**Limpieza base de datos investigadores:**

Para llevar a cabo la depuración de caracteres y de información por medio del SMBD SQL se procede a crear la base de datos llamada BASE\_DATOS donde se carga la tabla de investigadores con el nombre INVE\_AJUST; a continuación, la sintaxis requerida para estandarizar cada campo:

**--BASE INICIAL SIN LIMPIAR PARA EL CAMPO: DEPARTAMENTO**

SELECT DEPARTAMENTO, count(\*) FROM BASE\_DATOS.INVE\_AJUST GROUP BY DEPARTAMENTO ORDER BY 1 ASC

Captura de pantalla de computadora

Descripción generada automáticamenteInterfaz de usuario gráfica, Texto

Descripción generada automáticamenteInterfaz de usuario gráfica, Texto

Descripción generada automáticamente

**--LIMPIEZA PARA EL CAMPO: DEPARTAMENTO**

UPDATE A

SET A.DEPARTAMENTO ='No\_Informa'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('',' ')

UPDATE A

SET A.DEPARTAMENTO ='Alta Verapaz'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('Alta Verapaz','Alvara Verapaz','Alta Verapaz ')

UPDATE A

SET A.DEPARTAMENTO ='Antigua Guatemala'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('Antigua Guatemala','Antigua Guatemala,Sacatepequez')

UPDATE A

SET A.DEPARTAMENTO ='Guatemala'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('Guatemala','Guatemala ','Guatemalteca')

UPDATE A

SET A.DEPARTAMENTO ='Guatemala'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('Guatemala','Guatemala ','Guatemalteca')

UPDATE A

SET A.DEPARTAMENTO ='Guatemala'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('Guatemala','Guatemala ','Guatemalteca')

UPDATE A

SET A.DEPARTAMENTO ='Huehuetenango'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('Huehuetenango','Huhuetenango')

UPDATE A

SET A.DEPARTAMENTO ='Poptun Peten'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('Peten','Poptun, Peten')

UPDATE A

SET A.DEPARTAMENTO ='Quetzaltenango'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('Qietzaltenango','Quetzaltenango')

UPDATE A

SET A.DEPARTAMENTO ='San Andres Solola'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('San Andres, Solola.')

UPDATE A

SET A.DEPARTAMENTO ='San Benito Peten'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('San Benito, Peten')

UPDATE A

SET A.DEPARTAMENTO ='San Marcos'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('San Marcos','San Marcos ')

UPDATE A

SET A.DEPARTAMENTO ='Zacapa'

from BASE\_DATOS.inve\_ajust as a

where a.DEPARTAMENTO in ('Zacapa','Zacapa ')

***--Resultado***

SELECT DEPARTAMENTO, count(\*) FROM BASE\_DATOS.INVE\_AJUST GROUP BY DEPARTAMENTO ORDER BY 1 ASC

Interfaz de usuario gráfica, Texto, Word

Descripción generada automáticamenteUna captura de pantalla de una computadora

Descripción generada automáticamenteUna captura de pantalla de una computadora

Descripción generada automáticamente

**--LIMPIEZA PARA EL CAMPO: ESTADO CIVIL**

SELECT estado\_civil, count(\*) FROM BASE\_DATOS.INVE\_AJUST GROUP BY estado\_civil ORDER BY 1 ASC

Una captura de pantalla de una computadora

Descripción generada automáticamenteUna captura de pantalla de una computadora

Descripción generada automáticamente

UPDATE A

SET A.estado\_civil ='casado\_a'

from BASE\_DATOS.inve\_ajust as a

where a.estado\_civil in ('casada','casadi','casado','casado ','casadp','casda')

UPDATE A

SET A.estado\_civil ='divorciado\_a'

from BASE\_DATOS.inve\_ajust as a

where a.estado\_civil in ('divorciada','divorciado')

UPDATE A

SET A.estado\_civil ='No\_Informa'

from BASE\_DATOS.inve\_ajust as a

where a.estado\_civil in (' ','','guatemala','no disponible')

UPDATE A

SET A.estado\_civil ='otro'

from BASE\_DATOS.inve\_ajust as a

where a.estado\_civil in ('otro','otro ','otros')

UPDATE A

SET A.estado\_civil ='separado\_a'

from BASE\_DATOS.inve\_ajust as a

where a.estado\_civil in ('separada','separado')

UPDATE A

SET A.estado\_civil ='soltero\_a'

from BASE\_DATOS.inve\_ajust as a

where a.estado\_civil in ('solero','soletra','solotera','solotero','solteo','soltera','soltera ','solterea','soltero','soltero ','solteroq','sotera')

UPDATE A

SET A.estado\_civil ='Union\_Libre'

from BASE\_DATOS.inve\_ajust as a

where a.estado\_civil in ('unida','unido')

UPDATE A

SET A.estado\_civil ='Viudo\_a'

from BASE\_DATOS.inve\_ajust as a

where a.estado\_civil in ('viuda','viudo')

***--Resultado***

SELECT estado\_civil, count(\*) FROM BASE\_DATOS.INVE\_AJUST GROUP BY estado\_civil ORDER BY 2 DESC

Una captura de pantalla de una computadora

Descripción generada automáticamenteUna captura de pantalla de una computadora

Descripción generada automáticamente

**--LIMPIEZA PARA EL CAMPO AREAS**

SELECT areas, count(\*) FROM BASE\_DATOS.INVE\_AJUST GROUP BY areas ORDER BY 1 ASC

Una captura de pantalla de una computadora

Descripción generada automáticamenteUna captura de pantalla de una computadora

Descripción generada automáticamenteUna captura de pantalla de una computadora

Descripción generada automáticamente

Se crean os campos para csda area para homologar los nombres y unificar

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD AREAS\_1 VARCHAR (50)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD AREAS\_2 VARCHAR (50)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD AREAS\_3 VARCHAR (50)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD AREAS\_4 VARCHAR (50)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD AREAS\_5 VARCHAR (50)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD AREAS\_6 VARCHAR (50)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD no\_inf VARCHAR (50)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD AREAS\_7 VARCHAR (50)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD AREAS\_DEF VARCHAR (200)

--HOMOLOGA CADA ÁREA QUE APLIQUE

UPDATE A

SET A.AREAS\_1 ='Ciencias Agricolas '

from BASE\_DATOS.inve\_ajust as a

where a.areas like '%agr%'

UPDATE A

SET A.AREAS\_2 ='Humanidades '

from BASE\_DATOS.inve\_ajust as a

where a.areas like '%Hum%'

UPDATE A

SET A.AREAS\_3 ='Ciencias Medicas '

from BASE\_DATOS.inve\_ajust as a

where a.areas like '%Med%'

UPDATE A

SET A.AREAS\_4 ='Ciencias Sociales '

from BASE\_DATOS.inve\_ajust as a

where a.areas like '%Soc%'

UPDATE A

SET A.AREAS\_5 ='Ciencias Medicas '

from BASE\_DATOS.inve\_ajust as a

where a.areas like '%Med%'

UPDATE A

SET A.AREAS\_6 ='Ciencias Naturales y Exactas '

from BASE\_DATOS.inve\_ajust as a

where a.areas like '%Nat%'

UPDATE A

SET A.no\_inf ='No\_Informa'

from BASE\_DATOS.inve\_ajust as a

where a.areas like '%dis%'

UPDATE A

SET A.AREAS\_7 ='Ingenieria y Tecnologia '

from (select \* from BASE\_DATOS.inve\_ajust where no\_inf is null) as a

where a.areas like '%ing%'

UPDATE A

SET A.AREAS\_DEF =CONCAT(A.AREAS\_1,A.AREAS\_2,A.AREAS\_3,A.AREAS\_4,A.AREAS\_5,A.AREAS\_6,A.AREAS\_7, a.no\_inf)

from BASE\_DATOS.inve\_ajust as a

***--Resultado***

SELECT AREAS\_DEF, count(\*) FROM BASE\_DATOS.INVE\_AJUST GROUP BY AREAS\_DEF ORDER BY 2 DESC

Una captura de pantalla de una computadora

Descripción generada automáticamente

**--LIMPIEZA PARA EL IDIOMA**

SELECT IDIOMAS, count(\*) FROM BASE\_DATOS.INVE\_AJUST GROUP BY IDIOMAS ORDER BY 1 ASC

Una captura de pantalla de una computadora

Descripción generada automáticamente

Se crean las variables para cada idioma, se homologa y se consolida:

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_1 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_2 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_3 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_4 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_5 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_6 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_7 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_8 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_9 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_10 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMA\_11 VARCHAR (20)

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD IDIOMAS\_DEF VARCHAR (200)

UPDATE A SET A.IDIOMA\_1 ='castellano ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%cast%'

UPDATE A SET A.IDIOMA\_2 ='aleman ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%ale%'

UPDATE A SET A.IDIOMA\_3 ='portugues ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%por%'

UPDATE A SET A.IDIOMA\_4 ='frances ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%fra%'

UPDATE A SET A.IDIOMA\_5 ='italiano ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%ita%'

UPDATE A SET A.IDIOMA\_6 ='maya ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%may%'

UPDATE A SET A.IDIOMA\_7 ='ingles ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%ing%'

UPDATE A SET A.IDIOMA\_8 ='japones ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%jap%'

UPDATE A SET A.IDIOMA\_9 ='mandarin ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%man%'

UPDATE A SET A.IDIOMA\_10 ='griego ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%gri%'

UPDATE A SET A.IDIOMA\_11 ='ruso ' from BASE\_DATOS.inve\_ajust as a where A.IDIOMAS LIKE '%rus%'

UPDATE A

SET A.IDIOMAS\_DEF =CONCAT(A.idioma\_1,A.idioma\_2,A.idioma\_3,A.idioma\_4,A.idioma\_5,A.idioma\_6,A.idioma\_7, A.idioma\_8, A.idioma\_9, A.idioma\_10, A.idioma\_11)

from BASE\_DATOS.inve\_ajust as a

***--Resultado***

**SELECT IDIOMAS\_DEF, count(\*) FROM BASE\_DATOS.INVE\_AJUST GROUP BY IDIOMAS\_DEF ORDER BY 2 DESC**

Una captura de pantalla de una computadora

Descripción generada automáticamenteInterfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**--LIMPIEZA PARA EL CAMPO GRADO**

UPDATE A SET A.GRADO ='Estudios\_en\_Progreso' from BASE\_DATOS.inve\_ajust as a where A.GRADO ='Estudios en Progreso'

UPDATE A SET A.GRADO ='Pensum\_Cerrado' from BASE\_DATOS.inve\_ajust as a where A.GRADO ='Pensum Cerrado'

UPDATE A SET A.GRADO ='PostDoctorado' from BASE\_DATOS.inve\_ajust as a where A.GRADO ='Post-Doctorado'

UPDATE A SET A.GRADO ='PostGrado' from BASE\_DATOS.inve\_ajust as a where A.GRADO ='Post-Grado'

UPDATE A SET A.GRADO ='No\_Informa' from BASE\_DATOS.inve\_ajust as a where A.GRADO ='Seleccione uno'

**RESULTADO**

Captura de pantalla de computadora

Descripción generada automáticamenteCaptura de pantalla de computadora

Descripción generada automáticamente

**--LIMPIEZA PARA EL CAMPO PAIS\_ESPECIALIDAD**

ALTER TABLE BASE\_DATOS.INVE\_AJUST ADD pais\_especialidad\_def VARCHAR (20)

UPDATE A SET A.pais\_especialidad\_def ='Alemania' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ('Alemania','Alemania Federal')

UPDATE A SET A.pais\_especialidad\_def ='Bolivia' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ('Bolivia','Cochabamba - Bolivia')

UPDATE A SET A.pais\_especialidad\_def ='Brasil' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ('Brasil','Brazil')

UPDATE A SET A.pais\_especialidad\_def ='Chile' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ('Chile','Chile ')

UPDATE A SET A.pais\_especialidad\_def ='Colombia' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ('Bogota, Colombia','Colombia','Colombia ')

UPDATE A SET A.pais\_especialidad\_def ='Corea' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ( 'Corea','Corea del Sur' )

UPDATE A SET A.pais\_especialidad\_def ='Costa\_Rica' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ( 'Costa Rica','Costa Rica ' )

UPDATE A SET A.pais\_especialidad\_def ='Cuba' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ( 'Cuba','La Habana, Cuba' )

UPDATE A SET A.pais\_especialidad\_def ='El\_Salvador' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ('El Salvador' )

UPDATE A SET A.pais\_especialidad\_def ='Espana' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ( 'España','Madrid' )

UPDATE A SET A.pais\_especialidad\_def ='Estados\_Unidos' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ('Boston, Estados Unidos','Boston, Mass, USA','Boston, USA','Colorado, USA','EEUU','Estados Unidos','Estados Unidos ','Estados Unidos de America','EUA','Madison, WI, USA','USA')

UPDATE A SET A.pais\_especialidad\_def ='Gran\_Bretana' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ( 'Gran Bretaña' )

UPDATE A SET A.pais\_especialidad\_def ='Guatemala' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ( 'Guatemala','Guatemala - España','Guatemala/EUA' )

UPDATE A SET A.pais\_especialidad\_def ='Holanda' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ( 'Holanda','Holanda (Paises Bajos)' )

UPDATE A SET A.pais\_especialidad\_def ='Mexico' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ( 'Mexico','Mexico ' )

UPDATE A SET A.pais\_especialidad\_def ='No\_Informa' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ( '','ninguno' ,'no disponible')

UPDATE A SET A.pais\_especialidad\_def ='Puerto\_Rico' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ('Puerto Rico')

UPDATE A SET A.pais\_especialidad\_def ='Reino\_Unido' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ( 'Reino Unido' )

UPDATE A SET A.pais\_especialidad\_def ='Republica\_Checa' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ('Republica CHECA' )

UPDATE A SET A.pais\_especialidad\_def ='Republica\_Dominicana' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ( 'Republica Dominicana' )

UPDATE A SET A.pais\_especialidad\_def ='Suecia' from BASE\_DATOS.inve\_ajust as a where A.pais\_especialidad in ( 'Suecia','Suecia ' )

UPDATE A SET A.pais\_especialidad\_def ='Taiwan' from BASE\_DATOS.inve\_ajust as a

where A.pais\_especialidad in ( 'Hsinchu, Taiwan','Taiwan' )

***--Resultado***

SELECT pais\_especialidad\_def, count(\*) FROM BASE\_DATOS.INVE\_AJUST GROUP BY pais\_especialidad\_def ORDER BY 2 DESC

Una captura de pantalla de una computadora

Descripción generada automáticamente

Normalización de la tabla INVESTIGADORES:

Se crea tabla temporal con los campos requeridos para eliminar duplicidad:

drop table ##tempo\_invest

SELECT codigo\_registro

,genero

,anio\_nacimiento

,departamento

,estado\_civil

,etnia

,anio\_ingreso

,categoria

,grado

,AREAS\_DEF as area

,IDIOMAS\_DEF as idiomas

,pais\_especialidad\_def as pais\_especialidad

into ##tempo\_invest

from BASE\_DATOS.INVE\_AJUST

Validamos cambios en la volumetria:

select count(\*) from ##tempo\_invest--2998

select distinct count (\*) from ##tempo\_invest—2998

No hay duplicidad, se tiene 2.998 registros únicos. A continuación, una imagen de la información contenida en la tabla investigadores dado el proceso de limpieza:

Captura de pantalla de computadora

Descripción generada automáticamente