CONSTRUCCIÓN BODEGA DE DATOS

SABERPRO

GUIA 3

Limpieza de Datos

Crear la tabla saberpro_limpio a partir de la tabla saberpro_2012_2014

Realizar los siguientes procesos de limpieza de datos:

1. Reemplazar los valores ‘EK20123’, EK20133’,’EK20142’,’EK20143 y ‘EKO2014’ del atributo *estu_cod_aplicacion* por ‘EK20122’,’EK20132’,’EK20141’, EK20142’ y EK20141 respectivamente:

2. Reemplazar los valores ‘20123’,20133’,’20142’,’20143 y ‘O2014’ del atributo *prueba* por ‘20122’,’20132’,’20141’,20142’ y 20141 respectivamente:
3. Reemplazar los valores nulos del atributo *estu_genero* por la moda de este atributo
4. Reemplazar los valores nulos del atributo estu_nacimiento_dia por la moda de este atributo en la tabla saberpro_2012_2014.
5. Reemplazar el valor ‘5’ del atributo *estu_nacimiento_dia* por ‘05’
6. Reemplazar los valores nulos del atributo *estu_nacimiento_mes* por la moda de este atributo en la tabla saberpro_2012_2014.
7. Reemplazar los valores nulos del atributo *estu_nacimiento_anno por* la moda de este atributo en los respectivos años de las pruebas 2012, 2013 y 2014.
8. Reemplazar los valores nulos del atributo *estu_pais_reside* por ‘COLOMBIA’
9. Reemplazar los valores ‘CO’ del atributo *estu_pais_reside* por ‘COLOMBIA’
10. Reemplazar los valores nulos del atributo *estu_estado_civil* por la moda de este atributo.

11. Reemplazar los valores nulos del atributo *estu_reside_codmpio* por los valores del atributo *estu_exam_codmpio_presentacion.*
12. Estandarizar el valor del atributo *estu_exam_mpio_presentacion* de “BOGOTÁ D.C.” por el valor “BOGOTA D.C”.
13. Estandarizar el valor del atributo *estu_exam_dpto_presentacion* de “BOGOTÁ” por el valor BOGOTA, "BOYACÁ" por BOYACA, CAQUETÁ por CAQUETA,

"ATLÁNTICO" por ATLANTICO, “CHOCÓ” por CHOCO, “CÓRDOBA” por CORDOBA,” SAN ANDRÉS” por SAN ANDRES, “VAUPÉS” por VAUPES, “QUINDÍO” por QUINDIO, “GUAINÍA” por GUANIA,” BOLÍVAR” por BOLIVAR,

14. Reemplazar los valores nulos del atributo *estu_exam_cod* por 141 para la prueba 20141 y por 142 para la prueba 20142.
15. Reemplazar los valores nulos del atributo *estu_exam_nombre* por EXAMEN SABER PRO 2014-1 para la prueba 20141 y por EXAMEN SABER PRO 2014-2 para la prueba 20142.
16. Estandarizar los nombres de las instituciones del atributo inst_nombre_institucion, el origen inst_origen el caracter academico inst_caracter_academico a partir de los nombres, origen y caracter academico de las instituciones de la tabla IES teniendo en cuenta los códigos de las instituciones
17. Reemplazar los valores nulos del atributo *estu_nivel_prgm_academico* por el valor no nulo del atributo *estu_nivel_prgm_academico* de los programas iguales de la misma institución
18. Reemplazar los valores nulos del atributo *estu_nivel_prgm_academico* por “UNIVERSITARIA” para aquellas instituciones cuyo nombre *inst_nombre_institucion* inicie con “UNIVERSIDAD”
19. Reemplazar los valores nulos del atributo *estu_nivel_prgm_academico* por “UNIVERSITARIA” para aquellos programas cuyo nombre *estu_prgm_academico* contengan las palabras INGENIERIA, DERECHO, MEDICINA, ODONTOLOGIA, CONTADURIA, LICENCIATURA,PSICOLOGIA ARQUITECTURA, ECOLOGIA, FISIOTERAPIA, ZOOTECNIA, ADMINISTRACIÓN, DISEÑO, SOCIAL, PUBLICIDAD, INTERNACIONAL, ARTES, PERIODISMO, GASTRONOMÍA, CIENCIA, TERAPIA, SALUD
20. Reemplazar los valores del atributo *estu_nivel_prgm_academico* por “TECNOLOGIA” para aquellos programas cuyo nombre *estu_prgm_academico* inicie con la palabra ‘TECNOLOGÍA’ o ‘TECNOLOGIA’

21. Reemplazar los valores del atributo *estu_nivel_prgm_academico* por “TECNICO” para aquellos programas cuyo nombre *estu_prgm_academico* inicie con la palabra ‘TECNICO’, ‘TÉCNICO’,’TÉCNICA’,’TECNICA’

22. Reemplazar los valores nulos del atributo *estu_metodo_prgm* por el valor no nulo del atributo *estu_metodo_prgm* de la institución con igual código (*inst_cod_institucion)* o nombre (*inst_nombre_institucion)*, y de los programas del mismo código(*estu_prac_id_prgrm_academico)* o nombre (*estu_prgm_academico)*
23. Reemplazar los valores nulos del atributo *estu_metodo_prgm* por los valores del atributo *metodología* de la tabla *programas_ies* teniendo en cuenta los códigos de las instituciones y de los programas
24. Reemplazar los 2 restantes valores nulos del atributo *estu_metodo_prgm* por la moda 'PRESENCIAL'
25. Estandarizar los valores del atributo *estu_metodo_prgm* en mayúsculas;

26. Reemplazar los valores nulos del atributo *dipo_codigomunicipio* por los valores del atributo *estu_exam_codmpio_presentacion*:
27. Reemplazar los valores nulos del atributo *inst_cod_jornada* por la moda "1" para los nombres de instituciones que contengan la palabra 'NORMAL'.
28. Reemplazar los valores nulos del atributo *inst_cod_jornada* por la moda del resto de instituciones que es '12'
29. Reemplazar los valores nulos del atributo *estu_area_conoc de saberpro_limpio,* por el valor no nulo del atributo *estu_area_conoc* de los programas del mismo nombre (*estu_prgm_academico)* de la tabla saberpro_2012_2014
30. Reemplazar los valores nulos del atributo *estu_nucleo_pregrado de saberpro_limpio*, por el valor no nulo del atributo *estu_nucleo_pregrado* de los programas del mismo nombre (*estu_prgm_academico)* de la tabla saberpro_2012_2014
31. Reemplazar los valores nulos del atributo *estu_cod_grupo_ref de saberpro_limpio*, por el valor no nulo del atributo *estu_cod_grupo_ref* de los grupos de referencia con el nombre (*estu_grupo_referencia)* de la tabla saberpro_2012_2014
32. Actualizar en la tabla *saberpro_limpio*, el atributo *estu_cod_grupo_ref utilizando* en algunos casos caracteres comodines "%" y "_" asi: (utilice select case when)

    BELLAS ARTES Y DISEÑO por el código 1,
    CIENCIAS NATURALES Y EXACTAS por el código 2,
    CIENCIAS SOCIALES por el código 3,
    HUMANIDADES por el código 4,
    DERECHO por el código 5,

COMUNICACI_N, PERIODISMO Y PUBLICIDAD por el código 6,
CIENCIAS MILITARES Y NAVALES por el código 7,
%CIENCIAS AGROPECUARIAS% por el código 8,
%ADMINISTRACI_N% por el código 9,
EDUCACI_N por el código 10,
ARQUITECTURA Y URBANISMO por el código 11,
INGENIER_A por el código 12,
%SALUD% por el código 13,
MEDICINA por el código 14,
%INGENIER_A, INDUSTRIA Y MINAS% por el código 15
%TIC% por el código 17,
%ARTES - DISEÑO – COMUNICACI_N% por el código 19,
%CIENCIAS AGROPECUARIAS% por el código 20,
NORMALES SUPERIORES por el código 27,
%JUDICIAL% por el código 28;
%MILITAR Y POLICIAL% por el código 29
RECREACI_N Y DEPORTES por el código 30,
%ECONOM% por el código 31,
%CONTADUR_A% por el código 32,
PSICOLOG_A por el código 33,
ENFERMER_A por el código 34;
GRUPO REFERENCIA NACIONAL% por el código 42,

33. Reemplazar los nulos del atributo *estu_semestre_cursa* por la moda (de las NORMALES) del atributo *estu_semestre_cursa* de la tabla *saberpro_2012_2014*, para los nombres de instituciones que contengan la palabra 'NORMAL'.
34. Reemplazar los valores nulos del atributo *estu_semestre_cursa* de *saberpro_limpio,* por el valor no nulo del atributo *estu_semestre_cursa* de la tabla saberpro_2012_2014 de igual instituciones con los mismos programas
35. Reemplazar los nulos del atributo *estu_pje_creditos* por el código '0' para aquellas instituciones cuyo valor del atributo *estu_nivel_prgm_academico* es 'NORMALISTA' o el nombre de la institución del atributo *inst_nombre_institucion* contiene la sigla SENA.
36. Reemplazar los valores nulos del atributo *estu_pje_creditos* por el valor no nulo del

atributo *estu_pje_creditos* de la institución con igual código (*inst_cod_institucion)* o nombre (*inst_nombre_institucion)*, y de los programas del mismo código(*estu_prac_id_prgrm_academico)* o nombre (*estu_prgm_academico)*

37. Reemplazar los valores nulos del atributo *inst_vlr_matricula_ant* por la moda (de las NORMALES) del atributo *inst_vlr_matricula_ant* de la tabla *saberpro_2012_2014*, para los nombres de instituciones que contengan la palabra 'NORMAL'.
38. Reemplazar los valores nulos del atributo *inst_vlr_matricula_ant* de *saberpro_limpio,* por el valor no nulo del atributo *inst_vlr_matricula_ant de* la tabla saberpro_2012_2014 de igual instituciones con los mismos programas
39. Reemplazar los valores nulos del atributo *estu_titulo_bto* por la moda (de las NORMALES) del atributo *estu_titulo_bto* de la tabla *saberpro_2012_2014*, para los nombres de instituciones que contengan la palabra 'NORMAL'.
40. Reemplazar los valores nulos del atributo *estu_titulo_bto* al resto de instituciones por la moda del atributo *estu_titulo_bto* de la tabla *saberpro_2012_2014*
41. Reemplazar los valores nulos del atributo *estu_hogar_actual* por la moda del atributo *estu_hogar_actual* de la tabla *saberpro_2012_2014*
42. Reemplazar los valores nulos del atributo *fami_num_pers_grup_fam* por la media (redondeada a entera) de los valores no nulos de *fami_num_pers_grup_fam* de la tabla saberpro_2012_2014;
43. Reemplazar los valores nulos del atributo *estu_sn_cabeza_fmlia* por la moda del atributo *estu_sn_cabeza_fmlia* de la tabla saberpro_2012_2014.
44. Reemplazar los valores nulos del atributo *fami_num_pers_cargo* por la media (redondeada a entera) de los valores no nulos de *fami_num_pers_cargo* de la tabla saberpro_2012_2014;
45. Reemplazar los valores nulos del atributo *fami_cod_educa_padre* por el código '99'
46. Reemplazar los valores 1,2,3,4,5,6,7,8 del atributo *fami_cod_educa_padre* por el código '9'
47. Reemplazar los valores nulos del atributo *fami_cod_educa_madre* por el código '99'
48. Reemplazar los valores 1,2,3,4,5,6,7,8 del atributo *fami_cod_educa_madre* por el código '9'
49. Reemplazar los valores nulos del atributo *fami_cod_ocup_madre* por el código '99'
55. Reemplazar los valores 01,02,03,04,05,06,07,08,09,10,11,12 del atributo *fami_cod_ocup_madre* por la moda del atributo *fami_cod_ocup_madre* de la tabla

saberpro_2012_2014

56. Reemplazar los valores nulos del atributo *estu_estrato* por la moda del atributo *estu_estrato* de la tabla *saberpro_2012_2014*
57. Reemplazar los valores nulos del atributo *fami_nivel_sisben* por la moda del atributo *fami_nivel_sisben* de la tabla *saberpro_2012_2014*
58. Reemplazar los valores nulos del atributo *econ_material_pisos* por la moda del atributo *econ_material_pisos* de la tabla *saberpro_2012_2014*
59. Reemplazar los valores nulos del atributo *econ_sn_telefonia* por la moda del atributo *econ_sn_telefonia* de la tabla *saberpro_2012_2014*
60. Reemplazar los valores nulos del atributo *econ_sn_internet* por la moda del atributo *econ_sn_internet* de la tabla *saberpro_2012_2014*
61. Reemplazar los valores nulos del atributo *econ_sn_servicio_tv* por la moda del atributo *econ_sn_servicio_tv* de la tabla *saberpro_2012_2014*
62. Reemplazar los valores nulos del atributo *econ_sn_computador* por la moda del atributo *econ_sn_computador* de la tabla *saberpro_2012_2014*
63. Actualizar el atributo *econ_sn_computador por 1* para aquellos valores del atributo *econ_sn_computador>0.*
64. Reemplazar los valores nulos del atributo *econ_sn_celular* y los de *econ_sn_celular* código *2* por la moda del atributo *econ_sn_celular* de la tabla *saberpro_2012_2014*
65. Reemplazar los valores nulos del atributo *econ_sn_dvd* y los de *econ_sn_dvd* código *2* por la moda del atributo *econ_sn_dvd* de la tabla *saberpro_2012_2014*
66. Reemplazar los valores nulos del atributo *econ_sn_lavadora* por la moda del atributo *econ_sn_lavadora* de la tabla *saberpro_2012_2014*
67. Reemplazar los valores nulos del atributo *econ_sn_microondas* por la moda del atributo *econ_sn_microondas* de la tabla *saberpro_2012_2014*
68. Reemplazar los valores nulos del atributo *econ_sn_automovil* por la moda del atributo *econ_sn_automovil* de la tabla *saberpro_2012_2014*
69. Reemplazar los valores nulos del atributo *econ_sn_horno* por la moda del atributo *econ_sn_horno* de la tabla *saberpro_2012_2014*

70. Reemplazar los valores nulos del atributo *econ_sn_nevera* por la moda del atributo *econ_sn_nevera* de la tabla *saberpro_2012_2014*
71. Reemplazar los valores nulos del atributo *infa_dormitorios* por la moda del atributo *infa_dormitorios* de la tabla *saberpro_2012_2014*

72. Reemplazar los valores nulos del atributo *fami_ing_fmliar_mensual* por la moda del atributo *fami_ing_fmliar_mensual* de la tabla *saberpro_2012_2014*
73. Reemplazar los valores nulos del atributo *estu_trabaja y los valores de estu_trabaja* códigos *1,6,7* por la moda del atributo *estu_trabaja* de la tabla *saberpro_2012_2014*
74. Reemplazar los valores nulos del atributo *estu_horas_trabajo* por 0 para los estudiantes que no trabajan.
75. Reemplazar los valores nulos del atributo *estu_horas_trabajo* por la media entera de los valores del atributo *estu_horas_trabajo* de los estudiantes cuyo atributo *estu_trabaja* es de código '3' de la tabla *saberpro_2012_2014*. Actualizar únicamente para los estudiantes que trabajan con código '3'
76. Reemplazar los valores nulos del atributo *estu_horas_trabajo* por la media entera de los valores del atributo *estu_horas_trabajo* de los estudiantes cuyo atributo *estu_trabaja* es de código '4' de la tabla *saberpro_2012_2014*. Actualizar únicamente para los estudiantes que trabajan con código '4'
77. Reemplazar los valores nulos del atributo *estu_horas_trabajo* por la media entera de los valores del atributo *estu_horas_trabajo* de los estudiantes cuyo atributo *estu_trabaja* es de código '5' de la tabla *saberpro_2012_2014*. Actualizar únicamente para los estudiantes que trabajan con código '5'
78. Reemplazar la coma por punto decimal a los valores del atributo *mod_lectura_critica*
79. Reemplazar los valores nulos del atributo *mod_lectura_critica* por 'NP'
80. Reemplazar la coma por punto decimal a los valores del atributo *mod_comunica_escrita_punt*
81. Reemplazar los valores nulos del atributo *mod_comunica_escrita_punt* por 'NP'
82. Estandarizar a punto decimal los valores del atributo *mod_razona_cuantitativo_punt*
83. Reemplazar los valores nulos del atributo *mod_razona_cuantitativo_punt* por 0
84. Estandarizar a punto decimal los valores del atributo *mod_ingles_punt.*
85. Reemplazar los valores nulos del atributo *mod_ingles_punt* por 0
86. Estandarizar a punto decimal los valores del atributo *mod_comp_ciudadanas_punt.*
87. Reemplazar los valores nulos del atributo *mod_comp_ciudadanas_punt* por 0