MAESTRÍA EN ESTADÍSTICA APLICADA/ MANEJO DE DATOS

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

*3.4. Funciones de Agregado*

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") #instalación de paquete RODBC  
library(RODBC) #referencia a la librería RODBC  
con<-odbcConnect("MySQLConexion\_world", uid = "root", pwd="marce") #Establecer la conexión con el DBMS

En esta sección nuevamente se trabajará con la función sqldf( ), mediante la cual es posible utilizar las sintaxis convencional de SQL en el entorno de R. Para el uso de la función sqldf, vamos a requerir previamente que los datos se almacenen en un Dataframe, se trabajará con los registros de la tabla country y city de la base de datos “world” de MySQL.

library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

df\_country<-sqlQuery(con,"Select \* from country") #Ejecutar una sentencia SQL  
df\_city<-sqlQuery(con,"Select \* from city")

**Funciones de agregación**

Una función de agregación es una función que resume las filas de un grupo en un solo valor. COUNT, MIN y MAX son ejemplos de funciones de agregación.

*COUNT: devuelve el número total de filas seleccionadas por la consulta.* MIN: devuelve el valor mínimo del campo que especifiquemos. *MAX: devuelve el valor máximo del campo que especifiquemos.* SUM: suma los valores del campo que especifiquemos. Sólo se puede utilizar en columnas numéricas. \*AVG: devuelve el valor promedio del campo que especifiquemos. Sólo se puede utilizar en columnas numéricas.

*Aplicación directa para resumir* Todas estas funciones se aplican a una sola columna, que especificaremos entre paréntesis, excepto la función COUNT, que se puede aplicar a una columna o indicar un “\*”.

df\_res18<-sqldf("select min(population) 'min', max(population) 'max', count(\*) 'count', sum(population) 'sum', avg(population) 'avg', variance(population) 'var' from df\_country", connection=NULL)  
df\_res18

## min max count sum avg var  
## 1 0 1277558000 239 6078749450 25434098 1.19552e+16

*Aplicación agrupando registro y resumiendo por grupo* La cláusula GROUP BY unida a un SELECT permite agrupar filas según las columnas que se indiquen como parámetros, y se suele utilizar en conjunto con las funciones de agrupación, para obtener datos resumidos y agrupados por las columnas que se necesiten.

Hemos visto en el ejemplo anterior que obteníamos sólo una fila con los datos indicados correspondientes a toda la tabla. Ahora vamos a ver con otro ejemplo cómo obtener datos correspondientes a diversos grupos de filas, concretamente agrupados.

df\_res19=sqldf("select Continent, min(population) as 'min' , max(population) as 'max', sum(population) as 'sum', avg(population) as 'avg' , variance(population) as 'var'   
 , sqrt(variance(population)) as 'sd' from df\_country group by Continent")  
df\_res19

## Continent min max sum avg var sd  
## 1 Africa 0 111506000 784475000 13525431 4.028821e+14 20071923  
## 2 Antarctica 0 0 0 0 0.000000e+00 0  
## 3 Asia 286000 1277558000 3705025700 72647563 5.055770e+16 224850388  
## 4 Europe 1000 146934000 730074600 15871187 7.713927e+14 27773958  
## 5 North America 7000 278357000 482993000 13053865 2.290992e+15 47864304  
## 6 Oceania 0 18886000 30401150 1085755 1.343720e+13 3665679  
## 7 South America 2000 170115000 345780000 24698571 1.946689e+15 44121303

**Actividad 1:** *Considerando la tabla ciudad(df\_city), agrupe por el código país y determine sobre el campo población las siguientes medidas max, sum, avg, variance y desviación estándar.*

**Actividad 2:** *Considerando la tabla país(df\_country), agrupe por el Sistema de Gobierno (GovernmentForm) y determine sobre el campo GNP las siguientes medidas max, sum, avg, variance y desviación estándar.*

**Actividad 3:** *Determine aquellos lenguajes (idiomas) que se hablan en más de 10 países*

**Actividad 4:** *Determine para cada lenguaje(idioma), la cantidad de población que habla dicho idioma. Nota: Recuerde considerar el porcentaje de habla asociado al idioma*