MAESTRÍA EN ESTADÍSTICA APLICADA/ MANEJO DE DATOS

Unidad 3: Recopilación, integración y manipulación y almacenamiento de datos

3.2. Selección y recuperación de datos (Queries SQL) de una sola tabla, por medio de comandos del software estadístico

En la sección 3.1 se vio como establecer una conexión con el DBMS de MySQL, para la manipulación y procesamiento de los datos mediante SQL, utilizaremos nuevamente la conexión por usuario DSN.

Sintaxis de conexión:

```
#install.packages("RODBC")                                          #inst
alación de paquete RODBC
library(RODBC)                                                      #refe
rencia a la librería RODBC
con<-odbcConnect("MySQLconexion1", uid = "root", pwd="Matias2710")  #Esta
blecer la conexión con el DBMS
```

En esta sección se revisará la función sqldf( ) única función del paquete {sqldf}, mediante la cual es posible utilizar la sintaxis convencional de SQL en el entorno de R. Para el uso de la función sqldf, vamos a requerir previamente que los datos se almacenan en un Dataframe, se trabajará con los registros de la tabla country de la base de datos "world" de MySQL.

```
df_country<-sqlQuery(con,"Select * from country")                   #
Ejecutar una sentencia SQL
```

Conozcamos el tipo de objeto y la estructura de df_country:

```
class(df_country)    #class nos permite conocer el tipo de objeto

## [1] "data.frame"

str(df_country)      #str nos indica la estructura del objeto

## 'data.frame':    239 obs. of  15 variables:
##  $ Code          : Factor w/ 239 levels "ABW","AFG","AGO",..: 1 2 3 4
5 6 7 8 9 10 ...
##  $ Name          : Factor w/ 239 levels "Afghanistan",..: 12 1 6 7 2 5
150 223 10 11 ...
##  $ Continent     : Factor w/ 7 levels "Africa","Antarctica",..: 5 3 1
5 4 4 5 3 7 3 ...
##  $ Region        : Factor w/ 25 levels "Antarctica","Australia and New
Zealand",..: 5 22 6 5 23 23 5 14 19 14 ...
##  $ SurfaceArea   : num  193 652090 1246700 96 28748 ...
##  $ IndepYear     : int  NA 1919 1975 NA 1912 1278 NA 1971 1816 1991 ..
.
```

```
##  $ Population    : int  103000 22720000 12878000 8000 3401200 78000 21
7000 2441000 37032000 3520000 ...
##  $ LifeExpectancy: num  78.4 45.9 38.3 76.1 71.6 ...
##  $ GNP           : num  828 5976 6648 63.2 3205 ...
##  $ GNPOld        : num  793 NA 7984 NA 2500 ...
##  $ LocalName     : Factor w/ 239 levels "Â–","Ã–sterreich",..: 22 6 18
19 188 17 143 10 21 90 ...
##  $ GovernmentForm: Factor w/ 35 levels "Administrated by the UN",..: 2
2 15 29 9 29 26 22 11 12 29 ...
##  $ HeadOfState   : Factor w/ 179 levels "","Ã“lafur Ragnar GrÃmsson",.
.: 31 127 96 49 147 1 31 178 53 150 ...
##  $ Capital       : int  129 1 56 62 34 55 33 65 69 126 ...
##  $ Code2         : Factor w/ 238 levels "AD","AE","AF",..: 15 3 9 5 6
1 8 2 11 7 ...
```

Para utilizar la función sqldf() se requiere previamente la instalación del paquete {sqldf}:

```
#install.packages("sqldf")
library(sqldf)
```

Order by

En SQL podemos utilizar la instrucción Order by para ordenar las observaciones de un dataset, para lo cual se requiere establecer el campo y la orientación del ordenamiento (ascendente o descendente). De esta forma, pongamos un ejemplo, ordenar los datos de países por el campo población de forma descendente. El resultado se almacena en un nuevo Dataframe.

```
df_res1<-sqldf("select * from df_country order by population desc", conne
ction=NULL)
head(df_res1,8)
```

```
##   Code               Name     Continent                    Region Surf
aceArea
## 1  CHN              China          Asia              Eastern Asia
9572900
## 2  IND              India          Asia Southern and Central Asia
3287263
## 3  USA      United States North America             North America
9363520
## 4  IDN          Indonesia          Asia            Southeast Asia
1904569
## 5  BRA             Brazil South America             South America
8547403
## 6  PAK           Pakistan          Asia Southern and Central Asia
796095
## 7  RUS Russian Federation        Europe            Eastern Europe    1
7075400
```

```
## 8  BGD         Bangladesh          Asia Southern and Central Asia
143998
##   IndepYear Population LifeExpectancy     GNP  GNPOld     LocalName
## 1     -1523 1277558000           71.4  982268  917719      Zhongquo
## 2      1947 1013662000           62.5  447114  430572  Bharat/India
## 3      1776  278357000           77.1 8510700 8110900 United States
## 4      1945  212107000           68.0   84982  215002     Indonesia
## 5      1822  170115000           62.9  776739  804108        Brasil
## 6      1947  156483000           61.1   61289   58549      Pakistan
## 7      1991  146934000           67.2  276608  442989       Rossija
## 8      1971  129155000           60.2   32852   31966    Bangladesh
##     GovernmentForm              HeadOfState Capital Code2
## 1 People'sRepublic              Jiang Zemin    1891    CN
## 2 Federal Republic  Kocheril Raman Narayanan    1109    IN
## 3 Federal Republic            George W. Bush    3813    US
## 4         Republic         Abdurrahman Wahid     939    ID
## 5 Federal Republic Fernando Henrique Cardoso     211    BR
## 6         Republic      Mohammad Rafiq Tarar    2831    PK
## 7 Federal Republic            Vladimir Putin    3580    RU
## 8         Republic         Shahabuddin Ahmad     150    BD
```

Distinct

La instrucción Distinct en SQL nos devuelve las observaciones, eliminando las posibles repeticiones. Veamos en el siguiente ejemplo su uso. Ejemplo: Determine los continentes y las diferentes regiones en donde se posee registros de países, indique la cantidad total del resultado.

```
df_res2<-sqldf("select Distinct Continent, Region from df_country", conne
ction=NULL)
df_res2
```

```
##        Continent                  Region
## 1  North America               Caribbean
## 2           Asia Southern and Central Asia
## 3         Africa           Central Africa
## 4         Europe          Southern Europe
## 5           Asia              Middle East
## 6  South America            South America
## 7        Oceania                Polynesia
## 8     Antarctica               Antarctica
## 9        Oceania Australia and New Zealand
## 10        Europe           Western Europe
## 11        Africa           Eastern Africa
## 12        Africa           Western Africa
## 13        Europe           Eastern Europe
## 14 North America          Central America
## 15 North America            North America
```

```
## 16          Asia            Southeast Asia
## 17        Africa           Southern Africa
## 18          Asia              Eastern Asia
## 19        Europe          Nordic Countries
## 20        Africa           Northern Africa
## 21        Europe          Baltic Countries
## 22       Oceania                 Melanesia
## 23       Oceania               Micronesia
## 24        Europe           British Islands
## 25       Oceania      Micronesia/Caribbean
```

```
print("La cantidad total es")
```

```
## [1] "La cantidad total es"
```

```
nrow(df_res2)
```

```
## [1] 25
```

***Observación**: Si no se hubiese aplicado la instrucción Distinc, el resultado de la consulta presentaría registros repetidos, dado que existen en la tabla, países con ubicación geográfica en la misma región del continente.*

Where

La instrucción WHERE nos permite establecer algún criterio de selección a la hora de extraer información del dataframe. Es una de las principales instrucciones del DML para la extracción y procesamiento de datos. Ejemplo1: Determinar los datos de aquellos países que poseen más de 100 millones de habitantes. Presente el listado de forma descendente.

```
df_res3<-sqldf("select * from df_country where population>100000000 order
by population desc", connection=NULL)
df_res3
```

```
##    Code               Name     Continent                    Region Sur
faceArea
## 1   CHN              China          Asia              Eastern Asia
9572900
## 2   IND              India          Asia Southern and Central Asia
3287263
## 3   USA      United States North America             North America
9363520
## 4   IDN          Indonesia          Asia            Southeast Asia
1904569
## 5   BRA             Brazil South America             South America
8547403
## 6   PAK           Pakistan          Asia Southern and Central Asia
796095
```

```
## 7   RUS Russian Federation        Europe            Eastern Europe
17075400
## 8   BGD         Bangladesh          Asia Southern and Central Asia
143998
## 9   JPN              Japan          Asia              Eastern Asia
377829
## 10  NGA            Nigeria        Africa            Western Africa
923768
##    IndepYear Population LifeExpectancy     GNP  GNPOld     LocalName
## 1      -1523 1277558000           71.4  982268  917719      Zhongquo
## 2       1947 1013662000           62.5  447114  430572  Bharat/India
## 3       1776  278357000           77.1 8510700 8110900 United States
## 4       1945  212107000           68.0   84982  215002     Indonesia
## 5       1822  170115000           62.9  776739  804108        Brasil
## 6       1947  156483000           61.1   61289   58549      Pakistan
## 7       1991  146934000           67.2  276608  442989       Rossija
## 8       1971  129155000           60.2   32852   31966    Bangladesh
## 9       -660  126714000           80.7 3787042 4192638  Nihon/Nippon
## 10      1960  111506000           51.6   65707   58623       Nigeria
##            GovernmentForm               HeadOfState Capital Code2
## 1         People'sRepublic               Jiang Zemin    1891    CN
## 2         Federal Republic  Kocheril Raman Narayanan    1109    IN
## 3         Federal Republic            George W. Bush    3813    US
## 4                 Republic         Abdurrahman Wahid     939    ID
## 5         Federal Republic Fernando Henrique Cardoso     211    BR
## 6                 Republic      Mohammad Rafiq Tarar    2831    PK
## 7         Federal Republic            Vladimir Putin    3580    RU
## 8                 Republic         Shahabuddin Ahmad     150    BD
## 9  Constitutional Monarchy                   Akihito    1532    JP
## 10        Federal Republic         Olusegun Obasanjo    2754    NG
```

Where ... And/Or

En ocasiones cuando se requiere definir algún filtro, se necesita establecer condiciones múltiples, en donde se utilizan los operadores lógicos para conectar las mismas.

Ejemplo: Determine los datos de los países que cuya población sea superior a 100 millones de habitantes y su índice de habitantes sobre superficie sea menor a 50 habitante por m2.

```
df_res4<-sqldf("SELECT CODE, NAME, continent, population/surfacearea AS '
Index' FROM df_country WHERE population>100000000 AND population/surfacea
rea<50", connection=NULL)
df_res4
```

```
##   Code               Name     Continent     Index
## 1  BRA             Brazil South America 19.902536
## 2  RUS Russian Federation        Europe  8.605011
## 3  USA      United States North America 29.727816
```

Actividad 2: *Determine el listado de países del continente South America con población mayor a 50 millones, que tengan expectativa de vida menor a 60 años.*

Actividad 3: *Determine el listado de países con Gobierno Republicano o Federal Republicano, con un GNP(PIB) mayor a 10000.*

Where ...IN/NOT IN

La instrucción IN/NOT IN nos permite condicionar la pertenencia o no de un elemento a una colección (conjunto de datos). Internamente se realiza un comparación si un elemento pertenece a un conjunto de datos, retornando el valor de verdad. Por ejemplo: Se desea listar los países de los continentes de África, Asia u Oceanía, con población superior a 50 millones de habitantes.

```
df_res5<-sqldf("SELECT * from df_country where Continent in ('Africa','As
ia','Oceania') and population>50000000",connection=NULL)
df_res5
```

```
##    Code                                  Name Continent
## 1   BGD                            Bangladesh      Asia
## 2   CHN                                 China      Asia
## 3   COD Congo, The Democratic Republic of the    Africa
## 4   EGY                                 Egypt    Africa
## 5   ETH                              Ethiopia    Africa
## 6   IDN                             Indonesia      Asia
## 7   IND                                 India      Asia
## 8   IRN                                  Iran      Asia
## 9   JPN                                 Japan      Asia
## 10  NGA                               Nigeria    Africa
## 11  PAK                              Pakistan      Asia
## 12  PHL                           Philippines      Asia
## 13  THA                              Thailand      Asia
## 14  TUR                                Turkey      Asia
## 15  VNM                               Vietnam      Asia
##                       Region SurfaceArea IndepYear Population LifeExpe
ctancy
## 1  Southern and Central Asia      143998      1971  129155000
60.2
## 2               Eastern Asia     9572900     -1523 1277558000
71.4
## 3            Central Africa     2344858      1960   51654000
48.8
## 4           Northern Africa     1001449      1922   68470000
63.3
## 5            Eastern Africa     1104300     -1000   62565000
45.2
## 6            Southeast Asia     1904569      1945  212107000
68.0
```

```
## 7  Southern and Central Asia     3287263      1947 1013662000
62.5
## 8  Southern and Central Asia     1648195      1906   67702000
69.7
## 9               Eastern Asia      377829      -660  126714000
80.7
## 10            Western Africa      923768      1960  111506000
51.6
## 11 Southern and Central Asia      796095      1947  156483000
61.1
## 12            Southeast Asia      300000      1946   75967000
67.5
## 13            Southeast Asia      513115      1350   61399000
68.6
## 14               Middle East      774815      1923   66591000
71.0
## 15            Southeast Asia      331689      1945   79832000
69.3
##        GNP  GNPOld                           LocalName          Governm
entForm
## 1    32852   31966                           Bangladesh               R
epublic
## 2   982268  917719                             Zhongquo        People'sR
epublic
## 3     6964    2474 RÃ©publique DÃ©mocratique du Congo               R
epublic
## 4    82710   75617                                Misr               R
epublic
## 5     6353    6180                          YeItyopÂ´iya               R
epublic
## 6    84982  215002                             Indonesia               R
epublic
## 7   447114  430572                          Bharat/India        Federal R
epublic
## 8   195746  160151                                  Iran        Islamic R
epublic
## 9  3787042 4192638                          Nihon/Nippon Constitutional M
onarchy
## 10   65707   58623                               Nigeria        Federal R
epublic
## 11   61289   58549                              Pakistan               R
epublic
## 12   65107   82239                             Pilipinas               R
epublic
## 13  116416  153907                          Prathet Thai Constitutional M
onarchy
## 14  210721  189122                              TÃ¼rkiye               R
epublic
## 15   21929   22834                             ViÃªt Nam    Socialistic R
epublic
```

```
##                      HeadOfState Capital Code2
## 1              Shahabuddin Ahmad     150    BD
## 2                    Jiang Zemin    1891    CN
## 3                  Joseph Kabila    2298    CD
## 4                  Hosni Mubarak     608    EG
## 5                 Negasso Gidada     756    ET
## 6              Abdurrahman Wahid     939    ID
## 7       Kocheril Raman Narayanan    1109    IN
## 8  Ali Mohammad Khatami-Ardakani    1380    IR
## 9                         Akihito    1532    JP
## 10             Olusegun Obasanjo    2754    NG
## 11          Mohammad Rafiq Tarar    2831    PK
## 12       Gloria Macapagal-Arroyo     766    PH
## 13            Bhumibol Adulyadej    3320    TH
## 14            Ahmet Necdet Sezer    3358    TR
## 15               TrÃ¢n Duc Luong    3770    VN
```

Actividad 4 *Liste los países que no pertenezcan al continente de Europa o Asia con población mayor a 50 millones de personas.*

```
df_res6<-sqldf("SELECT * from df_country where Continent not in ('Europe'
,'Asia') and population>50000000",connection=NULL)
df_res6
```

```
##   Code                                    Name     Continent          Re
gion
## 1  BRA                                  Brazil South America   South Ame
rica
## 2  COD Congo, The Democratic Republic of the        Africa  Central Af
rica
## 3  EGY                                   Egypt        Africa Northern Af
rica
## 4  ETH                                 Ethiopia        Africa  Eastern Af
rica
## 5  MEX                                  Mexico North America Central Ame
rica
## 6  NGA                                 Nigeria        Africa  Western Af
rica
## 7  USA                           United States North America   North Ame
rica
##   SurfaceArea IndepYear Population LifeExpectancy     GNP  GNPOld
## 1     8547403      1822  170115000           62.9  776739  804108
## 2     2344858      1960   51654000           48.8    6964    2474
## 3     1001449      1922   68470000           63.3   82710   75617
## 4     1104300     -1000   62565000           45.2    6353    6180
## 5     1958201      1810   98881000           71.5  414972  401461
## 6      923768      1960  111506000           51.6   65707   58623
## 7     9363520      1776  278357000           77.1 8510700 8110900
##                             LocalName   GovernmentForm              He
adOfState
## 1                             Brasil Federal Republic Fernando Henriqu
```

```
e Cardoso
## 2 RÃ©publique DÃ©mocratique du Congo         Republic             Jose
ph Kabila
## 3                                Misr         Republic             Hosn
i Mubarak
## 4                        YeItyopÂ´iya         Republic            Negas
so Gidada
## 5                             MÃ©xico Federal Republic       Vicente Fo
x Quesada
## 6                             Nigeria Federal Republic         Olusegun
Obasanjo
## 7                       United States Federal Republic            Georg
e W. Bush
##   Capital Code2
## 1     211    BR
## 2    2298    CD
## 3     608    EG
## 4     756    ET
## 5    2515    MX
## 6    2754    NG
## 7    3813    US
```

Where ...Like/Not Like

La instrucción like nos permite establecer filtros en patrones sobre campos string, de tal forma, se pueda examinar el contenido parcial o total de una cadena de datos. Los formatos utilizados para los filtros se muestran a continuación: [cadena]% : inicia con la cadena de chars. %[cadena] : finaliza con la cadena de chars. %[cadena]% : contiene la cadena de chars. _[char]% : el segundo elemento es char.

Por ejemplo: Presentar todos el listado de países cuyo nombre inicie con letra E.

```
df_res7<-sqldf("SELECT * from df_country where name like 'E%'",connection
=NULL)
df_res7
```

```
##   Code              Name     Continent           Region SurfaceArea In
depYear
## 1  ECU           Ecuador South America    South America      283561
1822
## 2  EGY             Egypt        Africa  Northern Africa     1001449
1922
## 3  ERI           Eritrea        Africa   Eastern Africa      117600
1993
## 4  EST           Estonia        Europe Baltic Countries       45227
1991
## 5  ETH          Ethiopia        Africa   Eastern Africa     1104300
-1000
## 6  GNQ Equatorial Guinea        Africa   Central Africa       28051
```

```
1968
## 7  SLV       El Salvador North America  Central America       21041
1841
## 8  TMP        East Timor          Asia   Southeast Asia       14874
NA
##   Population LifeExpectancy   GNP GNPOld         LocalName
## 1   12646000           71.1 19770  19769           Ecuador
## 2   68470000           63.3 82710  75617              Misr
## 3    3850000           55.8   650    755             Ertra
## 4    1439200           69.5  5328   3371             Eesti
## 5   62565000           45.2  6353   6180      YeItyopÂ´iya
## 6     453000           53.6   283    542 Guinea Ecuatorial
## 7    6276000           69.7 11863  11203       El Salvador
## 8     885000           46.0     0     NA       Timor Timur
##            GovernmentForm                        HeadOfState Capital Co
de2
## 1               Republic            Gustavo Noboa Bejarano     594
EC
## 2               Republic                     Hosni Mubarak     608
EG
## 3               Republic  Isayas Afewerki [Isaias Afwerki]     652
ER
## 4               Republic                     Lennart Meri    3791
EE
## 5               Republic                    Negasso Gidada     756
ET
## 6               Republic     Teodoro Obiang Nguema Mbasogo    2972
GQ
## 7               Republic Francisco Guillermo Flores PÃ©rez     645
SV
## 8 Administrated by the UN           JosÃ© Alexandre GusmÃ£o    1522
TP
```

Group by

La instrucción group by permite formar grupos de registros acorde a un elemento o categoría especificada. Por ejemplo: Agrupar los países por continente y determine la cantidad correspondiente en cada grupo.

```
df_res8<-sqldf("SELECT Continent, count(*) as 'Count' from df_country gro
up by Continent",connection=NULL)
df_res8
```

```
##       Continent Count
## 1        Africa    58
## 2    Antarctica     5
## 3          Asia    51
## 4        Europe    46
## 5 North America    37
```

```
## 6       Oceania    28
## 7 South America    14
```

Actividad 5 *De acuerdo a los datos de los países, determine la cantidad de países sistema de gobierno "GovernmentForm"*

Having

En ocasiones cuando se tiene datos agrupados, se requiere establecer una condición sobre una medida de resumen del grupo. La instrucción Having nos permite establecer condiciones sobre grupos. Por ejemplo, Determine aquellos Sistemas de Gobierno que son aplicados en más de 20 países.

```
df_res9<-sqldf("SELECT GovernmentForm, count(*) as 'Count' from df_countr
y group by GovernmentForm having count(*)>20 ",connection=NULL)
df_res9
```

```
##            GovernmentForm Count
## 1 Constitutional Monarchy    29
## 2                Republic   122
```

Case When…then..end

La instrucción Case when..then nos permite establecer casos y acciones de respuestas cuando se cumplan. El resultado de la instrucción se proyecta en la vista de datos. Por ejemplo, presente un listado con el nombre del país, continente, población y grupo. Para la cuarta columna considere los siguientes criterios: Grupo 1: Población < 50 millones de habitantes, Grupo 2: Población entre 50 y 600 millones de habitantes, y Grupo 3: Población > 600 millones de habitantes.

```
df_res10<-sqldf("SELECT Name, Continent, Population, case when Population
<50000000 then 'G1'
                when Population>=50000000 and Population<=600000000 then
'G2'
                When Population>600000000 then 'G3' end As 'Group' from d
f_country ",connection=NULL)
df_res10
```

```
##                                          Name     Continent Populat
ion Group
## 1                                       Aruba North America     103
000    G1
## 2                                 Afghanistan          Asia   22720
000    G1
## 3                                      Angola        Africa   12878
```

```
000    G1
## 4                                      Anguilla North America       8
000    G1
## 5                                       Albania        Europe    3401
200    G1
## 6                                       Andorra        Europe      78
000    G1
## 7                          Netherlands Antilles North America     217
000    G1
## 8                          United Arab Emirates          Asia    2441
000    G1
## 9                                    Argentina South America   37032
000    G1
## 10                                      Armenia          Asia    3520
000    G1
## 11                               American Samoa       Oceania      68
000    G1
```