PRACTICA 2: LIMPIEZA Y VALIDACIÓN DE LOS DATOS

Carles Maggi Gómez - Joan A. Maggi Gómez

29 de mayo 2019

Partim de tres fitxers que cal unificar en un sol per poder ser tratats

```
# Lectura del fitxer - read.csv()

categories <-read.csv("https://raw.githubusercontent.com/joanmaggigo/practica2/master/CSV/CategoriesXComentariBooking.csv",h
eader=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)


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")
```

1 Descripcó dels dataset.

Podem trobar el dataset en el següent repositori :

https://github.com/joanmaggigo/practica2/CSV (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.

El que pretenem fer és estudiar les relacions entre les categories i la nota de l'hotel. Fem una primera inspecció visual ens adonem que el que està catalogat com categoria no són les categories sinó els valors que poden pendre les categories.

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)

```
head(categories)
```

```
##   idHotel idIteracio index              Categoria
## 1   90587          1     1             Viajedeocio
## 2   90587          1     1                 Pareja
## 3   90587          1     1 HabitaciónDobleSuperior
## 4   90587          1     1        Estanciade1noche
## 5   90587          1     1         Enviadopormóvil
## 6   90587          1     2                  Grupo
```

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

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

```
## [1] "Urquinaona"                  "Habitat Apartments Bail<e9>n"
## [3] "Habitat Apartments Alibei"   "My Address in Barcelona"
## [5] "Barceloneta"                 "Ronda Sant Pere with Terrace"
```

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

```
# 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"))
# 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),]

#Veiem, per descrició 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="_")

#Anem a valorar quantes categories hi ha per comentaris, per veure si podem convertir en n variables descrivint un aspecte d
e l'hotel, com ara numero de nits, tipus d'estancia, etc.. primer mirem com es distribueixen per numero de 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="Total Valors categories per comentaris")
```

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 pero era un element diferenciat.

```
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 categories pensem que podem agrupar en 6 grups. veiem valors que podrien respondre a les categories: Nits,ProcedenciaComentaria,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)),20))
```

Categoria	total_cat

Categoria	total_cat
Viajedeocio	12546
Enviadopormóvil	11538
Pareja	7875
Estanciade1noche	6155
HabitaciónDoble-1o2camas	5274
Estanciade2noches	4203
Personaqueviajasola	3261
Estanciade3noches	3073
Familiaconniñospequeños	2992
Viajedenegocios	2098
Grupo	1732
Estanciade4noches	1669
HabitaciónDoble-2camas	1248
HabitaciónDoble	1189
HabitaciónIndividual	1188
HabitaciónDobleEstándar-1o2camas	1013
HabitaciónDobleSuperior-1o2camas	849
Estanciade5noches	643
HabitaciónTriple	588
HabitaciónDobleDeluxe-1o2camas	587

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

1.1 Creació noves varaibles agrupacio_categoria, id_persona

```
#La agrupació de categoria la inicialitzo amb la agrupació altres/darrer
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("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(head(aux2,20), caption=as.character(cols))
}
```

```
## # A tibble: 2 x 2
##   Categoria       total_cat
##   <chr>               <int>
## 1 Viajedenegocios      2098
## 2 Viajedeocio         12546
## # A tibble: 6 x 2
##   Categoria               total_cat
##   <chr>                       <int>
## 1 Familiaconniñosmayores          2
## 2 Familiaconniñospequeños      2992
## 3 Grupo                        1732
## 4 Grupodeamigos                 422
## 5 Pareja                       7875
## 6 Personaqueviajasola          3261
## # A tibble: 20 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 DobleEstándar                                              1
##  8 Habitación(1o2adultos)-1o2camas                          291
##  9 HabitaciónCompartida(4adultos)                            44
## 10 Habitacióncompartida(6adultos)                            14
## 11 Habitaciónconcamadobleoextragrandeyvistaspanorámicas       5
## 12 Habitaciónconcamaextragrande.                           148
## 13 Habitaciónconcamaextragrandeyvistasalapiscina             7
## 14 Habitaciónconcamaextragrandeyvistasalmar                  6
## 15 Habitacióncuádruple                                     162
## 16 HabitaciónCuádrupleconbalcón                              5
## 17 HabitaciónCuádrupleconvistas                             27
## 18 HabitaciónCuádrupleFamiliar                               8
## 19 HabitaciónDeluxecon2camasgrandes                          8
## 20 HabitaciónDeluxeconbañeradehidromasaje                    5
## # A tibble: 17 x 2
##    Categoria          total_cat
##    <chr>                  <int>
##  1 Estanciade10noches        10
##  2 Estanciade11noches         9
##  3 Estanciade12noches         6
##  4 Estanciade13noches         5
##  5 Estanciade14noches         5
##  6 Estanciade15noches         2
##  7 Estanciade1noche        6155
##  8 Estanciade22noches         1
##  9 Estanciade27noches         1
## 10 Estanciade2noches       4203
## 11 Estanciade3noches       3073
## 12 Estanciade4noches       1669
## 13 Estanciade5noches        643
## 14 Estanciade6noches        289
## 15 Estanciade7noches        133
## 16 Estanciade8noches         52
## 17 Estanciade9noches         20
## # A tibble: 1 x 2
##   Categoria       total_cat
##   <chr>               <int>
## 1 Enviadopormóvil     11538
## # A tibble: 1 x 2
##   Categoria     total_cat
##   <chr>             <int>
## 1 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.

```
#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))>7) ,"ReCategoria"]<-"Mesde7"

#A les habitacions les recategoritzem en funció de la descripció i del que sembla raonable per volums.
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

1.2 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")]
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","num_cat","TipusHabitacio","Nits","Pro
cedenciaComentari","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            num_cat
##  Length:16284       Min.   :1.00   Length:16284       Min.   :2.000
##  Class :character   1st Qu.:2.00   Class :character   1st Qu.:4.000
##  Mode  :character   Median :3.00   Mode  :character   Median :5.000
##                     Mean   :3.11                      Mean   :4.609
##                     3rd Qu.:4.00                      3rd Qu.:5.000
##                     Max.   :5.00                      Max.   :6.000
##  TipusHabitacio         Nits           ProcedenciaComentari
##  Length:16284       Length:16284       Length:16284
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  TipusViatge        Acompanyament      ViajaConMascota
##  Length:16284       Length:16284       Length:16284
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

2 Neteja de dades

2.1 Tipus de dades

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)
```

```
##   idPersona           nomHotel          notapersona
##  Length:16284       Length:16284       Min.   : 2.500
##  Class :character   Class :character   1st Qu.: 7.500
##  Mode  :character   Mode  :character   Median : 8.800
##                                        Mean   : 8.558
##                                        3rd Qu.:10.000
##                                        Max.   :10.000
##       data              estrelles      notaHotel        num_cat
##  Min.   :2017-04-13   Min.   :1.00   Min.   :7.900   Min.   :2.000
##  1st Qu.:2017-11-25   1st Qu.:2.00   1st Qu.:8.300   1st Qu.:4.000
##  Median :2018-05-13   Median :3.00   Median :8.500   Median :5.000
##  Mean   :2018-05-06   Mean   :3.11   Mean   :8.535   Mean   :4.609
##  3rd Qu.:2018-11-04   3rd Qu.:4.00   3rd Qu.:8.700   3rd Qu.:5.000
##  Max.   :2019-04-13   Max.   :5.00   Max.   :9.400   Max.   :6.000
##  TipusHabitacio         Nits           ProcedenciaComentari
##  Length:16284       Length:16284       Length:16284
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  TipusViatge        Acompanyament      ViajaConMascota
##  Length:16284       Length:16284       Length:16284
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

2.2 Eliminacó si cal de valors outliers i fringeliers

Mirarem els valors 0 de les notes mitjana d'hotel i les notes de les persons. 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 donade 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 nforça normals a la puntuació própia del hotel.

```
summary(data)
```

```
##   idPersona           nomHotel          notapersona
##  Length:16284       Length:16284       Min.   : 2.500
##  Class :character   Class :character   1st Qu.: 7.500
##  Mode  :character   Mode  :character   Median : 8.800
##                                        Mean   : 8.558
##                                        3rd Qu.:10.000
##                                        Max.   :10.000
##       data              estrelles      notaHotel        num_cat
##  Min.   :2017-04-13   Min.   :1.00   Min.   :7.900   Min.   :2.000
##  1st Qu.:2017-11-25   1st Qu.:2.00   1st Qu.:8.300   1st Qu.:4.000
##  Median :2018-05-13   Median :3.00   Median :8.500   Median :5.000
##  Mean   :2018-05-06   Mean   :3.11   Mean   :8.535   Mean   :4.609
##  3rd Qu.:2018-11-04   3rd Qu.:4.00   3rd Qu.:8.700   3rd Qu.:5.000
##  Max.   :2019-04-13   Max.   :5.00   Max.   :9.400   Max.   :6.000
##  TipusHabitacio         Nits           ProcedenciaComentari
##  Length:16284       Length:16284       Length:16284
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  TipusViatge        Acompanyament      ViajaConMascota
##  Length:16284       Length:16284       Length:16284
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
dataaux<-data
noudata <- data.frame(Doubles=double(),
                 Ints=integer(),
                 Factors=factor(),
                 Logicals=logical(),
                 Characters=character(),
                 stringsAsFactors=FALSE)
{
  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,]

    un.hotel<-un.hotel[remove_outliers(un.hotel$notapersona,3)==FALSE,]
    noudata<-rbind(noudata,un.hotel)
  }
}
hotels.outliers
```

```
##  [1] "Hostal Balkonis"
##  [2] "Atlantis by Atbcn"
##  [3] "Bcn Urban Hotels Gran Rosellon"
##  [4] "Hotel Àmbit Barcelona"
##  [5] "Hotel Turin"
##  [6] "Hotel Arts Barcelona"
##  [7] "Hotel Casa Bonay"
##  [8] "Upper Diagonal"
##  [9] "Negresco Princess 4* Sup"
## [10] "Hotel El Call"
## [11] "Hostal Fernando"
## [12] "The Corner Hotel"
## [13] "Hostal Central Barcelona"
## [14] "Grupotel Gran Via 678"
## [15] "Hostal Benidorm"
## [16] "Golden Tulip Barcelona"
## [17] "Cuatro Naciones"
## [18] "Hotel Lloret Ramblas"
## [19] "H10 Madison 4* Sup"
## [20] "Hotel Rec Barcelona - Adults Only"
## [21] "Habitat Apartments Eixample Balconies."
## [22] "Hostal Q Barcelona"
## [23] "Eurostars Ramblas"
## [24] "Exe Plaza Catalunya"
## [25] "Room Mate Pau"
## [26] "Hotel Barcelona 1882"
## [27] "Hotel Center Gran Via"
## [28] "Catalonia Gran Via BCN"
## [29] "Catalonia Passeig de Gràcia 4* Sup"
## [30] "Hesperia Barcelona Ramblas"
## [31] "H10 Urquinaona Plaza"
## [32] "Olivia Balmes Hotel"
## [33] "H10 Port Vell 4* Sup"
## [34] "Fairmont Rey Juan Carlos I"
## [35] "Hotel Ginebra"
## [36] "Hotel Astoria"
## [37] "Hotel Balmes"
## [38] "HCC Taber"
## [39] "HCC Regente"
## [40] "Grupotel Gravina"
## [41] "HCC Montblanc"
## [42] "Paral·lel"
## [43] "Acta Atrium Palace"
## [44] "Boutique Hotel H10 Montcada"
## [45] "Melia Barcelona Sarriá"
## [46] "Acta Splendid"
## [47] "Hotel Jazz"
## [48] "Catalonia La Pedrera"
## [49] "Catalonia Born"
## [50] "Catalonia Sagrada Familia"
## [51] "NH Collection Barcelona Podium"
## [52] "Park Hotel"
## [53] "Catalunya"
## [54] "NH Barcelona Stadium"
## [55] "Hesperia Barcelona Del Mar"
## [56] "Hotel Lleó"
## [57] "Cram"
## [58] "Hotel Roger de Llúria"
## [59] "Tryp Barcelona Apolo Hotel"
## [60] "Hesperia Barcelona Barri Gòtic"
## [61] "Hotel America Barcelona"
## [62] "Ciutat de Barcelona"
## [63] "Hotel Regina"
## [64] "Hotel Cortes"
## [65] "Occidental Atenea Mar - Adults Only"
## [66] "Barceló Sants"
## [67] "Pestana Arena Barcelona"
## [68] "Condes de Barcelona"
## [69] "Hotel Villa Emilia"
## [70] "NH Collection Barcelona Constanza"
## [71] "Acta BCN 40"
## [72] "Pullman Barcelona Skipper"
## [73] "Ayre Hotel Gran Vía"
## [74] "ICON BCN by Petit Palace"
## [75] "Hotel Market"
## [76] "Hotel 54 Barceloneta"
## [77] "Petit Palace Museum"
## [78] "Onix Liceo"
## [79] "Catalonia Avinyo"
## [80] "Ciutat Vella"
## [81] "Hilton Diagonal Mar Barcelona"
## [82] "Guest House Center Inn"
## [83] "Grand Hotel Central"
```

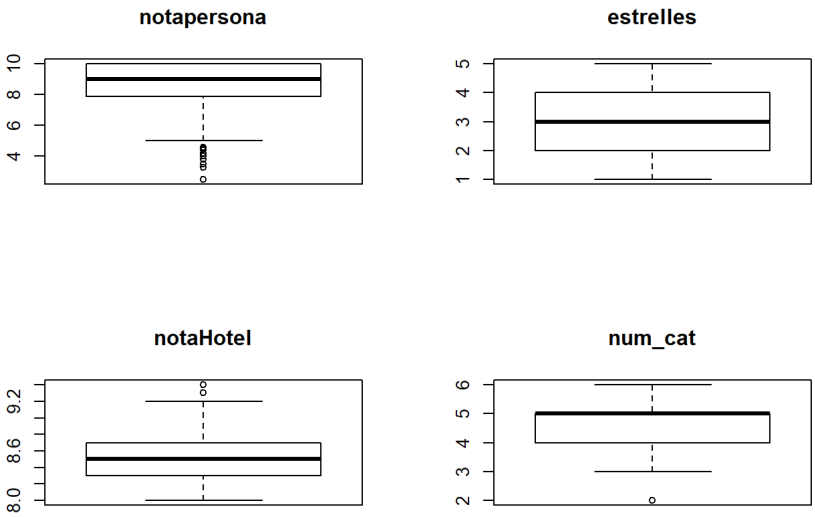
```
summary(noudata)
```

```
##   idPersona           nomHotel          notapersona
##  Length:16091       Length:16091       Min.   : 2.500
##  Class :character   Class :character   1st Qu.: 7.900
##  Mode  :character   Mode  :character   Median : 9.000
##                                        Mean   : 8.617
##                                        3rd Qu.:10.000
##                                        Max.   :10.000
##                                        NA's   :1
##       data              estrelles      notaHotel        num_cat
##  Min.   :2017-04-13   Min.   :1.00   Min.   :8.000   Min.   :2.000
##  1st Qu.:2017-11-25   1st Qu.:2.00   1st Qu.:8.300   1st Qu.:4.000
##  Median :2018-05-13   Median :3.00   Median :8.500   Median :5.000
##  Mean   :2018-05-06   Mean   :3.11   Mean   :8.534   Mean   :4.608
##  3rd Qu.:2018-11-04   3rd Qu.:4.00   3rd Qu.:8.700   3rd Qu.:5.000
##  Max.   :2019-04-13   Max.   :5.00   Max.   :9.400   Max.   :6.000
##  NA's   :1            NA's   :1      NA's   :1       NA's   :1
##  TipusHabitacio         Nits           ProcedenciaComentari
##  Length:16091       Length:16091       Length:16091
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  TipusViatge        Acompanyament      ViajaConMascota
##  Length:16091       Length:16091       Length:16091
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
data<-noudata
```

2.3 Mirem si bloxplot de les variables numèriques

```
par(mfrow=c(2,2))
for(i in 1:ncol(data)) {
  if (is.numeric(data[,i])){
    boxplot(data[,i], main = colnames(data)[i], width = 100)
  }
}
```




```
par(mfrow=c(1,1))
```

```
#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:1209        Min.   :8.000   Min.   :1.00   Min.   : 2.50
##  Class :character   1st Qu.:8.300   1st Qu.:3.00   1st Qu.: 6.30
##  Mode  :character   Median :8.500   Median :4.00   Median : 7.50
##                     Mean   :8.557   Mean   :3.32   Mean   : 7.43
##                     3rd Qu.:8.700   3rd Qu.:4.00   3rd Qu.: 8.80
##                     Max.   :9.400   Max.   :5.00   Max.   :10.00
##                     NA's   :1       NA's   :1      NA's   :1
```

```
#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.
```

2.4 Tractament dels valors na

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

```
##            idPersona             nomHotel          notapersona
##                    1                    1                    1
##                 data            estrelles             notaHotel
##                    1                    1                    1
##              num_cat       TipusHabitacio                 Nits
##                    1                    8                    8
## ProcedenciaComentari          TipusViatge        Acompanyament
##                 4702                 1621                    1
##      ViajaConMascota
##                16060
```

```
#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),]
```

```
##          idPersona                  nomHotel notapersona       data
## 1645   173142_3_40             Hotel El Call         9.6 2018-01-02
## 3215  2311290_4_22     Golden Tulip Barcelona        10.0 2017-07-30
## NA            <NA>                      <NA>          NA       <NA>
## 7849    90480_2_15     Melia Barcelona Sarriá         8.8 2018-08-08
## 7878    90480_2_41     Melia Barcelona Sarriá         2.5 2018-04-17
## 10336   91120_2_24 Hesperia Barcelona Del Mar         9.6 2018-08-29
## 10892   91493_1_46 Tryp Barcelona Apolo Hotel         3.3 2019-02-22
## 14135   93879_1_14  Pullman Barcelona Skipper        10.0 2019-02-13
##       estrelles notaHotel num_cat TipusHabitacio Nits ProcedenciaComentari
## 1645          1       8.3       3           <NA> <NA>      Enviadopormóvil
## 3215          4       8.7       2           <NA> <NA>        EnviadoporWeb
## NA           NA        NA      NA           <NA> <NA>        EnviadoporWeb
## 7849          5       8.2       3           <NA> <NA>      Enviadopormóvil
## 7878          5       8.2       3           <NA> <NA>      Enviadopormóvil
## 10336         4       8.2       2           <NA> <NA>        EnviadoporWeb
## 10892         4       8.6       3           <NA> <NA>      Enviadopormóvil
## 14135         5       8.5       2           <NA> <NA>        EnviadoporWeb
##           TipusViatge          Acompanyament ViajaConMascota
## 1645      Viajedeocio                 Pareja               N
## 3215      Viajedeocio Familiaconniñospequeños               N
## NA      SenseInformar                   <NA>               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)
```

```
##   idPersona           nomHotel          notapersona
##  Length:16083       Length:16083       Min.   : 2.500
##  Class :character   Class :character   1st Qu.: 7.900
##  Mode  :character   Mode  :character   Median : 9.000
##                                        Mean   : 8.617
##                                        3rd Qu.:10.000
##                                        Max.   :10.000
##       data              estrelles      notaHotel        num_cat
##  Min.   :2017-04-13   Min.   :1.00   Min.   :8.000   Min.   :3.000
##  1st Qu.:2017-11-25   1st Qu.:2.00   1st Qu.:8.300   1st Qu.:4.000
##  Median :2018-05-13   Median :3.00   Median :8.500   Median :5.000
##  Mean   :2018-05-06   Mean   :3.11   Mean   :8.534   Mean   :4.609
##  3rd Qu.:2018-11-04   3rd Qu.:4.00   3rd Qu.:8.700   3rd Qu.:5.000
##  Max.   :2019-04-13   Max.   :5.00   Max.   :9.400   Max.   :6.000
##  TipusHabitacio         Nits           ProcedenciaComentari
##  Length:16083       Length:16083       Length:16083
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  TipusViatge        Acompanyament      ViajaConMascota
##  Length:16083       Length:16083       Length:16083
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

2.5 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$Nits<-as.factor(data$Nits)
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)
```

```
##   idPersona                                    nomHotel
##  Length:16083       Catalonia Sagrada Familia        :  911
##  Class :character   Hotel Lloret Ramblas              :  834
##  Mode  :character   Ayre Hotel Gran Vía               :  775
##                     Barceló Sants                    :  627
##                     Hotel Cortes                     :  540
##                     NH Collection Barcelona Constanza:  463
##                     (Other)                          :11933
##   notapersona          data            estrelles   notaHotel
##  Min.   : 2.500   Min.   :2017-04-13   1:1364    Min.   :8.000
##  1st Qu.: 7.900   1st Qu.:2017-11-25   2:3068    1st Qu.:8.300
##  Median : 9.000   Median :2018-05-13   3:4415    Median :8.500
##  Mean   : 8.617   Mean   :2018-05-06   4:6906    Mean   :8.534
##  3rd Qu.:10.000   3rd Qu.:2018-11-04   5: 330    3rd Qu.:8.700
##  Max.   :10.000   Max.   :2019-04-13             Max.   :9.400
##
##     num_cat                   TipusHabitacio                Nits
##  Min.   :3.000   Altres              : 1166   Estanciade1noche :6071
##  1st Qu.:4.000   Habitacio Doble     :12511   Estanciade2noches:4146
##  Median :5.000   Habitacio Individual: 1411   Estanciade3noches:3044
##  Mean   :4.609   Habitacio quadruple :  199   Estanciade4noches:1654
##  3rd Qu.:5.000   Habitacio Triple    :  678   Estanciade5noches: 640
##  Max.   :6.000   Suite               :  118   Estanciade6noches: 286
##                                               (Other)          : 242
##       ProcedenciaComentari          TipusViatge
##  Enviadopormóvil:11385     SenseInformar  : 1620
##  EnviadoporWeb  : 4698     Viajedenegocios: 2066
##                            Viajedeocio    :12397
##
##
##
##
##                  Acompanyament  ViajaConMascota
##  Familiaconniñosmayores :   2   N:16052
##  Familiaconniñospequeños:2961   S:   31
##  Grupo                  :1711
##  Grupodeamigos          : 419
##  Pareja                 :7781
##  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")
```

2.6 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)
```

```
##   idPersona                                    nomHotel
##  Length:16083       Catalonia Sagrada Familia        :  911
##  Class :character   Hotel Lloret Ramblas              :  834
##  Mode  :character   Ayre Hotel Gran Vía               :  775
##                     Barceló Sants                    :  627
##                     Hotel Cortes                     :  540
##                     NH Collection Barcelona Constanza:  463
##                     (Other)                          :11933
##   notapersona          data            estrelles   notaHotel
##  Min.   : 2.500   Min.   :2017-04-13   1:1364    Min.   :8.000
##  1st Qu.: 7.900   1st Qu.:2017-11-25   2:3068    1st Qu.:8.300
##  Median : 9.000   Median :2018-05-13   3:4415    Median :8.500
##  Mean   : 8.617   Mean   :2018-05-06   4:6906    Mean   :8.534
##  3rd Qu.:10.000   3rd Qu.:2018-11-04   5: 330    3rd Qu.:8.700
##  Max.   :10.000   Max.   :2019-04-13             Max.   :9.400
##
##     num_cat                  TipusHabitacio                Nits
##  Min.   :3.000   Altres              : 1166   Estanciade1noche :6071
##  1st Qu.:4.000   Habitacio Doble     :12511   Estanciade2noches:4146
##  Median :5.000   Habitacio Individual: 1411   Estanciade3noches:3044
##  Mean   :4.609   Habitacio quadruple :  199   Estanciade4noches:1654
##  3rd Qu.:5.000   Habitacio Triple    :  678   Estanciade5noches: 640
##  Max.   :6.000   Suite               :  118   Estanciade6noches: 286
##                                               (Other)          : 242
##       ProcedenciaComentari          TipusViatge
##  Enviadopormóvil:11385     SenseInformar  : 1620
##  EnviadoporWeb  : 4698     Viajedenegocios: 2066
##                            Viajedeocio    :12397
##
##
##
##
##             Acompanyament  ViajaConMascota              Season
##  Familiaconniños    :2963   N:16052         PRIMER TRIMESTRE:4428
##  Grupo              :1711   S:   31         QUART TRIMESTRE :4079
##  Grupodeamigos      : 419                   SEGON TRIMESTRE :3770
##  Pareja             :7781                   TERCER TRIMESTRE:3806
##  Personaqueviajasola:3209
##
##
```

```
#Anem a veure si la diferencia en general de les notes de les persones pels hotels difereix entre el primer trimestre i el t
ercer trimetre. Per això, apliquem un anova.
hotels_1er_3er<-data %>% filter(Season=='PRIMER TRIMESTRE' | Season=='TERCER TRIMESTRE')
aov.hotels_1er_3er=aov(notapersona~Season,hotels_1er_3er)
summary(aov.hotels_1er_3er)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Season         1    120  119.55   61.21 5.77e-15 ***
## Residuals   8232  16078    1.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kable(hotels_1er_3er%>% group_by(Season) %>% summarise(n=n(),mean=mean(notapersona),sd=sd(notapersona)))
```

Season	n	mean	sd
PRIMER TRIMESTRE	4428	8.733695	1.349286
TERCER TRIMESTRE	3806	8.492013	1.451689

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.

Ara volem veure, d'aquells hotels que tinguin una mostra superior a 30 tant per l'estiu com per l'hivern tenen una nota significativament diferent entre els dos trimestres.

```
#busco els hotels que tinguin com a minim 30 elements tant en el primer trimestre com en el tercer.
hotels_comparables<-data %>% group_by(nomHotel,Season)  %>% filter(Season=='PRIMER TRIMESTRE' | Season=='TERCER TRIMESTRE')
 %>% summarise(n=n()) %>% filter(n>30) %>% group_by(nomHotel) %>% summarise(n=n()) %>% filter(n==2)
data.hotels_comparables <- data[data$nomHotel %in% hotels_comparables$nomHotel,]
data.hotels_comparables<-data.hotels_comparables %>% filter(Season=='PRIMER TRIMESTRE' | Season=='TERCER TRIMESTRE')

#Anem a fer anova hotel per hotel i creant el resultat en forma de taula
llista_hotels=unique(data.hotels_comparables$nomHotel)
result_aov<-NULL
for (cols in llista_hotels)
{
  hotel_actual<-data.hotels_comparables[data.hotels_comparables$nomHotel==cols,]
  aov.hotel_actual=aov(notapersona~Season,hotel_actual)
  if (is.null(result_aov))
  {
    aux<-hotel_actual %>% group_by(nomHotel,Season) %>% summarise(n=n(),mean=mean(notapersona),sd=sd(notapersona))
    aux.1<-aux[aux$Season=='PRIMER TRIMESTRE',]
    aux.2<-aux[aux$Season=='TERCER TRIMESTRE',]
    aux.merge<-merge(aux.1,aux.2,by.x=c("nomHotel"),by.y=c("nomHotel"))
    result_aov<-cbind(aux.merge,cols,summary(aov.hotel_actual)[[1]][["Pr(>F)"]][1])
  }
  else
  {
    aux<-hotel_actual %>% group_by(nomHotel,Season) %>% summarise(n=n(),mean=mean(notapersona),sd=sd(notapersona))
    aux.1<-aux[aux$Season=='PRIMER TRIMESTRE',]
    aux.2<-aux[aux$Season=='TERCER TRIMESTRE',]
    aux.merge<-merge(aux.1,aux.2,by.x=c("nomHotel"),by.y=c("nomHotel"))
    aux.merge<-cbind(aux.merge,cols,summary(aov.hotel_actual)[[1]][["Pr(>F)"]][1])
    result_aov<-rbind(result_aov,aux.merge)
  }
}

result_aov$diferenciaNota<-FALSE
result_aov[(result_aov$`summary(aov.hotel_actual)[[1]][["Pr(>F)"]][1]`<0.05),"diferenciaNota"]<-TRUE
result_aov$diferencia<-result_aov$mean.x-result_aov$mean.y
kable(result_aov)
```

nomHotel	Season.x	n.x	mean.x	sd.x	Season.y	n.y	mean.y	sd.y	cols	summary(aov.hotel_actual) [[1]][[“Pr(>F)”]][1]	diferenciaNota
Atlantis by Atbcn	PRIMER TRIMESTRE	69	8.879710	1.1141575	TERCER TRIMESTRE	78	8.797436	0.9559077	Atlantis by Atbcn	0.6306356	FALSE
Bcn Urban Hotels Gran Rosellon	PRIMER TRIMESTRE	74	8.571622	1.4409396	TERCER TRIMESTRE	50	8.128000	1.5141563	Bcn Urban Hotels Gran Rosellon	0.1020065	FALSE
Upper Diagonal	PRIMER TRIMESTRE	33	9.130303	1.2885275	TERCER TRIMESTRE	42	8.969048	0.9991431	Upper Diagonal	0.5432973	FALSE
Hostal Fernando	PRIMER TRIMESTRE	118	8.738983	1.2965460	TERCER TRIMESTRE	81	8.370370	1.2641642	Hostal Fernando	0.0479331	TRUE
Grupotel Gran Via 678	PRIMER TRIMESTRE	94	8.880851	1.1365177	TERCER TRIMESTRE	39	8.902564	1.2055763	Grupotel Gran Via 678	0.9216623	FALSE
Hostal Benidorm	PRIMER TRIMESTRE	104	8.964423	1.2026603	TERCER TRIMESTRE	97	8.541237	1.4297255	Hostal Benidorm	0.0238960	TRUE
Golden Tulip Barcelona	PRIMER TRIMESTRE	86	8.832558	1.1844524	TERCER TRIMESTRE	61	8.868852	1.2437230	Golden Tulip Barcelona	0.8579625	FALSE
Cuatro Naciones	PRIMER TRIMESTRE	54	8.701852	1.1018836	TERCER TRIMESTRE	109	8.848624	1.1813845	Cuatro Naciones	0.4465321	FALSE
Hotel Lloret Ramblas	PRIMER TRIMESTRE	187	8.270588	1.5592048	TERCER TRIMESTRE	216	8.087500	1.6073035	Hotel Lloret Ramblas	0.2482383	FALSE
Hostal Q Barcelona	PRIMER TRIMESTRE	52	8.707692	1.4447805	TERCER TRIMESTRE	98	8.191837	1.4921564	Hostal Q Barcelona	0.0434232	TRUE
Room Mate Pau	PRIMER TRIMESTRE	40	9.362500	0.8285568	TERCER TRIMESTRE	35	9.222857	1.0535893	Room Mate Pau	0.5230305	FALSE
Hotel Ginebra	PRIMER TRIMESTRE	50	8.696000	1.4356908	TERCER TRIMESTRE	45	8.706667	1.4141171	Hotel Ginebra	0.9710293	FALSE
Hotel Astoria	PRIMER TRIMESTRE	55	8.752727	1.1783313	TERCER TRIMESTRE	35	8.917143	1.3162973	Hotel Astoria	0.5391781	FALSE
Grupotel Gravina	PRIMER TRIMESTRE	98	8.921429	1.2136042	TERCER TRIMESTRE	83	8.025301	1.7342211	Grupotel Gravina	0.0000697	TRUE

nomHotel	Season.x	n.x	mean.x	sd.x	Season.y	n.y	mean.y	sd.y	cols	summary(aov.hotel_actual)[[1]][["Pr(>F)"]][1]	diferenciaNota
Acta Atrium Palace	PRIMER TRIMESTRE	53	8.586792	1.4228949	TERCER TRIMESTRE	37	8.818919	1.2281056	Acta Atrium Palace	0.4232002	FALSE
Acta Splendid	PRIMER TRIMESTRE	51	8.629412	1.2392408	TERCER TRIMESTRE	54	8.053704	1.8183253	Acta Splendid	0.0622607	FALSE
Catalonia La Pedrera	PRIMER TRIMESTRE	61	8.457377	1.7123531	TERCER TRIMESTRE	58	8.432759	1.3326386	Catalonia La Pedrera	0.9306490	FALSE
Catalonia Born	PRIMER TRIMESTRE	37	8.789189	1.1932178	TERCER TRIMESTRE	45	9.002222	1.3344200	Catalonia Born	0.4529497	FALSE
Catalonia Sagrada Familia	PRIMER TRIMESTRE	293	8.456997	1.5616313	TERCER TRIMESTRE	188	8.106383	1.6548621	Catalonia Sagrada Familia	0.0193328	TRUE
Catalunya	PRIMER TRIMESTRE	81	7.945679	1.8143904	TERCER TRIMESTRE	78	7.852564	1.4815510	Catalunya	0.7240415	FALSE
NH Barcelona Stadium	PRIMER TRIMESTRE	87	8.159770	1.5827897	TERCER TRIMESTRE	64	8.031250	1.5652856	NH Barcelona Stadium	0.6210621	FALSE
Hotel Lleó	PRIMER TRIMESTRE	48	8.547917	1.5442398	TERCER TRIMESTRE	39	8.787180	1.5317997	Hotel Lleó	0.4727031	FALSE
Hesperia Barcelona Barri Gòtic	PRIMER TRIMESTRE	41	8.678049	1.4490535	TERCER TRIMESTRE	36	8.002778	1.7208779	Hesperia Barcelona Barri Gòtic	0.0655062	FALSE
Ciutat de Barcelona	PRIMER TRIMESTRE	75	8.929333	1.1303878	TERCER TRIMESTRE	37	8.732432	1.0852685	Ciutat de Barcelona	0.3816601	FALSE
Hotel Regina	PRIMER TRIMESTRE	45	9.111111	1.1162482	TERCER TRIMESTRE	51	8.468628	1.4168261	Hotel Regina	0.0163565	TRUE
Hotel Cortes	PRIMER TRIMESTRE	109	8.376147	1.2706005	TERCER TRIMESTRE	150	8.210667	1.4534452	Hotel Cortes	0.3414629	FALSE
Occidental Atenea Mar - Adults Only	PRIMER TRIMESTRE	36	8.450000	1.5529695	TERCER TRIMESTRE	38	7.684210	1.8197901	Occidental Atenea Mar - Adults Only	0.0560326	FALSE
Barceló Sants	PRIMER TRIMESTRE	100	8.793000	1.1516571	TERCER TRIMESTRE	214	8.575701	1.3571816	Barceló Sants	0.1671281	FALSE
Pestana Arena Barcelona	PRIMER TRIMESTRE	53	8.647170	1.3460203	TERCER TRIMESTRE	41	8.951220	1.1097572	Pestana Arena Barcelona	0.2447737	FALSE
Condes de Barcelona	PRIMER TRIMESTRE	67	9.056716	1.1118131	TERCER TRIMESTRE	76	8.918421	1.2022990	Condes de Barcelona	0.4783109	FALSE
NH Collection Barcelona Constanza	PRIMER TRIMESTRE	137	9.210219	0.8674069	TERCER TRIMESTRE	134	9.076119	1.0680725	NH Collection Barcelona Constanza	0.2570832	FALSE
Acta BCN 40	PRIMER TRIMESTRE	44	8.372727	1.3074676	TERCER TRIMESTRE	39	8.274359	1.5518610	Acta BCN 40	0.7548109	FALSE
Ayre Hotel Gran Vía	PRIMER TRIMESTRE	244	8.961885	1.1661119	TERCER TRIMESTRE	160	8.956875	1.0817306	Ayre Hotel Gran Vía	0.9653631	FALSE
Hotel Market	PRIMER TRIMESTRE	119	7.973109	1.5534924	TERCER TRIMESTRE	88	7.877273	1.6728891	Hotel Market	0.6715462	FALSE
Petit Palace Museum	PRIMER TRIMESTRE	51	8.890196	1.2734606	TERCER TRIMESTRE	43	8.616279	1.3575065	Petit Palace Museum	0.3160811	FALSE
Catalonia Avinyo	PRIMER TRIMESTRE	38	8.905263	1.3805497	TERCER TRIMESTRE	40	9.042500	1.2518474	Catalonia Avinyo	0.6465981	FALSE
Ciutat Vella	PRIMER TRIMESTRE	47	8.568085	1.6869198	TERCER TRIMESTRE	43	8.232558	1.2876189	Ciutat Vella	0.2951072	FALSE

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 logistica per predicció de nota per cada hotel. Comprovarem el grau de significancia de la regressió. Un cop fet això 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+ProcedenciaComentari+TipusViatge+Acompanyament+Season, data)
summary(model.hotel.nota)
```

```
##
## Call:
## lm(formula = notapersona ~ TipusHabitacio + Nits + ProcedenciaComentari +
##     TipusViatge + Acompanyament + Season, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3423 -0.8859  0.3741  1.1682  2.2003
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        8.792116   0.059170 148.590  < 2e-16
## TipusHabitacioHabitacio Doble      0.025870   0.043130   0.600 0.548640
## TipusHabitacioHabitacio Individual -0.233941   0.061204  -3.822 0.000133
## TipusHabitacioHabitacio quadruple   0.189064   0.106979   1.767 0.077196
## TipusHabitacioHabitacio Triple      0.082383   0.067426   1.222 0.221795
## TipusHabitacioSuite                 0.307168   0.134113   2.290 0.022013
## NitsEstanciade2noches              -0.031861   0.028032  -1.137 0.255725
## NitsEstanciade3noches               0.032869   0.031114   1.056 0.290792
## NitsEstanciade4noches              -0.010501   0.038778  -0.271 0.786552
## NitsEstanciade5noches               0.017327   0.057717   0.300 0.764018
## NitsEstanciade6noches              -0.002793   0.083816  -0.033 0.973420
## NitsEstanciade7noches               0.180331   0.121293   1.487 0.137102
## NitsMesde7                          0.320379   0.133748   2.395 0.016614
## ProcedenciaComentariEnviadoporWeb  -0.098422   0.024368  -4.039 5.40e-05
## TipusViatgeViajedenegocios         -0.301196   0.047317  -6.366 2.00e-10
## TipusViatgeViajedeocio              0.051187   0.037006   1.383 0.166616
## AcompanyamentGrupo                 -0.051617   0.043156  -1.196 0.231698
## AcompanyamentGrupodeamigos          0.132669   0.072773   1.823 0.068315
## AcompanyamentPareja                -0.056432   0.032162  -1.755 0.079342
## AcompanyamentPersonaqueviajasola   -0.016479   0.043529  -0.379 0.705011
## SeasonQUART TRIMESTRE              -0.026826   0.030048  -0.893 0.371995
## SeasonSEGON TRIMESTRE              -0.218786   0.030808  -7.102 1.28e-12
## SeasonTERCER TRIMESTRE             -0.275367   0.030831  -8.932  < 2e-16
##
## (Intercept)                        ***
## TipusHabitacioHabitacio Doble
## TipusHabitacioHabitacio Individual ***
## TipusHabitacioHabitacio quadruple  .
## TipusHabitacioHabitacio Triple
## TipusHabitacioSuite                *
## NitsEstanciade2noches
## NitsEstanciade3noches
## NitsEstanciade4noches
## NitsEstanciade5noches
## NitsEstanciade6noches
## NitsEstanciade7noches
## NitsMesde7                         *
## ProcedenciaComentariEnviadoporWeb  ***
## 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: 1.382 on 16060 degrees of freedom
## Multiple R-squared:  0.02056,    Adjusted R-squared:  0.01922
## F-statistic: 15.32 on 22 and 16060 DF,  p-value: < 2.2e-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.

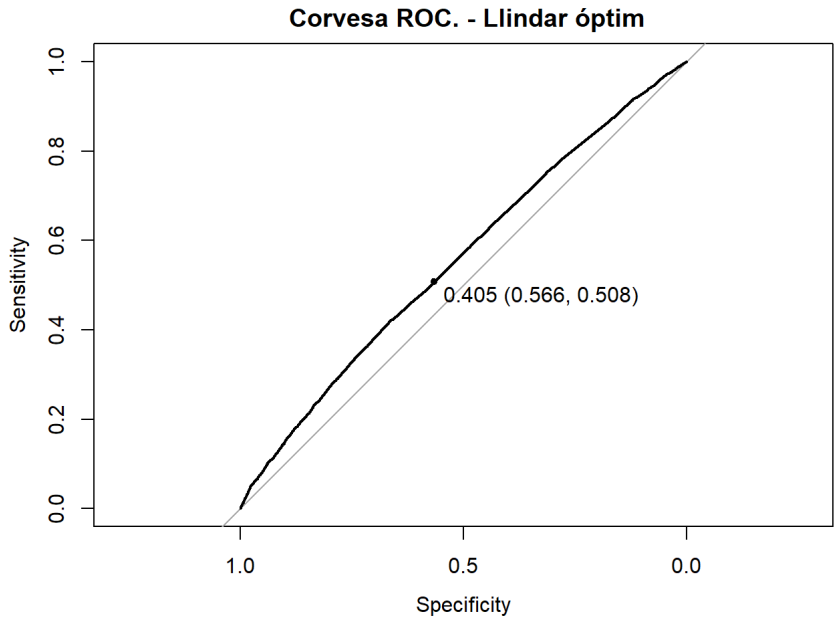
Anem a plantejar una altra estrategia, en aquest cas determinem que una persona recomana un hotel X si la nota que ha posat (notapersona) és superior a la nota mitjana de l'hotel (notaHOtel), per fer això, crearem una variable binària (Recomana amb valors S/N i mirarem de treballar amb una regressió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

```
data.regressio<-data
data.regressio$recomanar<-ifelse(data.regressio$notaHotel<data.regressio$notapersona,0,1)

model.recomanancio<-glm(recomanar~TipusHabitacio +Nits+  ProcedenciaComentari +
    TipusViatge  + Acompanyament , data = data.regressio)
summary(model.recomanancio)
```

```
##
## Call:
## glm(formula = recomanar ~ TipusHabitacio + Nits + ProcedenciaComentari +
##     TipusViatge + Acompanyament, data = data.regressio)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6316  -0.4052  -0.3748   0.5787   0.7273
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       0.386068   0.020145  19.165  < 2e-16
## TipusHabitacioHabitacio Doble    -0.013403   0.015295  -0.876  0.38091
## TipusHabitacioHabitacio Individual  0.061891   0.021704   2.852  0.00436
## TipusHabitacioHabitacio quadruple  -0.085846   0.037939  -2.263  0.02366
## TipusHabitacioHabitacio Triple   -0.032829   0.023913  -1.373  0.16982
## TipusHabitacioSuite              -0.096451   0.047563  -2.028  0.04259
## NitsEstanciade2noches             0.024696   0.009939   2.485  0.01297
## NitsEstanciade3noches             0.012925   0.011020   1.173  0.24087
## NitsEstanciade4noches             0.027951   0.013718   2.038  0.04162
## NitsEstanciade5noches             0.023099   0.020442   1.130  0.25850
## NitsEstanciade6noches             0.011496   0.029719   0.387  0.69889
## NitsEstanciade7noches            -0.074902   0.043009  -1.742  0.08160
## NitsMesde7                       -0.058071   0.047418  -1.225  0.22073
## ProcedenciaComentariEnviadoporWeb   0.043057   0.008637   4.985 6.26e-07
## TipusViatgeViajedenegocios        0.098732   0.016774   5.886 4.04e-09
## TipusViatgeViajedeocio           -0.009329   0.013113  -0.711  0.47682
## AcompanyamentGrupo                0.017118   0.015298   1.119  0.26317
## AcompanyamentGrupodeamigos       -0.032539   0.025811  -1.261  0.20744
## AcompanyamentPareja               0.014937   0.011393   1.311  0.18986
## AcompanyamentPersonaqueviajasola   -0.007257   0.015424  -0.471  0.63800
##
## (Intercept)                      ***
## TipusHabitacioHabitacio Doble
## TipusHabitacioHabitacio Individual **
## TipusHabitacioHabitacio quadruple  *
## TipusHabitacioHabitacio Triple
## TipusHabitacioSuite              *
## NitsEstanciade2noches            *
## NitsEstanciade3noches
## NitsEstanciade4noches            *
## NitsEstanciade5noches
## NitsEstanciade6noches
## NitsEstanciade7noches            .
## NitsMesde7
## ProcedenciaComentariEnviadoporWeb  ***
## TipusViatgeViajedenegocios       ***
## TipusViatgeViajedeocio
## AcompanyamentGrupo
## AcompanyamentGrupodeamigos
## AcompanyamentPareja
## AcompanyamentPersonaqueviajasola
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2403366)
##
##     Null deviance: 3903.3  on 16082  degrees of freedom
## Residual deviance: 3860.5  on 16063  degrees of freedom
## AIC: 22734
##
## Number of Fisher Scoring iterations: 2
```

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




```
auc(data.regressio$recomanar, pred)
```

```
## [1] 0.5567048
```

```
#Anem a ferun regressió logistica per hotel i creant el resultat en forma de taula
llista_hotels=unique(data.hotels_comparables$nomHotel)
result_roc<-NULL
for (cols in llista_hotels)
{

  data.regressio<-data[data$nomHotel==cols,]
  data.regressio$recomanar<-ifelse(data.regressio$notaHotel<data.regressio$notapersona,0,1)

  try(model.recomanacio<-glm(recomanar~TipusHabitacio+Nits+ProcedenciaComentari+TipusViatge+Acompanyament+Season , data = da
ta.regressio),silent=T)
  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
  list.model.recomanacio<-model.recomanacio

  if (is.null(result_roc))
  {
    result_roc<-cbind(cols,area,optim_point,model.recomanacio)

  }
  else
  {
    aux.merge<-cbind(cols,area,optim_point,model.recomanacio)
    result_roc<-rbind(result_roc,aux.merge)

  }


}

# kable(result_roc[,c(1:3)])
```