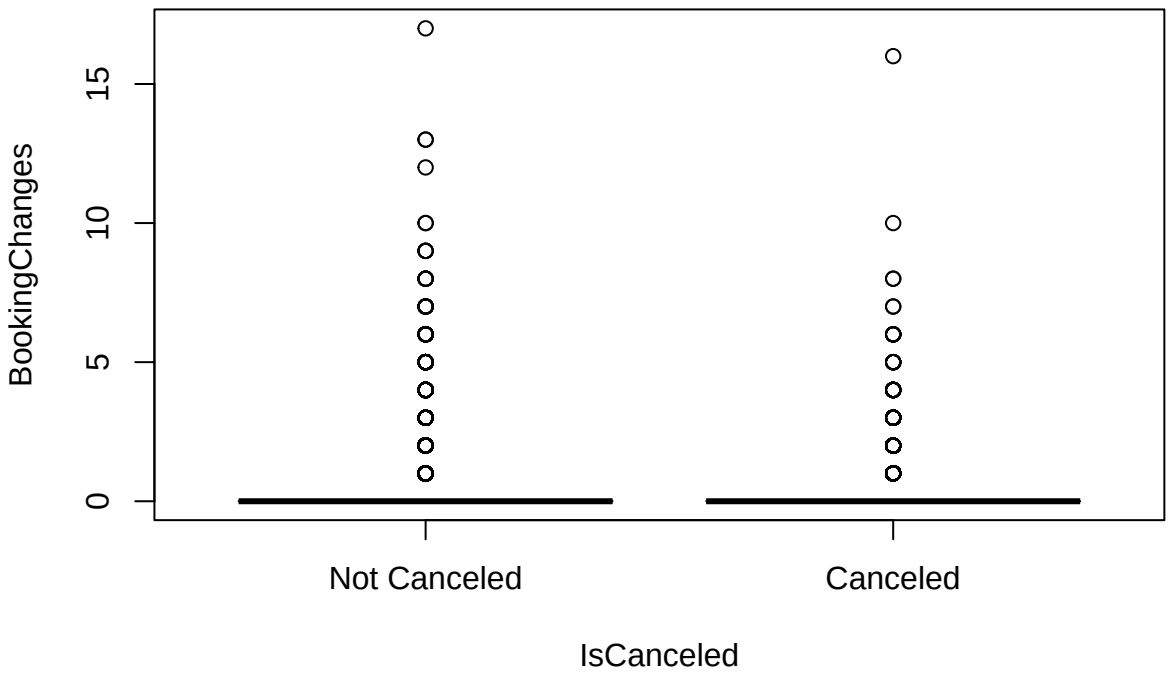

```
boxplot(PreviousCancellations ~ IsCanceled, data = numdata) #PreviousCancellations
```



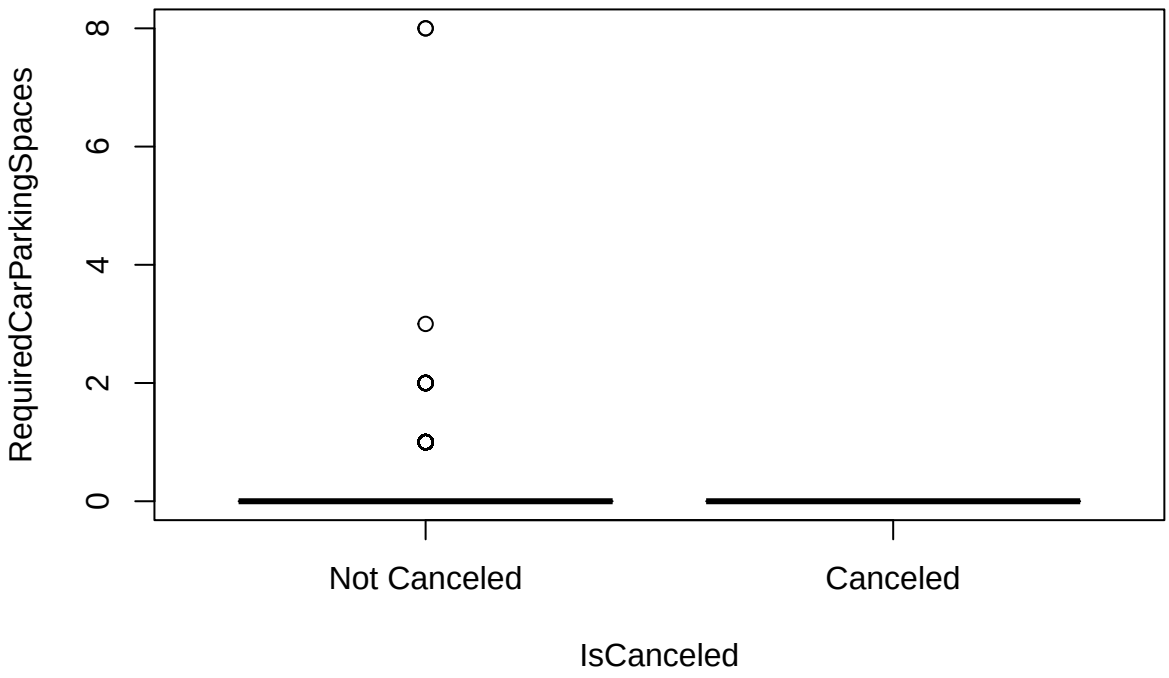

```
boxplot(PreviousBookingsNotCanceled ~ IsCanceled, data = numdata) #PreviousBookingsNotCanceled
```




```
boxplot(BookingChanges ~ IsCanceled, data = numdata) #BookingChanges
```

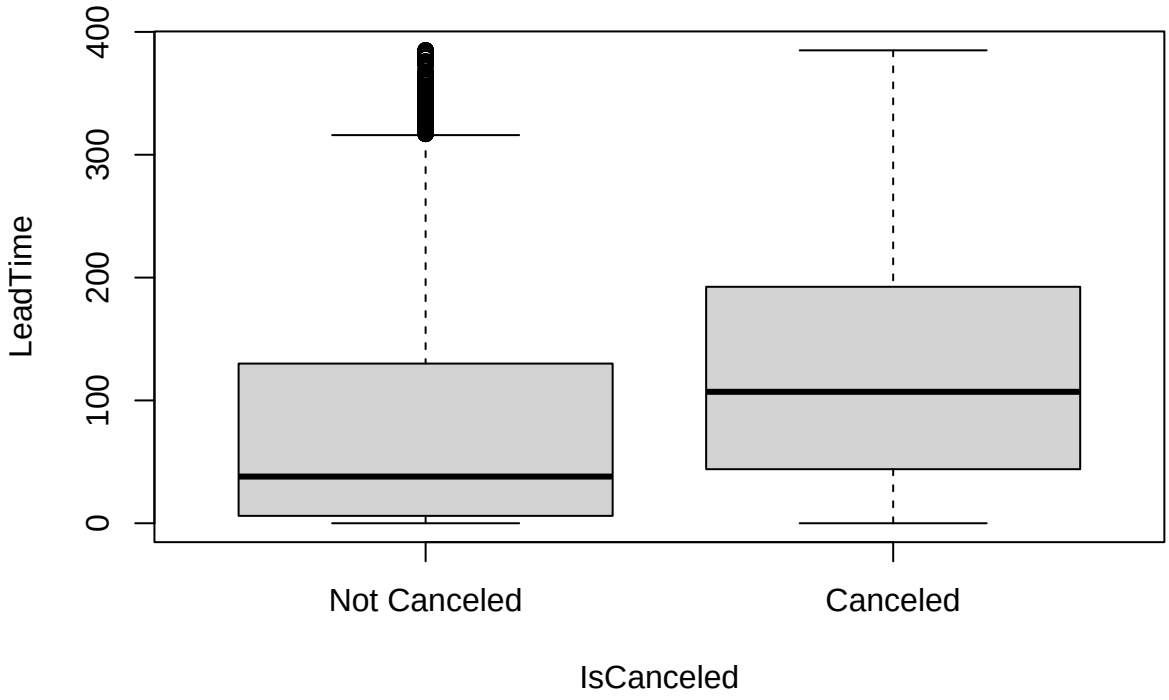



```
boxplot(RequiredCarParkingSpaces ~ IsCanceled, data = numdata) #RequiredParkingSpaces
```

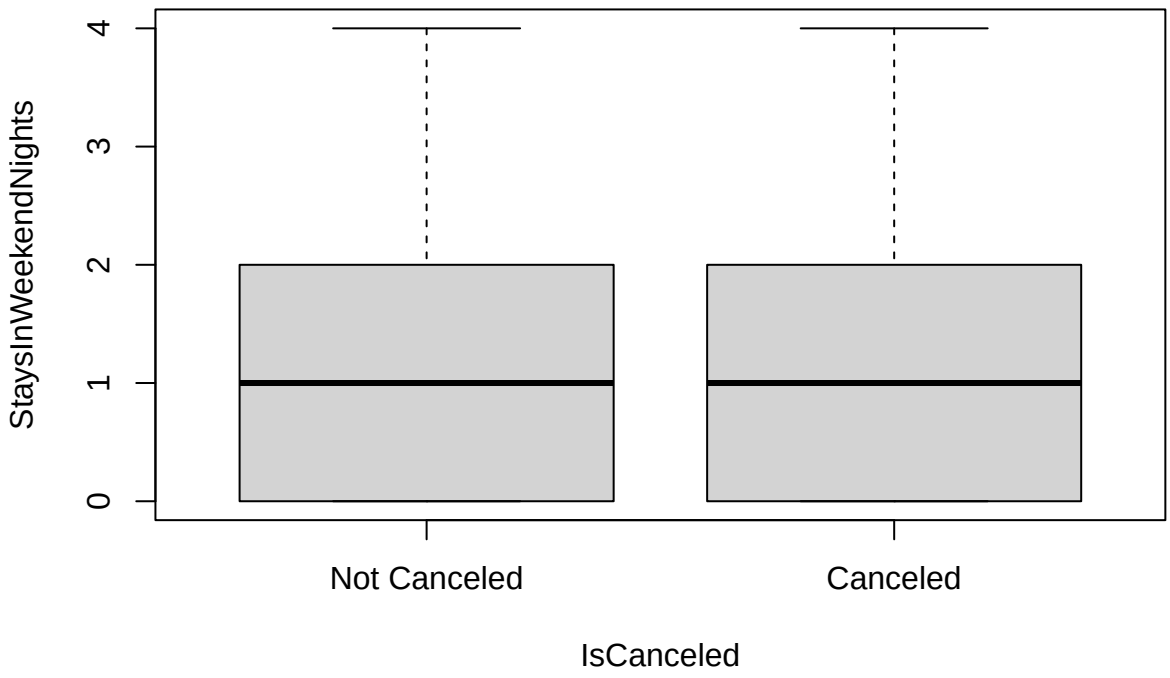


Modified box plots to Get a Better Idea of Variable Distribution (temporily remove outliers 3 standard deviations away from mean)

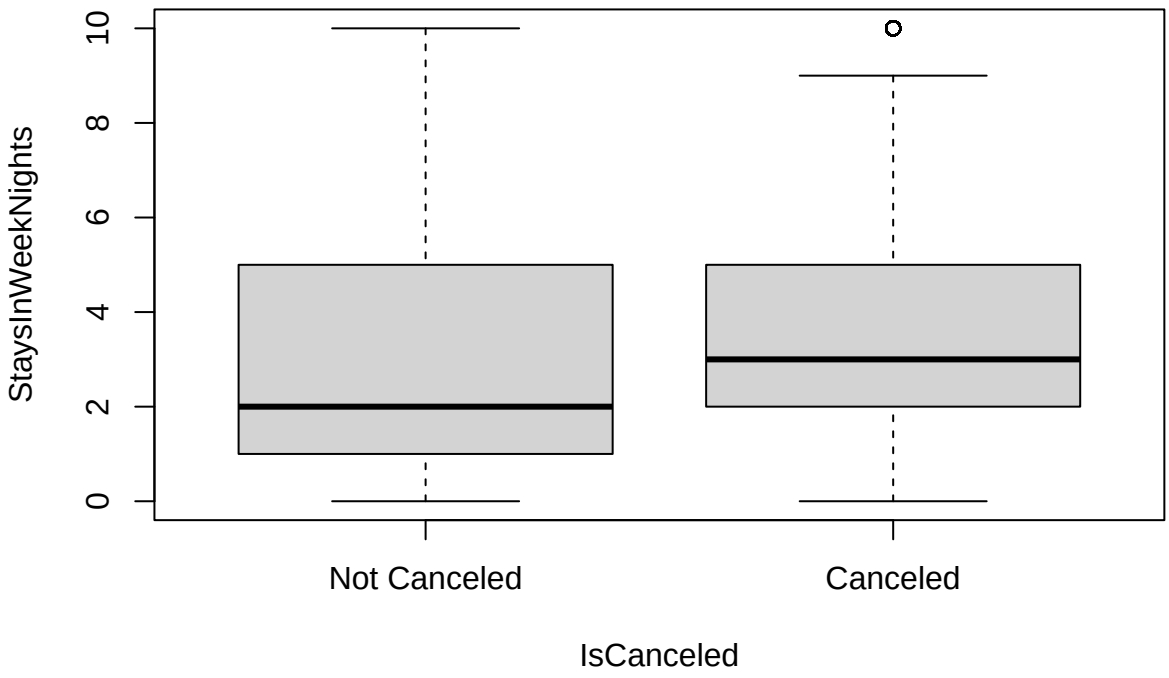
```
boxplot(LeadTime ~ IsCanceled, data = numdata[numdata$LeadTime<mean(numdata$LeadTime)+3*sd(numdata$Lead
```



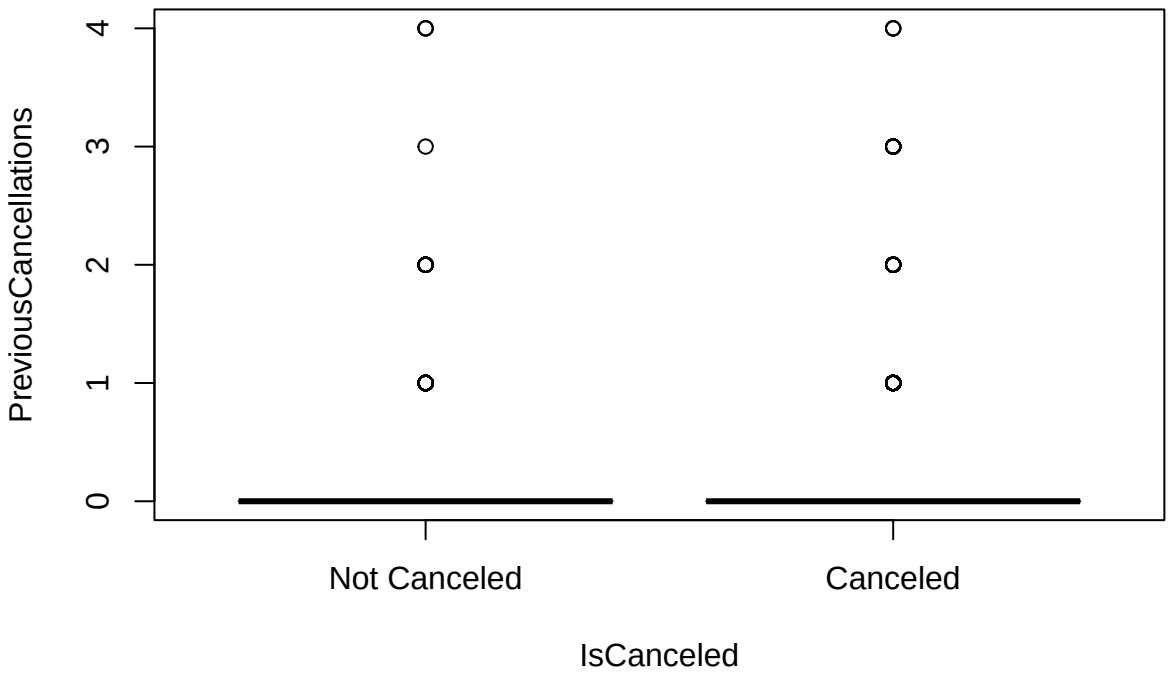

```
boxplot(StaysInWeekendNights ~ IsCanceled, data = numdata[numdata$StaysInWeekendNights < mean(numdata$S
                                                          3*sd(numdata$StaysInWeekendNights),]) #Stay
```



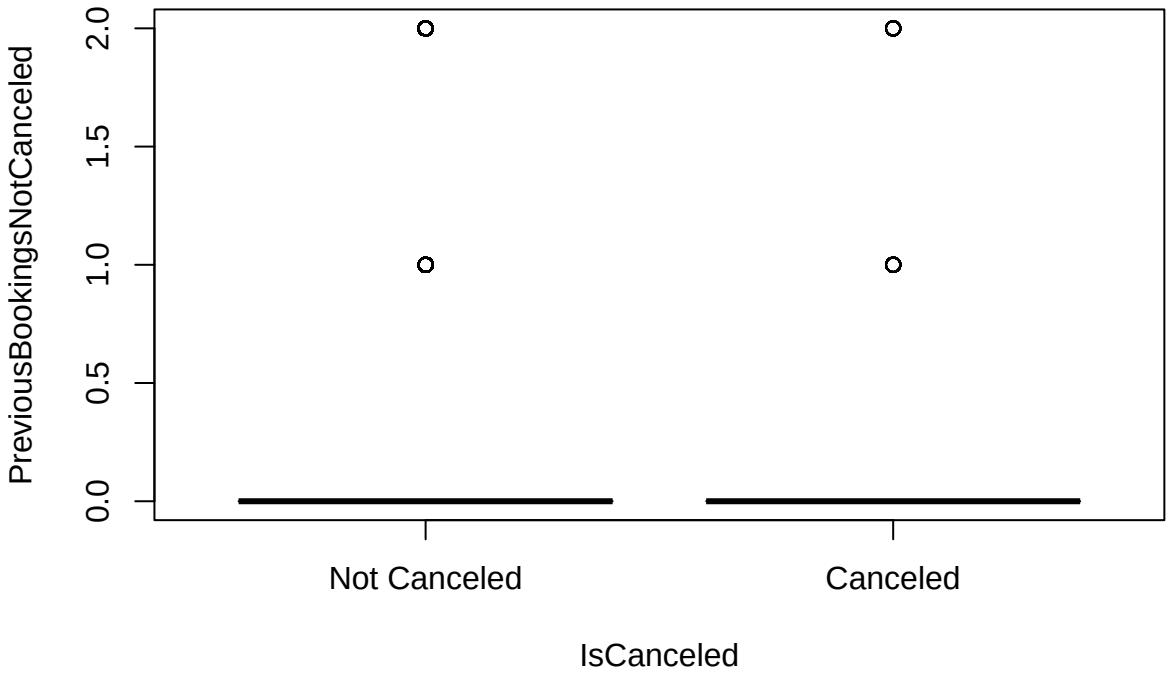

```
boxplot(StaysInWeekNights ~ IsCanceled, data = numdata[numdata$StaysInWeekNights < mean(numdata$StaysInW
                                                       3*sd(numdata$StaysInWeekNights),]) #StaysInWee
```



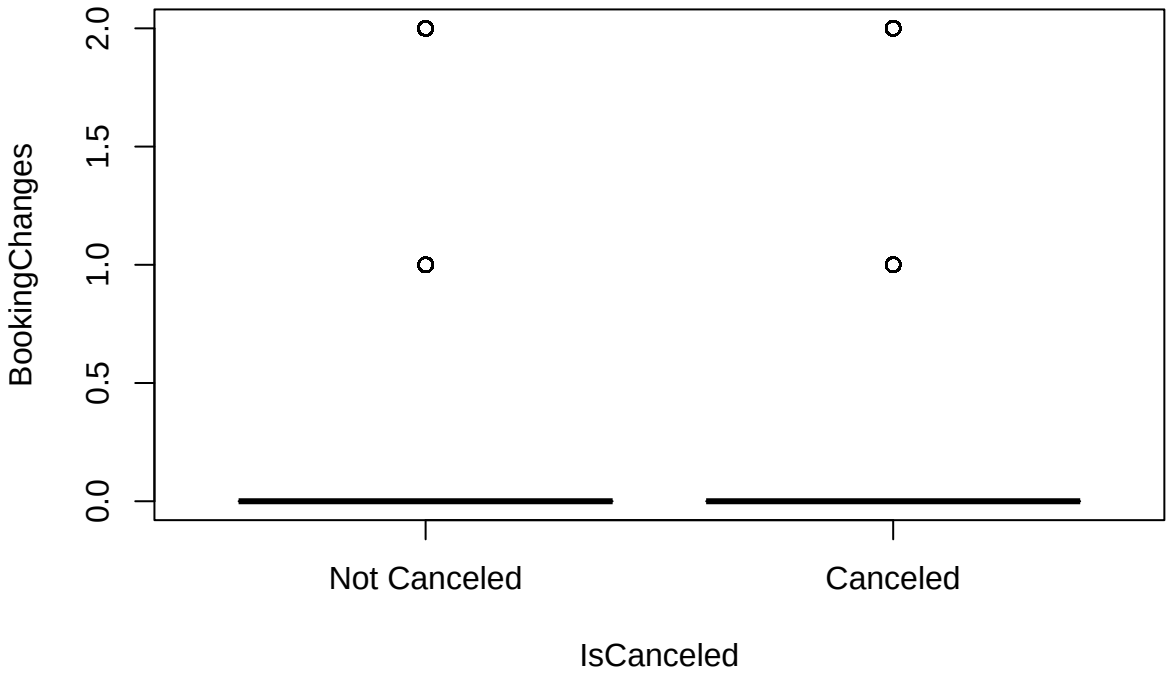

```
boxplot(PreviousCancellations ~ IsCanceled, data = numdata[numdata$PreviousCancellations < mean(numdata$
                                                          3*sd(numdata$PreviousCancellations),]) #Pr
```



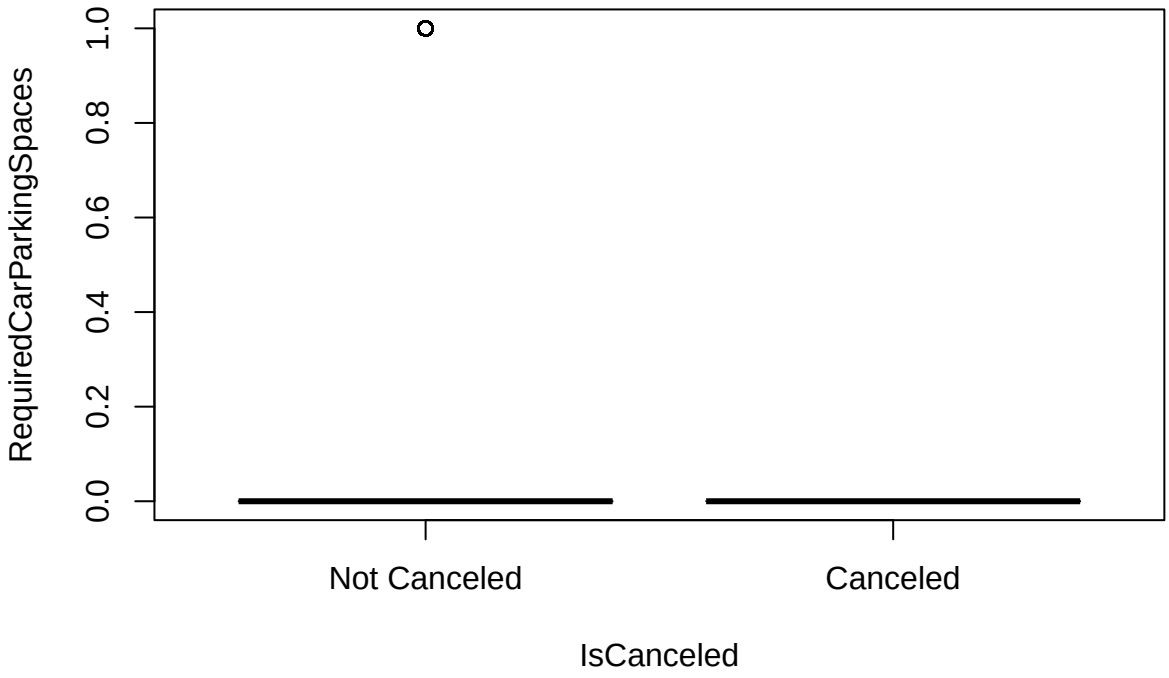

```
boxplot(PreviousBookingsNotCanceled ~ IsCanceled, data = numdata[numdata$PreviousBookingsNotCanceled < m
                                                                 3*sd(numdata$PreviousBookingsNotCanc
```




```
boxplot(BookingChanges ~ IsCanceled, data = numdata[numdata$BookingChanges < mean(numdata$BookingChanges
                                                    3*sd(numdata$BookingChanges),]) #BookingChanges
```



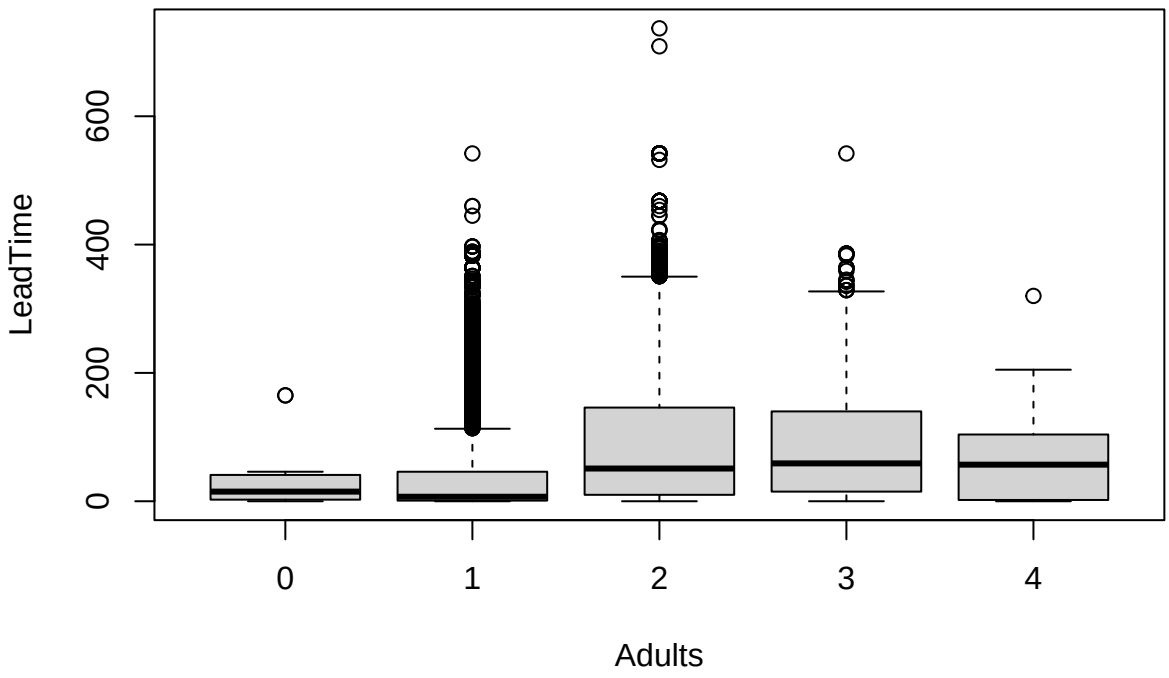

```
boxplot(RequiredCarParkingSpaces ~ IsCanceled, data = numdata[numdata$RequiredCarParkingSpaces <
                                                              mean(numdata$RequiredCarParkingSpaces)+
                                                              3*sd(numdata$RequiredCarParkingSpaces),]
```



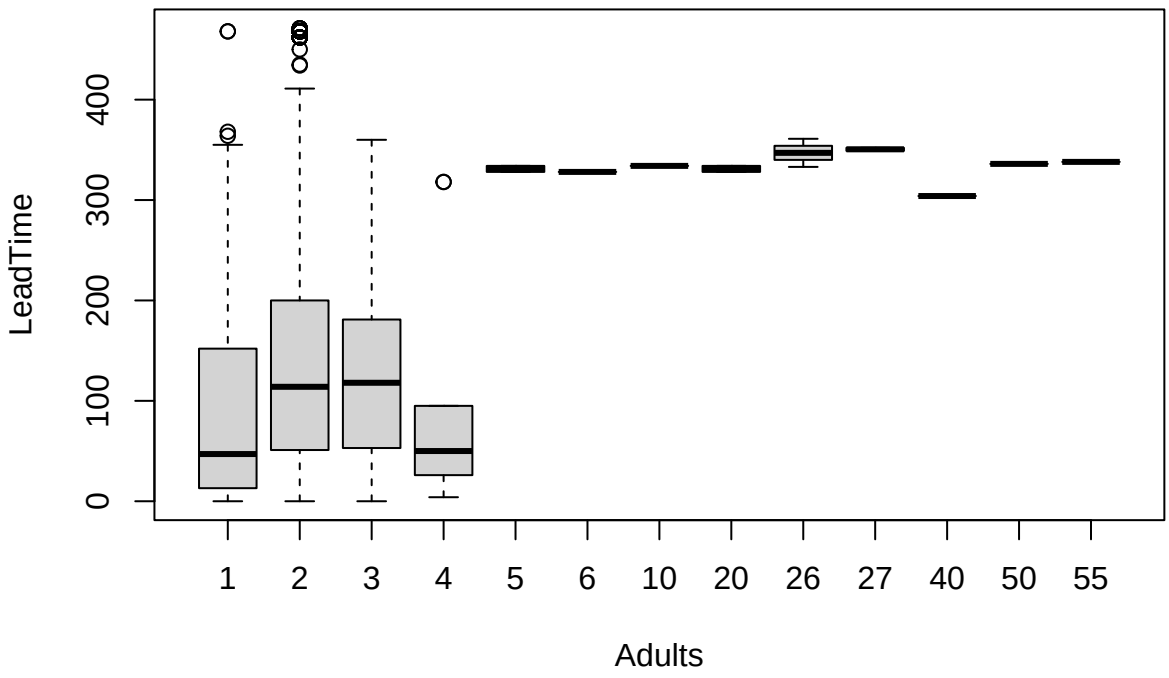
Looking at Boxplots of LeadTime in Correspondance with other factors

```
stayfull <- hotels[hotels$IsCanceled == 0,][,-20][,-1] #data set with all of the non-cancellations
cancelfull <- hotels[hotels$IsCanceled == 1,][,-20][,-1] #data set with all of the cancellations
```

```
boxplot(LeadTime ~ Adults, data = stayfull) #stays
```




```
boxplot(LeadTime ~ Adults, data = cancelfull) #cancels
```



8. Running some models

```
linmod <- lm(as.numeric(IsCanceled) ~ ., data = numdata) #running a linear model. Won't be super accura
summary(linmod)
```

```
##
## Call:
## lm(formula = as.numeric(IsCanceled) ~ ., data = numdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7371 -0.2951 -0.2334  0.4952  1.9468
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.259e+00  3.906e-03 322.227  < 2e-16 ***
## LeadTime                     9.028e-04  2.403e-05  37.574  < 2e-16 ***
## StaysInWeekendNights         5.537e-03  2.651e-03   2.089 0.036726 *
## StaysInWeekNights           -4.399e-03  1.272e-03  -3.460 0.000542 ***
## PreviousCancellations        2.935e-02  1.587e-03  18.494  < 2e-16 ***
## PreviousBookingsNotCanceled -2.254e-02  2.283e-03  -9.875  < 2e-16 ***
## BookingChanges              -7.041e-02  2.929e-03 -24.035  < 2e-16 ***
## RequiredCarParkingSpaces    -2.614e-01  6.148e-03 -42.525  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4211 on 39588 degrees of freedom
## Multiple R-squared:   0.12,  Adjusted R-squared:  0.1198
## F-statistic: 771.2 on 7 and 39588 DF,  p-value: < 2.2e-16
```

```
logmod <- glm(IsCanceled ~ ., family = binomial(link="logit"), data = numdata[-13920,]) #major outlier
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logmod)
```

```
##
## Call:
## glm(formula = IsCanceled ~ ., family = binomial(link = "logit"),
##     data = numdata[-13920, ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1655  -0.8241  -0.6386   1.0656   5.1439
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.034e+00  2.213e-02 -46.720  < 2e-16 ***
## LeadTime                     3.946e-03  1.322e-04  29.844  < 2e-16 ***
## StaysInWeekendNights         2.886e-02  1.537e-02   1.878  0.06036 .
## StaysInWeekNights           -2.011e-02  7.272e-03  -2.766  0.00568 **
## PreviousCancellations        3.599e+00  1.640e-01  21.942  < 2e-16 ***
## PreviousBookingsNotCanceled -1.220e+00  7.272e-02 -16.783  < 2e-16 ***
## BookingChanges              -5.280e-01  2.424e-02 -21.786  < 2e-16 ***
## RequiredCarParkingSpaces    -1.744e+01  8.424e+01  -0.207  0.83596
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46936  on 39594  degrees of freedom
## Residual deviance: 39243  on 39587  degrees of freedom
## AIC: 39259
##
## Number of Fisher Scoring iterations: 17
```

```
logmod2 <- glm(IsCanceled ~ LeadTime, family = binomial(link="logit"), data = numdata) #just checking o
summary(logmod2)
```

```
##
## Call:
## glm(formula = IsCanceled ~ LeadTime, family = binomial(link = "logit"),
##     data = numdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1386  -0.7844  -0.6639   1.2458   1.8223
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.4495134  0.0168945  -85.80   <2e-16 ***
## LeadTime     0.0049242  0.0001131   43.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46938  on 39595  degrees of freedom
## Residual deviance: 44998  on 39594  degrees of freedom
## AIC: 45002
##
## Number of Fisher Scoring iterations: 4
```

```
model_performance(logmod) #used to find psuedo r-squared of logistics
```

```
## # Indices of model performance
##
## AIC       |       BIC | Tjur's R2 |  RMSE | Sigma | Log_loss | Score_log | Score_spherical |   PCP
## ------------------------------------------------------------------------------------------------
## 39258.574 | 39327.266 |     0.161 | 0.411 | 0.996 |    0.496 |      -Inf |           0.002 | 0.662
```

```
anova(logmod, test = "Chisq") #analyzing chi-squared results
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: IsCanceled
##
## Terms added sequentially (first to last)
##
##
##                              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                         39594      46936
## LeadTime                      1  1940.86     39593      44995 < 2.2e-16 ***
## StaysInWeekendNights          1     2.35     39592      44993 0.1248955
## StaysInWeekNights             1    12.20     39591      44981 0.0004786 ***
## PreviousCancellations         1  1098.29     39590      43882 < 2.2e-16 ***
## PreviousBookingsNotCanceled   1   928.79     39589      42953 < 2.2e-16 ***
## BookingChanges                1   762.10     39588      42191 < 2.2e-16 ***
## RequiredCarParkingSpaces      1  2948.77     39587      39243 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#vif(linmod) #no colinarity!
#vif(logmod) #none here either
```

Predictions:

```r
linout <- predict(linmod, numdata[,-1]) #prediction data with linear model
linout <- data.frame(linout) #merging it into data frame
linout$data <- 0 #creating new column of 0's
#linout$data[linout$linout > as.numeric(median(linout$linout))] <- 1
linout$data[linout$linout > 1.5] <- 1 #1.5 is the halfway point in the factored IsCanceled data; make t
linout$data <- factor(linout$data, levels = c(0,1), labels = c('Not Canceled', 'Canceled')) #changing n

confusionMatrix(as.factor(linout$data),numdata$IsCanceled) #creating a confusion matrix to analyze resu
```

```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     Not Canceled Canceled
##   Not Canceled        27717     9824
##   Canceled              802     1253
##
##                Accuracy : 0.7316
##                  95% CI : (0.7272, 0.736)
##     No Information Rate : 0.7202
##     P-Value [Acc > NIR] : 2.051e-07
##
##                   Kappa : 0.1132
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9719
##             Specificity : 0.1131
##          Pos Pred Value : 0.7383
##          Neg Pred Value : 0.6097
##              Prevalence : 0.7202
##          Detection Rate : 0.7000
##    Detection Prevalence : 0.9481
##       Balanced Accuracy : 0.5425
##
##        'Positive' Class : Not Canceled
##
```

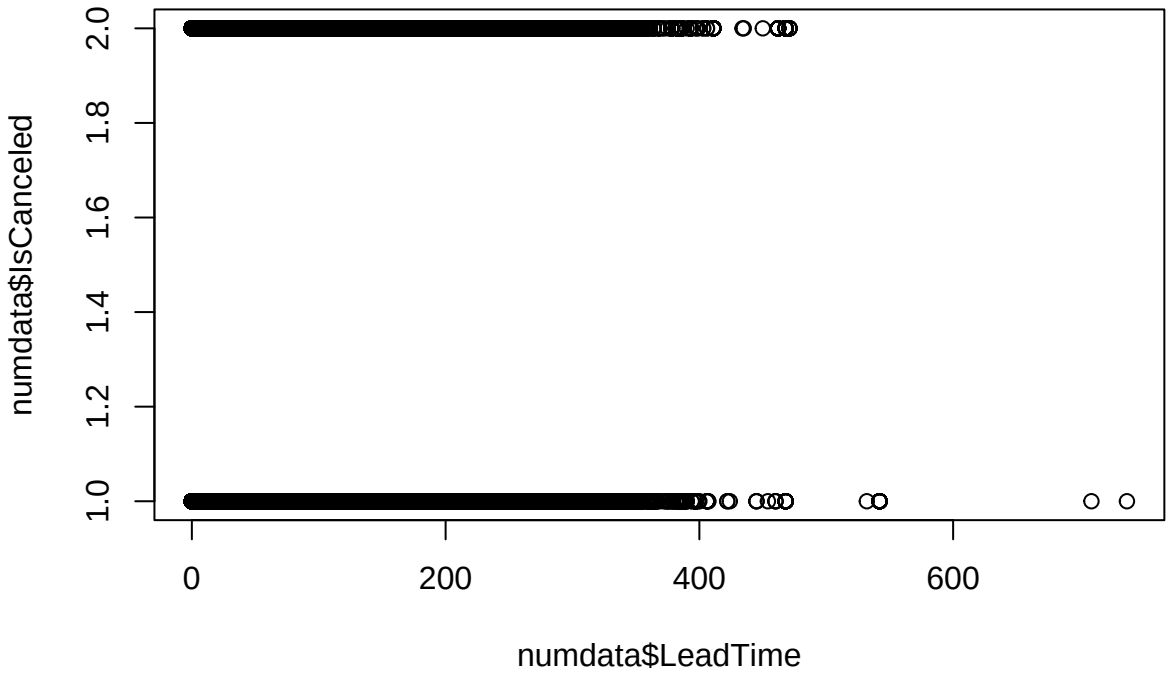
```r
logout <- predict(logmod, numdata[,-1]) #prediction data with logistic model
logout <- data.frame(logout) #merging it into data frame
logout$data <- 0 #creating new column of 0's
#logout$data[logout$logout > as.numeric(median(logout$logout))] <- 1
logout$data[logout$logout > 1.5] <- 1 #1.5 is halfway point in the factored IsCanceled data; make the t
logout$data <- factor(logout$data, levels = c(0,1), labels = c('Not Canceled', 'Canceled')) #changing n

confusionMatrix(as.factor(logout$data),numdata$IsCanceled) #creating confusion matrix to analyze result
```

```
## Confusion Matrix and Statistics
```

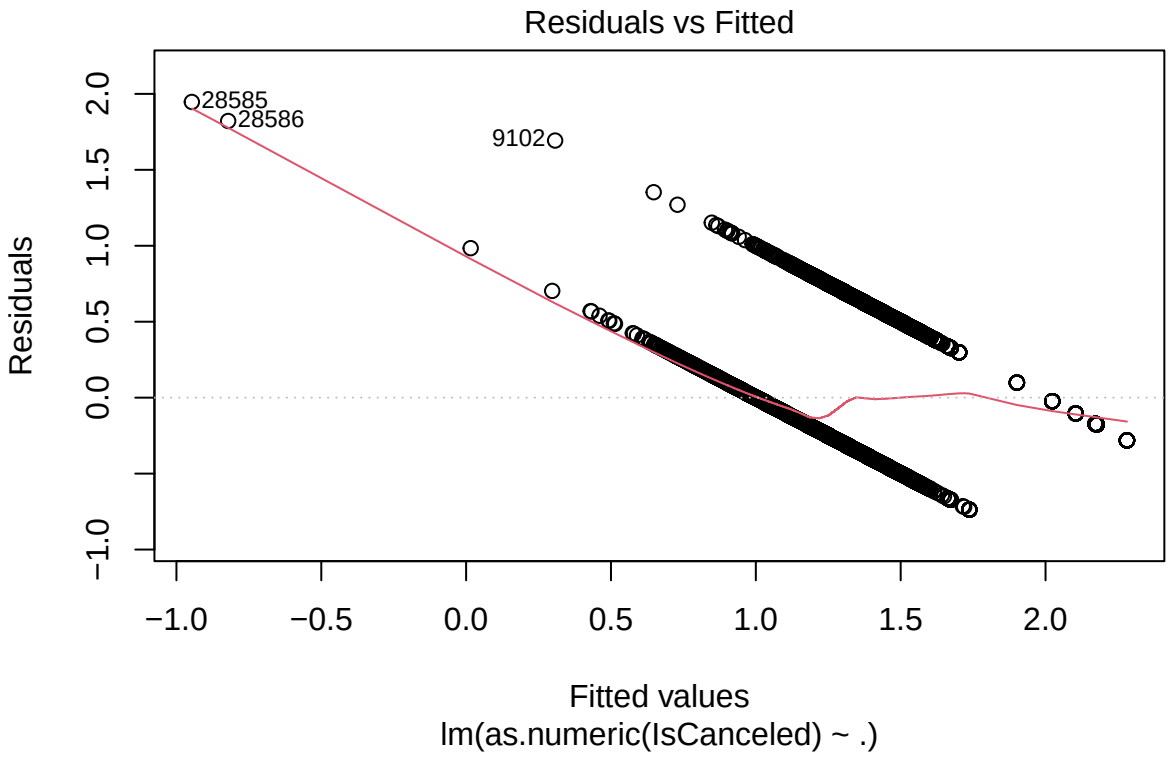
```
##
##               Reference
## Prediction     Not Canceled Canceled
##   Not Canceled        28507    10204
##   Canceled               12      873
##
##                Accuracy : 0.742
##                  95% CI : (0.7377, 0.7463)
##     No Information Rate : 0.7202
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1091
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.99958
##             Specificity : 0.07881
##          Pos Pred Value : 0.73641
##          Neg Pred Value : 0.98644
##              Prevalence : 0.72025
##          Detection Rate : 0.71995
##    Detection Prevalence : 0.97765
##       Balanced Accuracy : 0.53920
##
##        'Positive' Class : Not Canceled
##
```

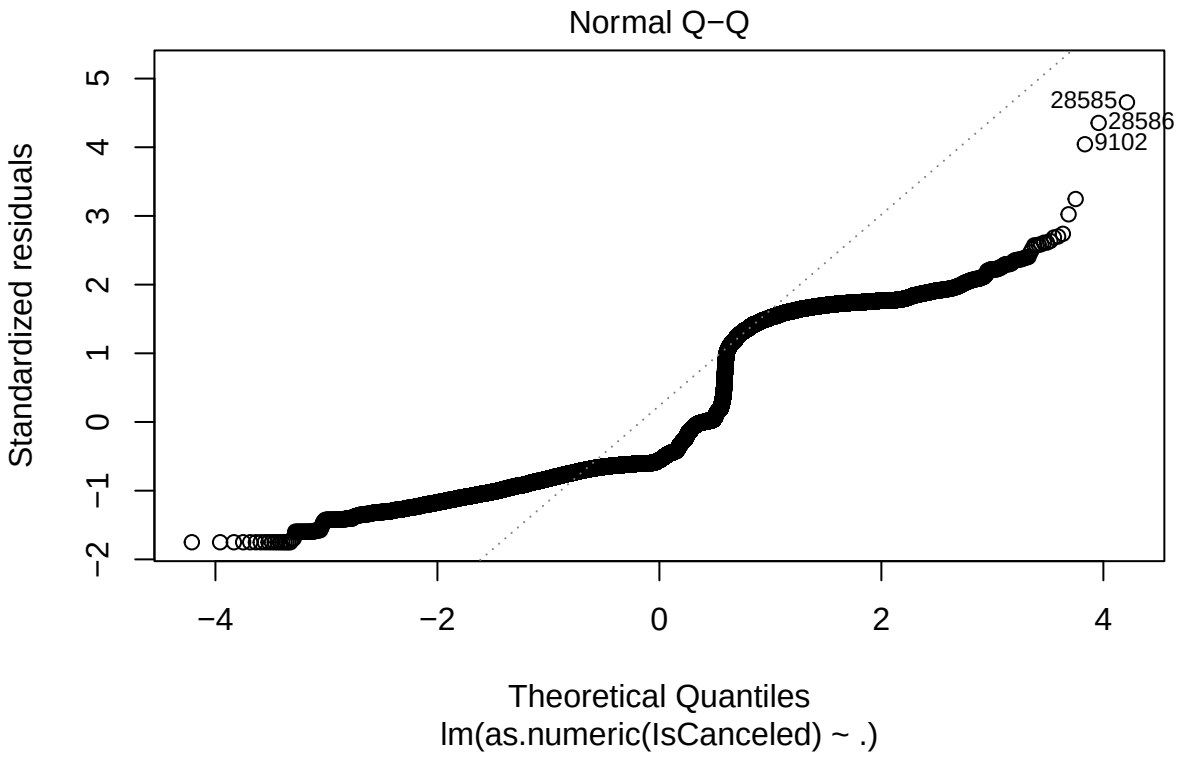
Analytical Plots

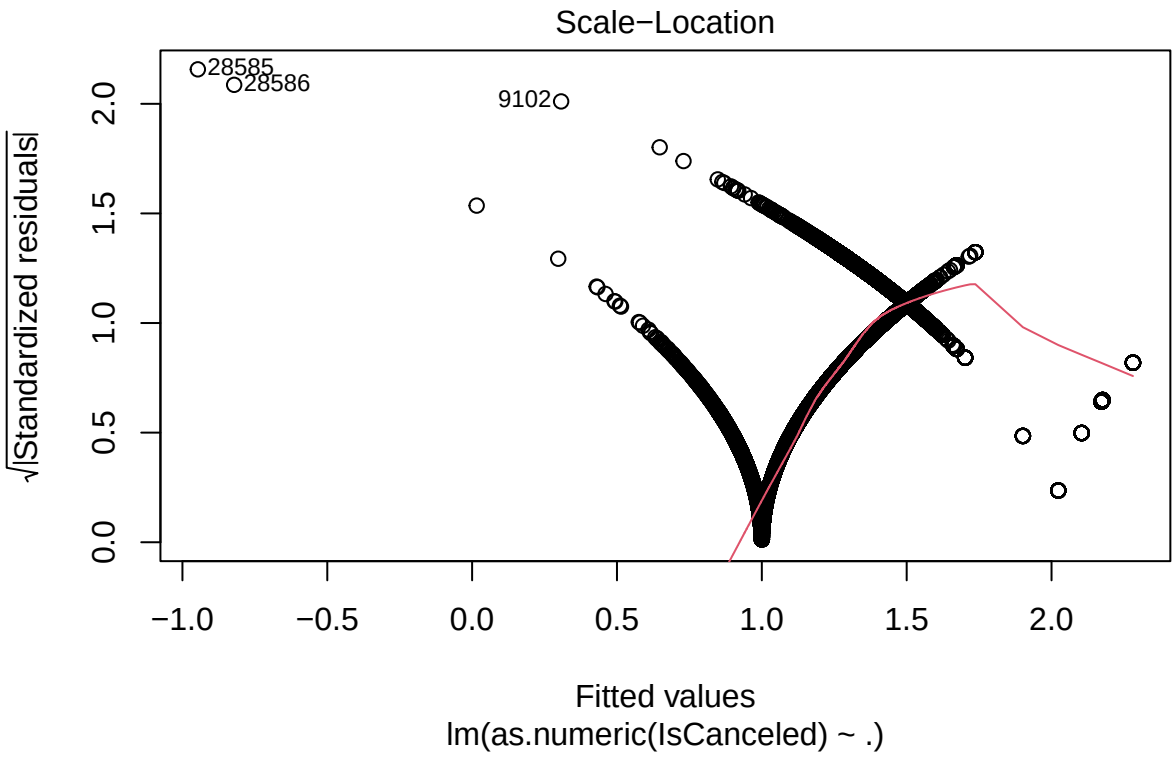
```
plot(x = numdata$LeadTime, y = numdata$IsCanceled) #ploting canceled vs leadtime
```

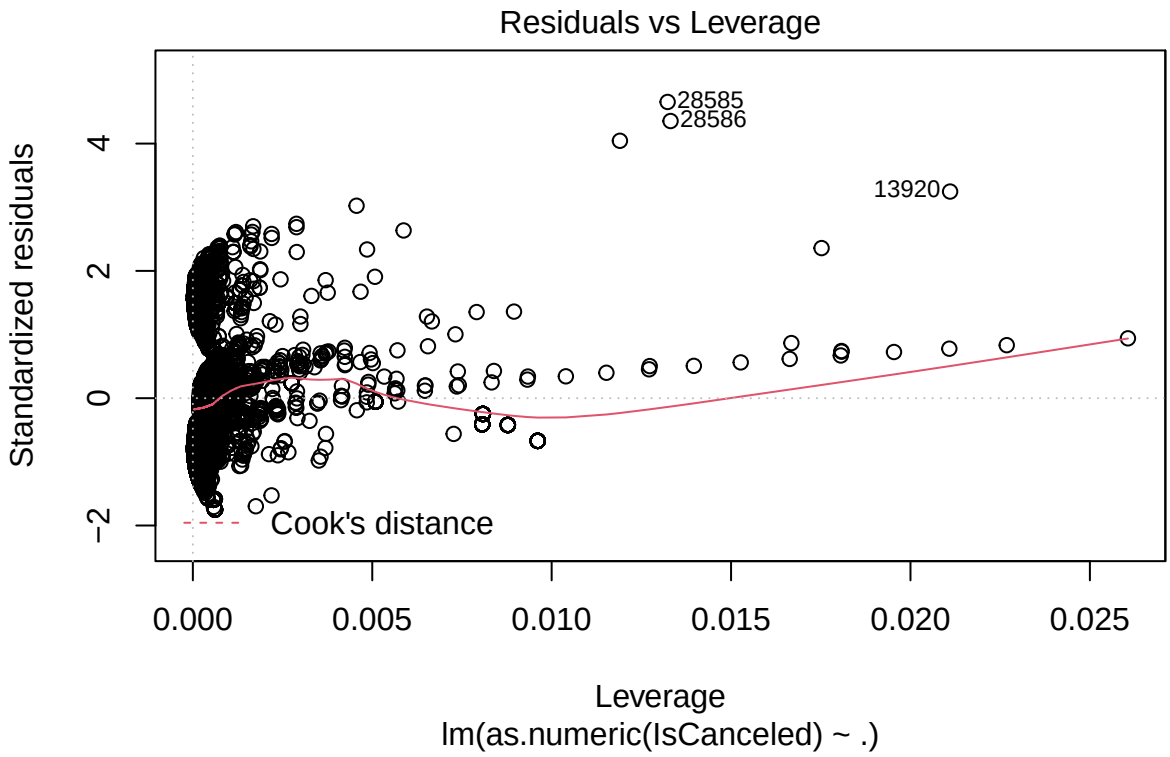



```
plot(linmod) #plotting linear model residuals
```

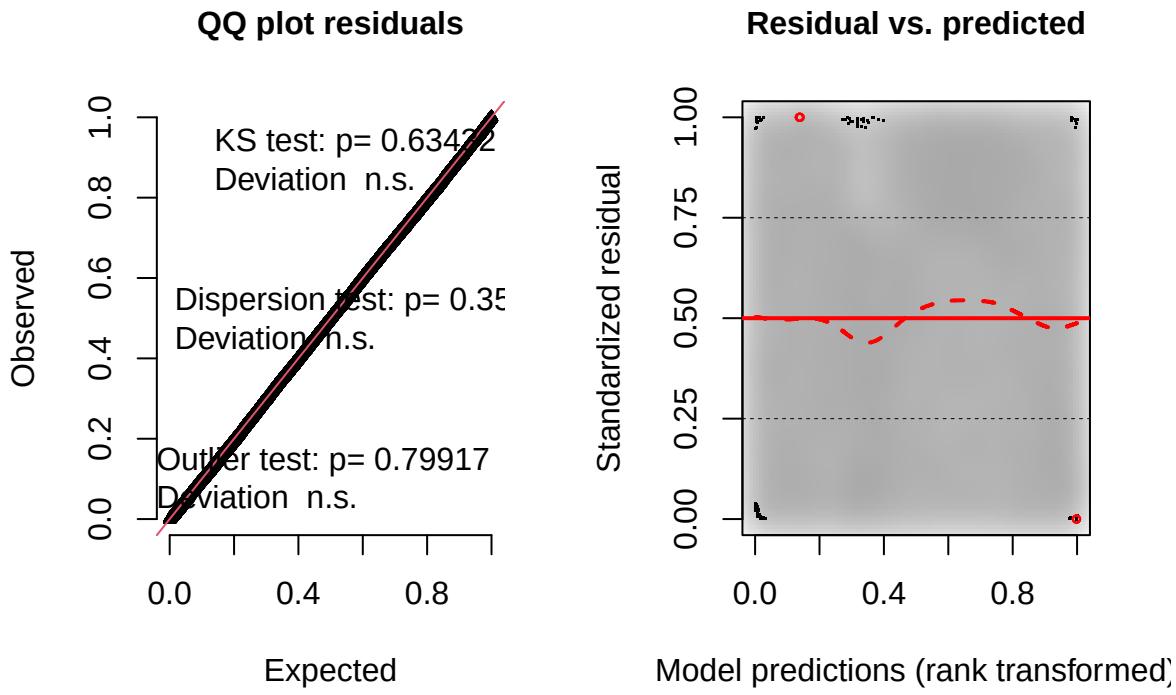

Residuals vs Fitted
Residuals
Fitted values
lm(as.numeric(IsCanceled) ~ .)
28585
28586
9102
−1.0
0.0
0.5
1.0
1.5
2.0
−0.5


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(as.numeric(IsCanceled) ~ .)
28585
28586
9102
−2
−1
0
1
2
3
4
5
−4
−2
0
2
4

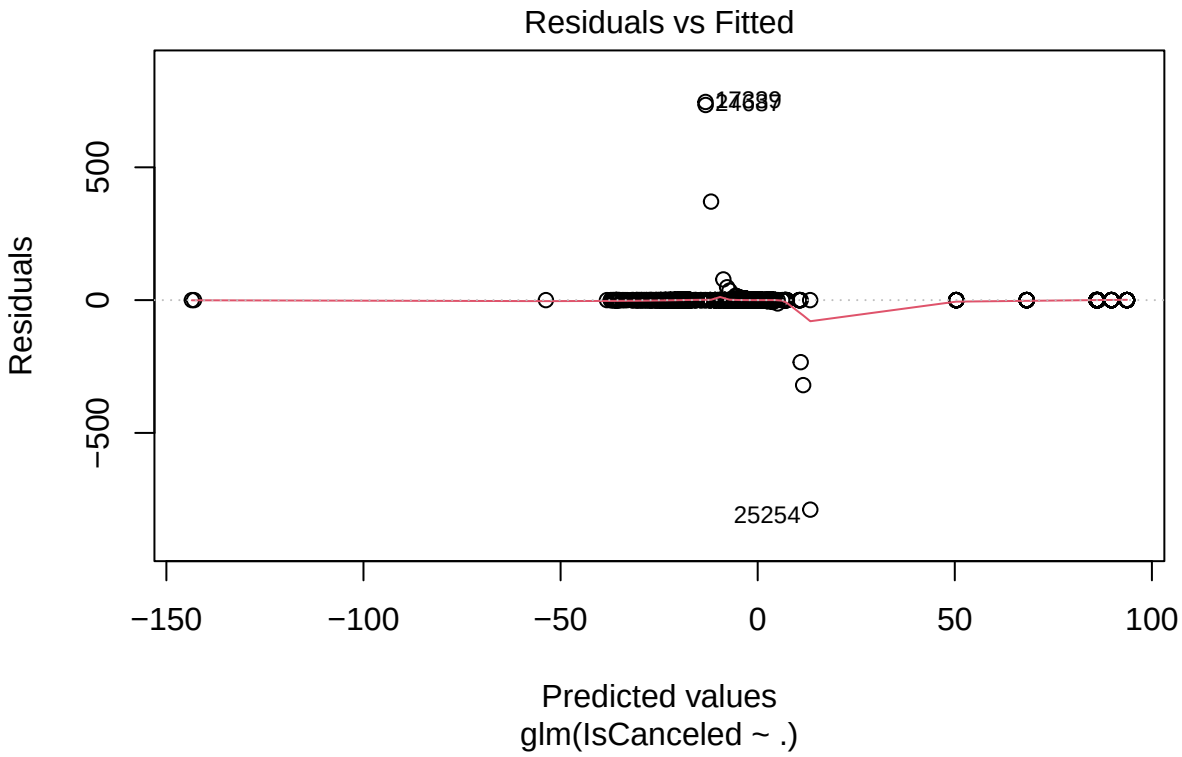

Scale−Location
28585
28586
9102
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Fitted values
lm(as.numeric(IsCanceled) ~ .)

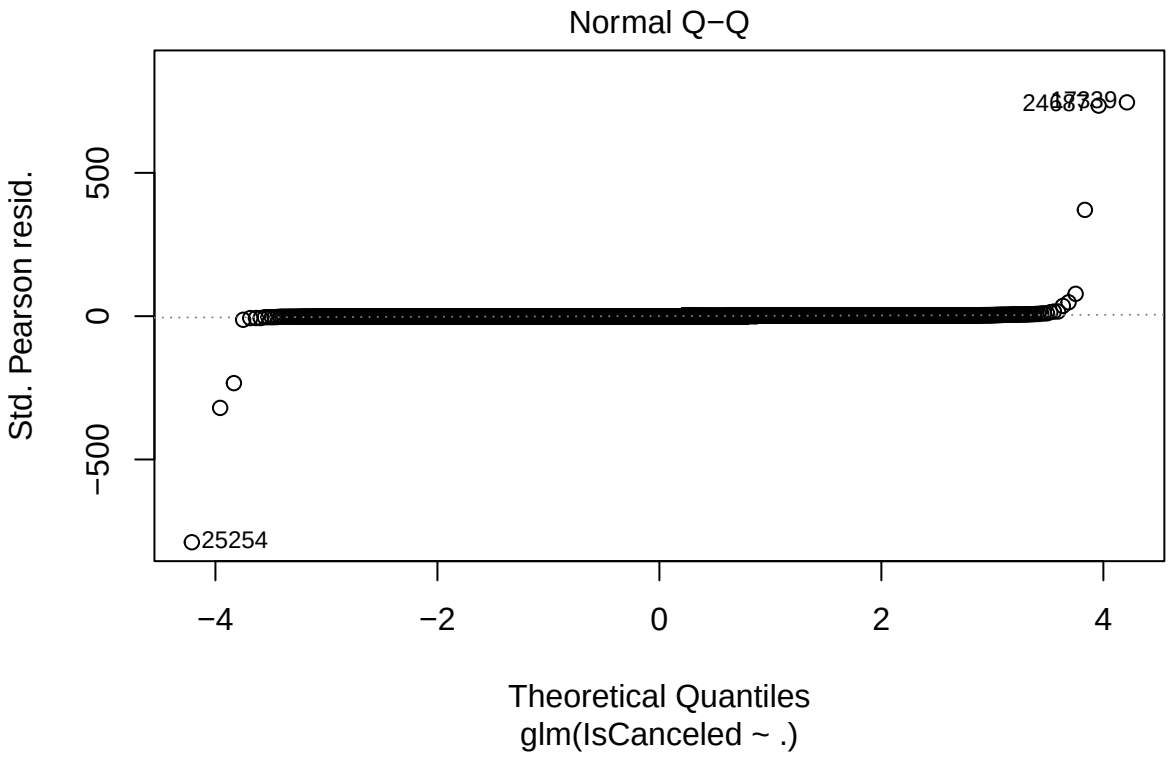


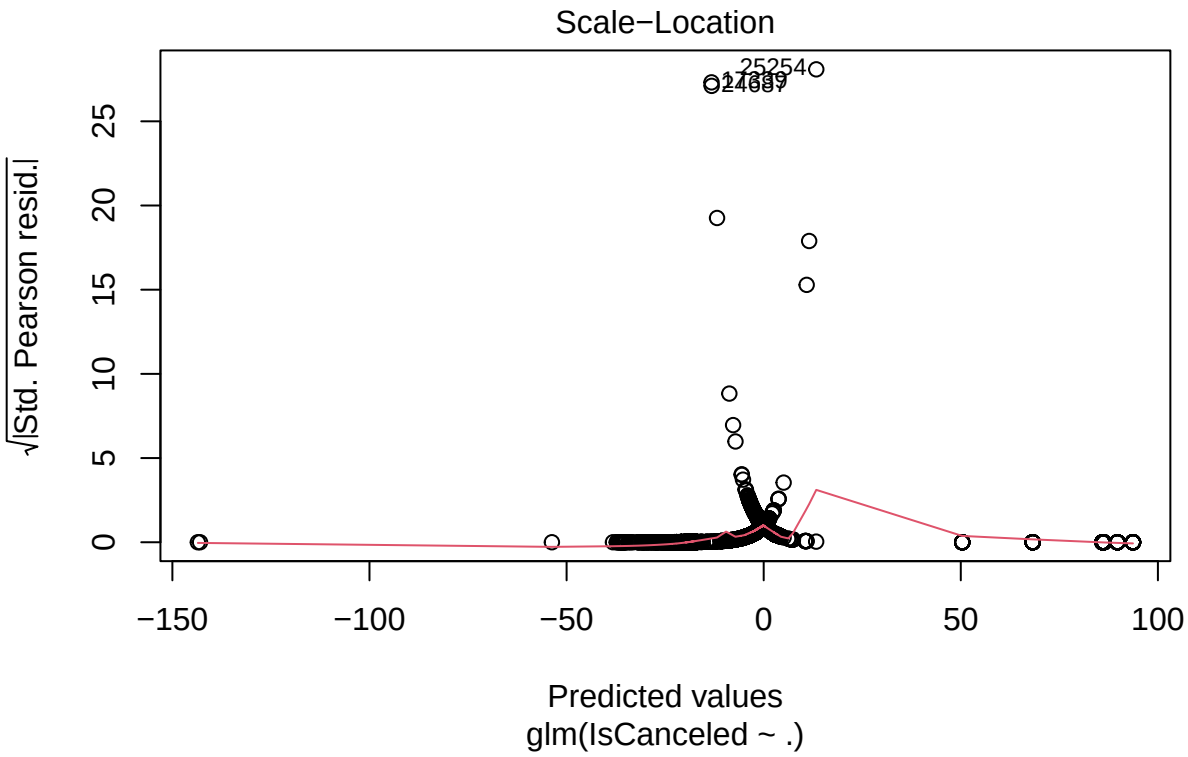

```
simulationOutput <- simulateResiduals(fittedModel = logmod, n = 250) #simulaiing the logarithmic residu
plot(simulationOutput) #ploting the simulation
```

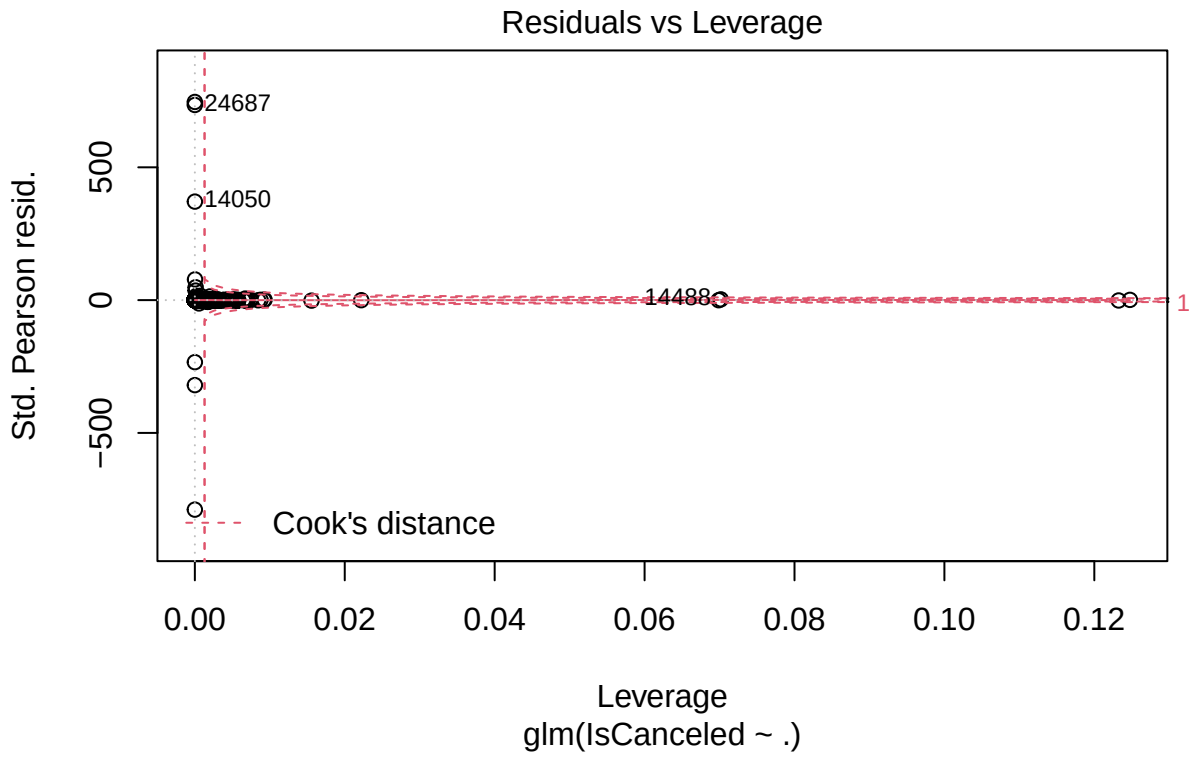



```
plot(logmod) #outlier at 13920; removed to better view graphic
```


Residuals vs Fitted
Residuals
500
0
−500
25254
−150
−100
−50
0
50
100
Predicted values
glm(IsCanceled ~ .)


Normal Q−Q
Std. Pearson resid.
500
0
−500
25254
−4
−2
0
2
4
Theoretical Quantiles
glm(IsCanceled ~ .)


Scale−Location
√|Std. Pearson resid.|
0
5
10
15
20
25
−150
−100
−50
0
50
100
Predicted values
glm(IsCanceled ~ .)



Looking at Variable Interactions

```
linintmod <- lm(as.numeric(IsCanceled) ~ .+StaysInWeekendNights:StaysInWeekNights+PreviousCancellations
summary(linintmod)
```

```
##
## Call:
## lm(formula = as.numeric(IsCanceled) ~ . + StaysInWeekendNights:StaysInWeekNights +
##     PreviousCancellations:PreviousBookingsNotCanceled, data = numdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8013 -0.2967 -0.2295  0.4955  1.9529
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                     1.2537004  0.0045927 272.974
## LeadTime                                        0.0008881  0.0000245  36.246
## StaysInWeekendNights                            0.0086388  0.0029673   2.911
## StaysInWeekNights                              -0.0020531  0.0015926  -1.289
## PreviousCancellations                           0.0289544  0.0015887  18.225
## PreviousBookingsNotCanceled                    -0.0302210  0.0027975 -10.803
## BookingChanges                                 -0.0704364  0.0029282 -24.054
## RequiredCarParkingSpaces                       -0.2610647  0.0061459 -42.478
## StaysInWeekendNights:StaysInWeekNights         -0.0007580  0.0003095  -2.449
## PreviousCancellations:PreviousBookingsNotCanceled  0.0148034  0.0029872   4.956
```

```
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## LeadTime                                        < 2e-16 ***
## StaysInWeekendNights                             0.0036 **
## StaysInWeekNights                                0.1973
## PreviousCancellations                           < 2e-16 ***
## PreviousBookingsNotCanceled                     < 2e-16 ***
## BookingChanges                                  < 2e-16 ***
## RequiredCarParkingSpaces                        < 2e-16 ***
## StaysInWeekendNights:StaysInWeekNights           0.0143 *
## PreviousCancellations:PreviousBookingsNotCanceled 7.24e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.421 on 39586 degrees of freedom
## Multiple R-squared:  0.1207, Adjusted R-squared:  0.1205
## F-statistic: 603.7 on 9 and 39586 DF,  p-value: < 2.2e-16
```

```
logintmod <- glm(IsCanceled ~  .+StaysInWeekendNights:StaysInWeekNights, family = binomial(link="logit"
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logintmod)
```

```
##
## Call:
## glm(formula = IsCanceled ~ . + StaysInWeekendNights:StaysInWeekNights,
##     family = binomial(link = "logit"), data = numdata[-13920,
##         ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1619  -0.8284  -0.6400   1.0715   5.1306
##
## Coefficients:
##                                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -1.101e+00  2.817e-02 -39.093  < 2e-16
## LeadTime                                 3.838e-03  1.349e-04  28.458  < 2e-16
## StaysInWeekendNights                     6.624e-02  1.821e-02   3.637 0.000276
## StaysInWeekNights                        6.147e-03  9.906e-03   0.620 0.534940
## PreviousCancellations                    3.599e+00  1.637e-01  21.990  < 2e-16
## PreviousBookingsNotCanceled             -1.212e+00  7.260e-02 -16.699  < 2e-16
## BookingChanges                          -5.306e-01  2.425e-02 -21.880  < 2e-16
## RequiredCarParkingSpaces                -1.744e+01  8.423e+01  -0.207 0.835954
## StaysInWeekendNights:StaysInWeekNights -8.894e-03  2.395e-03  -3.714 0.000204
##
## (Intercept)                             ***
## LeadTime                                ***
## StaysInWeekendNights                    ***
## StaysInWeekNights
## PreviousCancellations                   ***
## PreviousBookingsNotCanceled             ***
## BookingChanges                          ***
```

```
## RequiredCarParkingSpaces
## StaysInWeekendNights:StaysInWeekNights ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46936  on 39594  degrees of freedom
## Residual deviance: 39226  on 39586  degrees of freedom
## AIC: 39244
##
## Number of Fisher Scoring iterations: 17
```

```
linintout <- predict(linintmod, numdata[,-1]) #interaction predictions; same method as before
linintout <- data.frame(linintout)
linintout$data <- 0
#linout$data[linout$linout > as.numeric(median(linout$linout))] <- 1
linintout$data[linintout$linintout > 1.5] <- 1 #1.5 is the halfway point in the factored IsCanceled dat
linintout$data <- factor(linintout$data, levels = c(0,1), labels = c('Not Canceled', 'Canceled')) #chan

confusionMatrix(as.factor(linintout$data),numdata$IsCanceled)
```

```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     Not Canceled Canceled
##   Not Canceled        27707     9848
##   Canceled              812     1229
##
##                Accuracy : 0.7308
##                  95% CI : (0.7264, 0.7351)
##     No Information Rate : 0.7202
##     P-Value [Acc > NIR] : 1.432e-06
##
##                   Kappa : 0.1099
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9715
##             Specificity : 0.1110
##          Pos Pred Value : 0.7378
##          Neg Pred Value : 0.6022
##              Prevalence : 0.7202
##          Detection Rate : 0.6997
##    Detection Prevalence : 0.9485
##       Balanced Accuracy : 0.5412
##
##        'Positive' Class : Not Canceled
##
```

```
logintout <- predict(logintmod, numdata[,-1]) #interaction predictions; same method as before
logintout <- data.frame(logintout)
logintout$data <- 0
```

```r
#logout$data[logout$logout > as.numeric(median(logout$logout))] <- 1
logintout$data[logintout$logintout > 1.5] <- 1
logintout$data <- factor(logintout$data, levels = c(0,1), labels = c('Not Canceled', 'Canceled')) #chan

confusionMatrix(as.factor(logintout$data),numdata$IsCanceled)
```

```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     Not Canceled Canceled
##   Not Canceled        28507    10204
##   Canceled               12      873
##
##                Accuracy : 0.742
##                  95% CI : (0.7377, 0.7463)
##     No Information Rate : 0.7202
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1091
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.99958
##             Specificity : 0.07881
##          Pos Pred Value : 0.73641
##          Neg Pred Value : 0.98644
##              Prevalence : 0.72025
##          Detection Rate : 0.71995
##    Detection Prevalence : 0.97765
##       Balanced Accuracy : 0.53920
##
##        'Positive' Class : Not Canceled
##
```

9. Moving on to Factors

```r
#changing hotel categorical variables into factored variables
hotels$IsCanceled <- as.factor(hotels$IsCanceled)
hotels$Adults <- as.factor(hotels$Adults)
hotels$Children <- as.factor(hotels$Children)
hotels$Babies <- as.factor(hotels$Babies)
hotels$Meal <- as.factor(hotels$Meal)
hotels$Country <- as.factor(hotels$Country)
hotels$MarketSegment <- as.factor(hotels$MarketSegment)
hotels$IsRepeatedGuest <- as.factor(hotels$IsRepeatedGuest)
hotels$ReservedRoomType <- as.factor(hotels$ReservedRoomType)
hotels$AssignedRoomType <- as.factor(hotels$AssignedRoomType)
hotels$DepositType <- as.factor(hotels$DepositType)
hotels$CustomerType <- as.factor(hotels$CustomerType)

facdata <- data.frame(hotels[,-20][,-19][,-16][,-13][,-12][,-4][,-3][,-2], #subsetting the numerics and
                      PreviousCancellations = factor(hotels$PreviousCancellations > 0, levels = c(FA
                      PreviousBookingsNotCanceled = factor(hotels$PreviousBookingsNotCanceled > 0, le
```

```
                        BookingChanges = factor(hotels$BookingChanges > 0, levels = c(FALSE, TRUE), la
                        RequiredParkingSpaces = factor(hotels$RequiredCarParkingSpaces > 0, levels = c
                        TotalOfSpecialRequests = factor(hotels$TotalOfSpecialRequests > 0, levels = c(
```

10. SVM

```
#svm part 1
set.seed(100) #setting random number generator
trainList <- createDataPartition(y=facdata$IsCanceled,p=.60,list=FALSE) #partitioning the data
summary(trainList)
```

```
##    Resample1
##  Min.   :    1
##  1st Qu.: 9904
##  Median :19861
##  Mean   :19840
##  3rd Qu.:29787
##  Max.   :39595
```

```
str(trainList)
```

```
##  int [1:23759, 1] 1 2 3 4 7 8 9 10 13 17 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr "Resample1"
```

```
#trainList
trainSet<- facdata[trainList, ] #making training data set with partitioned indices
testSet<- facdata[-trainList, ] #making testing data set with remaining observations
testSet<- data.frame(testSet) #converting test set into data frame
#str(trainSet)
#str(testSet)

#svm part2
svmModel<- ksvm(IsCanceled~., data=trainSet, C = 5, cross=3, prob.model=TRUE) #running svm model
```

```
## line search fails -1.470582 0.2611903 3.808992e-05 -2.191104e-05 -1.09683e-08 5.896208e-09 -5.469737
```

```
svmModel
```

```
## Support Vector Machine object of class "ksvm"
##
## SV type: C-svc  (classification)
##  parameter : cost C = 5
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.170454545454545
##
## Number of Support Vectors : 8314
##
```

```
## Objective Function Value : -34164.32
## Training error : 0.130519
## Cross validation error : 0.149627
## Probability model included.
```

```
predOut<- predict(svmModel, testSet) #getting predictions from svm model using test data set

confusionMatrix(predOut,testSet$IsCanceled) #analyzing results in a confusion matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 10329  1178
##          1  1078  3252
##
##                Accuracy : 0.8575
##                  95% CI : (0.852, 0.863)
##     No Information Rate : 0.7203
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.644
##
##  Mcnemar's Test P-Value : 0.03713
##
##             Sensitivity : 0.9055
##             Specificity : 0.7341
##          Pos Pred Value : 0.8976
##          Neg Pred Value : 0.7510
##              Prevalence : 0.7203
##          Detection Rate : 0.6522
##    Detection Prevalence : 0.7266
##       Balanced Accuracy : 0.8198
##
##        'Positive' Class : 0
##
```

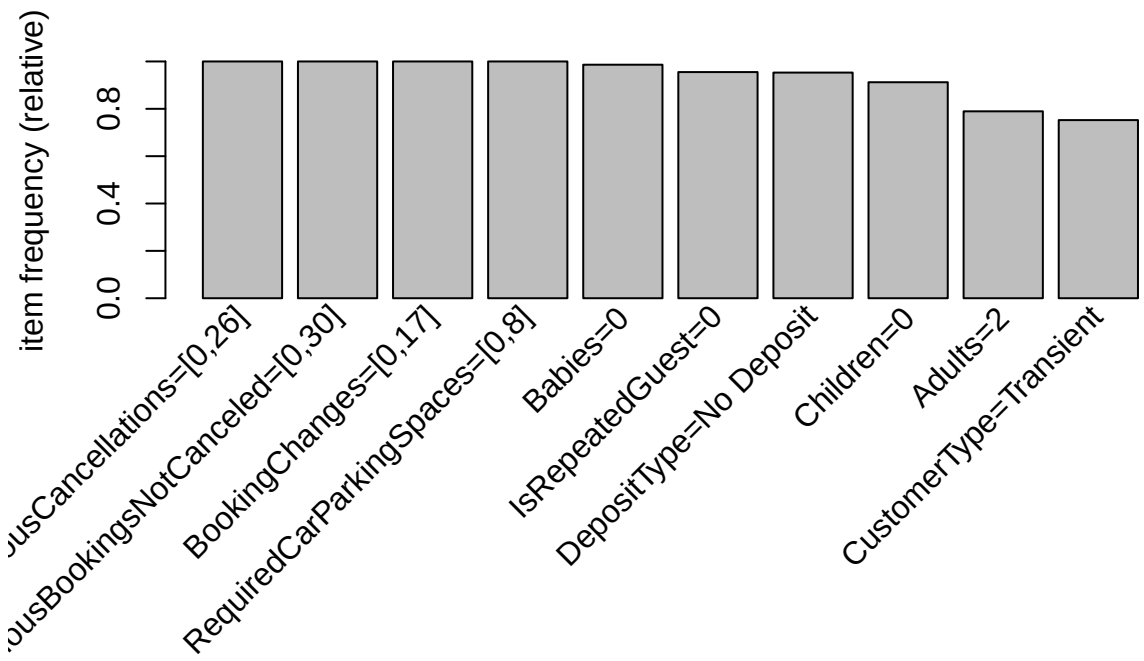
11. Association Rule Mining

```
hotels$IsCanceled <- as.factor(hotels$IsCanceled) #repositioning IsCanceled to remain consistent with f
hotelsX <- as(hotels, Class = 'transactions') #making a hotels data set with transactions
```

```
## Warning: Column(s) 2, 3, 4, 12, 13, 16, 19, 20 not logical or factor. Applying
## default discretization (see '? discretizeDF').

## Warning in discretize(x = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, : The calculated breaks are
##   Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.

## Warning in discretize(x = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, : The calculated breaks are
##   Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.

## Warning in discretize(x = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, : The calculated breaks are
##   Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.
```

```
## Warning in discretize(x = c(3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, : The calculated breaks are
##   Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.

## Warning in discretize(x = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, : The calculated breaks are
##   Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.

## Warning in discretize(x = c(0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 3, 1, 0, : The calculated breaks are
##   Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.
```

```
itemFrequencyPlot(hotelsX, topN = 10) #making an item frequency plot; not necessarily needed
```




```
freq <- itemFrequency(hotelsX) #creating item frequency
str(freq)
```

```
##  Named num [1:202] 0.72 0.28 0.332 0.334 0.334 ...
##  - attr(*, "names")= chr [1:202] "IsCanceled=0" "IsCanceled=1" "LeadTime=[0,23)" "LeadTime=[23,117)"
```

```
items <- sort(freq) #sorting data in ascending :sort(-item)#Descending

head(items) # least bought items
```

```
##     Adults=6    Adults=10    Adults=40    Adults=50    Adults=55  Children=10
## 2.525508e-05 2.525508e-05 2.525508e-05 2.525508e-05 2.525508e-05 2.525508e-05
```

```
tail(items)
```

```
##                   IsRepeatedGuest=0                                Babies=0
##                           0.9551975                               0.9862360
##       PreviousCancellations=[0,26] PreviousBookingsNotCanceled=[0,30]
##                           1.0000000                               1.0000000
##              BookingChanges=[0,17]     RequiredCarParkingSpaces=[0,8]
##                           1.0000000                               1.0000000
```

```
hotel_rule <- apriori(hotelsX, #apriori transactions
parameter=list(supp=0.008, conf=0.9), #0.8 \% of instances with 90% confidence
control=list(verbose=F),
appearance=list(default="lhs",rhs=("IsCanceled=1"))) #looking for canceled reservations
```

```
inspectDT(hotel_rule) #inspecting rule
```

```
## Warning in instance$preRenderHook(instance): It seems your data is too big
## for client-side DataTables. You may consider server-side processing: https://
## rstudio.github.io/DT/server.html
```

Show 10 entries Search:

	LHS	RHS	support	confidence	coverage	lift	count
	All	All	All	All	All	All	All
[1]	{DepositType=Non Refund}	{IsCanceled=1}	0.042	0.960	0.043	3.431	1,650.000
[2]	{MarketSegment=Groups,DepositType=Non Refund}	{IsCanceled=1}	0.037	0.962	0.038	3.437	1,454.000
[3]	{Meal=HB,DepositType=Non Refund}	{IsCanceled=1}	0.011	0.903	0.013	3.229	448.000
[4]	{StaysInWeekNights=[2,4),DepositType=Non Refund}	{IsCanceled=1}	0.022	0.987	0.022	3.529	855.000
[5]	{LeadTime=[23,117),DepositType=Non Refund}	{IsCanceled=1}	0.012	0.989	0.012	3.536	456.000
[6]	{LeadTime=[117,737],DepositType=Non Refund}	{IsCanceled=1}	0.029	0.964	0.030	3.445	1,145.000
[7]	{StaysInWeekNights=[4,40],DepositType=Non Refund}	{IsCanceled=1}	0.010	0.998	0.010	3.566	407.000
[8]	{StaysInWeekendNights=[2,16],DepositType=Non Refund}	{IsCanceled=1}	0.015	0.994	0.016	3.551	612.000
[9]	{AssignedRoomType=A,DepositType=Non Refund}	{IsCanceled=1}	0.035	0.959	0.036	3.428	1,375.000
[10]	{Country=PRT,DepositType=Non Refund}	{IsCanceled=1}	0.036	0.986	0.037	3.525	1,434.000

Showing 1 to 10 of 6,910 entries Previous 1 2 3 4 5 … 691 Next