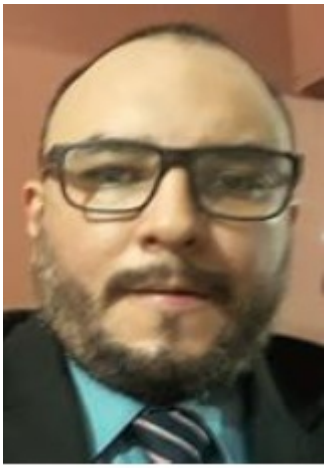
RNotebook: Resolución Prueba DGA

Contents

Resumen: El presente código fue desarrollado por **Elias Preza** para optar a la plaza de Senior Data Scientist de la Dirección General de Aduana **DGA**, en su contenido se resuelve sintaxis para el ejercicio de la **parte: 1 de Manejo de Datos** y la para la **parte 2: la Construcción de un Modelo de Machine Learning**.



Link hacia perfil de Linkedin

Preparación de Librerias

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

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

```
library(readr)
library(sjmisc)
```

```
##
## Attaching package: 'sjmisc'
```

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

```
## The following object is masked from 'package:tidyr':
##
##     replace_na
```

```
## The following object is masked from 'package:tibble':
##
##     add_case
```

Preparación de bases de datos

Data Set: BX_Books

```
BX_Books<- read_delim("BX-Books.csv", ";",
                      escape_double = FALSE, trim_ws = TRUE)
```

```
## Parsed with column specification:
## cols(
##   ISBN = col_character(),
##   `Book-Title` = col_character(),
##   `Book-Author` = col_character(),
##   `Year-Of-Publication` = col_double(),
##   Publisher = col_character(),
##   `Image-URL-S` = col_character(),
##   `Image-URL-M` = col_character(),
##   `Image-URL-L` = col_character()
## )
```

```
dplyr::glimpse(BX_Books)
```

```
## Rows: 271,379
## Columns: 8
## $ ISBN                 <chr> "0195153448", "0002005018", "0060973129", "03...
## $ `Book-Title`         <chr> "Classical Mythology", "Clara Callan", "Decis...
## $ `Book-Author`        <chr> "Mark P. O. Morford", "Richard Bruce Wright",...
## $ `Year-Of-Publication` <dbl> 2002, 2001, 1991, 1999, 1999, 1991, 2000, 199...
## $ Publisher            <chr> "Oxford University Press", "HarperFlamingo Ca...
## $ `Image-URL-S`        <chr> "http://images.amazon.com/images/P/0195153448...
## $ `Image-URL-M`        <chr> "http://images.amazon.com/images/P/0195153448...
## $ `Image-URL-L`        <chr> "http://images.amazon.com/images/P/0195153448...
```

```
head(BX_Books,5)
```

```
## # A tibble: 5 x 8
##   ISBN  `Book-Title` `Book-Author` `Year-Of-Public~ Publisher `Image-URL-S`
##   <chr> <chr>        <chr>                    <dbl> <chr>     <chr>
## 1 0195~ Classical M~ Mark P. O. M~             2002 Oxford U~ http://image~
## 2 0002~ Clara Callan Richard Bruc~             2001 HarperFl~ http://image~
## 3 0060~ Decision in~ Carlo D'Este              1991 HarperPe~ http://image~
## 4 0374~ Flu: The St~ Gina Bari Ko~             1999 Farrar S~ http://image~
## 5 0393~ The Mummies~ E. J. W. Bar~             1999 W. W. No~ http://image~
## # ... with 2 more variables: `Image-URL-M` <chr>, `Image-URL-L` <chr>
```

Data Set: BX__Book__Ratings

```
BX_Book_Ratings<-read_delim("BX-Book-Ratings.csv", ";",
                            escape_double = FALSE, trim_ws = TRUE)

dplyr::glimpse(BX_Book_Ratings)
```

```
## Rows: 1,149,780
## Columns: 3
## $ `User-ID`     <dbl> 276725, 276726, 276727, 276729, 276729, 276733, 27673...
## $ ISBN          <chr> "034545104X", "0155061224", "0446520802", "052165615X...
## $ `Book-Rating` <dbl> 0, 5, 0, 3, 6, 0, 8, 6, 7, 10, 0, 0, 0, 0, 0, 0, 9, 0...
```

```
head(BX_Book_Ratings,5)
```

```
## # A tibble: 5 x 3
```

```
##   `User-ID` ISBN       `Book-Rating`
##       <dbl> <chr>              <dbl>
## 1    276725 034545104X             0
## 2    276726 0155061224             5
## 3    276727 0446520802             0
## 4    276729 052165615X             3
## 5    276729 0521795028             6
```

Data Set: BX_Users

```
BX_Users<-read_delim("BX-Users.csv", ";",
                             escape_double = FALSE, trim_ws = TRUE)

dplyr::glimpse(BX_Users)
```

```
## Rows: 278,858
## Columns: 3
## $ `User-ID` <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17...
## $ Location  <chr> "nyc, new york, usa", "stockton, california, usa", "mosco...
## $ Age       <chr> "NULL", "18", "NULL", "17", "NULL", "61", "NULL", "NULL",...
```

```
head(BX_Users,5)
```

```
## # A tibble: 5 x 3
##   `User-ID` Location                           Age
##       <dbl> <chr>                              <chr>
## 1         1 nyc, new york, usa                 NULL
## 2         2 stockton, california, usa          18
## 3         3 moscow, yukon territory, russia    NULL
## 4         4 porto, v.n.gaia, portugal          17
## 5         5 farnborough, hants, united kingdom NULL
```

PARTE 1: Manejo de Datos

Como sabrás, en el trabajo de un científico de datos es necesario saber manejar datos de fuentes diversas para crear datasets para el entrenamiento de modelos de aprendizaje de máquina. En esta parte de la prueba descargaras el dataset: Book-Crossing Dataset. Un dataset para la creación de un sistema de recomendación de libros recopilado por Cai-Nicolas Ziegler del Instituto para la Informática de la Universidad de Freiburg. Podés descargar el dataset como un Zip aquí.

Puedes usar R, Python o cualquier lenguaje de programación de tu preferencia para responder las siguientes preguntas, te pedimos que nos adjuntes el código que usaste para resolverlas.

Literal a

Te pedimos que agregues los archivos en la carpeta descargada para formar un dataset que pueda ser usado para entrenar un modelo de agrupación como KNN o un sistema de recomendaciones.

«En este agregado se fusionaron los tres Data-Set para contar con todas las variables claves que pueden servir para un análisis más profundo y aplicar un modelo de clasificación o de recomendación, porque puede perfilarse al usuario desde su edad, lugar de residencia, cantidad o frecuencia de lectura, su preferencia de libros, autores e editoriales; con transformaciones en variables cualitativas ya estandarizadas a númericas se podría factiblemente correr modelos».

```
base1<-left_join(BX_Book_Ratings,BX_Books,by=c("ISBN"="ISBN"))
```

```
agregado<-left_join(base1,BX_Users,by=c("User-ID"="User-ID"))

dplyr::glimpse(agregado)
```

```
## Rows: 1,149,792
## Columns: 12
## $ `User-ID`             <dbl> 276725, 276726, 276727, 276729, 276729, 27673...
## $ ISBN                  <chr> "034545104X", "0155061224", "0446520802", "05...
## $ `Book-Rating`         <dbl> 0, 5, 0, 3, 6, 0, 8, 6, 7, 10, 0, 0, 0, 0, 0,...
## $ `Book-Title`          <chr> "Flesh Tones: A Novel", "Rites of Passage", "...
## $ `Book-Author`         <chr> "M. J. Rose", "Judith Rae", "Nicholas Sparks"...
## $ `Year-Of-Publication` <dbl> 2002, 2001, 1996, 1999, 2001, 1998, NA, NA, 2...
## $ Publisher             <chr> "Ballantine Books", "Heinle", "Warner Books",...
## $ `Image-URL-S`         <chr> "http://images.amazon.com/images/P/034545104X...
## $ `Image-URL-M`         <chr> "http://images.amazon.com/images/P/034545104X...
## $ `Image-URL-L`         <chr> "http://images.amazon.com/images/P/034545104X...
## $ Location              <chr> "tyler, texas, usa", "seattle, washington, us...
## $ Age                   <chr> "NULL", "NULL", "16", "16", "16", "37", "NULL...
```

```
head(agregado,10)
```

```
## # A tibble: 10 x 12
##    `User-ID` ISBN  `Book-Rating` `Book-Title` `Book-Author` `Year-Of-Public~
##        <dbl> <chr>         <dbl> <chr>        <chr>                    <dbl>
##  1    276725 0345~             0 Flesh Tones~ M. J. Rose                2002
##  2    276726 0155~             5 Rites of Pa~ Judith Rae                2001
##  3    276727 0446~             0 The Notebook Nicholas Spa~             1996
##  4    276729 0521~             3 Help!: Leve~ Philip Prowse             1999
##  5    276729 0521~             6 The Amsterd~ Sue Leather               2001
##  6    276733 2080~             0 Les Particu~ Michel Houel~             1998
##  7    276736 3257~             8 <NA>         <NA>                        NA
##  8    276737 0600~             6 <NA>         <NA>                        NA
##  9    276744 0385~             7 A Painted H~ JOHN GRISHAM              2001
## 10    276745 3423~            10 <NA>         <NA>                        NA
## # ... with 6 more variables: Publisher <chr>, `Image-URL-S` <chr>,
## #   `Image-URL-M` <chr>, `Image-URL-L` <chr>, Location <chr>, Age <chr>
```

Literal b

Ahora notaras que la ubicación del usuario está dada por ciudad, estado o región y país. Crea columnas separadas que contengan el país, ciudad y región de cada usuario.

« Se preparo una desconcatenación de la columna de Location posteriormente se procedió a extraer y renombrar unicamente a las variables de interes ».

```
ubicacion <- within(data=BX_Users, Location<-data.frame
                    (do.call('rbind',strsplit(as.character(Location),",",fixed=TRUE))))
```

```
## Warning in rbind(c("nyc", " new york", " usa"), c("stockton", " california", :
## number of columns of result is not a multiple of vector length (arg 19)
```

```
ubicacion<-ubicacion%>%
    dplyr::mutate(ciudad=ubicacion$Location$X1,región=ubicacion$Location$X2,país=ubicacion$Location$X3,
    dplyr::select(`User-ID`,país,región,ciudad,edad)

head(ubicacion,5)
```

```
## # A tibble: 5 x 5
##   `User-ID` país              región             ciudad      edad
##       <dbl> <chr>             <chr>              <chr>       <chr>
## 1         1 " usa"            " new york"        nyc         NULL
## 2         2 " usa"            " california"      stockton    18
## 3         3 " russia"         " yukon territory" moscow      NULL
## 4         4 " portugal"       " v.n.gaia"        porto       17
## 5         5 " united kingdom" " hants"           farnborough NULL
```

Literal c

¿Cuáles son los libros con más raitings?

« Se desarrollo una agregación por ISBN para obtener únicos ISBN y obtener las sumas de los raitings por libro, luego se fusiono con la de libros para pegar los detalles como nombre, autor editorial, etc ».

```
frq(agregado$`Book-Rating`) #--Solo para verificar la frecuencia del raiting
```

```
##
## x <numeric>
## # total N=1149792  valid N=1149792  mean=2.87  sd=3.85
##
## Value |      N | Raw % | Valid % | Cum. %
## --------------------------------------------
##     0 | 716117 | 62.28 |   62.28 |  62.28
##     1 |   1770 |  0.15 |    0.15 |  62.44
##     2 |   2759 |  0.24 |    0.24 |  62.68
##     3 |   5996 |  0.52 |    0.52 |  63.20
##     4 |   8904 |  0.77 |    0.77 |  63.97
##     5 |  50974 |  4.43 |    4.43 |  68.41
##     6 |  36925 |  3.21 |    3.21 |  71.62
##     7 |  76458 |  6.65 |    6.65 |  78.27
##     8 | 103737 |  9.02 |    9.02 |  87.29
##     9 |  67542 |  5.87 |    5.87 |  93.16
##    10 |  78610 |  6.84 |    6.84 | 100.00
##  <NA> |      0 |  0.00 |    <NA> |   <NA>
```

```
#---agregndo los raiting de la base BX_Book_Ratings para fusionarla con BX_Books

dplyr::glimpse(BX_Book_Ratings)
```

```
## Rows: 1,149,780
## Columns: 3
## $ `User-ID`     <dbl> 276725, 276726, 276727, 276729, 276729, 276733, 27673...
## $ ISBN          <chr> "034545104X", "0155061224", "0446520802", "052165615X...
## $ `Book-Rating` <dbl> 0, 5, 0, 3, 6, 0, 8, 6, 7, 10, 0, 0, 0, 0, 0, 0, 9, 0...
```

```
AgregadoRainting<-BX_Book_Ratings %>%
  dplyr::select(ISBN,`Book-Rating`)%>%
  dplyr::group_by(ISBN)%>%
  dplyr::summarise(RaitingsLibro=sum(`Book-Rating`))


#----fusionando la base para conocer el nombre de los libros con mayor raitings

base2<-left_join(AgregadoRainting,BX_Books,by=c("ISBN"="ISBN"))
```

```
dplyr::glimpse(base2)
```

```
## Rows: 340,554
## Columns: 9
## $ ISBN                  <chr> "'9607092856'", "'9607092910'", "\"\\\"009474...
## $ RaitingsLibro         <dbl> 0, 10, 7, 8, 9, 20, 8, 7, 0, 7, 10, 5, 10, 0,...
## $ `Book-Title`          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ `Book-Author`         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ `Year-Of-Publication` <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ Publisher             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ `Image-URL-S`         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ `Image-URL-M`         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ `Image-URL-L`         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
```

```
librosMasRaitings<-base2 %>%
  dplyr::arrange(desc(RaitingsLibro))

head(librosMasRaitings,10)
```

```
## # A tibble: 10 x 9
##    ISBN  RaitingsLibro `Book-Title` `Book-Author` `Year-Of-Public~ Publisher
##    <chr>         <dbl> <chr>        <chr>                    <dbl> <chr>
##  1 0316~          5787 The Lovely ~ Alice Sebold              2002 Little, ~
##  2 0385~          4108 The Da Vinc~ Dan Brown                 2003 Doubleday
##  3 0312~          3134 The Red Ten~ Anita Diamant             1998 Picador ~
##  4 0679~          2800 <NA>         <NA>                        NA <NA>
##  5 0590~          2798 Harry Potte~ J. K. Rowling             1999 Arthur A~
##  6 0142~          2595 The Secret ~ Sue Monk Kidd             2003 Penguin ~
##  7 0971~          2551 Wild Animus  Rich Shapero              2004 Too Far
##  8 0060~          2524 Divine Secr~ Rebecca Wells             1997 Perennial
##  9 0446~          2402 Where the H~ Billie Letts              1998 Warner B~
## 10 0452~          2219 Girl with a~ Tracy Cheval~             2001 Plume Bo~
## # ... with 3 more variables: `Image-URL-S` <chr>, `Image-URL-M` <chr>,
## #   `Image-URL-L` <chr>
```

Literal d

¿Cuál es el top 10 de libros con mejores ratings?

« Se retoma la base anterior de los raitings pero se ejecutaron los filtros con los mayores raitings, dejando los 10 primeros libros, se eliminaron los NA para la limpieza del top 10 ».

```
dplyr::glimpse(librosMasRaitings)
```

```
## Rows: 340,554
## Columns: 9
## $ ISBN                  <chr> "0316666343", "0385504209", "0312195516", "06...
## $ RaitingsLibro         <dbl> 5787, 4108, 3134, 2800, 2798, 2595, 2551, 252...
## $ `Book-Title`          <chr> "The Lovely Bones: A Novel", "The Da Vinci Co...
## $ `Book-Author`         <chr> "Alice Sebold", "Dan Brown", "Anita Diamant",...
## $ `Year-Of-Publication` <dbl> 2002, 2003, 1998, NA, 1999, 2003, 2004, 1997,...
## $ Publisher             <chr> "Little, Brown", "Doubleday", "Picador USA", ...
## $ `Image-URL-S`         <chr> "http://images.amazon.com/images/P/0316666343...
## $ `Image-URL-M`         <chr> "http://images.amazon.com/images/P/0316666343...
## $ `Image-URL-L`         <chr> "http://images.amazon.com/images/P/0316666343...
```

```
librosMasRaitingsNoNA<-na.omit(librosMasRaitings)#---se omite los NA para mayor limpieza

librosTop10Raitings<-librosMasRaitingsNoNA%>%
  dplyr::select(`Book-Title`,`Book-Author`,Publisher,RaitingsLibro)%>%
  dplyr::arrange(desc(RaitingsLibro))%>%
  dplyr::filter(RaitingsLibro>2062)

dplyr::glimpse(librosTop10Raitings)
```

```
## Rows: 10
## Columns: 4
## $ `Book-Title`  <chr> "The Lovely Bones: A Novel", "The Da Vinci Code", "Th...
## $ `Book-Author` <chr> "Alice Sebold", "Dan Brown", "Anita Diamant", "J. K. ...
## $ Publisher     <chr> "Little, Brown", "Doubleday", "Picador USA", "Arthur ...
## $ RaitingsLibro <dbl> 5787, 4108, 3134, 2798, 2595, 2551, 2524, 2402, 2219,...
```

```
librosTop10Raitings<-librosTop10Raitings %>%
  dplyr::rename(Libro=`Book-Title`,Autor=`Book-Author`,Editorial=Publisher,Raiting=RaitingsLibro)

librosTop10Raitings
```

```
## # A tibble: 10 x 4
##    Libro                                  Autor       Editorial        Raiting
##    <chr>                                  <chr>       <chr>              <dbl>
##  1 The Lovely Bones: A Novel              Alice Sebo~ Little, Brown       5787
##  2 The Da Vinci Code                      Dan Brown   Doubleday           4108
##  3 The Red Tent (Bestselling Backlist)    Anita Diam~ Picador USA         3134
##  4 Harry Potter and the Sorcerer's Stone (~ J. K. Rowl~ Arthur A. Levin~    2798
##  5 The Secret Life of Bees                Sue Monk K~ Penguin Books       2595
##  6 Wild Animus                            Rich Shape~ Too Far             2551
##  7 Divine Secrets of the Ya-Ya Sisterhood:~ Rebecca We~ Perennial          2524
##  8 Where the Heart Is (Oprah's Book Club (~ Billie Let~ Warner Books       2402
##  9 Girl with a Pearl Earring              Tracy Chev~ Plume Books         2219
## 10 Angels &amp Demons                     Dan Brown   Pocket Star         2179
```

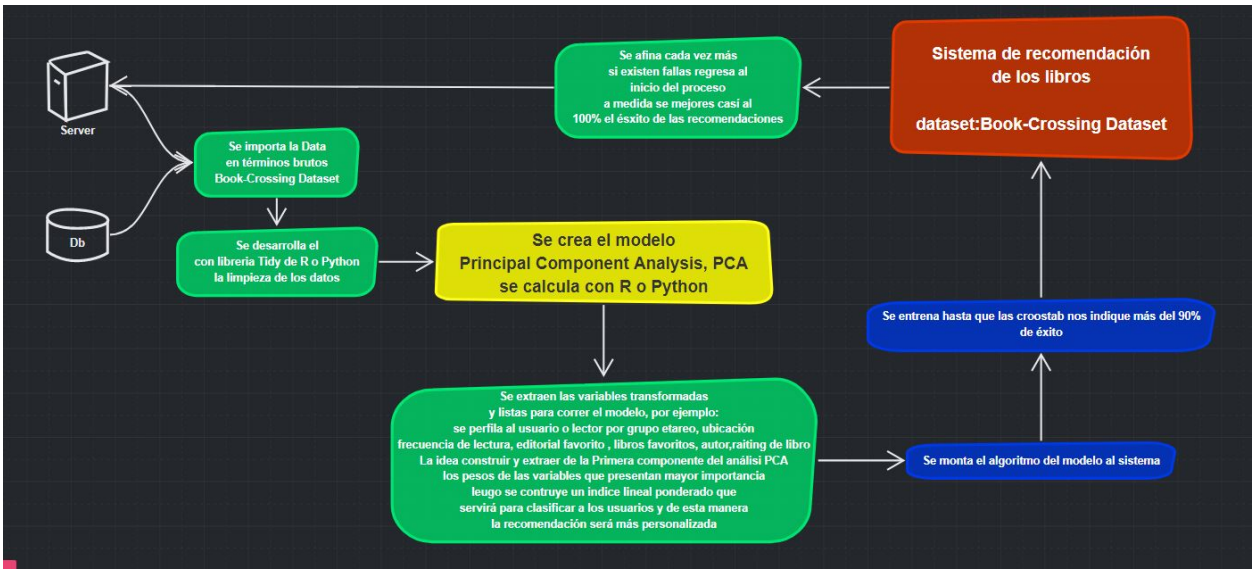
Literal e

¿Cómo diseñarías el sistema de recomendación? (No se necesita programarlo, puedes explicar conceptualmente el modelo/sistema que implementarías).

Prácticamente se propone un proceso normal de corrida de modelo, resumiendo algunos sub-procesos, los pasos generales puden englobarse de la forma siguiente:

1- Extracción e importación de las diferentes fuentes de datos del data set Book-Crossing 2- Luego este debe ser transformado, pasando por el proceso de limpieza y consistencia 3- Al tener preparada, se transforma para que algunas variables puedan comprender una forma cuantitativa para someterse a los modelos multivariantes que exige machine learning. 4- EL modelo propuesto es el de componentes principales PCA, para extraer las ponderaciones o pesos de la primera componete para construir un indicador o indice que permita desarrollar una clasificación de usarios y mejorar la campaña de recomendación. 5- Se monta el algoritmo a los sistemas y se entrena el modelo hasta que su presición logre tasas cercanas o mayores al 90%. 6- Luego en plena producción se verifica constantemente para que el proceso afine y mejore totalmente, hasta que el algoritmo alcance su máximo desarrollo y predicción en modo de producción.

El esquema del modelo conceptual muy general se observa en la siguiente figura:



PARTE 2: Construcción de un Modelo Machine Learning

Be-A-Host.com es un Marketplace bilateral que permite crear conexiones entre huéspedes y hospedajes. La plataforma funciona de la siguiente manera: un huésped encuentra un alojamiento disponible (listing) que le gusta y envía una solicitud al dueño del alojamiento. Hay dos formas de solicitar alojamiento, una es la de 'reserva' ('book_it') y otra la de 'reserva instantanea' (instant_book) que automáticamente hace la reservación. Al recibir la solicitud de 'reserva' el alojamiento puede decidir si aceptar o no la reservación.

Los alojamientos pueden rechazar un huésped por variar razones. Algunas pueden ser logísticas: las fechas no funcionan o personales: los huéspedes pueden ser riesgosos para el alojamiento. El objetivo de esta prueba es maximizar la probabilidad de los huéspedes de ser aceptados en el alojamiento que solicitan.

Preguntas

Con los archivos adjuntos en el correo que recibió responda las siguientes preguntas:

1-Be-A-Host necesita comprender por qué razón un huésped logra tener una solicitud de reserva exitosa y cuales son los factores que hacen que un huésped tenga mayores probabilidades de ser aceptado por los alojamientos. Construye un modelo que permita comprender la tasa de aceptación de solicitudes. Basado en este modelo, ¿qué adición podrías recomendar hacer al website para aumentar la probabilidad de aceptación de los huéspedes?. 2-Como un experimento Be-A-Host ha agregado un nuevo feature que obliga a los huéspedes a enviar un mensaje de no menos de 140 caracteres explicando por qué les interesa ese alojamiento en particular, las asignaciones están en el archivo assignments.csv, se corrió un experimento en que la mitad de los huéspedes se pusieron en un control. ¿Debería lanzarse ese cambio a la plataforma para todos los clientes?

Desarrollo del Modelo

Preparación de Librerias

```
library(ggplot2)
library(lubridate)
library(cluster)
library(mclust)
library(pheatmap)
library(clustertend)
library(eclust)
```

```
library(NbClust)
library(pheatmap)
library(d3heatmap)
library(rattle)
library(factoextra)
library(dendextend)
library(igraph)
library(clValid)
library(nortest)
library(magrittr)
library(ggpubr)
```

Preparación de Base de Datos

```
#--contacts
Contacts<- read_delim("contacts.csv", ",",
                      escape_double = FALSE, trim_ws = TRUE)

dplyr::glimpse(Contacts)
```

Base Contacts

```
## Rows: 23,143
## Columns: 18
## $ id_guest_anon                          <chr> "56d70d7c-1d0a-4594-a250-ed62f...
## $ id_host_anon                           <chr> "65a56b50-faf2-44a2-845f-0c467...
## $ id_listing_anon                        <chr> "4deeb033-183e-437c-b94c-851ff...
## $ ts_interaction_first                   <dttm> 2013-01-13 21:03:07, 2013-01-...
## $ ts_reply_at_first                      <dttm> 2013-01-14 21:19:42, 2013-01-...
## $ ts_accepted_at_first                   <dttm> 2013-01-14 21:19:42, 2013-01-...
## $ ts_booking_at                          <dttm> 2013-01-14 21:19:42, 2013-01-...
## $ ds_checkin_first                       <date> 2013-01-18, 2013-01-23, 2013-...
## $ ds_checkout_first                      <date> 2013-01-20, 2013-01-25, 2013-...
## $ m_guests_first                         <dbl> 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, ...
## $ m_interactions                         <dbl> 13, 5, 7, 5, 7, 9, 3, 4, 3, 4,...
## $ m_first_message_length_in_characters <dbl> 165, NA, 350, 426, 239, NA, 64...
## $ dim_contact_channel_first              <chr> "book_it", "book_it", "book_it...
## $ dim_room_type                          <chr> "Private room", "Private room"...
## $ dim_total_reviews                      <dbl> 78, 78, 78, 78, 78, 78, 78, 78...
## $ dim_person_capacity                    <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...
## $ dim_guest_language                     <lgl> NA, NA, NA, NA, NA, NA, NA, NA...
## $ dim_host_language                      <chr> "es", "es", "es", "es", "es", ...
```

```
head(Contacts,10)
```

```
## # A tibble: 10 x 18
##    id_guest_anon id_host_anon id_listing_anon ts_interaction_fir~
##    <chr>         <chr>        <chr>           <dttm>
##  1 56d70d7c-1d0~ 65a56b50-fa~ 4deeb033-183e-~ 2013-01-13 21:03:07
##  2 dcfb93c4-07c~ 65a56b50-fa~ 4deeb033-183e-~ 2013-01-15 23:43:49
##  3 aef63ee0-7fd~ 65a56b50-fa~ 4deeb033-183e-~ 2013-03-19 21:30:13
##  4 58788b6f-121~ 65a56b50-fa~ 4deeb033-183e-~ 2013-01-03 16:32:01
##  5 c3945c48-53b~ 65a56b50-fa~ 4deeb033-183e-~ 2013-04-13 04:31:33
##  6 5dd9d290-9c2~ 65a56b50-fa~ 4deeb033-183e-~ 2013-02-07 17:13:18
```

```
##  7 e0a44946-35a~ 65a56b50-fa~ 4deeb033-183e-~ 2013-12-25 07:14:56
##  8 6602fc3b-024~ 65a56b50-fa~ 4deeb033-183e-~ 2013-01-25 20:34:51
##  9 82b913a1-a81~ 65a56b50-fa~ 4deeb033-183e-~ 2013-02-10 16:52:51
## 10 6e8a1693-593~ 65a56b50-fa~ 4deeb033-183e-~ 2013-01-24 17:16:28
## # ... with 14 more variables: ts_reply_at_first <dttm>,
## #   ts_accepted_at_first <dttm>, ts_booking_at <dttm>, ds_checkin_first <date>,
## #   ds_checkout_first <date>, m_guests_first <dbl>, m_interactions <dbl>,
## #   m_first_message_length_in_characters <dbl>,
## #   dim_contact_channel_first <chr>, dim_room_type <chr>,
## #   dim_total_reviews <dbl>, dim_person_capacity <dbl>,
## #   dim_guest_language <lgl>, dim_host_language <chr>
```

```
#--assignments

assignments<- read_delim("assignments.csv", ",",
                      escape_double = FALSE, trim_ws = TRUE)

dplyr::glimpse(assignments)
```

Base Assigments

```
## Rows: 19,996
## Columns: 2
## $ id_user_anon <chr> "3e3b1bc7-9c46-4798-b955-8ed3f6bdf841", "99fef26b-5b8c...
## $ ab           <chr> "treatment", "treatment", "control", "treatment", "tre...
```

```
head(assignments,5)
```

```
## # A tibble: 5 x 2
##   id_user_anon                         ab
##   <chr>                                <chr>
## 1 3e3b1bc7-9c46-4798-b955-8ed3f6bdf841 treatment
## 2 99fef26b-5b8c-4d96-b6f4-552a951512d6 treatment
## 3 25886018-ed8c-4c40-9915-ead001e2c021 control
## 4 f7e16080-c7ae-4e46-8247-26f92d4495a6 treatment
## 5 9c2d741a-acfe-4728-ba3b-0571bfd82306 treatment
```

Fusión de bases

Se desarrolla la fusión de la base de **contacts y assigments** para empezar a prepararla y transformarla con todas sus variables, con la finalidad de empezar a desarrollar el modelo posteriormente a su transformación, además se debe saber las caracteristicas del grupo de control.

```
#---Se fusionan las bases para el análisis para conocer el grupo de control

DB_Modelo<-left_join(Contacts,assignments,by=c("id_guest_anon"="id_user_anon"))

head(DB_Modelo,5)
```

```
## # A tibble: 5 x 19
##   id_guest_anon id_host_anon id_listing_anon ts_interaction_fir~
##   <chr>         <chr>        <chr>           <dttm>
## 1 56d70d7c-1d0~ 65a56b50-fa~ 4deeb033-183e-~ 2013-01-13 21:03:07
## 2 dcfb93c4-07c~ 65a56b50-fa~ 4deeb033-183e-~ 2013-01-15 23:43:49
## 3 aef63ee0-7fd~ 65a56b50-fa~ 4deeb033-183e-~ 2013-03-19 21:30:13
## 4 58788b6f-121~ 65a56b50-fa~ 4deeb033-183e-~ 2013-01-03 16:32:01
```

```
## 5 c3945c48-53b~ 65a56b50-fa~ 4deeb033-183e-~ 2013-04-13 04:31:33
## # ... with 15 more variables: ts_reply_at_first <dttm>,
## #   ts_accepted_at_first <dttm>, ts_booking_at <dttm>, ds_checkin_first <date>,
## #   ds_checkout_first <date>, m_guests_first <dbl>, m_interactions <dbl>,
## #   m_first_message_length_in_characters <dbl>,
## #   dim_contact_channel_first <chr>, dim_room_type <chr>,
## #   dim_total_reviews <dbl>, dim_person_capacity <dbl>,
## #   dim_guest_language <lgl>, dim_host_language <chr>, ab <chr>
```

Preparación y Transformación de la base

Para aplicar el modelo, la base de datos deben de cumplir con un estandard de limpieza y transformación, como el modelo que se pretende aplicar es el de Clúster Jerárquico, la base debe de ser transformados a terminos númericos.

```
#----Se realiza un sut set para ir determinando la base

DB_Modelo_sub<-DB_Modelo%>%
  dplyr::select(id_guest_anon,ts_interaction_first,ts_accepted_at_first,ds_checkin_first,
                ds_checkout_first,m_guests_first,m_interactions,
                m_first_message_length_in_characters,dim_contact_channel_first,dim_room_type,
                dim_total_reviews,dim_person_capacity,ab)

dplyr::glimpse(DB_Modelo_sub)
```

Se ejecuta un sub-set

```
## Rows: 25,522
## Columns: 13
## $ id_guest_anon                       <chr> "56d70d7c-1d0a-4594-a250-ed62f...
## $ ts_interaction_first                <dttm> 2013-01-13 21:03:07, 2013-01-...
## $ ts_accepted_at_first                <dttm> 2013-01-14 21:19:42, 2013-01-...
## $ ds_checkin_first                    <date> 2013-01-18, 2013-01-23, 2013-...
## $ ds_checkout_first                   <date> 2013-01-20, 2013-01-25, 2013-...
## $ m_guests_first                      <dbl> 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, ...
## $ m_interactions                      <dbl> 13, 5, 7, 5, 7, 9, 3, 4, 3, 4,...
## $ m_first_message_length_in_characters <dbl> 165, NA, 350, 426, 239, NA, 64...
## $ dim_contact_channel_first           <chr> "book_it", "book_it", "book_it...
## $ dim_room_type                       <chr> "Private room", "Private room"...
## $ dim_total_reviews                   <dbl> 78, 78, 78, 78, 78, 78, 78, 78...
## $ dim_person_capacity                 <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...
## $ ab                                  <chr> "control", "control", "treatme...
```

```
head(DB_Modelo_sub,5)
```

```
## # A tibble: 5 x 13
##   id_guest_anon ts_interaction_fir~ ts_accepted_at_fir~ ds_checkin_first
##   <chr>         <dttm>              <dttm>              <date>
## 1 56d70d7c-1d0~ 2013-01-13 21:03:07 2013-01-14 21:19:42 2013-01-18
## 2 dcfb93c4-07c~ 2013-01-15 23:43:49 2013-01-16 09:04:37 2013-01-23
## 3 aef63ee0-7fd~ 2013-03-19 21:30:13 2013-03-20 12:50:49 2013-05-31
## 4 58788b6f-121~ 2013-01-03 16:32:01 2013-01-03 17:02:53 2013-02-09
## 5 c3945c48-53b~ 2013-04-13 04:31:33 2013-04-13 16:44:20 2013-06-06
## # ... with 9 more variables: ds_checkout_first <date>, m_guests_first <dbl>,
## #   m_interactions <dbl>, m_first_message_length_in_characters <dbl>,
```

```
## #   dim_contact_channel_first <chr>, dim_room_type <chr>,
## #   dim_total_reviews <dbl>, dim_person_capacity <dbl>, ab <chr>
```

```
#---Transformaciones de variables

#---Renombrando
DB_Modelo_sub<-DB_Modelo_sub%>%
  dplyr::rename(IdUser=id_guest_anon,PrimeraConsulta=ts_interaction_first,AceptaConsulta=ts_accepted_at,
                SelloSalida=ds_checkout_first,NumeroInvitados=m_guests_first,NumeroMensajes=m_interacti
                NumeroCaracteres=m_first_message_length_in_characters,TipoCanal=dim_contact_channel_firs
                TotalRevisiones=dim_total_reviews,CapacidadPersonas=dim_person_capacity,Control=ab)

dplyr::glimpse(DB_Modelo_sub)
```

Se ejecutan las transformaciones y recodificaciones

```
## Rows: 25,522
## Columns: 13
## $ IdUser            <chr> "56d70d7c-1d0a-4594-a250-ed62f7cf7ac4", "dcfb93c4...
## $ PrimeraConsulta   <dttm> 2013-01-13 21:03:07, 2013-01-15 23:43:49, 2013-0...
## $ AceptaConsulta    <dttm> 2013-01-14 21:19:42, 2013-01-16 09:04:37, 2013-0...
## $ SelloEntrada      <date> 2013-01-18, 2013-01-23, 2013-05-31, 2013-02-09, ...
## $ SelloSalida       <date> 2013-01-20, 2013-01-25, 2013-06-03, 2013-02-11, ...
## $ NumeroInvitados   <dbl> 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2...
## $ NumeroMensajes    <dbl> 13, 5, 7, 5, 7, 9, 3, 4, 3, 4, 4, 2, 9, 9, 17, 11...
## $ NumeroCaracteres  <dbl> 165, NA, 350, 426, 239, NA, 640, 487, 246, 528, 5...
## $ TipoCanal         <chr> "book_it", "book_it", "book_it", "book_it", "book...
## $ TipoHabitacion    <chr> "Private room", "Private room", "Private room", "...
## $ TotalRevisiones   <dbl> 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 7...
## $ CapacidadPersonas <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ Control           <chr> "control", "control", "treatment", "treatment", "...
```

```
#-------Transformando

#---Verificando algunas frecuencias para las transformaciones
frq(DB_Modelo_sub$TipoHabitacion)
```

```
##
## x <character>
## # total N=25522  valid N=25522  mean=1.34  sd=0.50
##
## Value           |     N | Raw % | Valid % | Cum. %
## ------------------------------------------------
## Entire home/apt | 17187 | 67.34 |   67.34 |  67.34
## Private room    |  7980 | 31.27 |   31.27 |  98.61
## Shared room     |   355 |  1.39 |    1.39 | 100.00
## <NA>            |     0 |  0.00 |    <NA> |   <NA>
```

```
frq(DB_Modelo_sub$TipoCanal)
```

```
##
## x <character>
## # total N=25522  valid N=25522  mean=1.14  sd=0.35
##
## Value          |     N | Raw % | Valid % | Cum. %
```

```
## ------------------------------------------------
## book_it        | 21909 | 85.84 |   85.84 |  85.84
## instant_booked |  3613 | 14.16 |   14.16 | 100.00
## <NA>           |     0 |  0.00 |    <NA> |   <NA>
```

```
frq(DB_Modelo_sub$Control)
```

```
##
## x <character>
## # total N=25522  valid N=25522  mean=1.50  sd=0.50
##
## Value     |     N | Raw % | Valid % | Cum. %
## ------------------------------------------
## control   | 12771 | 50.04 |   50.04 |  50.04
## treatment | 12751 | 49.96 |   49.96 | 100.00
## <NA>      |     0 |  0.00 |    <NA> |   <NA>
```

```
#---Tranasformaciones
DB_Transform<-DB_Modelo_sub%>%
  dplyr::mutate(TiempoEspera=(as.numeric(difftime(AceptaConsulta, PrimeraConsulta), units="secs")))%>%
  dplyr::mutate(TiempoSello=(as.numeric(difftime(SelloSalida, SelloEntrada), units="days")))%>%
  dplyr::mutate(Canal=(if_else(TipoCanal=="instant_booked",1,2))) %>%
  dplyr::mutate(Habitacion=(if_else(TipoHabitacion=="Entire home/apt",1,
                                    if_else(TipoHabitacion=="Private room",2,if_else(TipoHabitacion=="S
  dplyr::mutate(GrupoControl=if_else(Control=="control",1, 2),corr=seq(1,25522,by=1))

head(DB_Transform,5)
```

```
## # A tibble: 5 x 19
##   IdUser PrimeraConsulta     AceptaConsulta      SelloEntrada SelloSalida
##   <chr>  <dttm>              <dttm>              <date>       <date>
## 1 56d70~ 2013-01-13 21:03:07 2013-01-14 21:19:42 2013-01-18   2013-01-20
## 2 dcfb9~ 2013-01-15 23:43:49 2013-01-16 09:04:37 2013-01-23   2013-01-25
## 3 aef63~ 2013-03-19 21:30:13 2013-03-20 12:50:49 2013-05-31   2013-06-03
## 4 58788~ 2013-01-03 16:32:01 2013-01-03 17:02:53 2013-02-09   2013-02-11
## 5 c3945~ 2013-04-13 04:31:33 2013-04-13 16:44:20 2013-06-06   2013-06-08
## # ... with 14 more variables: NumeroInvitados <dbl>, NumeroMensajes <dbl>,
## #   NumeroCaracteres <dbl>, TipoCanal <chr>, TipoHabitacion <chr>,
## #   TotalRevisiones <dbl>, CapacidadPersonas <dbl>, Control <chr>,
## #   TiempoEspera <dbl>, TiempoSello <dbl>, Canal <dbl>, Habitacion <dbl>,
## #   GrupoControl <dbl>, corr <dbl>
```

Seleccionando las variables que entraran al modelo

```
#---Seleccionando las variables que entraran al modelo
dplyr::glimpse(DB_Transform)
```

```
## Rows: 25,522
## Columns: 19
## $ IdUser            <chr> "56d70d7c-1d0a-4594-a250-ed62f7cf7ac4", "dcfb93c4...
## $ PrimeraConsulta   <dttm> 2013-01-13 21:03:07, 2013-01-15 23:43:49, 2013-0...
## $ AceptaConsulta    <dttm> 2013-01-14 21:19:42, 2013-01-16 09:04:37, 2013-0...
## $ SelloEntrada      <date> 2013-01-18, 2013-01-23, 2013-05-31, 2013-02-09, ...
## $ SelloSalida       <date> 2013-01-20, 2013-01-25, 2013-06-03, 2013-02-11, ...
## $ NumeroInvitados   <dbl> 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2...
## $ NumeroMensajes    <dbl> 13, 5, 7, 5, 7, 9, 3, 4, 3, 4, 4, 2, 9, 9, 17, 11...
```

```
## $ NumeroCaracteres  <dbl> 165, NA, 350, 426, 239, NA, 640, 487, 246, 528, 5...
## $ TipoCanal         <chr> "book_it", "book_it", "book_it", "book_it", "book...
## $ TipoHabitacion    <chr> "Private room", "Private room", "Private room", "...
## $ TotalRevisiones   <dbl> 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 7...
## $ CapacidadPersonas <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ Control           <chr> "control", "control", "treatment", "treatment", "...
## $ TiempoEspera      <dbl> 87395, 33648, 55236, 1852, 43967, 81884, NA, 6828...
## $ TiempoSello       <dbl> 2, 2, 3, 2, 2, 2, 5, 3, 5, 4, 4, 5, 3, 3, 2, 4, 4...
## $ Canal             <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ Habitacion        <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ GrupoControl      <dbl> 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 2, 2, 2...
## $ corr              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...
```

```
bm<-DB_Transform %>%
  dplyr::select(TiempoEspera,TiempoSello,NumeroInvitados,NumeroMensajes,NumeroCaracteres,Canal,Habitaci
                TotalRevisiones,CapacidadPersonas,corr)

dplyr::glimpse(bm)
```

```
## Rows: 25,522
## Columns: 10
## $ TiempoEspera      <dbl> 87395, 33648, 55236, 1852, 43967, 81884, NA, 6828...
## $ TiempoSello       <dbl> 2, 2, 3, 2, 2, 2, 5, 3, 5, 4, 4, 5, 3, 3, 2, 4, 4...
## $ NumeroInvitados   <dbl> 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2...
## $ NumeroMensajes    <dbl> 13, 5, 7, 5, 7, 9, 3, 4, 3, 4, 4, 2, 9, 9, 17, 11...
## $ NumeroCaracteres  <dbl> 165, NA, 350, 426, 239, NA, 640, 487, 246, 528, 5...
## $ Canal             <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ Habitacion        <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ TotalRevisiones   <dbl> 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 78, 7...
## $ CapacidadPersonas <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ corr              <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...
```

```
head(bm,5)
```

```
## # A tibble: 5 x 10
##   TiempoEspera TiempoSello NumeroInvitados NumeroMensajes NumeroCaracteres Canal
##          <dbl>       <dbl>           <dbl>          <dbl>             <dbl> <dbl>
## 1        87395           2               2             13               165     2
## 2        33648           2               2              5                NA     2
## 3        55236           3               2              7               350     2
## 4         1852           2               2              5               426     2
## 5        43967           2               2              7               239     2
## # ... with 4 more variables: Habitacion <dbl>, TotalRevisiones <dbl>,
## #   CapacidadPersonas <dbl>, corr <dbl>
```

Se calcula una muestra representativa para correr el modelo que también será de entrenamiento

```
#----Se sacará una muestra representativa de la base que a la vez servirá para entrenamiento
#----la muestra se reduce por la eliminación de los NA de hecho por eso se sobremuestrea a 7000

bm_muestra<-bm %>%
  sample_n(size=7000,replace=FALSE)

bm_muestra<-na.omit(bm_muestra)
```

```r
bm<-bm_muestra

dplyr::glimpse(bm)
```

```
## Rows: 4,469
## Columns: 10
## $ TiempoEspera      <dbl> 33757, 8812651, 50053, 0, 672, 61754, 583, 2549, ...
## $ TiempoSello       <dbl> 4, 5, 5, 2, 4, 2, 2, 8, 2, 3, 3, 13, 30, 2, 3, 4,...
## $ NumeroInvitados   <dbl> 1, 4, 3, 2, 1, 2, 2, 1, 3, 2, 2, 2, 2, 3, 2, 2, 2...
## $ NumeroMensajes    <dbl> 12, 12, 9, 5, 7, 7, 17, 11, 14, 13, 4, 12, 4, 9, ...
## $ NumeroCaracteres  <dbl> 0, 198, 252, 280, 5, 284, 478, 144, 156, 326, 270...
## $ Canal             <dbl> 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2...
## $ Habitacion        <dbl> 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ TotalRevisiones   <dbl> 208, 10, 36, 67, 109, 76, 63, 16, 44, 94, 66, 17,...
## $ CapacidadPersonas <dbl> 3, 4, 3, 2, 2, 2, 4, 4, 2, 4, 2, 2, 4, 3, 2, 3, 5...
## $ corr              <dbl> 9491, 21648, 2665, 5460, 4823, 15965, 10949, 1817...
```

```r
head(bm,5)
```

```
## # A tibble: 5 x 10
##   TiempoEspera TiempoSello NumeroInvitados NumeroMensajes NumeroCaracteres Canal
##          <dbl>       <dbl>           <dbl>          <dbl>             <dbl> <dbl>
## 1        33757           4               1             12                 0     2
## 2      8812651           5               4             12               198     2
## 3        50053           5               3              9               252     2
## 4            0           2               2              5               280     1
## 5          672           4               1              7                 5     2
## # ... with 4 more variables: Habitacion <dbl>, TotalRevisiones <dbl>,
## #   CapacidadPersonas <dbl>, corr <dbl>
```

Desarrollo del modelo de Clúster Jerárquico

Al modelo entrarán 10 variables transformadas,algunas se han recodificadas y se han convertido a cuantitativas y se eliminaron los NA. para que las pruebas estadísticas del clúster no tengan problema, el set de entrenamiento o muestra es de 4,564, una muestra representativa de los 25 mil que resultaron de las fusiones.

```r
#----Modelo de cluster jerarquico
#---Escalando la base bm (normalizando o tipificando)

bm<-scale(bm)

head(bm,5)
```

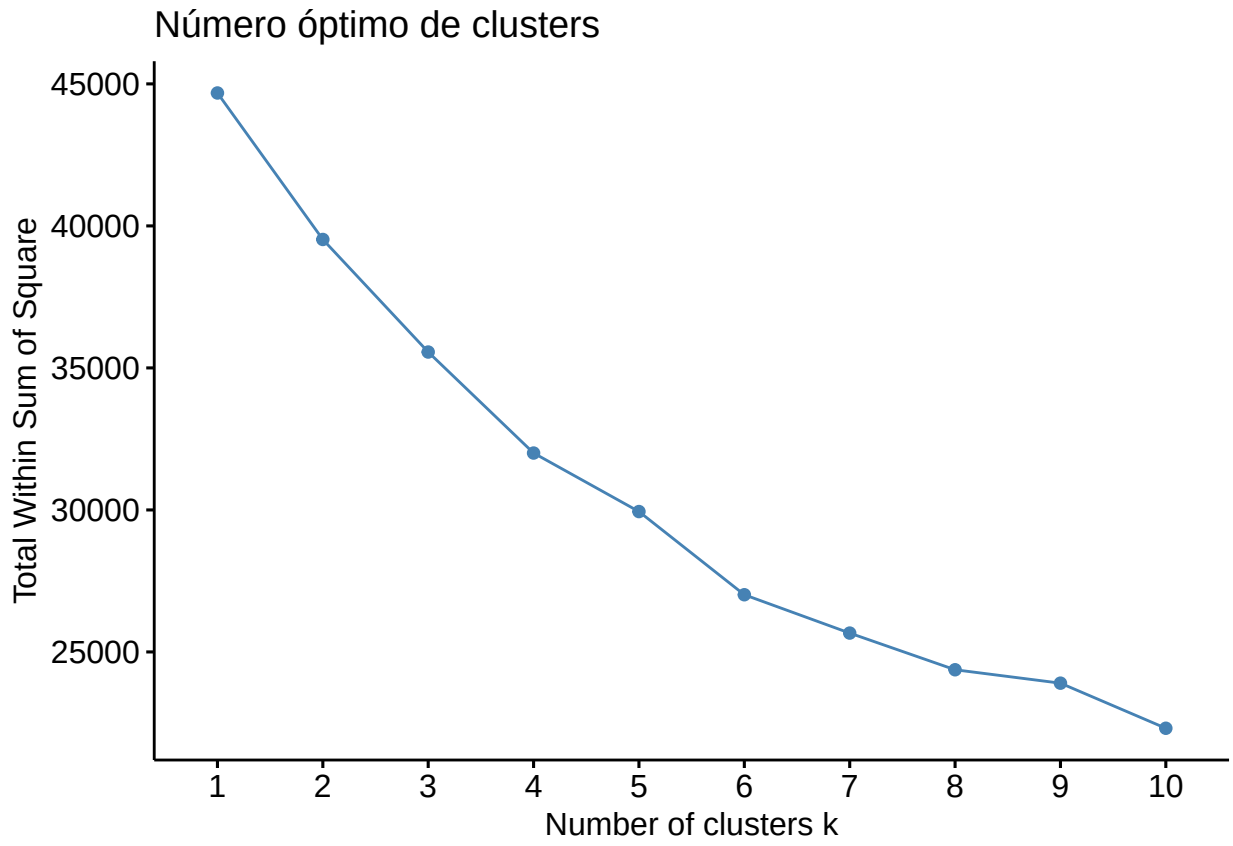
Escalando la base

```
##      TiempoEspera TiempoSello NumeroInvitados NumeroMensajes NumeroCaracteres
## [1,]   0.01690597 -0.04512473      -0.9768127     0.61199687      -1.53496524
## [2,]  28.34405588  0.08439512       1.2492842     0.61199687      -0.42275943
## [3,]   0.06948881  0.08439512       0.5072519     0.06965818      -0.11943057
## [4,]  -0.09201887 -0.30416441      -0.2347804    -0.65346007       0.03785106
## [5,]  -0.08985050 -0.04512473      -0.9768127    -0.29190094      -1.50687923
##           Canal Habitacion TotalRevisiones CapacidadPersonas       corr
## [1,]  0.4831727 -0.5872021      2.88995804        -0.3283825 -0.3964847
## [2,]  0.4831727 -0.5872021     -1.06596894         0.1285108  1.3394623
## [3,]  0.4831727 -0.5872021     -0.54650378        -0.3283825 -1.3711967
```

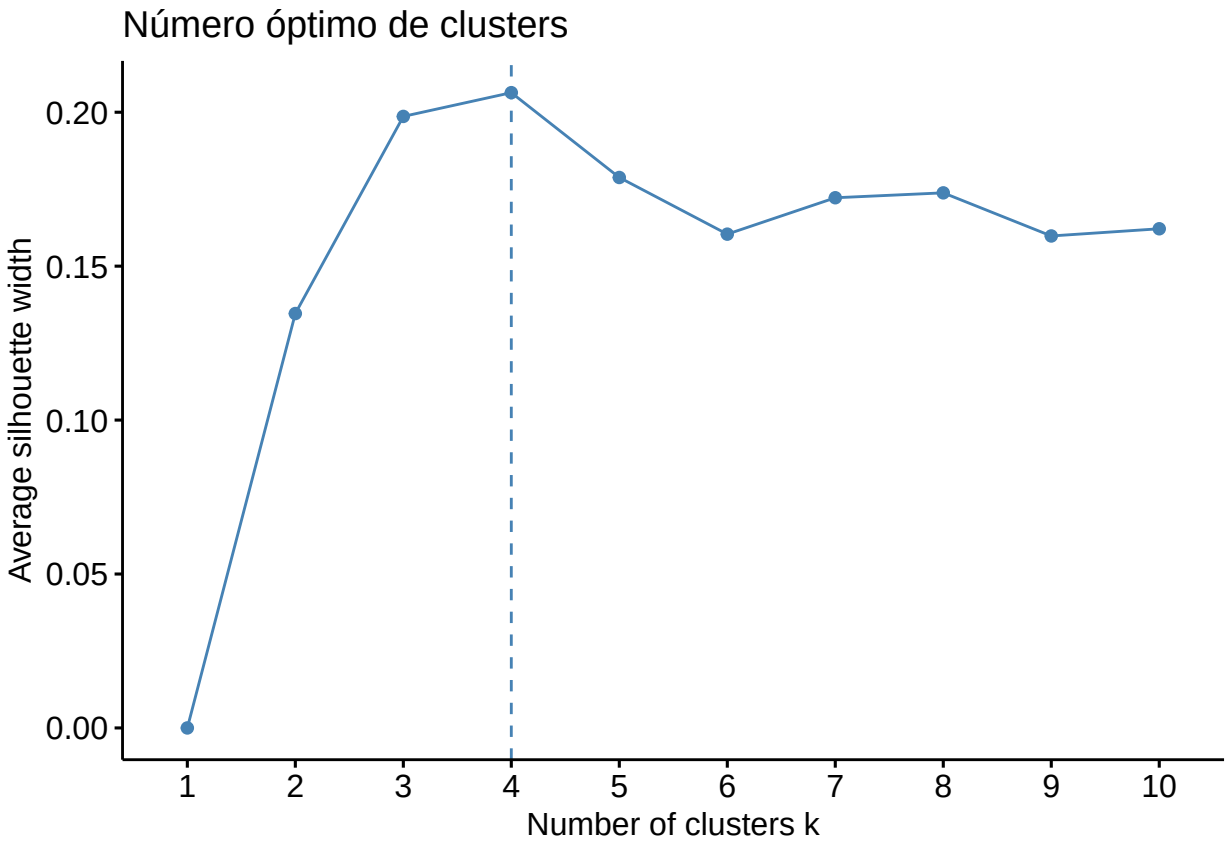
```
## [4,] -2.0691901 -0.5872021      0.07285853        -0.7852757 -0.9720874
## [5,]  0.4831727  1.5256854      0.91199455        -0.7852757 -1.0630472
```

Definiendo el número óptimo de clústers Las pruebas gráficas por el método de **ELBOW y el método de AVERAGE SILHOUETTE**, indican entre 4 y 3, clúster, decidí determinarlo en 3 clústers porque la prueba de calidad del clúster, que es el próximo ítems indica mejor establecimiento para 3 grupos.

```
#--------Número optimo de cluster---Método: ELBOW
fviz_nbclust(x = bm, FUNcluster = kmeans, method = "wss") +
  labs(title = "Número óptimo de clusters")
```



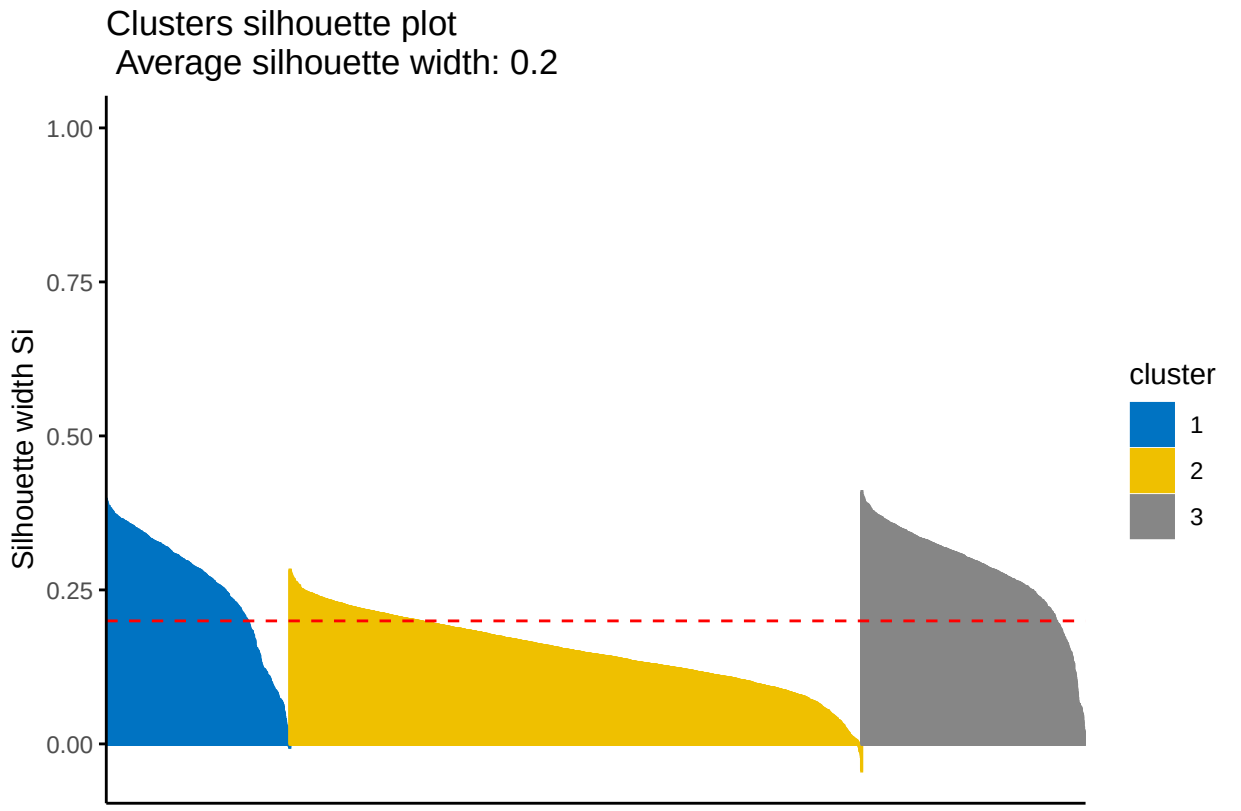

```
#--------Número optimo de cluster---Método: AVERAGE SILHOUETTE METHOD
fviz_nbclust(x = bm, FUNcluster = kmeans, method = "silhouette") +
  labs(title = "Número óptimo de clusters")
```



Calidad de los grupos del cluster Tal como se comento en la sección anterior la calidad del clúster indica mejor agrupamiento para 3 grupos.

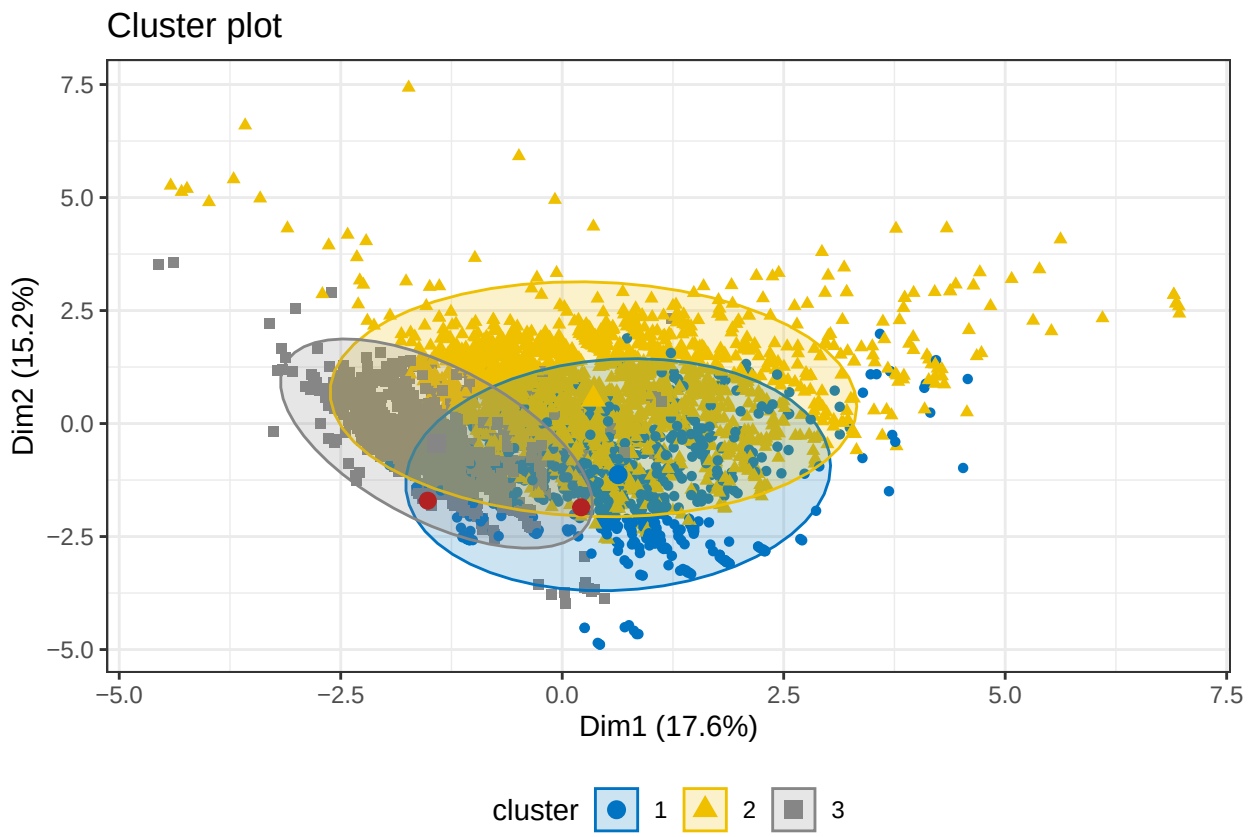
```
#------Gráfico de Silhouette width
km_clusters <- eclust(x = bm, FUNcluster = "kmeans", k = 3, seed = 123,
                      hc_metric = "euclidean", nstart = 50, graph = FALSE)
fviz_silhouette(sil.obj = km_clusters, print.summary = TRUE, palette = "jco",
                ggtheme = theme_classic())
```

```
##   cluster size ave.sil.width
## 1       1  838          0.26
## 2       2 2610          0.15
## 3       3 1021          0.28
```



Visualización gráfica de los grupos Se observa a continuación la representación gráfica de los clústers conformados:

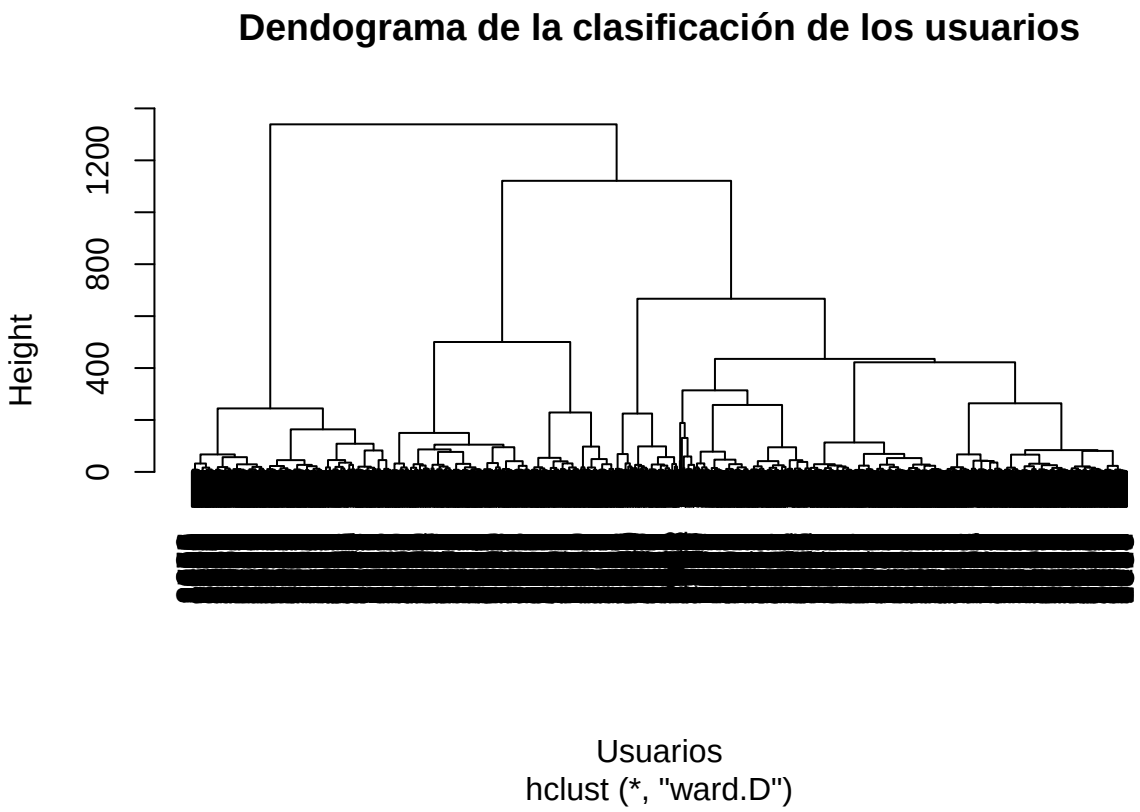
```
#----Gráfica del cluster
p <- fviz_cluster(object = km_clusters, geom = "point", ellipse.type = "norm",
                  palette = "jco")
p + geom_point(data = p$data[c(1, 3000),], colour = "firebrick", size = 2.5) +
  theme_bw() + theme(legend.position = "bottom")
```



Dendograma

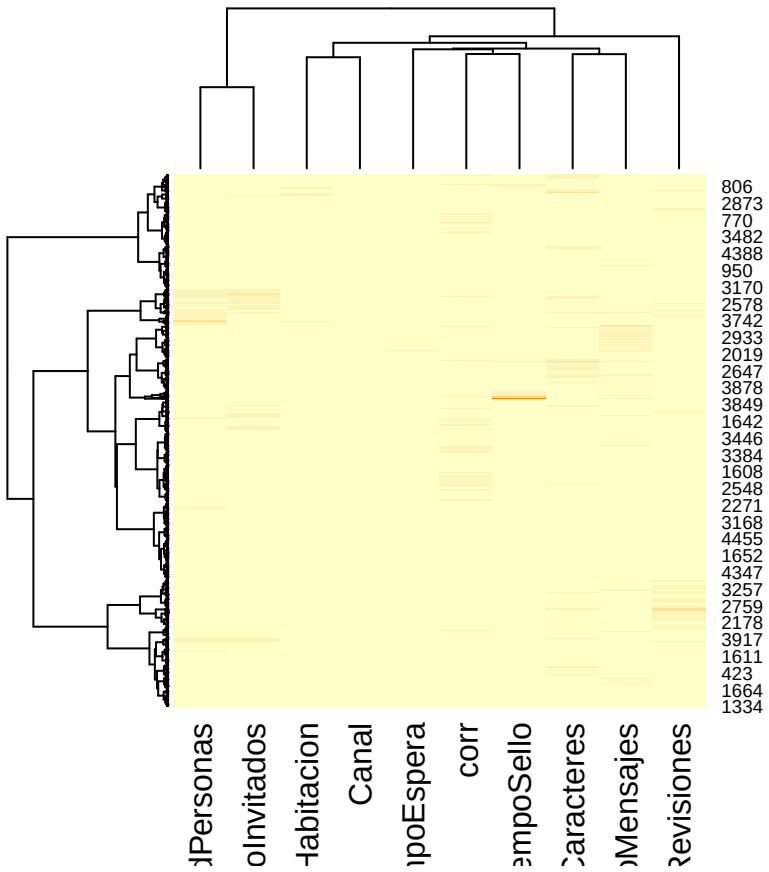
```
#----Dendograma
hc_completo <- bm %>% scale() %>% dist(method = "euclidean") %>% hclust(method = "ward.D")

plot(hc_completo, main="Dendograma de la clasificación de los usuarios",xlab="Usuarios")
```



Heatmap se construyé el heatmap:

```
#------heatmap
heatmap(x = bm, scale = "none", distfun = function(x){dist(x, method = "euclidean")},
        hclustfun = function(x){hclust(x, method = "ward.D")}, cexRow = 0.7)
```



Preparando la base para el análisis En este apartado se fusionan los cluster con la base de entrenamiento o muestra para desarrollar los análisis de recomendación para la respuestas de los númerales 1 y 2

```
#-----------Preparando la base

clusters <- cutree(tree = hc_completo, k = 3)

clusters
```

```
##    [1] 1 2 2 1 3 2 2 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 3 3 1 2 2 3 2 1 1 2 2 2 1 1 2
##   [38] 2 2 2 2 1 2 3 2 1 1 2 2 2 2 2 2 2 2 2 3 1 2 1 3 3 3 3 2 3 2 2 2 1 3 2 1 1
##   [75] 3 2 2 3 2 2 3 1 2 3 2 2 2 2 3 2 1 2 3 2 2 2 1 2 1 2 2 2 1 3 2 1 2 2 2 2 3
##  [112] 2 2 2 2 2 2 1 2 1 2 2 2 2 3 1 2 1 2 2 2 1 3 3 1 1 2 2 2 3 2 2 2 3 2 2 1 3 3
##  [149] 1 3 3 3 1 1 2 1 3 2 1 1 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 3 2 2 2 2 2 2 3 1 2
##  [186] 2 2 2 2 2 2 2 2 1 2 1 2 1 3 2 2 2 2 1 2 3 3 1 2 2 1 3 1 3 3 2 2 2 1 1 2 1
##  [223] 1 1 2 1 2 3 3 1 3 2 2 3 3 2 1 1 2 1 2 2 2 2 2 1 2 3 2 2 2 2 2 1 1 2 2 3 2
##  [260] 1 2 2 1 1 1 1 2 1 1 1 2 2 2 1 1 2 1 2 2 1 3 1 1 3 2 3 3 2 3 3 2 1 3 2 2 1
##  [297] 1 2 2 1 2 2 1 2 1 2 2 3 2 2 3 1 1 1 3 2 1 2 2 1 2 1 3 1 3 1 1 2 2 1 1 2 2
##  [334] 2 1 3 2 2 3 3 2 2 2 2 2 3 2 2 2 2 1 3 1 2 3 1 2 2 2 2 3 2 3 1 2 1 1 1 2 1
##  [371] 3 2 1 1 3 2 2 2 3 3 3 2 1 3 1 3 2 2 2 2 3 2 2 1 3 3 2 1 2 2 1 2 3 3 2 1 2
##  [408] 2 2 1 2 2 1 2 2 1 2 1 2 2 1 2 1 2 2 2 2 2 3 2 2 2 2 2 2 2 1 3 3 3 3 1 2 2
##  [445] 1 2 2 2 1 1 2 2 3 1 3 2 2 2 2 2 1 2 2 3 2 1 3 2 2 2 3 1 2 2 1 2 3 1 2 2 2
##  [482] 2 2 2 1 3 2 1 2 2 3 2 2 3 2 2 1 3 1 2 1 2 2 2 2 2 3 2 2 2 2 1 3 2 2 2 3 2
##  [519] 3 3 2 2 2 3 2 3 3 2 3 2 2 3 2 1 2 3 1 2 1 1 1 2 2 1 3 3 2 3 1 3 2 2 3 2 2
##  [556] 1 3 2 1 3 1 1 2 2 2 1 2 1 2 2 2 2 2 2 2 2 3 2 2 2 2 2 3 2 1 1 3 3 2 1 3 1
##  [593] 2 1 3 2 3 2 2 3 3 3 2 2 3 2 3 2 2 2 2 2 2 1 2 2 2 2 1 3 2 1 1 3 2 3 2 3 3
##  [630] 2 2 2 3 1 2 2 2 2 3 2 2 1 3 3 2 1 2 2 2 3 2 2 2 3 1 3 1 3 2 3 1 1 2 2 2 1
```

```
##  [667] 2 2 1 1 2 2 1 2 3 2 2 2 1 2 2 2 3 2 3 3 2 3 2 1 1 1 2 3 2 2 2 3 3 3 3 1 2
##  [704] 2 2 1 2 2 2 2 3 1 2 2 3 2 2 1 2 2 3 3 1 1 2 2 2 2 2 2 1 1 2 2 1 1 2 3 2 2
##  [741] 2 3 2 1 2 3 2 3 3 3 1 2 2 3 2 2 2 2 2 3 2 2 3 2 2 2 2 2 2 3 1 2 3 2 2 2 2
##  [778] 2 1 1 1 3 3 2 2 3 1 2 3 2 3 2 2 1 1 2 2 2 3 1 2 2 3 2 2 3 2 3 3 2 2 2 1 3
##  [815] 1 1 1 2 2 2 2 3 2 3 2 1 1 1 1 2 1 1 1 2 3 3 3 2 3 3 2 3 2 2 3 2 2 2 2 2 2
##  [852] 2 2 1 3 3 1 2 2 1 2 1 2 3 2 2 2 1 1 1 3 1 3 1 3 2 1 3 2 1 3 2 2 2 1 2 2 3
##  [889] 1 3 3 2 1 1 2 1 3 3 2 2 3 3 3 3 2 2 1 1 3 1 2 2 3 1 1 2 3 3 1 1 2 2 2 3 3
##  [926] 2 3 2 1 1 2 2 2 2 2 2 3 2 2 2 1 3 3 2 3 3 1 1 3 3 2 1 2 2 3 2 2 2 1 2 2 1
##  [963] 3 1 2 2 1 3 1 3 2 3 1 2 1 3 2 2 2 1 2 2 3 2 2 2 2 2 2 1 2 1 2 2 1 2 2 1 2
## [1000] 2 2 2 2 3 2 1 2 1 1 1 2 2 1 2 2 2 2 3 2 2 3 2 2 2 3 2 3 1 2 2 1 3 2 3 1 2
## [1037] 3 1 3 2 2 2 3 3 1 2 2 2 3 1 2 2 3 1 3 3 2 1 3 2 2 1 2 1 1 1 2 2 3 2 2 1 3
## [1074] 2 2 2 2 1 2 2 2 1 1 3 3 2 2 2 1 3 1 2 3 3 1 3 2 2 2 2 1 2 2 1 1 2 3 3 1 2
## [1111] 1 2 2 1 3 2 2 2 1 2 2 2 2 3 1 2 2 1 1 1 3 3 2 1 2 2 2 3 2 2 2 3 2 3 3 1 2
## [1148] 3 2 1 1 3 3 1 2 2 2 1 2 2 1 1 3 1 3 2 2 3 2 2 2 2 3 2 2 1 2 2 2 3 2 3 2 2
## [1185] 1 2 1 2 2 2 1 2 1 1 2 2 2 1 2 2 2 2 2 3 2 2 2 2 1 3 1 1 3 1 2 2 2 2 2 2 1
## [1222] 2 2 2 2 1 2 2 1 3 3 3 2 1 2 2 2 3 3 1 3 1 2 3 2 2 3 2 3 2 3 1 3 1 2 1 2 2
## [1259] 2 1 2 2 1 1 1 3 2 2 2 2 2 1 2 2 2 1 3 2 2 2 1 2 3 2 2 2 1 3 2 2 2 1 2 2 3
## [1296] 2 2 3 2 1 2 2 3 2 1 3 2 3 1 1 1 1 2 3 1 1 3 2 1 3 2 3 2 2 1 1 2 2 2 1 1 1
## [1333] 2 1 1 1 2 3 1 2 2 2 2 1 1 2 1 3 3 2 3 2 2 2 3 2 2 2 1 2 1 2 2 1 2 2 3 3 3
## [1370] 2 2 2 2 2 3 2 2 2 1 3 1 2 2 3 1 1 2 2 2 2 2 2 2 2 2 1 2 1 3 1 1 2 2 2 1 2
## [1407] 2 1 3 2 2 2 2 2 1 2 2 1 1 2 3 2 1 1 2 2 3 2 1 3 2 2 2 1 3 2 2 1 2 1 2 2 3
## [1444] 3 2 2 1 3 2 2 2 3 2 1 1 2 2 2 3 1 2 2 3 2 3 3 2 2 2 2 2 1 1 2 1 2 3 2 2 1
## [1481] 3 3 1 2 2 3 2 2 2 2 1 1 3 2 2 1 3 2 3 1 2 3 3 2 2 1 3 2 2 1 3 2 3 2 1 1 1
## [1518] 1 2 1 2 3 3 2 2 2 2 3 3 3 3 3 1 2 2 3 3 1 2 2 3 2 3 2 3 2 3 1 1 3 2 2 1 2
## [1555] 2 2 3 2 2 3 2 1 2 1 2 2 3 3 3 1 3 2 1 1 2 2 3 1 2 1 3 1 1 2 2 2 2 2 2 1 2
## [1592] 2 1 1 2 2 1 2 2 3 2 2 3 1 2 2 2 2 3 2 1 2 2 2 2 2 2 2 2 3 3 2 2 2 2 3 2 2
## [1629] 1 1 2 2 2 2 2 2 2 1 2 3 3 2 2 2 1 3 2 2 2 3 2 2 1 2 2 2 3 2 2 1 2 2 1 1 2
## [1666] 2 3 3 1 3 2 1 2 2 2 2 1 1 2 2 2 2 1 3 3 2 2 1 2 2 1 2 1 1 2 3 3 1 2 2 1 1
## [1703] 3 1 2 1 1 1 1 3 1 1 2 2 3 2 2 2 2 1 2 2 2 2 3 2 1 3 3 3 1 2 2 2 2 3 2 2 2
## [1740] 3 2 2 3 1 2 1 2 2 2 2 2 2 2 2 1 2 2 2 1 2 1 2 3 3 2 2 1 1 3 2 2 1 2 2 1 1
## [1777] 1 2 2 2 2 2 2 1 1 2 1 2 1 1 1 3 2 2 3 2 2 1 2 2 1 3 1 2 2 2 2 2 2 2 2 3 2
## [1814] 3 2 2 2 3 2 1 3 2 3 1 2 2 2 2 3 2 2 1 3 1 2 3 2 2 2 3 2 1 3 2 1 2 2 1 2 1
## [1851] 1 2 2 1 3 1 2 2 3 1 2 3 3 3 2 1 1 2 2 2 2 1 3 1 3 2 3 2 3 3 2 1 3 2 2 2 2
## [1888] 2 2 2 1 1 1 1 2 2 2 2 1 3 1 3 2 2 2 2 2 3 2 2 2 3 3 2 2 2 2 2 2 2 3 2 2 2
## [1925] 3 2 2 2 1 2 2 2 3 2 1 2 1 1 3 2 2 3 1 3 1 3 3 2 2 1 2 2 2 2 2 2 1 1 2 1 3 2
## [1962] 2 2 3 1 2 3 2 3 1 2 2 3 2 2 3 2 3 1 1 2 2 2 1 3 3 2 1 1 2 2 2 2 2 3 2 3 3
## [1999] 1 2 3 2 2 1 2 2 2 3 3 3 3 2 2 1 2 1 2 2 2 3 2 3 2 2 1 2 2 2 1 1 1 3 3 2 1
## [2036] 2 2 2 2 3 3 2 2 3 2 2 2 1 2 2 2 1 2 2 2 2 3 2 2 1 2 2 2 2 2 2 3 2 3 1 1 2
## [2073] 2 2 2 1 1 2 2 3 3 2 3 2 2 2 1 2 3 2 2 2 2 2 2 1 2 3 1 1 2 2 2 3 2 2 3 2 2
## [2110] 1 1 2 2 2 2 3 2 2 1 2 2 2 1 2 2 3 3 2 1 3 2 2 3 2 1 2 2 3 2 3 2 3 2 2 3 1
## [2147] 2 2 1 2 1 2 3 2 2 1 3 1 1 2 2 2 2 2 1 1 3 2 1 2 3 2 2 3 1 3 2 1 2 3 1 3 2
## [2184] 3 2 2 1 1 3 1 2 2 2 2 2 2 2 2 3 2 1 2 2 2 1 3 1 3 2 2 2 1 1 2 2 2 2 2 1 2
## [2221] 2 1 1 1 1 2 3 3 2 1 3 3 3 1 1 3 3 1 2 2 3 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 1
## [2258] 1 1 1 3 1 3 2 1 2 2 3 2 2 2 2 3 2 2 2 3 2 3 2 1 3 2 3 3 3 2 2 1 2 3 2 2 2
## [2295] 2 2 3 2 3 2 2 3 2 2 1 2 2 3 2 1 1 2 2 2 2 2 1 1 2 2 1 2 3 2 1 2 1 3 2 1 2 2
## [2332] 1 2 3 2 3 2 2 1 1 2 1 2 3 1 2 1 1 2 3 2 3 3 2 2 1 2 2 1 3 1 2 1 2 2 2 1 2
## [2369] 1 3 1 3 2 2 2 2 3 3 2 3 3 2 2 1 2 2 2 1 2 3 2 2 1 2 2 2 3 2 1 1 2 2 1 2 2
## [2406] 2 1 2 1 3 2 2 3 1 2 2 2 2 3 2 3 3 2 2 1 3 2 2 1 3 2 1 2 2 2 2 2 2 2 1 2 2
## [2443] 2 2 3 2 2 3 2 1 3 3 2 2 1 3 2 3 2 2 2 1 2 1 2 2 2 3 1 1 3 3 3 1 2 2 2 2 1
## [2480] 1 2 2 2 3 2 2 1 3 2 1 2 2 2 1 2 2 2 2 2 1 3 2 2 1 1 1 3 1 2 2 2 2 1 3 2 3
## [2517] 2 2 3 3 3 2 3 2 2 2 3 2 2 1 2 2 3 1 3 2 2 2 1 2 2 1 3 2 1 2 2 2 2 1 2 2 1
## [2554] 3 2 2 2 2 2 1 2 2 3 1 2 2 2 3 2 3 1 2 1 3 2 2 2 2 1 2 2 2 1 1 1 2 2 1 2 3
## [2591] 3 2 1 1 1 2 2 2 1 2 1 1 2 2 3 2 1 2 1 2 3 3 2 1 1 2 2 2 2 1 2 2 2 2 3 3 3
## [2628] 2 3 1 2 2 1 2 3 2 1 3 2 2 3 2 2 3 2 2 2 2 1 2 1 2 2 2 2 1 1 1 3 1 3 2 1 2
```

```
## [2665] 2 1 3 1 1 2 3 3 1 3 1 1 2 2 2 2 2 2 2 2 3 2 1 2 3 1 2 2 1 3 2 1 3 2 2 2 3
## [2702] 2 1 1 3 2 1 3 2 2 1 1 1 2 2 1 1 2 2 2 2 1 1 2 1 2 2 3 2 1 2 2 2 2 2 3 1 1
## [2739] 2 1 2 1 1 2 1 2 1 2 2 3 2 3 2 2 2 3 3 3 1 1 3 3 2 2 3 3 3 2 1 3 2 1 2 3 2
## [2776] 2 2 2 1 2 3 2 1 2 2 3 1 2 2 2 2 2 3 3 3 2 2 1 2 2 1 3 3 3 2 1 3 2 2 2 3 3 2
## [2813] 2 1 2 2 2 2 2 2 1 2 1 2 2 2 1 3 3 3 1 3 2 2 2 2 2 1 2 2 2 2 3 2 3 2 1 2 1
## [2850] 3 2 3 1 1 1 2 2 3 3 2 3 2 2 2 2 2 2 2 3 1 3 2 3 1 2 3 2 1 3 2 2 3 3 1 2 2
## [2887] 2 2 1 2 2 2 1 2 2 2 2 3 2 1 2 2 1 3 2 1 2 3 2 1 1 2 1 2 2 1 2 1 2 1 2 1 1
## [2924] 1 1 3 1 1 2 2 2 3 2 3 1 2 2 2 2 2 2 1 2 2 1 2 1 2 2 2 2 1 2 2 2 2 3 2 2 2 3 3
## [2961] 3 3 2 1 1 3 2 3 3 3 2 2 1 2 2 2 1 3 2 1 2 2 2 1 3 1 2 3 2 2 1 3 2 2 2 2 2
## [2998] 1 2 1 2 2 2 3 1 2 2 2 1 1 2 3 2 2 3 1 2 2 2 3 1 2 2 2 2 3 2 2 3 1 2 2 2 3
## [3035] 2 3 3 2 1 2 1 3 2 3 2 1 2 2 2 2 2 2 2 3 3 1 2 1 1 2 1 2 2 3 3 2 2 2 2 1 2 2
## [3072] 1 2 3 1 2 1 2 3 3 2 2 3 3 2 1 3 2 1 3 2 2 1 1 3 2 2 2 2 2 2 1 3 2 2 2 2 2
## [3109] 1 2 2 1 2 3 1 2 1 2 2 2 2 3 2 1 2 2 3 3 2 2 2 1 2 3 2 2 1 1 3 2 2 2 2 2 3
## [3146] 3 2 2 2 1 1 2 3 2 2 2 2 2 2 2 3 2 1 3 2 1 2 2 2 3 1 3 2 2 2 2 2 3 3 2 1 2
## [3183] 2 3 2 2 2 2 2 1 2 1 1 1 1 2 2 1 2 3 2 1 2 2 2 1 2 1 1 2 2 2 2 1 2 2 2 2 2
## [3220] 2 2 2 1 1 1 1 2 2 2 3 2 2 3 1 2 3 2 1 1 2 2 2 2 2 2 1 3 2 2 3 3 3 2 2 3 2 2
## [3257] 1 1 2 1 2 3 1 2 3 3 1 1 2 1 3 3 1 1 2 1 3 2 1 2 2 3 1 2 2 2 2 2 2 2 2 2 2
## [3294] 1 2 2 2 3 1 2 2 2 2 2 2 3 3 2 2 1 2 3 2 2 2 3 1 2 1 2 2 2 3 3 1 2 2 2 2 2
## [3331] 1 1 2 3 2 3 2 3 1 2 1 3 1 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 2 1 3 2 2 3 3 2 2
## [3368] 3 1 2 3 2 2 3 1 2 1 2 1 2 2 2 2 2 3 2 2 3 2 2 2 2 2 2 2 2 2 2 1 1 3 1 1 2
## [3405] 1 1 1 1 1 3 3 1 2 2 1 2 3 1 3 2 3 3 1 3 2 3 3 2 2 2 3 2 3 3 3 1 3 2 1 1 2
## [3442] 1 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 3 3 2 2 2 2 2 2 2 2 2 2 1 2 2 2 1 1 2 2
## [3479] 3 1 3 3 2 2 2 2 2 2 1 1 1 3 2 3 1 2 2 2 2 2 1 2 1 3 2 1 2 2 2 1 3 2 3 1 1 2 2
## [3516] 3 1 2 1 2 2 2 2 2 2 2 2 2 3 1 3 3 1 2 3 1 1 3 2 2 1 2 2 3 2 2 2 3 3 2 3 2 2
## [3553] 2 2 3 2 3 3 1 2 1 2 2 2 1 2 2 2 2 1 3 3 2 2 1 1 3 1 3 2 2 3 2 3 2 1 1 3 2 2
## [3590] 2 2 2 2 2 2 1 3 2 1 2 3 3 2 2 2 1 3 3 2 3 2 2 2 2 2 2 2 1 2 1 1 2 2 2 2 3 2 2
## [3627] 2 1 2 1 1 2 3 2 3 3 1 3 3 2 2 2 3 1 2 3 2 2 2 1 2 2 2 1 2 3 2 2 3 1 2 2 1
## [3664] 2 3 3 2 1 1 2 2 3 2 1 1 2 2 1 2 2 2 2 1 1 2 2 2 2 2 2 3 2 1 3 3 2 2 3 1 1 3 3
## [3701] 3 1 1 1 1 2 3 2 3 2 3 2 2 2 2 2 2 1 2 1 3 3 1 2 2 3 3 2 1 2 3 2 2 2 2 2 2 2
## [3738] 2 3 3 2 2 3 2 2 3 1 2 2 2 3 2 2 2 3 1 2 2 3 2 1 2 1 2 3 2 1 1 2 3 3 2 2 1
## [3775] 1 2 1 1 3 2 2 2 2 2 2 3 2 2 2 2 2 1 2 3 3 2 1 3 1 2 1 2 2 1 1 2 1 2 3 2 2
## [3812] 1 2 2 2 2 2 1 2 1 1 2 2 2 2 2 1 2 2 2 2 2 2 2 1 1 3 3 2 1 3 2 1 2 3 1 3 2
## [3849] 2 3 2 2 2 3 2 1 2 1 2 2 1 3 1 2 3 2 2 3 2 1 2 2 3 2 2 2 1 2 3 2 1 2 2 3 2
## [3886] 3 1 2 3 2 3 3 2 1 2 1 3 3 1 2 1 3 2 2 2 2 3 2 2 1 3 2 2 1 2 1 1 2 2 2 1 1
## [3923] 2 2 3 2 2 2 3 2 2 2 1 2 2 2 3 2 1 2 3 3 3 3 2 2 2 2 1 3 2 2 3 2 2 2 3 2 2
## [3960] 1 2 1 1 2 2 1 3 1 2 3 2 3 1 3 2 2 3 2 1 3 1 3 2 1 1 3 2 2 2 2 2 2 2 2 2 1
## [3997] 2 2 2 2 3 3 2 1 2 2 3 3 2 2 3 2 2 2 2 2 2 2 2 3 1 3 1 1 1 1 3 2 2 1 2 3 1
## [4034] 3 1 1 1 2 2 2 3 1 3 2 1 2 1 2 1 2 3 1 3 2 3 1 2 2 2 2 2 3 3 2 1 2 3 2 3 3
## [4071] 2 2 2 1 2 3 2 3 1 2 2 2 2 2 2 2 1 1 2 3 2 2 2 2 1 2 3 2 3 2 1 2 2 2 2 2 1
## [4108] 2 2 2 2 3 2 1 3 2 2 3 3 2 3 2 2 3 1 2 3 2 1 2 3 1 2 2 2 3 1 2 2 2 3 1 1 2
## [4145] 2 2 1 1 1 2 2 3 2 1 1 1 2 2 2 1 3 3 1 3 3 3 3 1 2 1 3 2 1 1 1 3 3 3 3 1 1
## [4182] 2 2 1 2 3 2 1 3 2 2 3 1 3 3 1 2 2 2 3 2 1 1 2 1 1 2 2 2 1 3 1 3 1 3 1 2 1
## [4219] 2 3 2 2 2 3 2 2 2 3 2 1 3 2 1 3 2 3 2 1 2 3 2 1 3 1 2 1 1 1 2 2 1 3 2 2 2
## [4256] 2 2 3 3 2 2 3 3 2 2 1 2 2 2 1 2 2 2 2 1 2 2 2 2 1 2 1 2 1 2 2 1 2 2 2 1 2
## [4293] 2 1 3 1 2 2 2 3 2 1 1 1 3 2 2 3 2 3 2 3 2 1 3 3 3 1 2 3 2 2 2 2 1 2 2 3 2
## [4330] 2 2 2 2 2 1 2 1 3 2 1 2 1 2 2 2 3 2 2 3 2 1 3 2 2 2 3 1 3 2 2 2 1 2 2 2 2
## [4367] 3 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 1 2 1 2 2 2 2 1 1 2 3 2 2 2 2
## [4404] 3 3 1 2 2 2 2 2 2 3 2 2 2 1 3 1 1 2 1 1 3 2 2 1 1 1 3 2 2 2 2 2 3 3 2 3 1
## [4441] 3 3 3 3 2 1 2 2 1 2 3 2 3 1 2 3 2 1 3 2 2 2 3 1 2 2 1 1 2
```

```
clusterBD<-cbind(clusters,bm_muestra)

#--join para agregar la columna control
```

```r
SubsetBD_Transf<- DB_Transform %>%
  dplyr::select(GrupoControl,corr)


dplyr::glimpse(clusterBD)
```

```
## Rows: 4,469
## Columns: 11
## $ clusters          <int> 1, 2, 2, 1, 3, 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2...
## $ TiempoEspera      <dbl> 33757, 8812651, 50053, 0, 672, 61754, 583, 2549, ...
## $ TiempoSello       <dbl> 4, 5, 5, 2, 4, 2, 2, 8, 2, 3, 3, 13, 30, 2, 3, 4,...
## $ NumeroInvitados   <dbl> 1, 4, 3, 2, 1, 2, 2, 1, 3, 2, 2, 2, 2, 3, 2, 2, 2...
## $ NumeroMensajes    <dbl> 12, 12, 9, 5, 7, 7, 17, 11, 14, 13, 4, 12, 4, 9, ...
## $ NumeroCaracteres  <dbl> 0, 198, 252, 280, 5, 284, 478, 144, 156, 326, 270...
## $ Canal             <dbl> 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2...
## $ Habitacion        <dbl> 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ TotalRevisiones   <dbl> 208, 10, 36, 67, 109, 76, 63, 16, 44, 94, 66, 17,...
## $ CapacidadPersonas <dbl> 3, 4, 3, 2, 2, 2, 4, 4, 2, 4, 2, 2, 4, 3, 2, 3, 5...
## $ corr              <dbl> 9491, 21648, 2665, 5460, 4823, 15965, 10949, 1817...
```

```r
dplyr::glimpse(SubsetBD_Transf)
```

```
## Rows: 25,522
## Columns: 2
## $ GrupoControl <dbl> 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 2, 2, 2, 2, ...
## $ corr         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,...
```

```r
clusterBD<-left_join(clusterBD,SubsetBD_Transf,by=c("corr"="corr"))


dplyr::glimpse(clusterBD)
```

```
## Rows: 4,469
## Columns: 12
## $ clusters          <int> 1, 2, 2, 1, 3, 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2...
## $ TiempoEspera      <dbl> 33757, 8812651, 50053, 0, 672, 61754, 583, 2549, ...
## $ TiempoSello       <dbl> 4, 5, 5, 2, 4, 2, 2, 8, 2, 3, 3, 13, 30, 2, 3, 4,...
## $ NumeroInvitados   <dbl> 1, 4, 3, 2, 1, 2, 2, 1, 3, 2, 2, 2, 2, 3, 2, 2, 2...
## $ NumeroMensajes    <dbl> 12, 12, 9, 5, 7, 7, 17, 11, 14, 13, 4, 12, 4, 9, ...
## $ NumeroCaracteres  <dbl> 0, 198, 252, 280, 5, 284, 478, 144, 156, 326, 270...
## $ Canal             <dbl> 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2...
## $ Habitacion        <dbl> 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ TotalRevisiones   <dbl> 208, 10, 36, 67, 109, 76, 63, 16, 44, 94, 66, 17,...
## $ CapacidadPersonas <dbl> 3, 4, 3, 2, 2, 2, 4, 4, 2, 4, 2, 2, 4, 3, 2, 3, 5...
## $ corr              <dbl> 9491, 21648, 2665, 5460, 4823, 15965, 10949, 1817...
## $ GrupoControl      <dbl> 1, 1, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1...
```

```r
frq(clusterBD$clusters)
```

```
##
## x <integer>
## # total N=4469  valid N=4469  mean=1.97  sd=0.67
##
## Value |    N | Raw % | Valid % | Cum. %
## --------------------------------------
##     1 | 1070 | 23.94 |   23.94 |  23.94
```

```
##     2 | 2444 | 54.69 |   54.69 |  78.63
##     3 |  955 | 21.37 |   21.37 | 100.00
##  <NA> |    0 |  0.00 |    <NA> |   <NA>
```

Analítica del modelo de Clúster Jerárquico construido

En esta sección se analizará los cruces de variables de la base **clusterBD** para tener elementos técnicos de respuesta a los numerales 1 y 2.

Respuesta al numeral 1

Con la interpretación muy general de los análisis de los resultados, que se obtuvieron con el modelo, podemos mencionar que la empresa Be-A-Host, tiene 3 nichos de mercado que puede atender y especializarse para cada uno de ellos, el primero representado por el clúster 1 encaminado a familias o grupos numerosos que prefieren casas completas y que muy probablemente pasen sus vacaciones, la interacción es un poco considerable, pero los tiempos de respuesta son los más altos en comparación al resto de clúster, dicho nicho representa el 60% de la cuota de mercado y por demandar mucho más espacio en número de personas pueden ser los de mayor capacidad adquisitiva; para el segundo nicho de mercado el clúster 2 una ponderación del 21%, prefieren el canal de reserva no inmediata similar al 1, se puede intuir que es para aquellas personas que viajan por trabajo o estudios, su demanda de personas es mucho menor al resto y prefieren predominantemente habitaciones privadas y pocas veces compartidas, su interacción es baja, los tiempos de respuesta son medios; el nicho o grupo de clúster 3, representa el 18.73% del mercado de Be-A-Host y es el que prefiere el canal de forma de reserva instantánea, sin embargo, este grupo es familiar, se podría decir o afirmar que demanda menos personas que el clúster 1, se componen muy probablemente por familias promedio, se reveló que prefieren bastantes interacción en las revisiones de los detalles con los huéspedes, sin embargo, lo descubierto que es muy interesante es que es el clúster con el mejor y excelente tiempo de respuesta entre ellos y el huésped, eso hace la diferencia para aplicar a un canal específico de reserva, por eso en este grupo predomina el instant_book.

En cuanto a que se puede recomendar de mejora al Website del sitio es que los tiempos de respuesta entre huéspedes y usuarios es clave para lograr reservas instantáneas, y que debe de establecerse una diferenciación de los nichos de mercados tal como lo revelan los clústeres.

Probabilidad de las reservas Se evidencia que los usuarios del clúster 3 tienen una alta probabilidad del 98% de reservar de forma instantánea **« instant_book »** pero solo representan el 18.73% del mercado de Be-A-Host, otro insight de los datos que se observa es que los usuarios del clúster 1 que representa el 60.21% del mercado tienen una probabilidad del 97% de casi ocupar el canal de alojamiento de reserva **« book_it »** y el clúster 2 tiende en un 100% a reservar en la modalidad de reserva book_it, también y representa el 21% del mercado para Be-A-Host.

```
cT<-clusterBD %>%
  dplyr::group_by(clusters,Canal) %>%
  dplyr::summarise(Reservas=n()) %>%
  dplyr::mutate(TotalReservas=sum(Reservas),ProbabilidadReserva=(Reservas/TotalReservas))
cT
```

```
## # A tibble: 5 x 5
## # Groups:   clusters [3]
##   clusters Canal Reservas TotalReservas ProbabilidadReserva
##      <int> <dbl>    <int>         <int>               <dbl>
## 1        1     1      817          1070              0.764
## 2        1     2      253          1070              0.236
## 3        2     1       29          2444              0.0119
## 4        2     2     2415          2444              0.988
## 5        3     2      955           955              1
```

Perfil general de los grupos de clústers

Clúster 1 El perfil de los usuarios que pertenecen al clúster 1 de forma general se resume en: canal preferido es el book_it, presentan una media en los tiempos de espera de 5,2171 segundos < menos de un día > y un máximo de 31,534,671 < un año > , en los tiempos de sello una media 4.93 días hasta un máximo de 145 días de espera, están por una media de 2.62 invitados y un máximo de 14, en el tipo de habitación prefieren las casas completas, una media de interacción de 9.14 mensajes y un máximo de 69, una media de 59 revisiones, prefieren una media de oferta de capacidad de personas por 4.29 y un máximo de 16, y un máximo de 2 reservas por gestión.

```
c1<-clusterBD %>%
  dplyr::group_by(clusters,Canal,TiempoEspera,TiempoSello,NumeroInvitados,
                  Habitacion,NumeroMensajes,TotalRevisiones,CapacidadPersonas)%>%
  dplyr::filter(clusters==1)%>%
  dplyr::summarise(Reservas=n()) %>%
  dplyr::mutate(TotalReservas=sum(Reservas),ProbabilidadReserva=(Reservas/TotalReservas))
c1
```

```
## # A tibble: 1,023 x 12
## # Groups:   clusters, Canal, TiempoEspera, TiempoSello, NumeroInvitados,
## #   Habitacion, NumeroMensajes, TotalRevisiones [1,018]
##    clusters Canal TiempoEspera TiempoSello NumeroInvitados Habitacion
##       <int> <dbl>        <dbl>       <dbl>           <dbl>      <dbl>
##  1        1     1            0           1               1          1
##  2        1     1            0           1               1          1
##  3        1     1            0           1               1          1
##  4        1     1            0           1               1          1
##  5        1     1            0           1               1          1
##  6        1     1            0           1               1          1
##  7        1     1            0           1               1          1
##  8        1     1            0           1               1          1
##  9        1     1            0           1               1          1
## 10        1     1            0           1               1          1
## # ... with 1,013 more rows, and 6 more variables: NumeroMensajes <dbl>,
## #   TotalRevisiones <dbl>, CapacidadPersonas <dbl>, Reservas <int>,
## #   TotalReservas <int>, ProbabilidadReserva <dbl>
```

```
summary(c1)
```

```
##     clusters     Canal        TiempoEspera       TiempoSello
##  Min.   :1   Min.   :1.000   Min.   :     0.0   Min.   : 1.000
##  1st Qu.:1   1st Qu.:1.000   1st Qu.:     0.0   1st Qu.: 2.000
##  Median :1   Median :1.000   Median :     0.0   Median : 3.000
##  Mean   :1   Mean   :1.238   Mean   :  2672.2   Mean   : 3.128
##  3rd Qu.:1   3rd Qu.:1.000   3rd Qu.:   138.5   3rd Qu.: 4.000
##  Max.   :1   Max.   :2.000   Max.   :188541.0   Max.   :10.000
##  NumeroInvitados   Habitacion  NumeroMensajes   TotalRevisiones
##  Min.   :1.000   Min.   :1.0   Min.   : 2.000   Min.   :  0.0
##  1st Qu.:2.000   1st Qu.:1.0   1st Qu.: 4.000   1st Qu.: 48.0
##  Median :2.000   Median :1.0   Median : 7.000   Median : 92.0
##  Mean   :2.175   Mean   :1.1   Mean   : 7.979   Mean   :100.7
##  3rd Qu.:2.000   3rd Qu.:1.0   3rd Qu.:10.000   3rd Qu.:149.0
##  Max.   :7.000   Max.   :2.0   Max.   :37.000   Max.   :321.0
##  CapacidadPersonas    Reservas     TotalReservas   ProbabilidadReserva
##  Min.   : 1.000    Min.   :1.000   Min.   :1.000   Min.   :0.5000
```

```
##  1st Qu.: 2.000    1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.0000
##  Median : 3.000    Median :1.000   Median :1.000   Median :1.0000
##  Mean   : 3.505    Mean   :1.046   Mean   :1.056   Mean   :0.9951
##  3rd Qu.: 4.000    3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :11.000    Max.   :3.000   Max.   :3.000   Max.   :1.0000
```

Clúster 2 El perfil de los usuarios que pertenecen al clúster 2 de forma general se resume en: canal preferido es el book_it, presentan una media en los tiempos de espera de 25,292 segundos < menos de un día > y un máximo de 71,2485 < casí un día > , en los tiempos de sello una media 3.326 días hasta un máximo de 17 días de espera, están por una media de 1.562 invitados y un máximo de 5, en el tipo de habitación prefieren las habitaciones reservadas y muy poco las compartidas, una media de interacción de 7.753 mensajes y un máximo de 17.000, una media de 51.47 revisiones, prefieren una media de oferta de capacidad de personas por 2.36 y un máximo de 9.00, y un máximo de 2 reservas por gestión.

```
c2<-clusterBD %>%
  dplyr::group_by(clusters,Canal,TiempoEspera,TiempoSello,NumeroInvitados,
                  Habitacion,NumeroMensajes,TotalRevisiones,CapacidadPersonas)%>%
  dplyr::filter(clusters==2)%>%
  dplyr::summarise(Reservas=n()) %>%
  dplyr::mutate(TotalReservas=sum(Reservas),ProbabilidadReserva=(Reservas/TotalReservas))
c2
```

```
## # A tibble: 2,378 x 12
## # Groups:   clusters, Canal, TiempoEspera, TiempoSello, NumeroInvitados,
## #   Habitacion, NumeroMensajes, TotalRevisiones [2,378]
##    clusters Canal TiempoEspera TiempoSello NumeroInvitados Habitacion
##       <int> <dbl>        <dbl>       <dbl>            <dbl>      <dbl>
##  1        2     1            0           1                1          2
##  2        2     1            0           2                1          2
##  3        2     1            0           2                1          2
##  4        2     1            0           2                2          2
##  5        2     1            0           2                4          2
##  6        2     1            0           2                7          2
##  7        2     1            0           3                1          2
##  8        2     1            0           3                1          2
##  9        2     1            0           3                1          3
## 10        2     1            0           3               10          1
## # ... with 2,368 more rows, and 6 more variables: NumeroMensajes <dbl>,
## #   TotalRevisiones <dbl>, CapacidadPersonas <dbl>, Reservas <int>,
## #   TotalReservas <int>, ProbabilidadReserva <dbl>
```

```
summary(c2)
```

```
##     clusters     Canal        TiempoEspera       TiempoSello
##  Min.   :2   Min.   :1.000   Min.   :       0   Min.   :  1.000
##  1st Qu.:2   1st Qu.:2.000   1st Qu.:     958   1st Qu.:  2.000
##  Median :2   Median :2.000   Median :    5132   Median :  3.000
##  Mean   :2   Mean   :1.989   Mean   :   41483   Mean   :  5.202
##  3rd Qu.:2   3rd Qu.:2.000   3rd Qu.:   24334   3rd Qu.:  5.000
##  Max.   :2   Max.   :2.000   Max.   :12788442   Max.   :160.000
##  NumeroInvitados    Habitacion    NumeroMensajes   TotalRevisiones
##  Min.   : 1.000   Min.   :1.000   Min.   : 2.000   Min.   :  0.0
##  1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 5.000   1st Qu.: 18.0
##  Median : 2.000   Median :1.000   Median : 8.000   Median : 46.0
##  Mean   : 2.677   Mean   :1.056   Mean   : 9.431   Mean   : 50.5
```

```
##  3rd Qu.: 3.000   3rd Qu.:1.000   3rd Qu.:12.000   3rd Qu.: 77.0
##  Max.   :12.000   Max.   :3.000   Max.   :48.000   Max.   :165.0
##  CapacidadPersonas    Reservas     TotalReservas   ProbabilidadReserva
##  Min.   : 1.000    Min.   :1.000   Min.   :1.000   Min.   :1
##  1st Qu.: 3.000    1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1
##  Median : 4.000    Median :1.000   Median :1.000   Median :1
##  Mean   : 4.363    Mean   :1.028   Mean   :1.028   Mean   :1
##  3rd Qu.: 5.000    3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1
##  Max.   :16.000    Max.   :3.000   Max.   :3.000   Max.   :1
```

Clúster 3 El perfil de los usuarios que pertenecen al clúster 3 de forma general se resume en: canal preferido es el instant book , presentan una media en los tiempos de espera de 2,392 segundos < minutos > y un máximo de 1,484,214 < un día >, en los tiempos de sello una media 3.083 días hasta un máximo de 7.00 días de espera, están por una media de 2.153 invitados y un máximo de 6.000, en el tipo de habitación prefieren es la casa completa, una media de interacción de 7.698 mensajes y un máximo de 28.000, una media de 87.12 revisiones, prefieren una media de oferta de capacidad de personas por 3.484 y un máximo de 11.000, y un máximo de 3 reservas por gestión.

```
c3<-clusterBD %>%
  dplyr::group_by(clusters,Canal,TiempoEspera,TiempoSello,NumeroInvitados,
                  Habitacion,NumeroMensajes,TotalRevisiones,CapacidadPersonas)%>%
  dplyr::filter(clusters==3)%>%
  dplyr::summarise(Reservas=n()) %>%
  dplyr::mutate(TotalReservas=sum(Reservas),ProbabilidadReserva=(Reservas/TotalReservas))
c3
```

```
## # A tibble: 934 x 12
## # Groups:   clusters, Canal, TiempoEspera, TiempoSello, NumeroInvitados,
## #   Habitacion, NumeroMensajes, TotalRevisiones [934]
##    clusters Canal TiempoEspera TiempoSello NumeroInvitados Habitacion
##       <int> <dbl>        <dbl>       <dbl>           <dbl>      <dbl>
##  1        3     2            0           1               1          2
##  2        3     2            0           1               2          2
##  3        3     2            0           1               2          2
##  4        3     2            0          11               2          2
##  5        3     2            0          14               1          2
##  6        3     2            1           1               1          2
##  7        3     2            1           2               1          2
##  8        3     2            1           2               1          2
##  9        3     2            2          27               1          2
## 10        3     2           61          11               1          2
## # ... with 924 more rows, and 6 more variables: NumeroMensajes <dbl>,
## #   TotalRevisiones <dbl>, CapacidadPersonas <dbl>, Reservas <int>,
## #   TotalReservas <int>, ProbabilidadReserva <dbl>
```

```
summary(c3)
```

```
##     clusters     Canal    TiempoEspera     TiempoSello     NumeroInvitados
##  Min.   :3   Min.   :2   Min.   :     0   Min.   : 1.000   Min.   :1.000
##  1st Qu.:3   1st Qu.:2   1st Qu.:  1547   1st Qu.: 2.000   1st Qu.:1.000
##  Median :3   Median :2   Median :  7928   Median : 3.000   Median :1.000
##  Mean   :3   Mean   :2   Mean   : 24444   Mean   : 3.547   Mean   :1.529
##  3rd Qu.:3   3rd Qu.:2   3rd Qu.: 31357   3rd Qu.: 4.000   3rd Qu.:2.000
##  Max.   :3   Max.   :2   Max.   :786756   Max.   :42.000   Max.   :5.000
##    Habitacion    NumeroMensajes   TotalRevisiones  CapacidadPersonas
```

```
##  Min.   :2.000   Min.   : 2.000   Min.   :  0.00   Min.   :1.000
##  1st Qu.:2.000   1st Qu.: 5.000   1st Qu.: 20.25   1st Qu.:2.000
##  Median :2.000   Median : 7.000   Median : 49.00   Median :2.000
##  Mean   :2.046   Mean   : 7.319   Mean   : 53.42   Mean   :2.279
##  3rd Qu.:2.000   3rd Qu.: 9.000   3rd Qu.: 78.00   3rd Qu.:2.000
##  Max.   :3.000   Max.   :23.000   Max.   :143.00   Max.   :8.000
##     Reservas     TotalReservas   ProbabilidadReserva
##  Min.   :1.000   Min.   :1.000   Min.   :1
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1
##  Median :1.000   Median :1.000   Median :1
##  Mean   :1.022   Mean   :1.022   Mean   :1
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1
##  Max.   :2.000   Max.   :2.000   Max.   :1
```

Respuesta al numeral 2

Grupo de control Se observa que el experimento del grupo de control, < los que fueron sometidos a 140 caracteres > , y que en la base son los que aparecen con el código 1, no muestra cambios significativos, al contrario la probabilidad de reserva se ve inferior en comparación a los que no están en control del experimento, la recomendación es no lanzar el cambio a toda la plataforma.

```
Control<-clusterBD %>%
  dplyr::group_by(clusters,Canal,GrupoControl) %>%
  dplyr::summarise(Reservas=n()) %>%
  dplyr::mutate(TotalReservas=sum(Reservas),ProbabilidadReserva=(Reservas/TotalReservas))
Control
```

```
## # A tibble: 10 x 6
## # Groups:   clusters, Canal [5]
##    clusters Canal GrupoControl Reservas TotalReservas ProbabilidadReserva
##       <int> <dbl>        <dbl>    <int>         <int>                <dbl>
##  1        1     1            1      431           817                0.528
##  2        1     1            2      386           817                0.472
##  3        1     2            1      122           253                0.482
##  4        1     2            2      131           253                0.518
##  5        2     1            1       17            29                0.586
##  6        2     1            2       12            29                0.414
##  7        2     2            1     1235          2415                0.511
##  8        2     2            2     1180          2415                0.489
##  9        3     2            1      490           955                0.513
## 10        3     2            2      465           955                0.487
```