PRACTICA 2: LIMPIEZA Y VALIDACIÓN DE LOS DATOS

Carles Maggi Gómez - Joan A. Maggi Gómez

10 de junio 2019

Partim de tres fitxers de la primera parts de la pràctica que cal unificar en un sol per poder ser tratats. L’objectiu es verure si hi ha diferencies de quailitat dels hotels en general i després en particular, segons les puntuacions dels comentaris desl usuaris, entre els periodes de hivern (‘primer trimestre’), i estiu(‘tercer trimestre’).

# Lectura del fitxer - read.csv()   
  
categories <-read.csv("https://raw.githubusercontent.com/joanmaggigo/practica2/master/CSV/CategoriesXComentariBooking.csv",header=F,sep=",",encoding = "UTF-8" ,stringsAsFactors = FALSE)  
comentaris <-read.csv("https://raw.githubusercontent.com/joanmaggigo/practica2/master/CSV/ComentarisXHotelsBarcelonaBooking.csv",header=F,sep=",",encoding = "UTF-8" ,stringsAsFactors = FALSE)  
estancies<-read.csv("https://raw.githubusercontent.com/joanmaggigo/practica2/master/CSV/HotelsBarcelonaBooking.csv",header=F,sep=",",encoding = "UTF-8" ,stringsAsFactors = FALSE)

Afegim els noms del atributs.

colnames(estancies)<-c("idHotel","estrelles","nota","nom","link")  
colnames(comentaris)<-c("idHotel","idIteracio","index","nomHotel","nota","comPositiu","comNegatiu","data")  
colnames(categories)<-c("idHotel","idIteracio","index","Categoria")

# Descripció dels dataset.

Podem trobar el datasets en el següent repositori :

<https://github.com/joanmaggigo/practica2/CSV>

El dataset d’hotels conté 300 registres i 5 variables. El dataset de comentaris conte 24874 registres i 8 variables. El dataset de categories per comentaries conte 114669 registres i 4 variables.

Comencem donant un cop d’ull al contingut del fitxers :

kable(head(categories,10),caption ="Inspecció visual categories")

Inspecció visual categories

|  |  |  |  |
| --- | --- | --- | --- |
| idHotel | idIteracio | index | Categoria |
| 90587 | 1 | 1 | Viajedeocio |
| 90587 | 1 | 1 | Pareja |
| 90587 | 1 | 1 | HabitaciónDobleSuperior |
| 90587 | 1 | 1 | Estanciade1noche |
| 90587 | 1 | 1 | Enviadopormóvil |
| 90587 | 1 | 2 | Grupo |
| 90587 | 1 | 2 | HabitaciónDoble-1o2camas |
| 90587 | 1 | 2 | Estanciade3noches |
| 90587 | 1 | 2 | Enviadopormóvil |
| 90587 | 1 | 3 | Viajedenegocios |

A tall d’exemple, veiem que el fitxer de categories té valors com ara ‘Pareja’ o ‘Grupo’ que entenem que són valors possibles dins la categoria d’acompanyament (o tipus d’acompanyament), així com ‘HabitaciónDobleSuperior’ o ‘HabitaciónDoble-1o2camas’, que serien valors de la categoria tipus d’habitació.

Mirem ara el fitxer dels hotels:

# head(estancies[(estancies$estrelles==0),"nom"])  
head(estancies,10)

## idHotel estrelles nota nom  
## 1 90587 3 8,2 Catalonia Sagrada Familia  
## 2 770635 0 9,3 Urquinaona  
## 3 1003967 0 7,9 Habitat Apartments Bail<e9>n  
## 4 584514 0 8,1 Habitat Apartments Alibei  
## 5 584538 0 8,4 My Address in Barcelona  
## 6 2934029 0 8,4 Barceloneta  
## 7 1565008 0 7,8 Ronda Sant Pere with Terrace  
## 8 291828 0 8,8 You Stylish City Centre Apartments  
## 9 730585 0 9,1 Bcn Paseo De Gracia Rocamora Apartment  
## 10 579277 0 7,5 AB Paseo de Gracia Apartments  
## link  
## 1 https://www.booking.com/hotel/es/cataloniaaragon.es.html?from=searchresults#hotelTmpl  
## 2 https://www.booking.com/hotel/es/chic-urquinaona.es.html?from=searchresults#hotelTmpl  
## 3 https://www.booking.com/hotel/es/habitat-apartments-bailen.es.html?from=searchresults#hotelTmpl  
## 4 https://www.booking.com/hotel/es/habitat-apartments-alibei-barcelona.es.html?from=searchresults#hotelTmpl  
## 5 https://www.booking.com/hotel/es/modernist-apartment-city-centre.es.html?from=searchresults#hotelTmpl  
## 6 https://www.booking.com/hotel/es/barceloneta-barcelona1.es.html?from=searchresults#hotelTmpl  
## 7 https://www.booking.com/hotel/es/ronda-sant-pere-i.es.html?from=searchresults#hotelTmpl  
## 8 https://www.booking.com/hotel/es/you-stylish-city-centre-apartments.es.html?from=searchresults#hotelTmpl  
## 9 https://www.booking.com/hotel/es/bcn-paseo-de-gracia-aparment.es.html?from=searchresults#hotelTmpl  
## 10 https://www.booking.com/hotel/es/apartment-barcelona-paseo-de-gracia.es.html?from=searchresults#hotelTmpl

D’altra banda, quan mirem el fitxer d’hotels, veiem que existeixen valors d’estrelles = 0 que mirant el nom de l’hotel veiem que no són hotels, sino que es refereixen a apartaments. Com que nosaltres ens centrarem en els Hotels tindrem que prescindir d’aquests últims.

Començarem a abordar el problema, fent el merge dels datasets, i analitzant el tema de les categories nomes amb els hotels tal i com hem dit, no tenint en compte els apartaments ( Estrelles = 0)

Primer ajuntarem en un sol dataset els comentaris i les categories. Aquest dataset estan relacionats pel “idHotel”,“idIteracio”,“index”, per tant utilitzarem aquet camps fer fusionar-los.

#A comentaris tenim dues columnes comentari possitiu i comentari negatiu que no farem res  
#de text mining per tant eliminem del procés el comentari possitiu i el conmentari negtiu  
comentaris<-comentaris [,-c(6,7)]  
# Creem un única dataset que fusiona tots els comentaries i categories  
comentaris.cat<-merge(comentaris,categories,by.x=c("idHotel","idIteracio","index"),by.y=c("idHotel","idIteracio","index"))

Tot seguit fusionarem el nou contingut de comentaris amb el hotel. En aquest cas els datsets estan relacionats per “idHotel”. Aprofitem per nomes incloure els que són hotels “estrelless <>0”, tal com hem dit que fariem abans. Al final mirem com ens queda el fitxer

# Fem el merge amb hotels  
hotels.comentaris.cat<-merge(comentaris.cat,estancies, by.x="idHotel", by.y="idHotel")  
hotels<-hotels.comentaris.cat[which(hotels.comentaris.cat$estrelles!=0),]  
kable(head(hotels,10), caption="Hotels Comentaris i Categories")

Hotels Comentaris i Categories

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | idHotel | idIteracio | index | nomHotel | nota.x | data | Categoria | estrelles | nota.y | nom | link |
| 412 | 13651 | 1 | 1 | Atlantis by Atbcn | 8,8 | 2019-04-13 | Viajedenegocios | 3 | 8,7 | Atlantis by Atbcn | <https://www.booking.com/hotel/es/atlantis.es.html?from=searchresults#hotelTmpl> |
| 413 | 13651 | 1 | 1 | Atlantis by Atbcn | 8,8 | 2019-04-13 | Grupodeamigos | 3 | 8,7 | Atlantis by Atbcn | <https://www.booking.com/hotel/es/atlantis.es.html?from=searchresults#hotelTmpl> |
| 414 | 13651 | 1 | 1 | Atlantis by Atbcn | 8,8 | 2019-04-13 | HabitaciónDoble-2camas | 3 | 8,7 | Atlantis by Atbcn | <https://www.booking.com/hotel/es/atlantis.es.html?from=searchresults#hotelTmpl> |
| 415 | 13651 | 1 | 1 | Atlantis by Atbcn | 8,8 | 2019-04-13 | Estanciade1noche | 3 | 8,7 | Atlantis by Atbcn | <https://www.booking.com/hotel/es/atlantis.es.html?from=searchresults#hotelTmpl> |
| 416 | 13651 | 1 | 1 | Atlantis by Atbcn | 8,8 | 2019-04-13 | Enviadopormóvil | 3 | 8,7 | Atlantis by Atbcn | <https://www.booking.com/hotel/es/atlantis.es.html?from=searchresults#hotelTmpl> |
| 417 | 13651 | 1 | 10 | Atlantis by Atbcn | 9,2 | 2019-01-30 | Viajedeocio | 3 | 8,7 | Atlantis by Atbcn | <https://www.booking.com/hotel/es/atlantis.es.html?from=searchresults#hotelTmpl> |
| 418 | 13651 | 1 | 10 | Atlantis by Atbcn | 9,2 | 2019-01-30 | Pareja | 3 | 8,7 | Atlantis by Atbcn | <https://www.booking.com/hotel/es/atlantis.es.html?from=searchresults#hotelTmpl> |
| 419 | 13651 | 1 | 10 | Atlantis by Atbcn | 9,2 | 2019-01-30 | HabitaciónDoble-2camas | 3 | 8,7 | Atlantis by Atbcn | <https://www.booking.com/hotel/es/atlantis.es.html?from=searchresults#hotelTmpl> |
| 420 | 13651 | 1 | 10 | Atlantis by Atbcn | 9,2 | 2019-01-30 | Estanciade1noche | 3 | 8,7 | Atlantis by Atbcn | <https://www.booking.com/hotel/es/atlantis.es.html?from=searchresults#hotelTmpl> |
| 421 | 13651 | 1 | 11 | Atlantis by Atbcn | 10 | 2019-01-20 | Viajedenegocios | 3 | 8,7 | Atlantis by Atbcn | <https://www.booking.com/hotel/es/atlantis.es.html?from=searchresults#hotelTmpl> |

#Veiem, per descripció de dataset que el identificador de la persona/comentari es la concateniació del idhotel, ititeració i index.  
hotels$idPersona <- paste(as.character(hotels$idHotel), as.character(hotels$idIteracio), as.character(hotels$index),sep="\_")  
  
#Cada comentari/persona té vinculada n descripcions de categories en el fitxer de categories. Volem veure cuantes descripcions de categories hi ha per cada comentari/persona (i el seu volumen) per poder saber sota quins conceptes s'agrupen les categories.  
categoria\_per\_comentari <- hotels %>% group\_by(idPersona) %>% summarize(total\_cat=n())  
comentaris\_per\_num\_categories<-categoria\_per\_comentari %>% group\_by(total\_cat) %>% summarize(total\_comentaris=n())  
kable(comentaris\_per\_num\_categories, caption="Distribució de total Valors categories per comentaris")

Distribució de total Valors categories per comentaris

|  |  |
| --- | --- |
| total\_cat | total\_comentaris |
| 2 | 4 |
| 3 | 465 |
| 4 | 5446 |
| 5 | 10352 |
| 6 | 17 |

En primer lloc vam pensar en agrupar només en 5 categories, tenint en compte que el comentaris que en tenen 6 eren pocs i ens semblaven irrellevants.Finalment, ens hem adonat que això tenia efectes sobre altres comentaris perquè ens hem trobat que per un mateixa categoria en un comentari hi havia dos valors possibles, pel que el primer procés d’agrupació ha sigut erroni i hem detectat on teniem l’error. L’error era sobre el valor “AmbMascota” que l’haviem considerat inicialment com tipus de companyia però era un element diferenciat. Per veure-ho clar, fem una selecció dels comentaris que tene 6 descripcions de la categoria i explorem visualment el contingut.

idpersona\_6categories<-unique(categoria\_per\_comentari[categoria\_per\_comentari$total\_cat==6,"idPersona"])  
kable(head(hotels[(hotels$idPersona %in% idpersona\_6categories$idPersona ),][c("idPersona","Categoria")],20))

|  |  |  |
| --- | --- | --- |
|  | idPersona | Categoria |
| 8157 | 90039\_2\_17 | Conunamascota |
| 8158 | 90039\_2\_17 | Viajedeocio |
| 8159 | 90039\_2\_17 | Familiaconniñospequeños |
| 8160 | 90039\_2\_17 | HabitaciónDobleSuperior-1o2camas |
| 8161 | 90039\_2\_17 | Estanciade1noche |
| 8162 | 90039\_2\_17 | Enviadopormóvil |
| 8163 | 90039\_2\_18 | Conunamascota |
| 8164 | 90039\_2\_18 | Viajedeocio |
| 8165 | 90039\_2\_18 | Pareja |
| 8166 | 90039\_2\_18 | HabitaciónDobleExecutive-1o2camas |
| 8167 | 90039\_2\_18 | Estanciade4noches |
| 8168 | 90039\_2\_18 | Enviadopormóvil |
| 16103 | 90583\_3\_33 | Conunamascota |
| 16104 | 90583\_3\_33 | Viajedeocio |
| 16105 | 90583\_3\_33 | Pareja |
| 16106 | 90583\_3\_33 | HabitaciónDoble-1o2camas |
| 16107 | 90583\_3\_33 | Estanciade1noche |
| 16108 | 90583\_3\_33 | Enviadopormóvil |
| 16297 | 90586\_1\_25 | Conunamascota |
| 16298 | 90586\_1\_25 | Viajedeocio |

Estudiant visualment les descripcions de les categories pensem que podem agrupar en 6 grups. veiem valors que podrien respondre a les categories: Nits,ProcedenciaComentari,Habitació,TipusdeViatge,Acompanyament,ViatjeAmbMascota

#Estudiem el nombre de comentaris agrupats per categoria per començar a definir els tipus de categoria.  
  
kable(head(hotels %>% group\_by(Categoria) %>% summarize(total\_cat=n()) %>% arrange(desc(total\_cat)),10))

|  |  |
| --- | --- |
| Categoria | total\_cat |
| Viajedeocio | 12546 |
| Enviadopormóvil | 11538 |
| Pareja | 7875 |
| Estanciade1noche | 6155 |
| HabitaciónDoble-1o2camas | 5274 |
| Estanciade2noches | 4203 |
| Personaqueviajasola | 3261 |
| Estanciade3noches | 3073 |
| Familiaconniñospequeños | 2992 |
| Viajedenegocios | 2098 |

En funció de volum comencem a veure patrons per tal de poder crear l’agrupació de la categoria

## Creació noves varaibles agrupacio\_categoria, id\_persona

# Inicialitzem agrupacio\_catgoria amb el valor "Acompanyament" i anirem posant els difertns valors segons el contingut de la "Categoria"  
hotels$agrupacio\_categoria<-"Acompanyament"  
hotels[grep("stancia",hotels$Categoria),"agrupacio\_categoria"]<-"Nits"  
hotels[grep("Enviadopormóvil",hotels$Categoria),"agrupacio\_categoria"]<-"ProcedenciaComentari"  
hotels[grep("abitaci",hotels$Categoria),"agrupacio\_categoria"]<-"Habitacio"  
hotels[grep("uite",hotels$Categoria),"agrupacio\_categoria"]<-"Habitacio"  
hotels[grep("DobleEstándar",hotels$Categoria),"agrupacio\_categoria"]<-"Habitacio"  
hotels[grep("Apartamento",hotels$Categoria),"agrupacio\_categoria"]<-"Habitacio"  
hotels[grep("Viaje",hotels$Categoria),"agrupacio\_categoria"]<-"TipusViatge"  
hotels[grep("mascota",hotels$Categoria),"agrupacio\_categoria"]<-"ViajaConMascota"  
  
# Visualització de tots els grups de categrioes.  
  
agrupacio\_categoria<-unique(hotels$agrupacio\_categoria)  
for (cols in agrupacio\_categoria)   
{  
 aux<-hotels[which(hotels$agrupacio\_categoria==cols),]  
 aux2<-aux %>% group\_by(Categoria) %>% summarize(total\_cat=n())  
 print(kable(head(aux2,5), caption=as.character(cols)))  
}

##   
##   
## Table: TipusViatge  
##   
## Categoria total\_cat  
## ---------------- ----------  
## Viajedenegocios 2098  
## Viajedeocio 12546  
##   
##   
## Table: Acompanyament  
##   
## Categoria total\_cat  
## ------------------------ ----------  
## Familiaconniñosmayores 2  
## Familiaconniñospequeños 2992  
## Grupo 1732  
## Grupodeamigos 422  
## Pareja 7875  
##   
##   
## Table: Habitacio  
##   
## Categoria total\_cat  
## -------------------------- ----------  
## 10habitaciones 1  
## 2habitaciones 239  
## 3habitaciones 9  
## ApartamentoConfort 16  
## Apartamentode2dormitorios 1  
##   
##   
## Table: Nits  
##   
## Categoria total\_cat  
## ------------------- ----------  
## Estanciade10noches 10  
## Estanciade11noches 9  
## Estanciade12noches 6  
## Estanciade13noches 5  
## Estanciade14noches 5  
##   
##   
## Table: ProcedenciaComentari  
##   
## Categoria total\_cat  
## ---------------- ----------  
## Enviadopormóvil 11538  
##   
##   
## Table: ViajaConMascota  
##   
## Categoria total\_cat  
## -------------- ----------  
## Conunamascota 31

Veiem que en el cas de l’agrupació que ens explica el tipus d’habitació i el nombre de nits és interesant recategoritzar la variable, per què hi ha mases valors possibles i alguns amb pocs representats, pel que guanya força sentit.

#Ens guardem el nombre de nits en una variable nova.  
  
hotels$nits<-as.numeric(gsub("\\D", "",hotels$Categoria))  
#nomes guardem les nits si la categoria és Nits, els altres números que surten dels texots (tipus 2 habitacions) no apliquen, per tant, els passem a NA  
hotels[which(hotels$agrupacio\_categoria!="Nits"),"nits"]<-NA  
  
  
#Recategoritzem dues agrupacions de categories, en funció dels valors observats. Seran l'agrupació Habitació, y la agrupació Nits  
hotels$ReCategoria<-hotels$Categoria  
  
#A les estancies de mes de 7 dies diem que la nova categoria es mésde7dies  
hotels[which(hotels$agrupacio\_categoria=="Nits" & as.numeric(gsub("\\D", "",hotels$Categoria))==1 ) ,"ReCategoria"]<-"Una Nit"  
hotels[which(hotels$agrupacio\_categoria=="Nits" & as.numeric(gsub("\\D", "",hotels$Categoria))== 2),"ReCategoria"]<-"Dos nits"  
hotels[which(hotels$agrupacio\_categoria=="Nits" & (as.numeric(gsub("\\D", "",hotels$Categoria))> 2) & as.numeric(gsub("\\D", "",hotels$Categoria))< 6 ) ,"ReCategoria"]<-"De 3 a 5 Nits"  
hotels[which(hotels$agrupacio\_categoria=="Nits" & (as.numeric(gsub("\\D", "",hotels$Categoria))> 6) & as.numeric(gsub("\\D", "",hotels$Categoria))< 8 ) ,"ReCategoria"]<-"De 6 a 7 Nits"  
hotels[which(hotels$agrupacio\_categoria=="Nits" & as.numeric(gsub("\\D", "",hotels$Categoria))>7) ,"ReCategoria"]<-"Mesde7"  
  
  
#En el cas del tipus d'habitació, reagrupem seguint la següent lógica  
hotels[which(hotels$agrupacio\_categoria=="Habitacio"),"ReCategoria"]<-"Altres"  
hotels[hotels$agrupacio\_categoria=="Habitacio" & grepl("oble",hotels$Categoria),"ReCategoria"]<-"Habitacio Doble"  
hotels[hotels$agrupacio\_categoria=="Habitacio" & grepl("ndividual",hotels$Categoria),"ReCategoria"]<-"Habitacio Individual"  
hotels[hotels$agrupacio\_categoria=="Habitacio" & grepl("uite",hotels$Categoria),"ReCategoria"]<-"Suite"  
hotels[hotels$agrupacio\_categoria=="Habitacio" & grepl("riple",hotels$Categoria),"ReCategoria"]<-"Habitacio Triple"  
hotels[hotels$agrupacio\_categoria=="Habitacio" & grepl("druple",hotels$Categoria),"ReCategoria"]<-"Habitacio quadruple"  
  
hotels.Habitacio<-hotels[hotels$agrupacio\_categoria=="Habitacio",]  
hotels.Habitacio$ReCategoria<-as.factor(hotels.Habitacio$ReCategoria)  
  
agrupacio\_categoria\_habitacion<-unique(hotels.Habitacio$ReCategoria)  
for (cols in agrupacio\_categoria\_habitacion)   
{  
print(head(hotels.Habitacio[which(hotels.Habitacio$ReCategoria==cols),] %>% group\_by(Categoria) %>% summarize(total\_cat=n()),10))  
}

## # A tibble: 10 x 2  
## Categoria total\_cat  
## <chr> <int>  
## 1 DobleEstándar 1  
## 2 Habitaciónconcamadobleoextragrandeyvistaspanorámicas 5  
## 3 HabitaciónDoble 1189  
## 4 HabitaciónDoble-1o2camas 5274  
## 5 HabitaciónDoble-2camas 1248  
## 6 HabitaciónDoble(1-2adultos) 25  
## 7 HabitaciónDoble(1adulto)-1o2camas 4  
## 8 HabitaciónDoble(2adultos) 128  
## 9 HabitaciónDoble(2adultos+1niño) 38  
## 10 HabitaciónDoble2camasconvistasalapiscina 3  
## # A tibble: 10 x 2  
## Categoria total\_cat  
## <chr> <int>  
## 1 HabitaciónDobleparausoindividual 17  
## 2 HabitaciónIndividual 1188  
## 3 HabitaciónIndividual-Interior 24  
## 4 HabitaciónIndividualBásica 74  
## 5 HabitaciónIndividualconbañocompartido 4  
## 6 HabitaciónIndividualconvistas 43  
## 7 HabitaciónIndividualconvistasalaciudad 5  
## 8 HabitaciónIndividualEstándar 52  
## 9 HabitaciónIndividualExecutive 13  
## 10 HabitaciónIndividualSuperior 8  
## # A tibble: 10 x 2  
## Categoria total\_cat  
## <chr> <int>  
## 1 10habitaciones 1  
## 2 2habitaciones 239  
## 3 3habitaciones 9  
## 4 ApartamentoConfort 16  
## 5 Apartamentode2dormitorios 1  
## 6 ApartamentoEconómico 4  
## 7 Habitación(1o2adultos)-1o2camas 291  
## 8 HabitaciónCompartida(4adultos) 44  
## 9 Habitacióncompartida(6adultos) 14  
## 10 Habitaciónconcamaextragrande. 148  
## # A tibble: 4 x 2  
## Categoria total\_cat  
## <chr> <int>  
## 1 Habitacióncuádruple 162  
## 2 HabitaciónCuádrupleconbalcón 5  
## 3 HabitaciónCuádrupleconvistas 27  
## 4 HabitaciónCuádrupleFamiliar 8  
## # A tibble: 7 x 2  
## Categoria total\_cat  
## <chr> <int>  
## 1 HabitaciónTriple 588  
## 2 HabitaciónTriple(2adultos+1niño) 30  
## 3 HabitaciónTriple(3adultos) 9  
## 4 HabitaciónTripleClásica 4  
## 5 HabitaciónTripleconterraza 2  
## 6 HabitaciónTripleEstándar 23  
## 7 HabitaciónTripleSuperior 28  
## # A tibble: 10 x 2  
## Categoria total\_cat  
## <chr> <int>  
## 1 Suite 23  
## 2 Suiteconbañeradehidromasaje 21  
## 3 Suiteconpiscinaprivada 1  
## 4 Suiteconterraza 1  
## 5 Suitede1dormitorioconvistasalaciudad 2  
## 6 SuiteDúplex 1  
## 7 SuiteExecutive 1  
## 8 SuiteImperial 2  
## 9 SuiteJunior 62  
## 10 SuiteJunior-ExperienciaImmersiveArt 2

La recategorització sembla raonable respecte els valors observats, per tant ja podem procedir a construir el dataset de partida per evaluar els objectius descrits

## Creació del Dataset de treball

#Anem a crear el Dataset de treball  
  
#Creem un identificador de persona  
categories\_per\_persona<-hotels %>% group\_by(idPersona) %>% summarize(total\_cat=n())  
  
#Creem el dataset de treball  
data<-unique(hotels[c("nomHotel","nota.x","data","estrelles","idPersona","nota.y")])  
#data<-merge(data,categories\_per\_persona,by.x="idPersona",by.y="idPersona")  
  
#Creem un dataset que ajunti el idcomentari amb l'agrupació de categoria  
data.Habitacio<-hotels[hotels$agrupacio\_categoria=="Habitacio",]  
data.Habitacio<-data.Habitacio[c("idPersona","ReCategoria")]  
data.Nits<-hotels[hotels$agrupacio\_categoria=="Nits",]  
data.Nits<-data.Nits[c("idPersona","ReCategoria","nits")]  
data.ProcedenciaComentari<-hotels[hotels$agrupacio\_categoria=="ProcedenciaComentari",]  
data.ProcedenciaComentari<-data.ProcedenciaComentari[c("idPersona","ReCategoria")]  
data.TipusViatge<-hotels[hotels$agrupacio\_categoria=="TipusViatge",]  
data.TipusViatge<-data.TipusViatge[c("idPersona","ReCategoria")]  
data.Acompanyament<-hotels[hotels$agrupacio\_categoria=="Acompanyament",]  
data.Acompanyament<-data.Acompanyament[c("idPersona","ReCategoria")]  
data.ViajaConMascota<-hotels[hotels$agrupacio\_categoria=="ViajaConMascota",]  
data.ViajaConMascota<-data.ViajaConMascota[c("idPersona","ReCategoria")]  
  
#fem els merges per crear el ddataset de treball  
data<-merge(data,data.Habitacio,by.x="idPersona",by.y="idPersona",all.x = T)  
data<-merge(data,data.Nits,by.x="idPersona",by.y="idPersona",all.x = T)  
data<-merge(data,data.ProcedenciaComentari,by.x="idPersona",by.y="idPersona",all.x = T)  
data<-merge(data,data.TipusViatge,by.x="idPersona",by.y="idPersona",all.x = T)  
data<-merge(data,data.Acompanyament,by.x="idPersona",by.y="idPersona",all.x = T)  
data<-merge(data,data.ViajaConMascota,by.x="idPersona",by.y="idPersona",all.x = T)  
  
  
colnames(data)<-c("idPersona","nomHotel","notapersona","data","estrelles","notaHotel","TipusHabitacio","DuradaEstancia","Nits","ProcedenciaComentari","TipusViatge","Acompanyament","ViajaConMascota")  
summary(data)

## idPersona nomHotel notapersona   
## Length:16284 Length:16284 Length:16284   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## data estrelles notaHotel TipusHabitacio   
## Length:16284 Min. :1.0 Length:16284 Length:16284   
## Class :character 1st Qu.:2.0 Class :character Class :character   
## Mode :character Median :3.0 Mode :character Mode :character   
## Mean :3.1   
## 3rd Qu.:4.0   
## Max. :5.0   
##   
## DuradaEstancia Nits ProcedenciaComentari TipusViatge   
## Length:16284 Min. : 1.0 Length:16284 Length:16284   
## Class :character 1st Qu.: 1.0 Class :character Class :character   
## Mode :character Median : 2.0 Mode :character Mode :character   
## Mean : 2.3   
## 3rd Qu.: 3.0   
## Max. :27.0   
## NA's :8   
## Acompanyament ViajaConMascota   
## Length:16284 Length:16284   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##

# Neteja de dades

## Tipus de dades

Eliminem l’identificador de la persona perquè ja no ens fa falta d’ara endevant.

data <- select(data, -idPersona)

Assignem el tipus de dades correcte, ja que les notes tenen una ‘,’ com a separador decimal i cal posar-hi un punt, i la data l’hem de passar com a tipus data.

#Veiem que s'ha de passar la nota del hotel a numèric  
data$notaHotel<-as.numeric(gsub(",",".",data$notaHotel))  
#veiem que s'ha de passar les notes de persona a numeric  
data$notapersona<-as.numeric(gsub(",",".",data$notapersona))  
#Hem de posar la data en format data  
data$data<-as.Date(data$data)  
summary(data)

## nomHotel notapersona data estrelles   
## Length:16284 Min. : 2.5 Min. :2017-04-13 Min. :1.0   
## Class :character 1st Qu.: 7.5 1st Qu.:2017-11-25 1st Qu.:2.0   
## Mode :character Median : 8.8 Median :2018-05-13 Median :3.0   
## Mean : 8.6 Mean :2018-05-06 Mean :3.1   
## 3rd Qu.:10.0 3rd Qu.:2018-11-04 3rd Qu.:4.0   
## Max. :10.0 Max. :2019-04-13 Max. :5.0   
##   
## notaHotel TipusHabitacio DuradaEstancia Nits   
## Min. :7.9 Length:16284 Length:16284 Min. : 1.0   
## 1st Qu.:8.3 Class :character Class :character 1st Qu.: 1.0   
## Median :8.5 Mode :character Mode :character Median : 2.0   
## Mean :8.5 Mean : 2.3   
## 3rd Qu.:8.7 3rd Qu.: 3.0   
## Max. :9.4 Max. :27.0   
## NA's :8   
## ProcedenciaComentari TipusViatge Acompanyament   
## Length:16284 Length:16284 Length:16284   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## ViajaConMascota   
## Length:16284   
## Class :character   
## Mode :character   
##   
##   
##   
##

#Comencem analitzant des de la perspectiva de l'hotel.  
aux.hotels<-unique(data[c("nomHotel","notaHotel","estrelles","notapersona")])  
summary(aux.hotels)

## nomHotel notaHotel estrelles notapersona   
## Length:1340 Min. :7.9 Min. :1.0 Min. : 2.5   
## Class :character 1st Qu.:8.3 1st Qu.:3.0 1st Qu.: 5.5   
## Mode :character Median :8.6 Median :4.0 Median : 7.1   
## Mean :8.6 Mean :3.3 Mean : 7.1   
## 3rd Qu.:8.8 3rd Qu.:4.0 3rd Qu.: 8.8   
## Max. :9.4 Max. :5.0 Max. :10.0

#Pel que fa a les estrelles veiem que té una distribució raonable amb valors, etc.. per tant ho donem per bo  
  
#Les notes mitjes del hotel semblen raonables.  
  
# Anteriorment ja hem eliminat registres que tenien la puntuacio de persona més gran que 3 vegades la desviació estandar, ja que pot ser puntual o molt subjectiu que hagin tingut una mala experiencia puntual i estiguin resentits, o be és un error.

## Tractament dels valors na

#veiem els NA  
colSums(is.na(data))

## nomHotel notapersona data   
## 0 0 0   
## estrelles notaHotel TipusHabitacio   
## 0 0 8   
## DuradaEstancia Nits ProcedenciaComentari   
## 8 8 4746   
## TipusViatge Acompanyament ViajaConMascota   
## 1640 0 16253

#Anem a pams, pel cas, ViajaConMascota, sabem que si no está informat, es raonable pensar que no viatge ambMascota. Fem una recategorització per transformar-ho en una binaria, els que tenen valors S els que no en tenen N  
  
data[!(is.na(data$ViajaConMascota)),"ViajaConMascota"]<-"S"  
data[(is.na(data$ViajaConMascota)),"ViajaConMascota"]<-"N"  
  
#Per tipus de viatge, el volum de NA es un 10% aprox, per tant pensem que té prou instancia per si sol com per afegir un valor més dins la categoria que sigui SENSEINFORMAR per veure impactes en la nota (o altres)  
  
data[(is.na(data$TipusViatge)),"TipusViatge"]<-"SenseInformar"  
  
#Procedencia comentari, només té un valor informat, per tant, entenem que l'altre valor és per web  
data[(is.na(data$ProcedenciaComentari)),"ProcedenciaComentari"]<-"EnviadoporWeb"  
  
#A tipusHabitacio I Nits, tenim la sospita que els que estan en NA son els mateixos.  
data[(is.na(data$Nits)) |is.na(data$TipusHabitacio),]

## nomHotel notapersona data estrelles  
## 23 Hostal Balkonis 3.8 2019-03-07 2  
## 1645 Hotel El Call 9.6 2018-01-02 1  
## 3215 Golden Tulip Barcelona 10.0 2017-07-30 4  
## 7849 Melia Barcelona Sarriá 8.8 2018-08-08 5  
## 7878 Melia Barcelona Sarriá 2.5 2018-04-17 5  
## 10336 Hesperia Barcelona Del Mar 9.6 2018-08-29 4  
## 10892 Tryp Barcelona Apolo Hotel 3.3 2019-02-22 4  
## 14135 Pullman Barcelona Skipper 10.0 2019-02-13 5  
## notaHotel TipusHabitacio DuradaEstancia Nits ProcedenciaComentari  
## 23 8.8 <NA> <NA> NA EnviadoporWeb  
## 1645 8.3 <NA> <NA> NA Enviadopormóvil  
## 3215 8.7 <NA> <NA> NA EnviadoporWeb  
## 7849 8.2 <NA> <NA> NA Enviadopormóvil  
## 7878 8.2 <NA> <NA> NA Enviadopormóvil  
## 10336 8.2 <NA> <NA> NA EnviadoporWeb  
## 10892 8.6 <NA> <NA> NA Enviadopormóvil  
## 14135 8.5 <NA> <NA> NA EnviadoporWeb  
## TipusViatge Acompanyament ViajaConMascota  
## 23 Viajedeocio Pareja N  
## 1645 Viajedeocio Pareja N  
## 3215 Viajedeocio Familiaconniñospequeños N  
## 7849 Viajedeocio Familiaconniñospequeños N  
## 7878 Viajedenegocios Personaqueviajasola N  
## 10336 Viajedeocio Grupo N  
## 10892 Viajedenegocios Personaqueviajasola N  
## 14135 Viajedenegocios Personaqueviajasola N

#Veiem que si per tant, sembla raonable pensar que és un error i que milor obviar la informació (eliminarla)  
data<-data[!(is.na(data$Nits)) & !(is.na(data$TipusHabitacio)),]  
  
# VIsualistació del contingut del nostre dataset  
  
summary(data)

## nomHotel notapersona data estrelles   
## Length:16276 Min. : 2.5 Min. :2017-04-13 Min. :1.0   
## Class :character 1st Qu.: 7.5 1st Qu.:2017-11-25 1st Qu.:2.0   
## Mode :character Median : 8.8 Median :2018-05-13 Median :3.0   
## Mean : 8.6 Mean :2018-05-06 Mean :3.1   
## 3rd Qu.:10.0 3rd Qu.:2018-11-04 3rd Qu.:4.0   
## Max. :10.0 Max. :2019-04-13 Max. :5.0   
## notaHotel TipusHabitacio DuradaEstancia Nits   
## Min. :7.9 Length:16276 Length:16276 Min. : 1.0   
## 1st Qu.:8.3 Class :character Class :character 1st Qu.: 1.0   
## Median :8.5 Mode :character Mode :character Median : 2.0   
## Mean :8.5 Mean : 2.3   
## 3rd Qu.:8.7 3rd Qu.: 3.0   
## Max. :9.4 Max. :27.0   
## ProcedenciaComentari TipusViatge Acompanyament   
## Length:16276 Length:16276 Length:16276   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## ViajaConMascota   
## Length:16276   
## Class :character   
## Mode :character   
##   
##   
##

## Eliminacó si cal de valors outliers i fringeliers

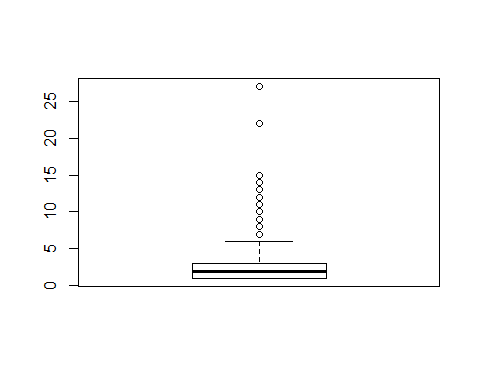
Mirarem els valors 0 de les notes mitjana d’hotel i les notes de les persones. Recordem que el valor 0 en les estrelles hem entés que significava que era un apartament i per tant no entra en el nostre estudi, ara mirem les notes mitja dels hotels i la nota de les persones.

Eliminem els registres que sobrepassen 3 desviacions estandars la nota donada per les persones.Ho fem a nivell d’hotel, i no pas a nivell general, perquè entenem que si un hotel és molt dolent potser tothom el puntua amb 1 o 2, que de segur serien outliers si els considressim en la totalitat de les puntuacions de tots els hotels, pero serien valors força normals a la puntuació própia del hotel.

dataaux<-data  
noudata <-NULL  
{  
 remove\_outliers <- function(x, limit = 3) {  
 mn <- mean(x, na.rm = T)  
 out <- limit \* sd(x, na.rm = T)  
 x < (mn - out) | x > (mn + out)  
 }  
 hotels.outliers<-unique(data$nomHotel)  
   
 for (cols in hotels.outliers){  
   
 un.hotel<-data[data$nomHotel==cols,]  
 if (NROW(un.hotel)>1) {  
 un.hotel<-un.hotel[remove\_outliers(un.hotel$notapersona,3)==FALSE,]  
 }  
   
 if (is.null(noudata) )  
 { noudata<-un.hotel}  
 else  
 {  
 noudata<-rbind(noudata,un.hotel)  
 }  
 }  
}  
  
summary(noudata)

## nomHotel notapersona data estrelles   
## Length:16082 Min. : 2.5 Min. :2017-04-13 Min. :1.0   
## Class :character 1st Qu.: 7.9 1st Qu.:2017-11-25 1st Qu.:2.0   
## Mode :character Median : 9.0 Median :2018-05-13 Median :3.0   
## Mean : 8.6 Mean :2018-05-06 Mean :3.1   
## 3rd Qu.:10.0 3rd Qu.:2018-11-04 3rd Qu.:4.0   
## Max. :10.0 Max. :2019-04-13 Max. :5.0   
## notaHotel TipusHabitacio DuradaEstancia Nits   
## Min. :7.9 Length:16082 Length:16082 Min. : 1.0   
## 1st Qu.:8.3 Class :character Class :character 1st Qu.: 1.0   
## Median :8.5 Mode :character Mode :character Median : 2.0   
## Mean :8.5 Mean : 2.3   
## 3rd Qu.:8.7 3rd Qu.: 3.0   
## Max. :9.4 Max. :27.0   
## ProcedenciaComentari TipusViatge Acompanyament   
## Length:16082 Length:16082 Length:16082   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## ViajaConMascota   
## Length:16082   
## Class :character   
## Mode :character   
##   
##   
##

data<-noudata  
  
#ANEm a mirar la distribució del nombre de nits  
boxplot(data$Nits)



outlier\_superior<-IQR(data$Nits)\*1.5+quantile(data$Nits)[4]  
#Anem si les estàncies llargues es concentren en uns hotels específics o no.  
kable(head(data %>% filter(Nits>outlier\_superior) %>% group\_by(nomHotel) %>% summarize(total\_hot=n()),10),caption="Holtes amb duració llarga")

Holtes amb duració llarga

|  |  |
| --- | --- |
| nomHotel | total\_hot |
| Acta BCN 40 | 1 |
| Atlantis by Atbcn | 5 |
| Ayre Hotel Gran Vía | 5 |
| Barceló Sants | 7 |
| Bcn Urban Hotels Gran Rosellon | 3 |
| Boutique Hotel H10 Montcada | 1 |
| Catalonia Avinyo | 6 |
| Catalonia Born | 3 |
| Catalonia Gran Via BCN | 7 |
| Catalonia La Pedrera | 4 |

Sembla que si que hi hagi uns hotels on el nombre de dies és més elevat, però no veiem cap disbarat en el nombre de nits, per tant, considerem que són valors vàlids i que no hem de fer cap tractament.

## Factorització de variables categòriques

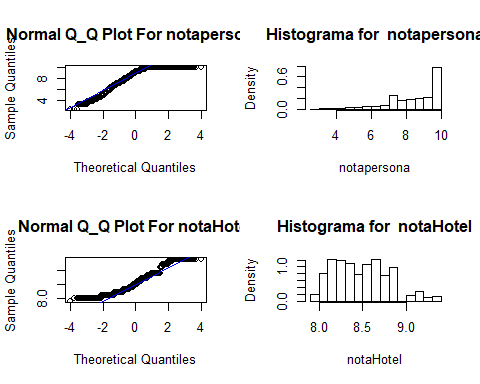
#Anem a transformar les variables categóriques as.factor  
data$ViajaConMascota<-as.factor(data$ViajaConMascota)  
data$Acompanyament<-as.factor(data$Acompanyament)  
data$TipusViatge<-as.factor(data$TipusViatge)  
data$ProcedenciaComentari<-as.factor(data$ProcedenciaComentari)  
data$DuradaEstancia<-as.factor(data$DuradaEstancia)  
data$TipusHabitacio<-as.factor(data$TipusHabitacio)  
data$estrelles<-as.factor(data$estrelles)  
  
# En el cas de estrelles , els hi donem un ordre, ja que la qualificació és el que indica.  
data$estrelles<-ordered(data$estrelles,levels=c("1","2","3","4","5"))  
data$nomHotel<-as.factor(data$nomHotel)  
summary(data)

## nomHotel notapersona   
## Catalonia Sagrada Familia : 911 Min. : 2.5   
## Hotel Lloret Ramblas : 834 1st Qu.: 7.9   
## Ayre Hotel Gran Vía : 775 Median : 9.0   
## Barceló Sants : 627 Mean : 8.6   
## Hotel Cortes : 540 3rd Qu.:10.0   
## NH Collection Barcelona Constanza: 463 Max. :10.0   
## (Other) :11932   
## data estrelles notaHotel TipusHabitacio   
## Min. :2017-04-13 1:1364 Min. :7.9 Altres : 1167   
## 1st Qu.:2017-11-25 2:3068 1st Qu.:8.3 Habitacio Doble :12509   
## Median :2018-05-13 3:4415 Median :8.5 Habitacio Individual: 1411   
## Mean :2018-05-06 4:6907 Mean :8.5 Habitacio quadruple : 199   
## 3rd Qu.:2018-11-04 5: 328 3rd Qu.:8.7 Habitacio Triple : 678   
## Max. :2019-04-13 Max. :9.4 Suite : 118   
##   
## DuradaEstancia Nits ProcedenciaComentari  
## De 3 a 5 Nits :5338 Min. : 1.0 Enviadopormóvil:11384   
## De 6 a 7 Nits : 133 1st Qu.: 1.0 EnviadoporWeb : 4698   
## Dos nits :4145 Median : 2.0   
## Estanciade6noches: 286 Mean : 2.3   
## Mesde7 : 109 3rd Qu.: 3.0   
## Una Nit :6071 Max. :27.0   
##   
## TipusViatge Acompanyament ViajaConMascota  
## SenseInformar : 1619 Familiaconniñosmayores : 2 N:16051   
## Viajedenegocios: 2066 Familiaconniñospequeños:2960 S: 31   
## Viajedeocio :12397 Grupo :1712   
## Grupodeamigos : 419   
## Pareja :7780   
## Personaqueviajasola :3209   
##

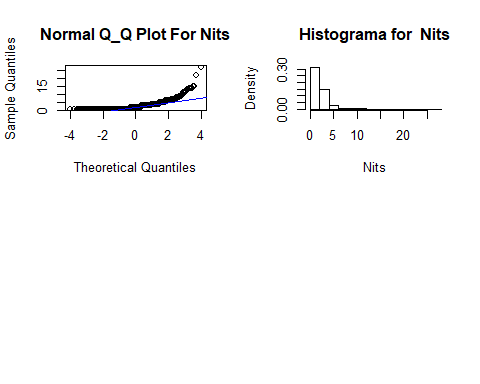
#Ens adonem que acompanyament te un factor que es 'Familia con niños mayores' i 'Familia con ninos pequeños', i el grup de 'Familia con niños mayores' es força petit, pel que agrupem els dos nivells a 'Familiaconniños'  
levels(data$Acompanyament) <- c("Familiaconniños","Familiaconniños","Grupo","Grupodeamigos","Pareja","Personaqueviajasola")

## Mirem les gràfiques quantile-quamtile plot i l’histograma per veure si hi ha variables susceptibles a la normalització

par(mfrow=c(2,2))  
for(i in 1:ncol(data)) {  
 if (is.numeric(data[,i])){  
 qqnorm(data[,i],main=paste("Normal Q\_Q Plot For",colnames(data)[i]))  
 qqline(data[,i],col="blue")  
 hist(data[,i],main=paste("Histograma for ",colnames(data)[i]), xlab=colnames(data)[i],freq=FALSE)  
 }  
}



par(mfrow=c(1,1))



La visualització dels gràfics no ens mostra normalitat en les dades, però per acabar-ho de concreta, aplicarem els test (no ho fem sobre notaHotel perquè pensem que no es pot estudiar en aquest dataset) Per acabar de comprobar, anem aplicar el test de Shapiro

set.seed(1000)  
shapiro.test(sample(data$notapersona,5000))

##   
## Shapiro-Wilk normality test  
##   
## data: sample(data$notapersona, 5000)  
## W = 0.9, p-value <2e-16

shapiro.test(sample(data$Nits,5000))

##   
## Shapiro-Wilk normality test  
##   
## data: sample(data$Nits, 5000)  
## W = 0.8, p-value <2e-16

Amb ambdos casos rebutjem la hipotésis nul·la i per tant no podem assumir normalitat, tal ic om haviem vist amb les gràfiques.

## Nous atributs

Ens decidim ara a fer les primeres hipotesis, i la primera és que volem saber si les notes que es donen en el primer trimestre de l’any son diferents a les que es donen a l’estiu

Per això, caldrà, que afegim una variable nova, que determini l’epoca de l’any en funció de la data del comentari (pot haver-hi un decalatge de dies, però assumin que els comentaris es fan aprop de la data en la que s’ha gaudit l’estancia de l’hotel)

data$Season<-"PRIMER TRIMESTRE"  
data[month(data$data)>=4 & month(data$data)<=6,"Season" ]<-"SEGON TRIMESTRE"  
data[month(data$data)>=7 & month(data$data)<=9,"Season" ]<-"TERCER TRIMESTRE"  
data[month(data$data)>=10 & month(data$data)<=12,"Season" ]<-"QUART TRIMESTRE"  
data$Season<-as.factor(data$Season)  
summary(data)

## nomHotel notapersona   
## Catalonia Sagrada Familia : 911 Min. : 2.5   
## Hotel Lloret Ramblas : 834 1st Qu.: 7.9   
## Ayre Hotel Gran Vía : 775 Median : 9.0   
## Barceló Sants : 627 Mean : 8.6   
## Hotel Cortes : 540 3rd Qu.:10.0   
## NH Collection Barcelona Constanza: 463 Max. :10.0   
## (Other) :11932   
## data estrelles notaHotel TipusHabitacio   
## Min. :2017-04-13 1:1364 Min. :7.9 Altres : 1167   
## 1st Qu.:2017-11-25 2:3068 1st Qu.:8.3 Habitacio Doble :12509   
## Median :2018-05-13 3:4415 Median :8.5 Habitacio Individual: 1411   
## Mean :2018-05-06 4:6907 Mean :8.5 Habitacio quadruple : 199   
## 3rd Qu.:2018-11-04 5: 328 3rd Qu.:8.7 Habitacio Triple : 678   
## Max. :2019-04-13 Max. :9.4 Suite : 118   
##   
## DuradaEstancia Nits ProcedenciaComentari  
## De 3 a 5 Nits :5338 Min. : 1.0 Enviadopormóvil:11384   
## De 6 a 7 Nits : 133 1st Qu.: 1.0 EnviadoporWeb : 4698   
## Dos nits :4145 Median : 2.0   
## Estanciade6noches: 286 Mean : 2.3   
## Mesde7 : 109 3rd Qu.: 3.0   
## Una Nit :6071 Max. :27.0   
##   
## TipusViatge Acompanyament ViajaConMascota  
## SenseInformar : 1619 Familiaconniños :2962 N:16051   
## Viajedenegocios: 2066 Grupo :1712 S: 31   
## Viajedeocio :12397 Grupodeamigos : 419   
## Pareja :7780   
## Personaqueviajasola:3209   
##   
##   
## Season   
## PRIMER TRIMESTRE:4429   
## QUART TRIMESTRE :4078   
## SEGON TRIMESTRE :3770   
## TERCER TRIMESTRE:3805   
##   
##   
##

Anem a veure si la diferencia en general de les notes de les persones pels hotels difereix entre el primer trimestre i el tercer trimetre. Per això,estudiarem la normalitat i la homocestadicitat de la nota. En un primer moment pel conjunt de la mostra i poc a poc anirem diferentciant pel conjunt de categoritzacions que aporten les dades

Amb tot això ja sabem que la nota persona no segueixen una distribució normal ni hi ha homocedasticitat entre els grups separats per season. Per tant, com que no es compleix, el que farem es aplicar un test parametric Mann-Whitney

fligner.test(notapersona ~ Season, data = data)

##   
## Fligner-Killeen test of homogeneity of variances  
##   
## data: notapersona by Season  
## Fligner-Killeen:med chi-squared = 100, df = 3, p-value <2e-16

#No hi hahomocedasticitat aplqieum una prova no paramètica (wolcox.test)  
wilcox.test(notapersona ~ Season, data = data, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: notapersona by Season  
## W = 9e+06, p-value = 8e-15  
## alternative hypothesis: true location shift is not equal to 0

#Creem un dataset per treure resultats resumits  
hotels\_1er\_3er<-data[which(data$Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE")),]  
#Visualitzem mitjanes  
kable(hotels\_1er\_3er%>% group\_by(Season) %>% summarise(n=n(),mean=mean(notapersona),sd=sd(notapersona)))

|  |  |  |  |
| --- | --- | --- | --- |
| Season | n | mean | sd |
| PRIMER TRIMESTRE | 4429 | 8.7 | 1.4 |
| TERCER TRIMESTRE | 3805 | 8.5 | 1.4 |

Existeix una forta evidència de diferència de les mitjanes de les notes entre el primer trimestre i el tercer. Sembla ser que cap a l’estiu la qualitat es percebuda pitjor que a l’hivern. Això vol dir o que els clients són més exigents o el personal té menys qualitat. Volem estudiar ara si això es igual per conjunt de categoriques. Començarem per diferencies en funció de les estrelles.

#Anem a iterar hotel per hotel per determinar quin test realitzar i veure si hi ha diferències dignificatives.  
resultat.estrelles <- data.frame(Doubles=double(),  
 Ints=integer(),  
 Factors=factor(),  
 Logicals=logical(),  
 Characters=character(),  
 stringsAsFactors=FALSE)  
llistat\_estrelles<-levels(data$estrelles)  
for(cols in llistat\_estrelles)  
{  
 aux<-data[which(data$estrelles==cols),]  
   
 sha<-shapiro.test(sample(aux$notapersona,min(nrow(aux),5000)))  
 homo<- fligner.test(notapersona ~ Season, data = aux)  
  
 test<-"No Parametric"  
 if ((sha$p.value>=0.05) & (homo$p.value>=0.05))  
 {  
 #tenim homocedasticitat  
 test<-"Student"  
 }  
   
 if (test=='No Parametric')   
 {  
 test.comparacio<-wilcox.test(notapersona ~ Season, data = aux, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))  
 }  
 else  
 {  
 test.comparacio<-t.test(notapersona ~ Season, data = aux, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))  
 }  
  
 actual<-data.frame(cols,sha$p.value,homo$p.value,test,test.comparacio$p.value,mean(aux[which(aux$Season=="PRIMER TRIMESTRE"),"notapersona"]),mean(aux[which(aux$Season=="TERCER TRIMESTRE"),"notapersona"]), NROW(aux))  
 colnames(actual)<-c("Estrelles","Shapiro","Fligner","Test Aplicat","Comparacio","Mitjana Primer Trimestre","Mitjana Tercer Trimestre","Nombre Comentaris")  
 #Construim la taula de resultats  
 resultat.estrelles<-rbind(resultat.estrelles,actual)  
   
}  
  
kable(resultat.estrelles,caption="Resum notes per estrelles")

Resum notes per estrelles

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Estrelles | Shapiro | Fligner | Test Aplicat | Comparacio | Mitjana Primer Trimestre | Mitjana Tercer Trimestre | Nombre Comentaris |
| 1 | 0 | 0.12 | No Parametric | 0.03 | 8.5 | 8.2 | 1364 |
| 2 | 0 | 0.01 | No Parametric | 0.00 | 8.7 | 8.4 | 3068 |
| 3 | 0 | 0.00 | No Parametric | 0.00 | 8.6 | 8.4 | 4415 |
| 4 | 0 | 0.01 | No Parametric | 0.00 | 8.9 | 8.7 | 6907 |
| 5 | 0 | 0.27 | No Parametric | 0.16 | 8.8 | 8.4 | 328 |

Veient els resultat, veiem que en el cas de les estrelles, són els de cinc que no hi ha diferències entre les notes mitjanes pero si entre els de 1 a 4 estrelles.

Ara volem comprovar per tipus d’habitació

#Per tipus de habitació  
  
resultat<- data.frame(Doubles=double(),  
 Ints=integer(),  
 Factors=factor(),  
 Logicals=logical(),  
 Characters=character(),  
 stringsAsFactors=FALSE)  
  
llistat<-levels(data$TipusHabitacio)  
for(cols in llistat)  
{  
 aux<-data[which(data$TipusHabitacio==cols),]  
   
 sha<-shapiro.test(sample(aux$notapersona,min(nrow(aux),5000)))  
 homo<- fligner.test(notapersona ~ Season, data = aux)  
  
 test<-"No Parametric"  
 if ((sha$p.value>=0.05) & (homo$p.value>=0.05))  
 {  
 #tenim homocedasticitat  
 test<-"Student"  
 }  
   
 if (test=='No Parametric')   
 {  
 test.comparacio<-wilcox.test(notapersona ~ Season, data = aux, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))  
 }  
 else  
 {  
 test.comparacio<-t.test(notapersona ~ Season, data = aux, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))  
 }  
 actual<-data.frame(cols,sha$p.value,homo$p.value,test,test.comparacio$p.value,mean(aux[which(aux$Season=="PRIMER TRIMESTRE"),"notapersona"]),mean(aux[which(aux$Season=="TERCER TRIMESTRE"),"notapersona"]), NROW(aux))  
 colnames(actual)<-c("Tipus Habitacio","Shapiro","Fligner","Test Aplicat","Comparacio","Mitjana Primer Trimestre","Mitjana Tercer Trimestre", "Nombre Comentaris")  
 #Construim la taula de resultats  
 resultat<-rbind(resultat,actual)  
   
}  
  
kable(resultat, caption="REsum per Tipus Habitació")

REsum per Tipus Habitació

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tipus Habitacio | Shapiro | Fligner | Test Aplicat | Comparacio | Mitjana Primer Trimestre | Mitjana Tercer Trimestre | Nombre Comentaris |
| Altres | 0 | 0.20 | No Parametric | 0.05 | 8.7 | 8.5 | 1167 |
| Habitacio Doble | 0 | 0.00 | No Parametric | 0.00 | 8.8 | 8.5 | 12509 |
| Habitacio Individual | 0 | 0.51 | No Parametric | 0.10 | 8.5 | 8.3 | 1411 |
| Habitacio quadruple | 0 | 0.05 | No Parametric | 0.01 | 9.2 | 8.6 | 199 |
| Habitacio Triple | 0 | 0.85 | No Parametric | 0.34 | 8.8 | 8.7 | 678 |
| Suite | 0 | 0.89 | No Parametric | 0.40 | 9.0 | 8.9 | 118 |

En el cas de tipus d’habitació, només veiem diferencia en els que han demanat una habitació quadruple i habitació doble.

Anem a estudiar per tipus de viatge

#Per DuradaEstancia  
  
resultat<- data.frame(Doubles=double(),  
 Ints=integer(),  
 Factors=factor(),  
 Logicals=logical(),  
 Characters=character(),  
 stringsAsFactors=FALSE)  
  
llistat<-levels(data$DuradaEstancia)  
for(cols in llistat)  
{  
 aux<-data[which(data$DuradaEstancia==cols),]  
   
 sha<-shapiro.test(sample(aux$notapersona,min(nrow(aux),5000)))  
 homo<- fligner.test(notapersona ~ Season, data = aux)  
  
 test<-"No Parametric"  
 if ((sha$p.value>=0.05) & (homo$p.value>=0.05))  
 {  
 #tenim homocedasticitat  
 test<-"Student"  
 }  
   
 if (test=='No Parametric')   
 {  
 test.comparacio<-wilcox.test(notapersona ~ Season, data = aux, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))  
 }  
 else  
 {  
 test.comparacio<-t.test(notapersona ~ Season, data = aux, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))  
 }  
 actual<-data.frame(cols,sha$p.value,homo$p.value,test,test.comparacio$p.value,mean(aux[which(aux$Season=="PRIMER TRIMESTRE"),"notapersona"]),mean(aux[which(aux$Season=="TERCER TRIMESTRE"),"notapersona"]),NROW(aux))  
 colnames(actual)<-c("Durada Estancia","Shapiro","Fligner","Test Aplicat","Comparacio","Mitjana Primer Trimestre","Mitjana Tercer Trimestre","Nombre Comentaris")  
 #Construim la taula de resultats  
 resultat<-rbind(resultat,actual)  
   
}  
  
kable(resultat, caption="Resum per durada Estancia")

Resum per durada Estancia

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Durada Estancia | Shapiro | Fligner | Test Aplicat | Comparacio | Mitjana Primer Trimestre | Mitjana Tercer Trimestre | Nombre Comentaris |
| De 3 a 5 Nits | 0 | 0.00 | No Parametric | 0.00 | 8.8 | 8.6 | 5338 |
| De 6 a 7 Nits | 0 | 0.15 | No Parametric | 0.61 | 9.0 | 8.9 | 133 |
| Dos nits | 0 | 0.00 | No Parametric | 0.00 | 8.7 | 8.5 | 4145 |
| Estanciade6noches | 0 | 0.68 | No Parametric | 0.13 | 8.9 | 8.5 | 286 |
| Mesde7 | 0 | 0.43 | No Parametric | 0.78 | 9.1 | 8.8 | 109 |
| Una Nit | 0 | 0.00 | No Parametric | 0.00 | 8.7 | 8.4 | 6071 |

En aquest cas veiem que com més llarga és l’estancia menys diferencia hi ha entre les notes d’estiu i d’hivern. A partir de 6 nits no hi ha evidència de diferencia de nota entre estiu i hivern.

#Per Acompanyament  
  
resultat<- data.frame(Doubles=double(),  
 Ints=integer(),  
 Factors=factor(),  
 Logicals=logical(),  
 Characters=character(),  
 stringsAsFactors=FALSE)  
  
llistat<-levels(data$Acompanyament)  
for(cols in llistat)  
{  
 aux<-data[which(data$Acompanyament==cols),]  
   
 sha<-shapiro.test(sample(aux$notapersona,min(nrow(aux),5000)))  
 homo<- fligner.test(notapersona ~ Season, data = aux)  
  
 test<-"No Parametric"  
 if ((sha$p.value>=0.05) & (homo$p.value>=0.05))  
 {  
 #tenim homocedasticitat  
 test<-"Student"  
 }  
   
 if (test=='No Parametric')   
 {  
 test.comparacio<-wilcox.test(notapersona ~ Season, data = aux, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))  
 }  
 else  
 {  
 test.comparacio<-t.test(notapersona ~ Season, data = aux, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))  
 }  
 actual<-data.frame(cols,sha$p.value,homo$p.value,test,test.comparacio$p.value,mean(aux[which(aux$Season=="PRIMER TRIMESTRE"),"notapersona"]),mean(aux[which(aux$Season=="TERCER TRIMESTRE"),"notapersona"]), NROW(aux))  
 colnames(actual)<-c("Acompanyament","Shapiro","Fligner","Test Aplicat","Comparacio","Mitjana Primer Trimestre","Mitjana Tercer Trimestre","Nombre Comentaris")  
 #Construim la taula de resultats  
 resultat<-rbind(resultat,actual)  
   
}  
  
kable(resultat, caption="Resum per Acompanyament")

Resum per Acompanyament

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Acompanyament | Shapiro | Fligner | Test Aplicat | Comparacio | Mitjana Primer Trimestre | Mitjana Tercer Trimestre | Nombre Comentaris |
| Familiaconniños | 0 | 0.00 | No Parametric | 0.00 | 8.8 | 8.6 | 2962 |
| Grupo | 0 | 0.18 | No Parametric | 0.01 | 8.6 | 8.4 | 1712 |
| Grupodeamigos | 0 | 0.01 | No Parametric | 0.05 | 9.0 | 8.7 | 419 |
| Pareja | 0 | 0.00 | No Parametric | 0.00 | 8.8 | 8.5 | 7780 |
| Personaqueviajasola | 0 | 0.00 | No Parametric | 0.00 | 8.6 | 8.3 | 3209 |

Només els que viatgem amb amics sembla que no percebin una diferencia entre la nota que posen a l’estiu i a l’hivern.

#Per TipusViatge  
  
resultat<- data.frame(Doubles=double(),  
 Ints=integer(),  
 Factors=factor(),  
 Logicals=logical(),  
 Characters=character(),  
 stringsAsFactors=FALSE)  
  
llistat<-levels(data$TipusViatge)  
for(cols in llistat)  
{  
 aux<-data[which(data$TipusViatge==cols),]  
   
 sha<-shapiro.test(sample(aux$notapersona,min(nrow(aux),5000)))  
 homo<- fligner.test(notapersona ~ Season, data = aux)  
  
 test<-"No Parametric"  
 if ((sha$p.value>=0.05) & (homo$p.value>=0.05))  
 {  
 #tenim homocedasticitat  
 test<-"Student"  
 }  
   
 if (test=='No Parametric')   
 {  
 test.comparacio<-wilcox.test(notapersona ~ Season, data = aux, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))  
 }  
 else  
 {  
 test.comparacio<-t.test(notapersona ~ Season, data = aux, subset = Season %in% c("PRIMER TRIMESTRE", "TERCER TRIMESTRE"))  
 }  
 actual<-data.frame(cols,sha$p.value,homo$p.value,test,test.comparacio$p.value,mean(aux[which(aux$Season=="PRIMER TRIMESTRE"),"notapersona"]),mean(aux[which(aux$Season=="TERCER TRIMESTRE"),"notapersona"]), NROW(aux))  
 colnames(actual)<-c("TipusViatge","Shapiro","Fligner","Test Aplicat","Comparacio","Mitjana Primer Trimestre","Mitjana Tercer Trimestre","Nombre Comentaris")  
 #Construim la taula de resultats  
 resultat<-rbind(resultat,actual)  
   
}  
  
kable(resultat, caption="Resum per TipusViatge")

Resum per TipusViatge

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TipusViatge | Shapiro | Fligner | Test Aplicat | Comparacio | Mitjana Primer Trimestre | Mitjana Tercer Trimestre | Nombre Comentaris |
| SenseInformar | 0 | 0 | No Parametric | 0 | 8.8 | 8.3 | 1619 |
| Viajedenegocios | 0 | 0 | No Parametric | 0 | 8.4 | 7.9 | 2066 |
| Viajedeocio | 0 | 0 | No Parametric | 0 | 8.8 | 8.6 | 12397 |

Per tipus de viatge, es percep que hi ha diferencia entre la nota que es posa a l’estiu respecte la nota que es posa a l’hiven.

Ara, ens proposem crear un algoritme que ens digui en funció de les categoriques quin és el millor hotel per nosaltres. Ens plantejem fer el següent. Crearem una regressió lineal per predicció de nota per cada hotel. Comprovarem el grau de significancia de la regressió. Un cop fet això, si veiem que podem prossegguir, aplicarem la regressió per cada hotel i obtindrem la llista d’hotels ordenat per nota ascendent, per donar-te la confiança i el grau de confiança de la recomenació (1-p\_value de la regressió)

model.hotel.nota<-lm(notapersona~TipusHabitacio+Nits+TipusViatge+Acompanyament+Season, data)  
summary(model.hotel.nota)

##   
## Call:  
## lm(formula = notapersona ~ TipusHabitacio + Nits + TipusViatge +   
## Acompanyament + Season, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.298 -0.884 0.383 1.183 2.115   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 8.74906 0.05950 147.03 < 2e-16  
## TipusHabitacioHabitacio Doble 0.02520 0.04308 0.59 0.55855  
## TipusHabitacioHabitacio Individual -0.23186 0.06115 -3.79 0.00015  
## TipusHabitacioHabitacio quadruple 0.17744 0.10686 1.66 0.09683  
## TipusHabitacioHabitacio Triple 0.08082 0.06736 1.20 0.23019  
## TipusHabitacioSuite 0.31517 0.13396 2.35 0.01865  
## Nits 0.01416 0.00745 1.90 0.05721  
## TipusViatgeViajedenegocios -0.31489 0.04722 -6.67 2.7e-11  
## TipusViatgeViajedeocio 0.04327 0.03691 1.17 0.24110  
## AcompanyamentGrupo -0.06894 0.04292 -1.61 0.10828  
## AcompanyamentGrupodeamigos 0.15566 0.07250 2.15 0.03181  
## AcompanyamentPareja -0.06075 0.03211 -1.89 0.05847  
## AcompanyamentPersonaqueviajasola -0.02677 0.04339 -0.62 0.53726  
## SeasonQUART TRIMESTRE -0.02949 0.03001 -0.98 0.32581  
## SeasonSEGON TRIMESTRE -0.22503 0.03073 -7.32 2.5e-13  
## SeasonTERCER TRIMESTRE -0.27684 0.03078 -8.99 < 2e-16  
##   
## (Intercept) \*\*\*  
## TipusHabitacioHabitacio Doble   
## TipusHabitacioHabitacio Individual \*\*\*  
## TipusHabitacioHabitacio quadruple .   
## TipusHabitacioHabitacio Triple   
## TipusHabitacioSuite \*   
## Nits .   
## TipusViatgeViajedenegocios \*\*\*  
## TipusViatgeViajedeocio   
## AcompanyamentGrupo   
## AcompanyamentGrupodeamigos \*   
## AcompanyamentPareja .   
## AcompanyamentPersonaqueviajasola   
## SeasonQUART TRIMESTRE   
## SeasonSEGON TRIMESTRE \*\*\*  
## SeasonTERCER TRIMESTRE \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.4 on 16066 degrees of freedom  
## Multiple R-squared: 0.0191, Adjusted R-squared: 0.0182   
## F-statistic: 20.9 on 15 and 16066 DF, p-value: <2e-16

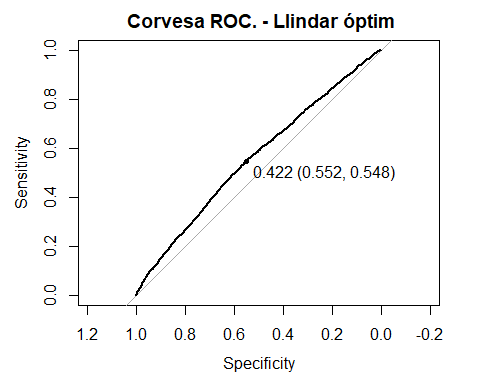
En termes generals la nota no es pot explicar a través d’aquestes dades perqué el valor d’R^2 é smolt baix, pel que no som capaços d’explicar tota la variabilitat, és a dir, ens falten informació (regressors) que ajudin a predir el comportament.

Anem a plantejar una altra estrategia, en aquest cas determinem que una persona recomana un hotel X si la nota que ha posat (notapersona) és superior a la nota mitjana de l’hotel (notaHOtel), per fer això, crearem una variable binària (Recomana amb valors S/N i mirarem de treballar amb una regressió logística). Mirarem de recomenar un hotel en funció de les variables explicatives.

#Creem la nova variable  
data$recomanar<-ifelse(data$notaHotel<data$notapersona,0,1)  
#Validem que hi hagi sempre més d'un cas  
hotel.excloure<-data %>% group\_by(nomHotel,recomanar) %>% group\_by(nomHotel) %>% summarise(n=n\_distinct(recomanar)) %>% filter(n<2)  
  
#excloem els que no tenen més d'un cas  
data<-data[which(!(data$nomHotel %in% hotel.excloure$nomHotel)),]  
  
#Mirem en termes generals els resultats de recomanació sense entrar hotel per hotel  
model.recomanacio<-glm(recomanar~TipusHabitacio+Nits+TipusViatge+Acompanyament+Season , data = data)  
summary(model.recomanacio)

##   
## Call:  
## glm(formula = recomanar ~ TipusHabitacio + Nits + TipusViatge +   
## Acompanyament + Season, data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.639 -0.425 -0.360 0.569 0.727   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.370282 0.021096 17.55 < 2e-16  
## TipusHabitacioHabitacio Doble -0.011322 0.015275 -0.74 0.459  
## TipusHabitacioHabitacio Individual 0.059547 0.021678 2.75 0.006  
## TipusHabitacioHabitacio quadruple -0.077846 0.037879 -2.06 0.040  
## TipusHabitacioHabitacio Triple -0.030655 0.023879 -1.28 0.199  
## TipusHabitacioSuite -0.096231 0.047485 -2.03 0.043  
## Nits -0.001327 0.002640 -0.50 0.615  
## TipusViatgeViajedenegocios 0.104878 0.016738 6.27 3.8e-10  
## TipusViatgeViajedeocio -0.010981 0.013084 -0.84 0.401  
## AcompanyamentGrupo 0.025375 0.015217 1.67 0.095  
## AcompanyamentGrupodeamigos -0.041209 0.025698 -1.60 0.109  
## AcompanyamentPareja 0.019388 0.011380 1.70 0.088  
## AcompanyamentPersonaqueviajasola 0.000352 0.015381 0.02 0.982  
## SeasonQUART TRIMESTRE 0.013427 0.010638 1.26 0.207  
## SeasonSEGON TRIMESTRE 0.065118 0.010893 5.98 2.3e-09  
## SeasonTERCER TRIMESTRE 0.080401 0.010911 7.37 1.8e-13  
##   
## (Intercept) \*\*\*  
## TipusHabitacioHabitacio Doble   
## TipusHabitacioHabitacio Individual \*\*   
## TipusHabitacioHabitacio quadruple \*   
## TipusHabitacioHabitacio Triple   
## TipusHabitacioSuite \*   
## Nits   
## TipusViatgeViajedenegocios \*\*\*  
## TipusViatgeViajedeocio   
## AcompanyamentGrupo .   
## AcompanyamentGrupodeamigos   
## AcompanyamentPareja .   
## AcompanyamentPersonaqueviajasola   
## SeasonQUART TRIMESTRE   
## SeasonSEGON TRIMESTRE \*\*\*  
## SeasonTERCER TRIMESTRE \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.24)  
##   
## Null deviance: 3902.6 on 16080 degrees of freedom  
## Residual deviance: 3851.1 on 16065 degrees of freedom  
## AIC: 22686  
##   
## Number of Fisher Scoring iterations: 2

pred<-predict(model.recomanacio,data,type="response")  
corba<-roc(data$recomanar,pred)  
plot(corba,print.thres="best", print.thres.best.method="closest.topleft", main=" Corvesa ROC. - Llindar óptim")



area<-auc(data$recomanar, pred)  
optim\_point<-optim.thresh(data$recomanar,pred)$min.ROC.plot.distance[1]  
  
  
# Creo vector amb les predccions segons el llibdar demanat.  
prediccions <- ifelse(test = model.recomanacio$fitted.values >= optim\_point,   
 yes = 1, no = 0)  
#Creem la matriu de confusió  
matriu\_confusio<- addmargins(table(data$recomanar, prediccions,  
 dnn = c("observacions", "prediccions")))  
  
#Matriu de confusio  
matriu\_confusio

## prediccions  
## observacions 0 1 Sum  
## 0 5112 4304 9416  
## 1 2955 3710 6665  
## Sum 8067 8014 16081

El model es força dolent, el punt òptim seria determinar 0.422 com a valor a partir del qual acceptem els possius. La matriu de confusions ens dona un % d’encert molt dolent. Assumint un valor de confiança 0.05 només són significatives, el tipus d’habitació individu, la quadruple i la suite. Pel tipus de viatge el de negocis. Acompanyament no és significatiu i el segon i el tercer Trimestre són significatius.

Assumint el que hem dit, el model ens diu que per recomanar o no un hotel comencem per una puntuació de 0.37. Tots els tipus d’habitació resten a excepció de l’habitació Individual i Altres. Les nits, resten 0.001 per cada nit. Els que viatge per negocis puntuene millor pels que els d’oci resten. Per grups resta la nota quan qui viatge es un grup d’amics i tant acompanyament com Quart Segon Tercer trimestre sumen a la nota.

Això, és un exercici per ajudar a entedre que fa que una persona recomani o no un hotel. Però per poder fer un model recomanació, el que ens plantegem fer, és un model de regressió per cada hotel, de manera que recomenem o no un hotel en funció de les variables descriptives disponibles en cada hotel.

#Anem a ferun regressió logistica per hotel i creant el resultat en forma de taula  
llista\_hotels=unique(data$nomHotel)  
result\_roc<-data.frame(Doubles=double(),  
 Ints=integer(),  
 Factors=factor(),  
 Logicals=logical(),  
 Characters=character(),  
 stringsAsFactors=FALSE)  
list.model.recomanacio<-NULL  
for (cols in llista\_hotels)  
{  
   
 data.regressio<-data[data$nomHotel==cols,]  
 data.regressio$recomanar<-ifelse(data.regressio$notaHotel<data.regressio$notapersona,0,1)  
 #Treiem els hotels que no tinguin com a mínim 2 nivells en les categoriques que estem estudiant  
 if ( NROW(levels(as.factor(as.character(data.regressio$TipusHabitacio))))>2 &  
 NROW(levels(as.factor(as.character(data.regressio$TipusViatge))))>2 &  
 NROW(levels(as.factor(as.character(data.regressio$Acompanyament))))>2 &  
 NROW(levels(as.factor(as.character(data.regressio$Season))))>2 )  
 {  
   
 model.recomanacio<-glm(recomanar~TipusHabitacio+Nits+TipusViatge+Acompanyament+Season , data = data.regressio)  
 pred<-predict(model.recomanacio,data.regressio,type="response")  
 area<-auc(data.regressio$recomanar, pred)  
 optim\_point<-optim.thresh(data.regressio$recomanar,pred)$min.ROC.plot.distance[1]  
  
 aux.merge<-data.frame(cols,area,as.numeric(optim\_point),data.regressio$notaHotel[1])  
 result\_roc<-rbind(result\_roc,aux.merge)  
 aux.list<-list(list("nomhotel"=cols,"model"=model.recomanacio))  
 list.model.recomanacio<-append(list.model.recomanacio,aux.list)  
  
 }  
}  
colnames(result\_roc)<-c("NomHotel","AUC","Optim\_point","nota")  
#treiem la taula dels punts òptims on donar el possitius per cada hotel  
kable(head(result\_roc[,c(1:4)],10),caption="Punt òptim d'acceptació de possitus per hotel")

Punt òptim d’acceptació de possitus per hotel

|  |  |  |  |
| --- | --- | --- | --- |
| NomHotel | AUC | Optim\_point | nota |
| Hostal Balkonis | 0.68 | 0.45 | 8.8 |
| Hotel Turin | 0.70 | 0.47 | 8.5 |
| Hotel Arts Barcelona | 0.75 | 0.37 | 8.6 |
| Hotel Casa Bonay | 0.83 | 0.33 | 8.7 |
| Upper Diagonal | 0.65 | 0.38 | 8.9 |
| Hotel El Call | 0.70 | 0.47 | 8.3 |
| Hostal Fernando | 0.63 | 0.43 | 8.5 |
| Grupotel Gran Via 678 | 0.56 | 0.35 | 8.6 |
| Hostal Benidorm | 0.64 | 0.42 | 8.5 |
| Golden Tulip Barcelona | 0.67 | 0.34 | 8.7 |

**Tenim més o menys precisió en funció de l’hotel, però de cara l’exercici dinàmic farem el model de predicció obviant que hi ha auc, arees molt baixes, per tant no serien uns grans models, en aquest casos. Per avançar en l’exercici, obviem mirar la significància de cadasquna de les variables així com si per cada hotel acabem tenint prou mostra, és a dir, diferents combinacions de descripcions de categòriques. S’hauria de fer per garantir un bon model però creiem interessant plantejar una primera iteració que quedaria oberta per posteriors refinaments**

recomanacio\_hotels<-function(entrada, list.model.recomanacio,result\_roc)  
{  
   
 recomanacio<-data.frame(Doubles=double(),  
 Ints=integer(),  
 Factors=factor(),  
 Logicals=logical(),  
 Characters=character(),  
 stringsAsFactors=FALSE)  
 roc\_actual<-data.frame(Doubles=double(),  
 Ints=integer(),  
 Factors=factor(),  
 Logicals=logical(),  
 Characters=character(),  
 stringsAsFactors=FALSE)  
 norecomenats<-data.frame(Doubles=double(),  
 Ints=integer(),  
 Factors=factor(),  
 Logicals=logical(),  
 Characters=character(),  
 stringsAsFactors=FALSE)  
 for (i in 1:NROW(list.model.recomanacio))  
 {  
 actual<-list.model.recomanacio[[i]]  
 predict.hotel<-predict(newdata=entrada,actual$model)  
 roc\_actual<-result\_roc[which(result\_roc$NomHotel==actual$nomhotel),]  
 if (round(predict.hotel,2)>=round(roc\_actual$Optim\_point,2))   
 {  
   
 result\_actual<-data.frame(as.character(roc\_actual$NomHotel),roc\_actual$nota,predict.hotel,roc\_actual$Optim\_point)  
 colnames(result\_actual)<-c("Nom Hotel","Nota","Valor\_Prediccio","Valor\_Optim")  
 recomanacio<-rbind(recomanacio,result\_actual)   
 }  
 else  
 {  
 result\_actual<-data.frame(as.character(roc\_actual$NomHotel),roc\_actual$nota,predict.hotel,roc\_actual$Optim\_point)  
 colnames(result\_actual)<-c("Nom Hotel","Nota","Valor\_Prediccio","Valor\_Optim")  
 norecomenats<-rbind(norecomenats,result\_actual)   
 }  
 }  
   
 recomanacio$diferencia<-recomanacio$Valor\_Prediccio-recomanacio$Valor\_Optim  
 norecomenats$diferencia<-norecomenats$Valor\_Prediccio-norecomenats$Valor\_Optim  
 recomanacio<-recomanacio %>% arrange(desc(diferencia,Nota))  
 norecomenats<-norecomenats %>% arrange(desc(diferencia,Nota))  
  
 return(list("recomenats"=recomanacio, "norecomenats"=norecomenats))  
}  
  
entrada<-data.frame("Habitacio Doble",3,"Viajedeocio","Pareja","PRIMER TRIMESTRE")  
#entrada<-data.frame("Habitacio Doble",1,"Viajedeocio","Grupo","SEGON TRIMESTRE")  
  
colnames(entrada)<-c("TipusHabitacio","Nits","TipusViatge","Acompanyament","Season")  
  
hotels.recomanats<-recomanacio\_hotels(entrada,list.model.recomanacio,result\_roc)  
  
kable(hotels.recomanats$recomenats, caption="Hotels recomenats pel dataframe Entrada")

Hotels recomenats pel dataframe Entrada

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nom Hotel | Nota | Valor\_Prediccio | Valor\_Optim | diferencia |
| H10 Port Vell 4\* Sup | 9.2 | 0.77 | 0.54 | 0.23 |
| Fairmont Rey Juan Carlos I | 8.7 | 0.63 | 0.47 | 0.16 |
| Hotel El Call | 8.3 | 0.60 | 0.47 | 0.13 |
| Catalonia Born | 8.9 | 0.46 | 0.36 | 0.10 |
| HCC Taber | 8.4 | 0.53 | 0.45 | 0.08 |
| Hesperia Barcelona Ramblas | 8.6 | 0.45 | 0.37 | 0.08 |
| Hotel Market | 8.1 | 0.58 | 0.51 | 0.07 |
| Cuatro Naciones | 8.4 | 0.48 | 0.42 | 0.06 |
| Acta Atrium Palace | 8.4 | 0.45 | 0.40 | 0.05 |
| Catalonia Avinyo | 8.6 | 0.36 | 0.31 | 0.05 |
| Pullman Barcelona Skipper | 8.5 | 0.49 | 0.44 | 0.05 |
| NH Barcelona Stadium | 8.3 | 0.56 | 0.52 | 0.04 |
| Hotel Lleó | 8.5 | 0.41 | 0.38 | 0.03 |
| HCC Montblanc | 8.3 | 0.41 | 0.41 | 0.00 |
| Grupotel Gran Via 678 | 8.6 | 0.35 | 0.35 | 0.00 |
| Paral·lel | 8.6 | 0.40 | 0.40 | 0.00 |
| Tryp Barcelona Apolo Hotel | 8.6 | 0.44 | 0.44 | 0.00 |

kable(hotels.recomanats$norecomenats, caption="Hotels no recomenats pel dataframe Entrada")

Hotels no recomenats pel dataframe Entrada

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nom Hotel | Nota | Valor\_Prediccio | Valor\_Optim | diferencia |
| Olivia Balmes Hotel | 9.4 | 0.22 | 0.23 | -0.01 |
| Ciutat de Barcelona | 8.6 | 0.41 | 0.42 | -0.01 |
| Exe Plaza Catalunya | 8.4 | 0.38 | 0.39 | -0.01 |
| Golden Tulip Barcelona | 8.7 | 0.32 | 0.34 | -0.02 |
| Catalunya | 8.0 | 0.49 | 0.51 | -0.02 |
| ICON BCN by Petit Palace | 9.1 | 0.23 | 0.25 | -0.02 |
| Hotel 54 Barceloneta | 8.3 | 0.44 | 0.47 | -0.03 |
| Ayre Hotel Gran Vía | 8.9 | 0.37 | 0.40 | -0.03 |
| Hostal Fernando | 8.5 | 0.40 | 0.43 | -0.03 |
| Hotel Lloret Ramblas | 8.1 | 0.40 | 0.43 | -0.03 |
| Hotel Ginebra | 8.6 | 0.33 | 0.37 | -0.04 |
| Acta BCN 40 | 8.3 | 0.52 | 0.56 | -0.04 |
| Hostal Balkonis | 8.8 | 0.41 | 0.45 | -0.04 |
| NH Collection Barcelona Podium | 8.9 | 0.27 | 0.31 | -0.04 |
| Hotel Cortes | 8.3 | 0.48 | 0.52 | -0.04 |
| Hotel Balmes | 8.6 | 0.35 | 0.40 | -0.05 |
| Pestana Arena Barcelona | 8.7 | 0.28 | 0.33 | -0.05 |
| NH Collection Barcelona Constanza | 8.8 | 0.31 | 0.37 | -0.06 |
| Barceló Sants | 8.7 | 0.33 | 0.39 | -0.06 |
| Catalonia Gran Via BCN | 8.8 | 0.35 | 0.41 | -0.06 |
| Guest House Center Inn | 8.4 | 0.34 | 0.40 | -0.06 |
| Hesperia Barcelona Barri Gòtic | 8.6 | 0.39 | 0.45 | -0.06 |
| Catalonia La Pedrera | 8.4 | 0.44 | 0.51 | -0.07 |
| Hotel Turin | 8.5 | 0.39 | 0.47 | -0.08 |
| Hotel Casa Bonay | 8.7 | 0.25 | 0.33 | -0.08 |
| Catalonia Sagrada Familia | 8.2 | 0.30 | 0.39 | -0.09 |
| Cram | 8.4 | 0.33 | 0.42 | -0.09 |
| Hostal Benidorm | 8.5 | 0.32 | 0.42 | -0.10 |
| Upper Diagonal | 8.9 | 0.28 | 0.38 | -0.10 |
| Petit Palace Museum | 8.7 | 0.31 | 0.42 | -0.11 |
| Melia Barcelona Sarriá | 8.2 | 0.33 | 0.44 | -0.11 |
| Acta Splendid | 8.2 | 0.28 | 0.42 | -0.14 |
| Boutique Hotel H10 Montcada | 8.9 | 0.51 | 0.66 | -0.15 |
| Ciutat Vella | 8.4 | 0.38 | 0.53 | -0.15 |
| Catalonia Passeig de Gràcia 4\* Sup | 9.2 | 0.19 | 0.34 | -0.15 |
| Hotel Regina | 8.7 | 0.23 | 0.40 | -0.17 |
| Grupotel Gravina | 8.2 | 0.22 | 0.40 | -0.18 |
| Park Hotel | 8.3 | 0.32 | 0.58 | -0.26 |
| Hotel Arts Barcelona | 8.6 | 0.10 | 0.37 | -0.27 |
| Hotel Roger de Llúria | 8.6 | 0.02 | 0.35 | -0.33 |
| H10 Urquinaona Plaza | 9.2 | 0.20 | 0.55 | -0.35 |

En el nostre cas recomenariem els hotels que surten al llista de recomenats. Hem endreçat en criteri descendent per la diferència que hi ha entre el valor predit i el valor\_optim i després per la nota mitjana de l’hotel.