1. DESARROLLO DEL PROYECTO

Para el desarrollo del proyecto y entendiendo que se trata de un proyecto basado en datos, se siguió la metodología de tratamiento de datos.

Leiva Vigorena plantea la metodología de tratamiento de datos como el conjunto de procesos, técnicas y herramientas utilizadas para manipular, analizar, interpretar y presentar datos de manera efectiva. Estas metodologías pueden variar dependiendo del tipo de datos, el contexto en el que se utilicen y los objetivos específicos del análisis.

Algunas de las metodologías comunes incluyen fases o etapas como:

* Recopilación de datos: Esto implica la recopilación o adquisición de datos de diversas fuentes, ya sea a través de encuestas, bases de datos existentes, sensores, registros históricos, entre otros.
* Limpieza de datos: es importante limpiar los datos, antes de realizar cualquier tipo de análisis para eliminar errores, valores atípicos o datos incompletos que puedan afectar la calidad del análisis.
* Transformación de datos: implica la manipulación de los datos para que estén en un formato adecuado para el análisis. Esta etapa puede incluir la normalización de datos, la creación de nuevas variables o la agregación de datos.
* Análisis exploratorio de datos (EDA): implica la utilización de técnicas estadísticas y visuales para explorar los datos y entender sus características principales, como tendencias, patrones, relaciones y posibles correlaciones.
* Análisis estadístico: Una vez que se ha explorado y comprendido los datos, se pueden aplicar técnicas estadísticas más avanzadas para extraer información significativa, como pruebas de hipótesis, análisis de regresión, análisis de varianza, entre otros.
* Modelado de datos: En esta etapa, se pueden construir modelos analíticos predictivos o descriptivos utilizando técnicas como el aprendizaje automático, la minería de datos o la modelización estadística para hacer predicciones o identificar patrones en los datos.
* Evaluación y validación: Es importante evaluar la calidad y la precisión de los resultados obtenidos a través del análisis de datos. Esto puede implicar la validación cruzada, la comparación con datos de referencia o la evaluación de métricas de rendimiento.
* Interpretación y presentación de resultados: Finalmente, los resultados del análisis de datos deben interpretarse y presentarse de manera clara y comprensible para los interesados, lo que puede implicar la creación de informes, visualizaciones de datos o presentaciones” [].

La fuente de los datos es el Sistema Optimizado para la Formación y el Aprendizaje Activo - Sofía Plus, de donde se realizó el proceso de adquisición.

Respetando la política de protección de datos, los datos extraídos se encuentran anonimizados.

A medida que se fueron explorando los datos, se plantearon los siguientes supuestos o imaginarios que se tienen comúnmente:

1. Las mujeres desertan más que los hombres.

2. Los aprendices que reciben algún tipo de apoyo (Fic o Regular) tienden a desertar menos que los que no los reciben.

Apoyo FIC es ..

Apoyo de sostenimiento regular es ..

3. Se presenta mayor nivel de deserción en los niveles de formación más altos.

PREGUNTAS:

1. ¿El tipo de población afecta la retención?

2. Los apoyos que los aprendices reciben son significativos para la retención, entendiendo como retención que la persona se certifica?

3. ¿Qué ciudades o regionales tiene mayor deserción?

4. ¿Existe diferencia en la retención por Genero y estado civil?

5. Las variables estrato, puntaje y número de hijos tienen relación en la deserción?

Primer acercamiento a los datos:

El objetivo del trabajo con este archivo era conocer la estructura de los datos, sus características, posibles valores y la generación de estadísticas sobre la presencia de los valores y su interpretación.

En ese sentido, la metodología de tratamiento de datos se acotó a las etapas de recopilación de datos, análisis exploratorio de datos e Interpretación y presentación de resultados.

Recopilación de datos:

En marzo de 2023 se recibe el archivo, “230303\_DATOS\_BASICOS\_UBICACION\_USUARIO.csv”

Este archivo contiene 100.000 instancias (registros) con 13 características.

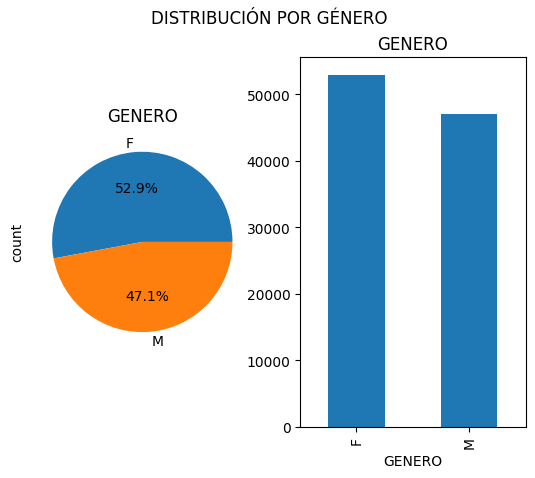
Las características observadas fueron:

|  |  |
| --- | --- |
| CARACTERÍSTICA | DESCRIPCIÓN |
| Nis | Número interno Sena que identifica de manera única a cada persona registrada en el Sistema Sofía Plus |
| Genero | Variable carácter con valores F o M, F-Femenino: M-Masculino |
| Fecha\_nacimiento | Fecha de nacimiento del aprendiz en formato DD/MM/AAAA |
| Estado\_civil | Variable carácter con valores S-Soltero, U-Unión Libre, C-Casado, V-Viudo, E-Separado, I-Indiferente, N-No Ingresado |
| Estrato | Estrato socioeconómico con valores enteros de 1 a 6 |
| Puntaje\_icfes | Resultado de la prueba Saber 11 |
| Sisben | Puntaje Sisbén del núcleo familiar |
| Dbu\_hijos | Número de hijos del aprendiz al momento del registro |
| Cod\_pais\_residencia | Código numérico que corresponde al país de residencia del aprendiz |
| Cod\_dpto\_residencia | Código numérico que corresponde al departamento de residencia del aprendiz |
| Cod\_municipio\_residencia | Código numérico que corresponde al municipio de residencia del aprendiz |
| Cod\_zona\_residencia | Código numérico que corresponde a la zona de residencia del aprendiz |
| Cod\_barrio\_residencia | Código numérico que corresponde al barrio de residencia del aprendiz |

Análisis exploratorio:

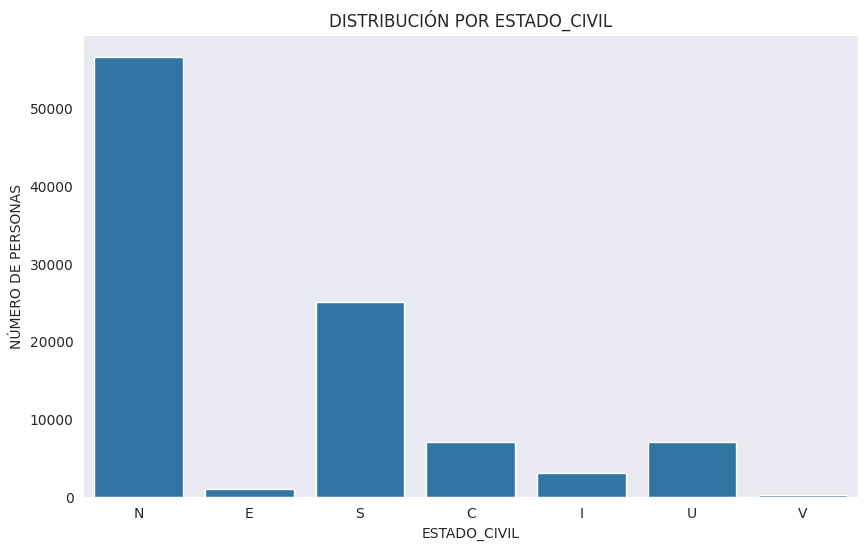
Algunas observaciones sobre los datos:



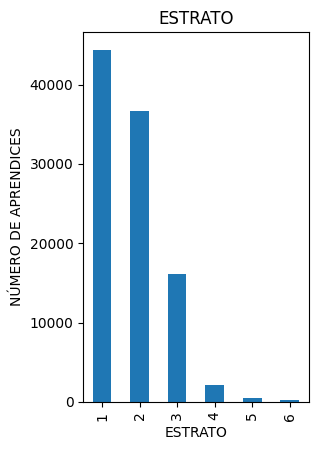


GENERO: Masculino 47072 (cerca del 47%), Femenino 52928 (cerca del 53%)

FECHA\_NACIMIENTO: Las fechas oscilaban entre 1.900 y 2.014, cerca del 25% de los aprendices nacieron entre 1.990 y 1.994. El hecho de que haya valores -> calidad de datos



ESTADO\_CIVIL: S-Soltero 25091 (cerca del 25%), U-Unión Libre 7056 (cerca del 7%), C-Casado 7067 (cerca del 7%), V-Viudo 200 (cerca del 0.2%), E-Separado 936 (cerca del 1%), I-Indiferente 3030 (cerca del 3%), N-No Ingresado 56620 (cerca del 57%)



ESTRATO: Estrato-1 44380 (cerca del 44%), estrato-2 36679 (cerca del 37%), estrato-3 16150 (cerca del 16%), estrato-4 2133 (cerca del 2%), estrato-5 485 (cerca del 0.5%), estrato-6 173 (cerca del 0.1%)

PUNTAJE\_ICFES: 98202 no tenían este valor (cerca del 98%), los puntajes se encuentran entre 1 y 999, y el puntaje que más se repite es 50.

SISBEN: Esta característica no tenía valor para el 100% de los datos

DBU\_HIJOS: Esta característica no tenía valor para el 100% de los datos

COD\_PAIS\_RESIDENCIA: Se encuentran 33 valores únicos para esta característica, 99.688 registros (más del 99%), tenían el valor 57.

COD\_DPTO\_RESIDENCIA: Se encontraron 164 valores únicos, cerca del 30% de los datos se concentran en dos códigos, así, 16866 (más del 16%) corresponden al código 57011, 12392 (más del 12%) corresponden al código 57005.

COD\_MUNICIPIO\_RESIDENCIA: Se encontraron 1248 valores únicos, cerca del 20% de los datos se concentran en dos códigos, así, 16862 (más del 16%) corresponden al código 57011001, 4745 (más del 4%) corresponden al código 57005001.

COD\_ZONA\_RESIDENCIA: Se encontraron 347 valores únicos, sin embargo 96323 (más del 96%) no traen valor.

COD\_BARRIO\_RESIDENCIA: Se encontraron 1643 valores diferentes, sin embargo 96796 (cerca del 97%) no traen valor.

Interpretación y presentación de resultados:

Se pudo observar que el archivo estaba compuesto por nueve (9) características numéricas y cuatro (4) categóricas.

En cuanto al género, femenino supera en casi 5% al masculino.

Analizando la característica COD\_PAIS\_RESIDENCIA más del 99% de los datos tenían el valor 57, que corresponde al código de país Colombia.

Así mismo, se encontraron características con alto porcentaje de valores nulos como PUNTAJE\_ICFES, SISBEN, DBU\_HIJOS, COD\_BARRIO\_RESIDENCIA y COD\_ZONA\_RESIDENCIA.

Como conclusión de este primer acercamiento a los datos, se pudo evaluar los posibles tipos de datos que se encontrarían y sus características, cumpliendo así con el objetivo propuesto y planteando realizar otro acercamiento con un nuevo archivo.

Segundo acercamiento a los datos:

El objetivo del trabajo con este archivo era ampliar la estructura de los datos y sus características, determinar los posibles valores y si se mantenía la ausencia de valores para varias características que podrían ser importantes como causa de un estado de abandono de la formación (deserción), así como la generación de estadísticas sobre la presencia de los valores y su interpretación.

En ese sentido, la metodología de tratamiento de datos se acotó a las etapas de recopilación de datos, análisis exploratorio de datos e Interpretación y presentación de resultados.

Recopilación de datos:

En marzo de 2023 se recibe el archivo, “230303\_formaciones.csv”

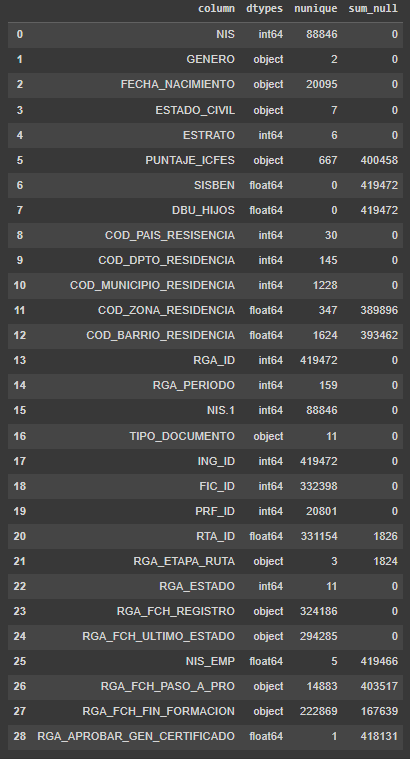
Este archivo contiene 419.472 instancias (registros) con 29 características.

Las características observadas fueron:

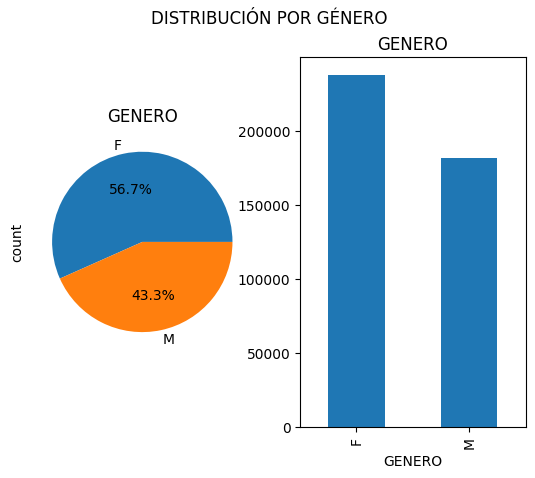
|  |  |
| --- | --- |
| CARACTERÍSTICA | DESCRIPCIÓN |
| Nis | Número interno Sena que identifica de manera única a cada persona registrada en el Sistema Sofía Plus |
| Genero | Variable carácter con valores F o M, F-Femenino: M-Masculino |
| Fecha\_nacimiento | Fecha de nacimiento del aprendiz en formato DD/MM/AAAA |
| Estado\_civil | Variable carácter con valores S-Soltero, U-Unión Libre, C-Casado, V-Viudo, E-Separado, I-Indiferente, N-No Ingresado |
| Estrato | Estrato socioeconómico con valores enteros de 1 a 6 |
| Puntaje\_icfes | Resultado de la prueba Saber 11 |
| Sisben | Puntaje Sisbén del núcleo familiar |
| Dbu\_hijos | Número de hijos del aprendiz al momento del registro |
| Cod\_pais\_residencia | Código numérico que corresponde al país de residencia del aprendiz |
| Cod\_dpto\_residencia | Código numérico que corresponde al departamento de residencia del aprendiz |
| Cod\_municipio\_residencia | Código numérico que corresponde al municipio de residencia del aprendiz |
| Cod\_zona\_residencia | Código numérico que corresponde a la zona de residencia del aprendiz |
| Cod\_barrio\_residencia | Código numérico que corresponde al barrio de residencia del aprendiz |
| Rga\_id | Valor numérico que identifica el registro académico del aprendiz |
| Rga\_periodo | Periodo de registro de la información en formato yyyymm |
| Nis | Número interno Sena que identifica de manera única a cada persona registrada en el Sistema Sofía Plus |
| Tipo\_documento | Tipo de documento con el que registró el aprendiz |
| Ing\_id | Consecutivo numérico con el que se realizó la inscripción del aspirante |
| Fic\_id | Valor numérico que corresponde al número consecutivo del curso |
| Prf\_id | Valor numérico que corresponde al número consecutivo del programa de formación |
| Rta\_id | Valor numérico que corresponde al número consecutivo de la ruta de aprendizaje |
| Rga\_etapa\_ruta | Variable carácter que indica la etapa en la que está la ruta de aprendizaje, con valores L-Lectiva, P-Práctica, M-Mixta |
| Rga\_estado | Variable numérica que indica el estado del aprendiz en su proceso de formación académica, tiene los valores 1 – Aplazado, 2 – Cancelado, 3 – Certificado, 4 – Condicionado, 5 – Reprobado, 6 - Retiro Voluntario, 7 – Formación, 8 – Inducción, 9 - No Apto, 10 – Reingresado, 11 – Traslado, 12 - Por Certificar, 13 - En Transito, 14 - Cancelamiento Virtual Complementaria y 15 - Deserción Virtual Complementaria |
| Rga\_fch\_registro | Fecha en la que se crea o registra el registro académico, formato dd/mm/aaaa |
| Rga\_fch\_ultimo\_estado | fecha del último estado del registro académico, formato dd/mm/aaaa |
| Nis\_emp | Nis de la empresa asociada al registro académico del aprendiz en la que presentará su etapa productiva |
| Rga\_fch\_paso\_a\_pro | Fecha en la que cambió de etapa lectiva a productiva |
| Rga\_fch\_fin\_formacion | Fecha en la cual el aprendiz finalizó todas sus competencias y resultados |
| Rga\_aprobar\_gen\_certificado | Código numérico que especifica si el aprendiz aprueba o no la generación de certificación del curso, con valores 0-no aprueba, 1-aprueba, null-el aprendiz no ha respondido |

Análisis exploratorio: (descriptivo)

Algunas observaciones sobre los datos:

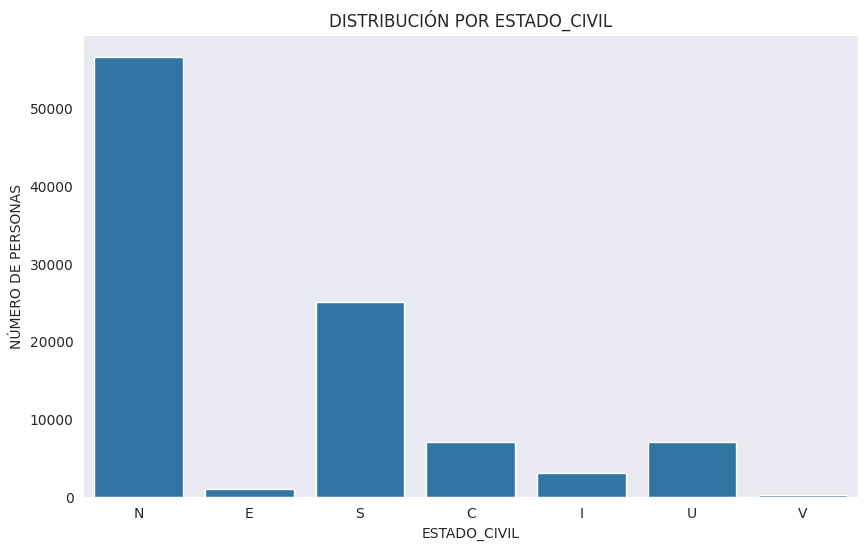


NIS: Se observa que no hay valores nulos y se encuentran 88846 valores diferentes, entendiendo que se trata del identificador único del aprendiz, se aprecia que hay aprendices con varias formaciones, como el NIS 11565308, que tiene 248 registros de formación, o el NIS 7599095 con 182 registros de formación.

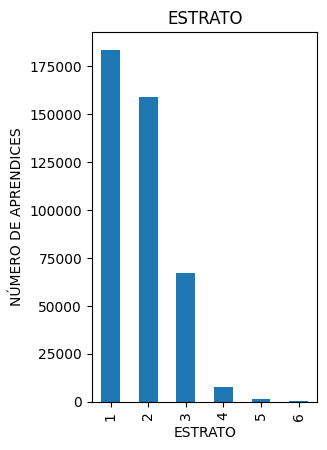


GENERO: Masculino 181788 (cerca del 43%), Femenino 237684 (cerca del 57%)

FECHA\_NACIMIENTO: Las fechas oscilaban entre 1.907 y 2.014, cerca del 28% de los aprendices nacieron entre 1.990 y 1.995.



ESTADO\_CIVIL: S-Soltero 191940 (cerca del 38%), U-Unión Libre 51095 (cerca del 12%), C-Casado 51071 (cerca del 12%), V-Viudo 1252 (cerca del 0.3%), E-Separado 6723 (cerca del 2%), I-Indiferente 15616 (cerca del 4%), N-No Ingresado 131775 (cerca del 31%)



ESTRATO: Estrato-1 183480 (cerca del 44%), estrato-2 158802 (cerca del 38%), estrato-3 67153 (cerca del 16%), estrato-4 7865 (cerca del 2%), estrato-5 1650 (cerca del 0.4%), estrato-6 522 (cerca del 0.12%)

PUNTAJE\_ICFES: 400458 no tenían este valor (cerca del 98%), los puntajes se encuentran entre 1 y 999

SISBEN: Esta característica no tenía valor para el 100% de los datos

DBU\_HIJOS: Esta característica no tenía valor para el 100% de los datos

COD\_PAIS\_RESIDENCIA: Se encuentran 30 valores únicos para esta característica, 418418 registros (más del 99%), tenían el valor 57.

COD\_DPTO\_RESIDENCIA: Se encontraron 145 valores únicos, cerca del 27% de los datos se concentran en dos códigos, así, 69570 (más del 16%) corresponden al código 57011, 45552 (cerca del 11%) corresponden al código 57005.

COD\_MUNICIPIO\_RESIDENCIA: Se encontraron 1288 valores únicos, cerca del 20% de los datos se concentran en dos códigos, así, 69564 (más del 16%) corresponden al código 57011001, 16640 (cerca del 4%) corresponden al código 57005001.

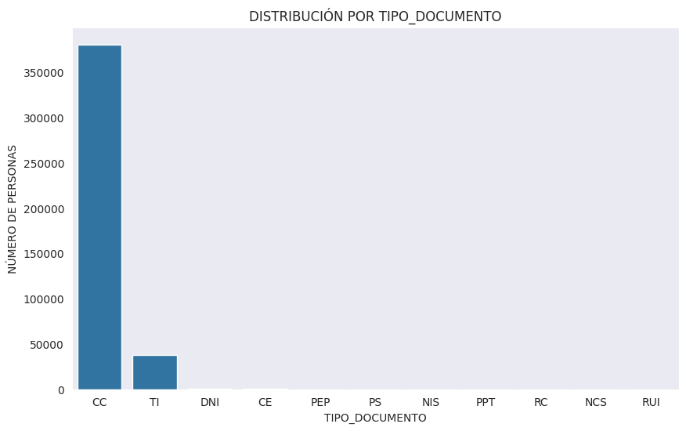
COD\_ZONA\_RESIDENCIA: Se encontraron 347 valores únicos, sin embargo 374324 (cerca del 90%) no traen valor.

COD\_BARRIO\_RESIDENCIA: Se encontraron 1624 valores diferentes, sin embargo 377556 (cerca del 90%) no traen valor.

RGA\_ID: Contiene un valor diferente para cada registro.

RGA\_PERIODO: Contiene 159 valores diferentes, el 10% de los registros están en los periodos 201103, 201105, 201001, 201104 y 201109.

NIS: corresponde a los mismos valores de la primera característica.



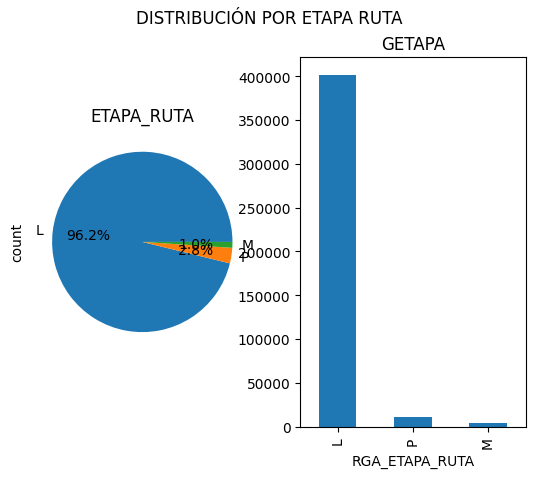
TIPO\_DOCUMENTO: Se encuentran 11 valores diferentes, CC 380723 (más del 90%), TI 37901 (cerca del 9%), CE 347 (cerca del 0.08%), DNI 256 (cerca del 0.06%), PEP 151 (cerca del 0.04%), PS 40 (cerca del 0.01%), los demás valores son PPP 22, NIS 17, RC 9, NCS 4, RUI 2.

ING\_ID: todos los valores son diferentes y no contiene valores nulos.

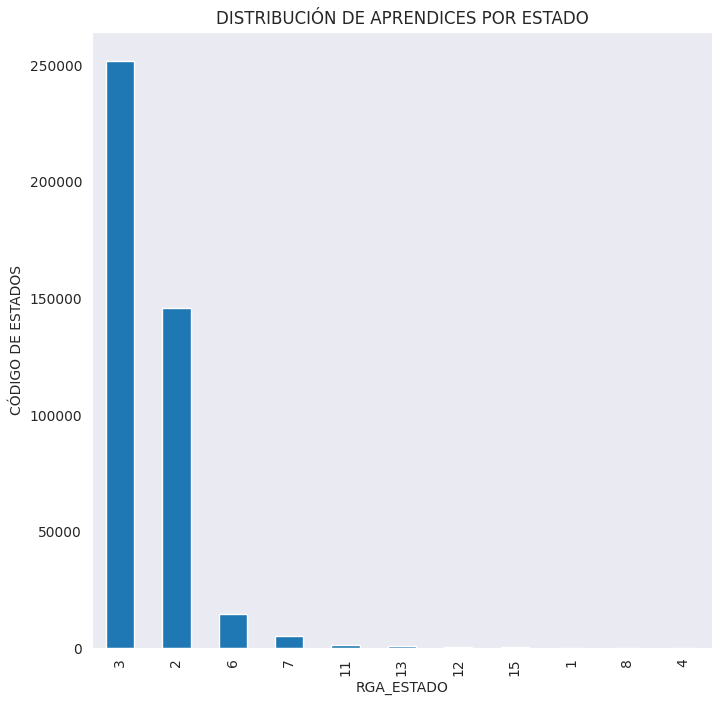
FIC\_ID: No contiene valores nulos y hay 332398 valores diferentes. Los FIC\_id con más registros corresponden a 174841 con 53 registros y 122664 con 45 registros, hay 274647 FIC\_ID con un solo registros.

PRF\_ID: No contiene valores nulos y hay 20801 valores diferentes. Los PRF\_ID con más registros corresponden a 74524 con 8035 registros y 73805 con 7943 registros.

RTA\_ID: Se observan 331154 valores diferentes y 1826 sin valor.



\*\* RGA\_ETAPA\_RUTA: contiene 3 valores diferentes, L 401682 (más del 95%), P 11590 (cerca del 3%), M 4376 (cerca del 1%)



RGA\_ESTADO: Estado-3 251431 (cerca del 60%), estado-2 145766 (cerca del 35%), estado-6 14425 (cerca del 3.5%), estado-7 4984 (cerca del 1.2%), estado-11 1406 (cerca del 0.34%), estado-13 819 (cerca del 0.2%), estado-12 339 (cerca del 0.08%), estado-15 211 (cerca del 0.05%), estado-1 43, estado-8 35 y estado-4 13

RGA\_FCH\_REGISTRO: No contiene valores nulos y hay 324186 valores diferentes. Las fechas se encuentran entre 05/04/2004 y 03/03/2023.

RGA\_FCH\_ULTIMO\_ESTADO: No contiene valores nulos y hay 294285 valores diferentes. Las fechas se encuentran entre 29/11/2004 y 03/03/2023.

NIS\_EMP: se encontraron 5 valores diferentes y 419466 registros sin valor (más del 99%).

RGA\_FCH\_PASO\_A\_PRO: se encontraron 14883 valores diferentes y 403517 registros sin valor (cerca del 96%). Las fechas se encuentran entre 08/03/2010 y 03/03/2023.

RGA\_FCH\_FIN\_FORMACION: se encontraron 222869 valores diferentes y 167639 registros sin valor (cerca del 40%). Las fechas se encuentran entre 11/02/2010 y 03/03/2023.

RGA\_APROBAR\_GEN\_CERTIFICADO: se encontraron 1341 registros con valor 1 y 418131 registros sin valor (99.6%).

Interpretación y presentación de resultados:

Se pudo observar que el archivo estaba compuesto por diez y nueve (19) características numéricas y diez (10) categóricas.

En este archivo, la cantidad de registros con género femenino superan en casi 13% al género masculino.

Analizando la característica COD\_PAIS\_RESIDENCIA más del 99% de los datos tenían el valor 57, que corresponde al código de país Colombia.

Así mismo, se encontraron características con alto porcentaje de valores nulos como PUNTAJE\_ICFES, SISBEN, DBU\_HIJOS, COD\_BARRIO\_RESIDENCIA y COD\_ZONA\_RESIDENCIA, NIS\_EMP, RGA\_FCH\_PASO\_A\_PRO, RGA\_FCH\_FIN\_FORMACION y RGA\_APROBAR\_GEN\_CERTIFICADO.

En este archivo se encontraron que cerca del 60% de los registros corresponde a aprendices que se certificaron en sus formaciones.

¿? Valdría la pena comparar algunos resultados de este archivo contra el archivo anterior?

Como conclusión de este segundo acercamiento a los datos se pudo determinar que el número de registros era alto, correspondía a una muestra representativa de la población y aumentaron las características, lo cual permite obtener más información. Sin embargo, nuevamente se pudo observar que el archivo contenía datos de aprendices en formaciones de titulada y complementaria, por esa razón y al querer analizar sólo la deserción en la formación titulada, los resultados iniciales podrían estar sesgados y no serían concluyentes.

Tercer acercamiento a los datos:

Luego de observar los resultados del análisis sobre el segundo archivo, la líder del proyecto SENA, determina que se debe obtener un tercer archivo con los aprendices matriculados en la primera oferta Titulada Presencial 2018 (0ferta 724) para tener una población homogénea, archivo que se recibe el 10 de marzo de 2023.

El objetivo del trabajo con este archivo era conocer si la estructura de los datos, sus características, posibles valores permitía identificar alguna interpretación con los estados asociados a la deserción en el SENA.

En ese sentido, la metodología de tratamiento de datos se acotó a las etapas de recopilación de datos, análisis exploratorio de datos e Interpretación y presentación de resultados.

Recopilación de datos:

En marzo de 2023 se recibe el archivo, “230310\_formaciones\_oferta\_724.csv”

Este archivo contiene 39.102 instancias (registros) con 47 características.

Las características observadas fueron:

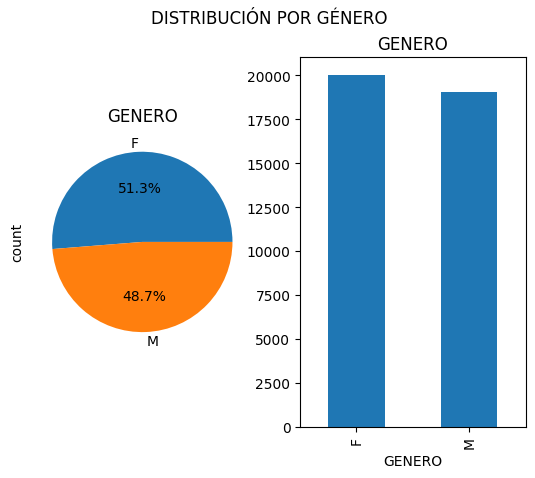
|  |  |
| --- | --- |
| CARACTERÍSTICA | DESCRIPCIÓN |
| Nis | Número interno Sena que identifica de manera única a cada persona registrada en el Sistema Sofía Plus |
| Genero | Variable carácter con valores F o M, F-Femenino: M-Masculino |
| Fecha\_nacimiento | Fecha de nacimiento del aprendiz en formato DD/MM/AAAA |
| Estado\_civil | Variable carácter con valores S-Soltero, U-Unión Libre, C-Casado, V-Viudo, E-Separado, I-Indiferente, N-No Ingresado |
| Estrato | Estrato socioeconómico con valores enteros de 1 a 6 |
| Puntaje\_icfes | Resultado de la prueba Saber 11 |
| Nivel\_sisben | Puntaje Sisbén del núcleo familiar |
| Dbu\_hijos | Número de hijos del aprendiz al momento del registro |
| Cod\_pais\_resisencia | Código numérico que corresponde al país de residencia del aprendiz |
| Pais\_residencia | Nombre de país de residencia del aprendiz |
| Cod\_dpto\_residencia | Código numérico que corresponde al departamento de residencia del aprendiz |
| Dpto\_residencia | Nombre del departamento de residencia del aprendiz |
| Cod\_municipio\_residencia | Código numérico que corresponde al municipio de residencia del aprendiz |
| Municipio\_residencia | Nombre del municipio de residencia del aprendiz |
| Cod\_zona\_residencia | Código numérico que corresponde a la zona de residencia del aprendiz |
| Cod\_barrio\_residencia | Código numérico que corresponde al barrio de residencia del aprendiz |
| Codigo\_estado | Variable numérica que indica el estado del aprendiz en su proceso de formación académica, tiene los valores 1 – Aplazado, 2 – Cancelado, 3 – Certificado, 4 – Condicionado, 5 – Reprobado, 6 - Retiro Voluntario, 7 – Formación, 8 – Inducción, 9 - No Apto, 10 – Reingresado, 11 – Traslado, 12 - Por Certificar, 13 - En Transito, 14 - Cancelamiento Virtual Complementaria y 15 - Deserción Virtual Complementaria |
| Estado\_aprendiz | Nombre del estado de formación del aprendiz |
| Cod\_nivel\_formacion | Código del nivel de formación, con valores 1:'auxiliar', 2:'técnico', 3:'ingeniero técnico', 4:'técnico profesional', 5:'especialización técnica', 6:'tecnólogo', 7:'especialización tecnológica', 8:'curso especial', 9:'evento de divulgación tecnológica', 10:'operario', 11:'complementaria virtual', 12:'ocupacion', 13:'técnico laboral', 100:'c.a.p. trab. calificado', 101:'diploma técnico', 102:'cert.form.esp.oficio', 103:'cert.especializacion', 104:'diploma técnico profesional', 105:'diploma tecnólogo', 107:'c.a.p. técnico', 108:'trabajador calificado', 109:'ayudante', 200:'tecnólogo interinstitucional', 201:'técnico profesional interinstitucional', 202:'formación interinstitucional', 223:'profundización técnica' |
| Nivel\_formacion | Nombre del nivel de formación del curso |
| Modalidad | Texto que corresponde a la Modalidad de formación |
| Cod\_programa | Valor numérico que corresponde al número consecutivo del programa de formación |
| Nombre\_programa | Nombre del programa de formación |
| Codigo\_red\_conocimiento | Código numérico de la red de conocimiento |
| Nom\_red\_conocimiento | Nombre de la red de conocimiento |
| Cod\_linea\_tecnológica | Código de la línea tecnológica, con valores 1:Tecnologías de la información y las comunicaciones gestión de la información, 2:Diseño, 3:Producción y transformación, 4:Materiales herramientas, 5:Cliente |
| Nom\_linea\_tecnológica | Nombre de la línea tecnológica |
| Código\_tecnología\_de\_red | Código de la tecnología de red, con valores 1: Tecnologías de la información, diseño y desarrollo de software, 2: Tecnologías de la información y las comunicaciones, 3: Diseño de joyas y objetos artesanales, 4: Diseño de modas y confecciones, 5: Diseño de producto, 6: Diseño de máquinas y equipos automatizados, 7: Explotación y transformación de minerales, 8: Tecnologías agrícolas, 9: Tecnologías de mantenimiento predictivo, preventivo y correctivo, 10: Tecnologías de producción industrial, 11: Tecnologías de producción limpia, 12: Tecnologías pecuarias, 13: Tecnologías agroindustriales, 14: Tecnologías de producción acuícola, 15: Biotecnología animal, 16: Biotecnología industrial, 17: Biotecnología vegetal, 18: Materiales para la construcción, 19: Materiales para la industria, 20: Investigación, planeación y desarrollo de mercados y medios de la comunicación comercial, 21: Ventas y comercialización, 22: Logística y transporte, 23: Tecnologías de servicios turísticos, 24: Tecnologías de servicios de salud, 25: Tecnologías de gestión administrativa y servicios financieros,  147: Red institucional de enseñanza de idiomas, 167: Pedagogía, 187: Red institucional de integralidad de la formación |
| Nombre\_tecnología\_de\_red | Nombre de la tecnología de red |
| Ficha\_num\_curso | Valor numérico que corresponde al número consecutivo del curso |
| Código\_regional | Código asociado a la regional en el SENA |
| Nombre\_regional | Nombre de la regional |
| Código\_centro | Código asociado al centro o sede de formación en el SENA |
| Sede | Nombre del centro o sede de formación |
| Edad | Edad en años del aprendiz al momento de hacer la extracción del archivo |
| Código\_tipo\_población | Código asociado al tipo de población al que corresponde el aprendiz en el SENA |
| Nombre\_tipo\_población | Nombre del tipo de población al que corresponde el aprendiz en el SENA |
| Apoyo\_sostnmto\_2018 | Valor numérico que indica 0-no recibió apoyo de sostenimiento en el año 2018, 1-si recibió apoyo |
| Apoyo\_sostnmto\_2019 | Valor numérico que indica 0-no recibió apoyo de sostenimiento en el año 2019, 1-si recibió apoyo |
| Apoyo\_sostnmto\_2020 | Valor numérico que indica 0-no recibió apoyo de sostenimiento en el año 2020, 1-si recibió apoyo |
| Apoyo\_sostnmto\_2021 | Valor numérico que indica 0-no recibió apoyo de sostenimiento en el año 2021, 1-si recibió apoyo |
| Apoyo\_sostnmto\_2022 | Valor numérico que indica 0-no recibió apoyo de sostenimiento en el año 2022, 1-si recibió apoyo |
| Apoyo\_sostnmto\_2023 | Valor numérico que indica 0-no recibió apoyo de sostenimiento en el año 2023, 1-si recibió apoyo |
| Apoyo\_fic\_2018 | Valor numérico que indica 0-no recibió apoyo FIC en el año 2018, 1-si recibió apoyo |
| Apoyo\_fic\_2019 | Valor numérico que indica 0-no recibió apoyo FIC en el año 2019, 1-si recibió apoyo |
| Apoyo\_fic\_2020 | Valor numérico que indica 0-no recibió apoyo FIC en el año 2020, 1-si recibió apoyo |
| Apoyo\_fic\_2021 | Valor numérico que indica 0-no recibió apoyo FIC en el año 2021, 1-si recibió apoyo |

Análisis exploratorio: no exploratorio

Algunas observaciones sobre los datos:

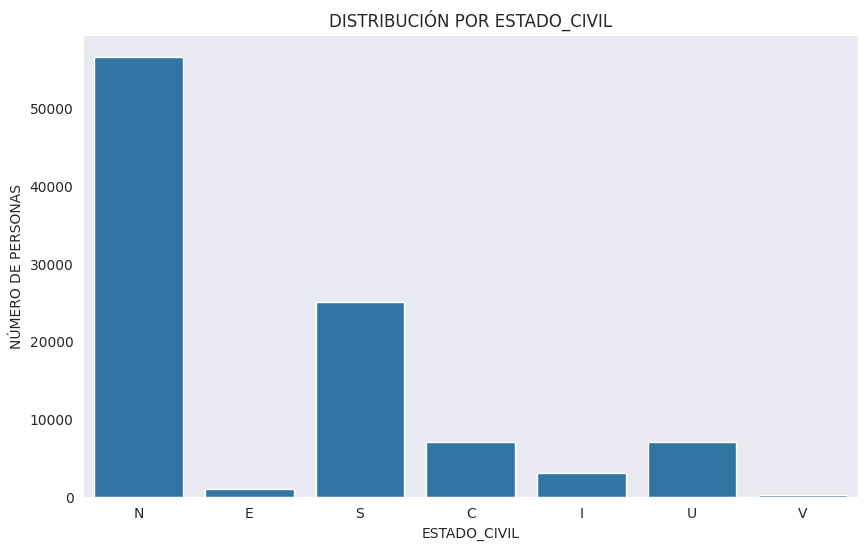


NIS: Se observa que no hay valores nulos y se encuentran 38054 valores diferentes, entendiendo que se trata del identificador único del aprendiz, se aprecia que hay aprendices con varias formaciones, como los NIS 26206428, 26160318, 26104737, 25775421, 26039723 y 24361021, tienen 4 registros de formación.

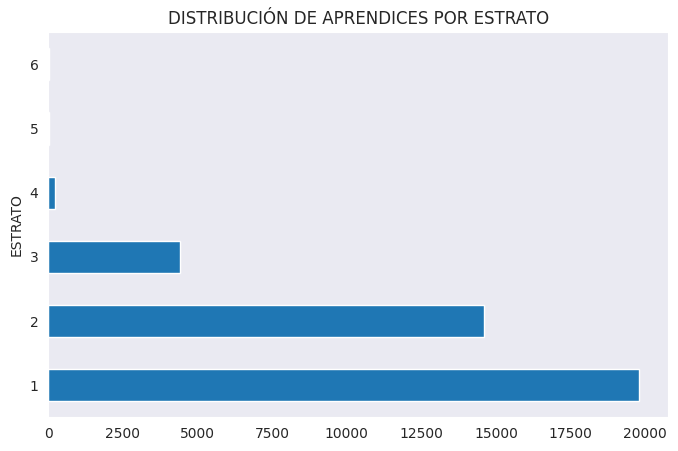


GENERO: Masculino 19058 (48.7%), Femenino 20044 (51.2%)

FECHA\_NACIMIENTO: Las fechas oscilaban entre 1.925 y 2.008, se encuentran registros de 74 años diferentes, cerca del 39% de los aprendices nacieron entre 1.997 y 2.000.



ESTADO\_CIVIL: S-Soltero 20606 (cerca del 52%), U-Unión Libre 4310 (cerca del 11%), C-Casado 2232 (cerca del 6%), V-Viudo 57 (cerca del 0.15%), E-Separado 317 (cerca del 0.8%), I-Indiferente 2562 (cerca del 6.6%), N-No Ingresado 9016 (cerca del 23%)



ESTRATO: Estrato-1 19781 (50.6%), estrato-2 14611 (37.4%), estrato-3 4422 (11.3%), estrato-4 238 (0.6%), estrato-5 38 (0.09%), estrato-6 12 (cerca del 0.03%)

PUNTAJE\_ICFES: 37196 no tenían este valor (cerca del 95%), los puntajes se encuentran entre 1 y 999

SISBEN: Se encuentran 5 valores para 6997 de los registros, 1-3205, 2-2138, 3-523, 4-4, 6-1127. Para 32105 registros (82%) esta característica no tenía valor.

DBU\_HIJOS: 94 registros tienen valor 0, 14 tienen valor 1, 7 tienen valor 2 y 38987 no tenían valor (99.7%)

COD\_PAIS\_RESIDENCIA: Se encuentran 15 valores únicos para esta característica, 39078 registros (99.9%), tenían el valor 57.

PAIS\_RESIDENCIA: Se encuentran 15 valores únicos para esta característica, 39078 registros (99.9%), tenían el valor COLOMBIA. Hay una correspondencia con la característica COD\_PAIS\_RESIDENCIA porque el código 57 corresponde al país COLOMBIA.

COD\_DPTO\_RESIDENCIA: Se encontraron 53 valores únicos, cerca del 34% de los datos se concentran en dos códigos, así, 9208 (más del 23.5%) corresponden al código 57011, 3953 (cerca del 10%) corresponden al código 57005.

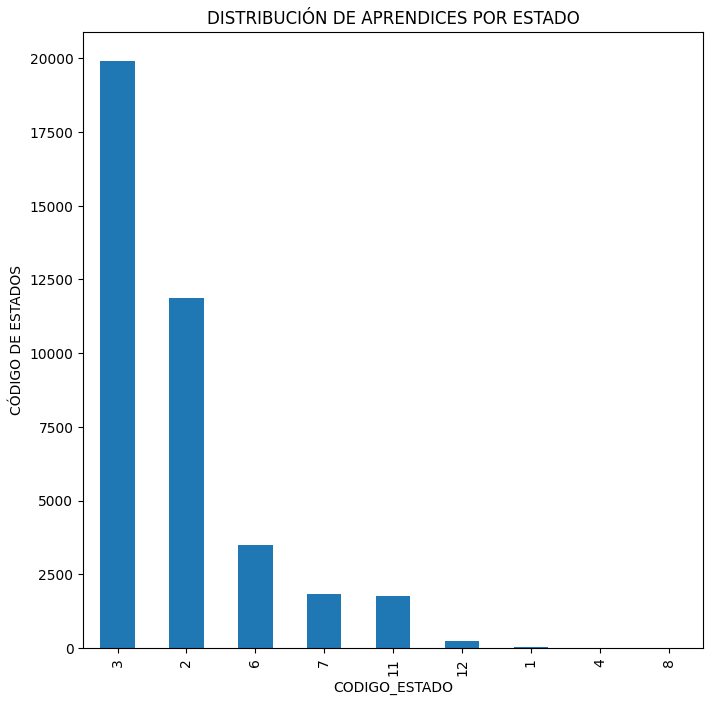
DPTO\_RESIDENCIA: Se encontraron 53 valores únicos, cerca del 34% de los datos se concentran en dos departamentos, así, 9208 (más del 23.5%) corresponden a BOGOTÁ D.C., 3953 (cerca del 10%) corresponden al departamento de ANTIOQUIA.

COD\_MUNICIPIO\_RESIDENCIA: Se encontraron 900 valores únicos, cerca del 30% de los datos se concentran en tres códigos, así, 9208 (más del 23%) corresponden al código 57011001, 1702 (cerca del 4.4%) corresponden al código 57005001 y 1068 (cerca del 3%) corresponden al código 57013001.

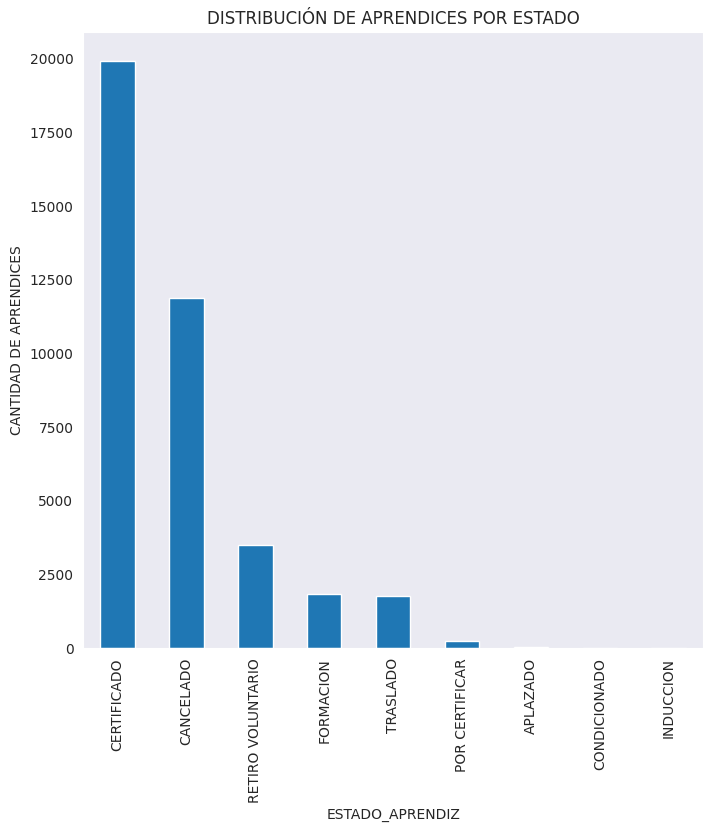
MUNICIPIO\_RESIDENCIA: Se encontraron 900 valores únicos, cerca del 30% de los datos se concentran en tres municipios, así, 9208 (más del 23%) corresponden a BOGOTÁ, 1702 (cerca del 4.4%) corresponden a MEDELLÍN y 1068 (cerca del 3%) corresponden a CARTAGENA.

COD\_ZONA\_RESIDENCIA: Se encontraron 274 valores únicos, sin embargo 36297 (cerca del 93%) no traen valor.

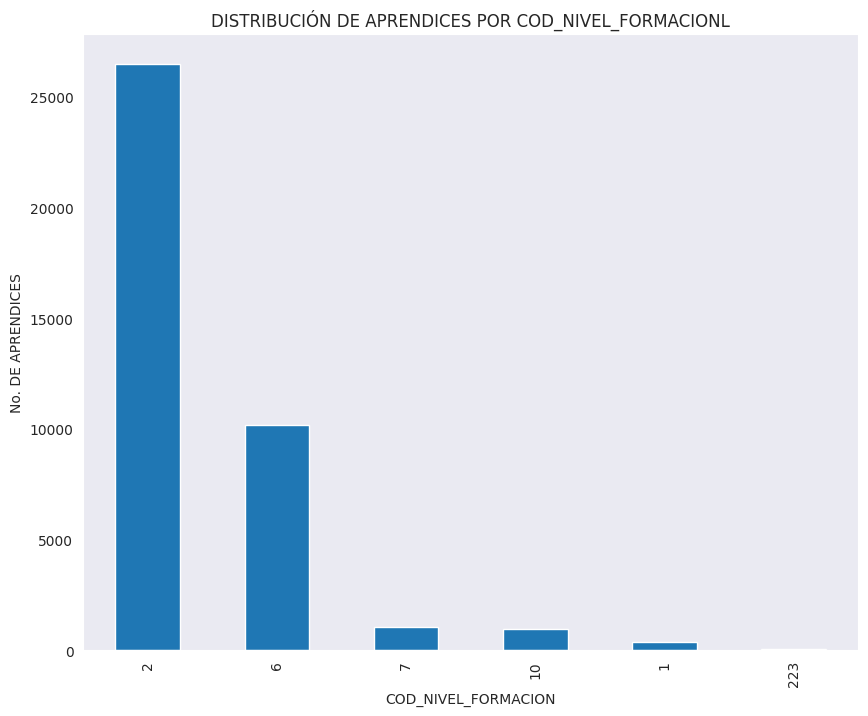
COD\_BARRIO\_RESIDENCIA: Se encontraron 1246 valores diferentes, sin embargo 36658 (cerca del 94%) no traen valor.



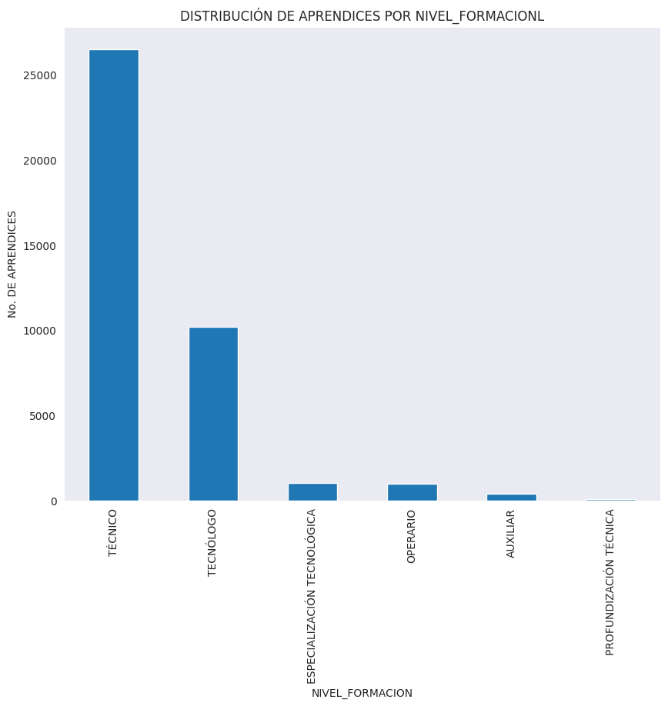
CODIGO\_ESTADO: Estado-3 19896 (cerca del 51%), estado-2 11859 (cerca del 30%), estado-6 3487 (cerca del 9%), estado-7 1819 (cerca del 5%), estado-11 1772 (cerca del 4.5%), estado-12 226 (cerca del 0.6%), estado-1 18 (cerca del 0.05%), estado-4 15 (cerca del 0.03%) y estado-8 10 (cerca del 0.2%)



ESTADO\_APRENDIZ: Aparece el estado CERTIFICADO 19896 (cerca del 51%), CANCELADO (cerca del 30%), RETIRO VOLUNTARIO 3487 (cerca del 9%), FORMACION 1819 (cerca del 5%), TRASLADO 1772 (cerca del 4.5%), POR CERTIFICAR 226 (cerca del 0.6%), APLAZADO 18 (cerca del 0.05%), CONDICIONADO 15 (cerca del 0.03%) y INDUCCION 10 (cerca del 0.2%)



COD\_NIVEL\_FORMACION: Se encontraron 6 niveles de formación, nivel 2 26480 (más del 67%), nivel 6 10162 (cerca del 26%), nivel 7 1038 (cerca del 2.65%), nivel 10 978 (cerca del 2.5%), nivel 1 379 (cerca del 0.96%) y nivel 223 65 (cerca del 1.16%).



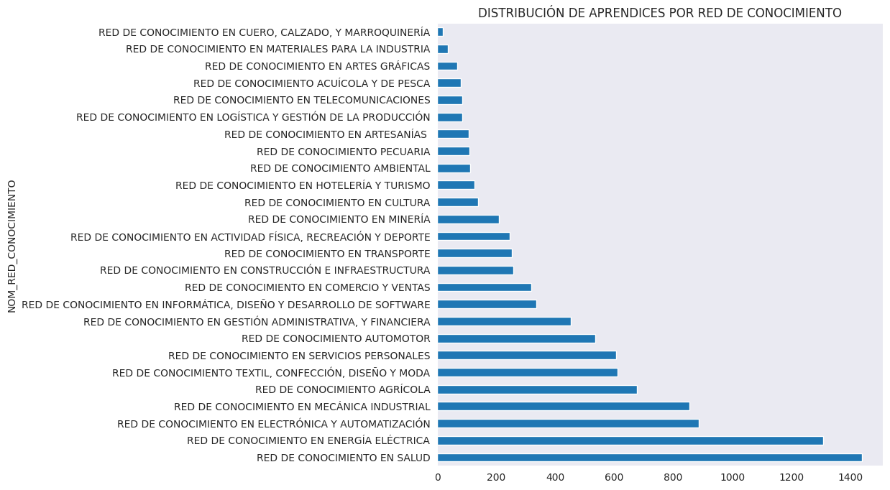
NIVEL\_FORMACION: Se encontraron 6 niveles de formación, TÉCNICO 26480 (más del 67%), TECNÓLOGO 10162 (cerca del 26%), ESPECIALIZACIÓN TECNOLÓGICA 1038 (cerca del 2.65%), OPERARIO 10 978 (cerca del 2.5%), AUXILIAR 379 (cerca del 0.96%) y PROFUNDIZACIÓN TÉCNICA 223 65 (cerca del 1.16%).

MODALIDAD: Todos los registros tenían el mismo valor PRESENCIAL, esto tiene sentido porque se trata de aprendices matriculados en la primera oferta Titulada Presencial 2018 (0ferta 724)

COD\_PROGRAMA: Se encontraron 215 códigos diferentes. Más del 20% de los aprendices está matriculados en los programas con códigos 122121 2361 (cerca del 6%), 228185 2341 (cerca del 6%), 112005 1478 (cerca del 3.8%) y 133146 1367 (cerca del 3.5%).

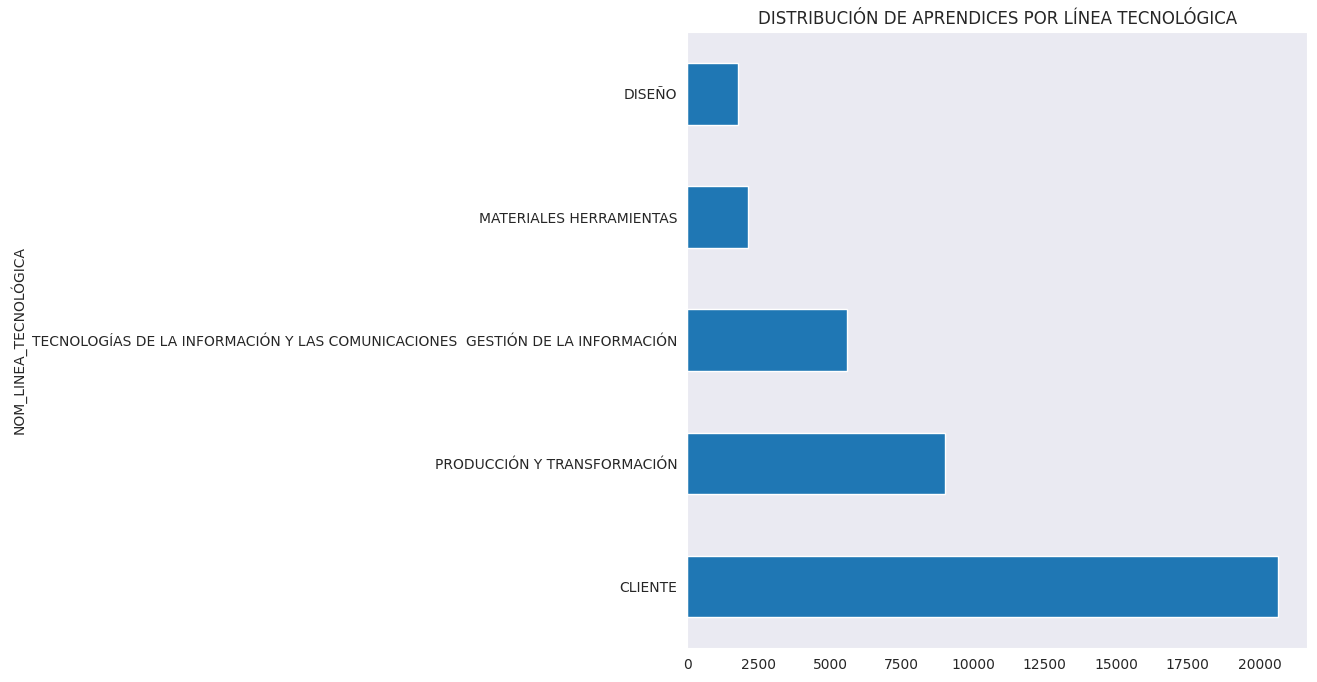
NOMBRE\_PROGRAMA: Se encontraron 215 códigos diferentes. Más del 20% de los aprendices está matriculados en los programas ASISTENCIA ADMINISTRATIVA 2361 (cerca del 6%), SISTEMAS 2341 (cerca del 6%), GESTIÓN DEL TALENTO HUMANO 1478 (cerca del 3.8%) y CONTABILIZACIÓN DE OPERACIONES COMERCIALES 1367 (cerca del 3.5%).

CODIGO\_RED\_CONOCIMIENTO: Se encontraron 26 códigos diferentes. 27.419 registros no traen valor (cerca del 70%). Los dos códigos que más se presentan son el código 26 con 1438 registros (3.7%) y el código 9 con 1308 registros (3.3%).



NOM\_RED\_CONOCIMIENTO: Se encontraron 26 valores diferentes. 27.419 registros no traen valor (cerca del 70%). Los valores que más se presentan son RED DE CONOCIMIENTO EN SALUD con 1438 registros (3.7%) y el RED DE CONOCIMIENTO EN ENERGÍA ELÉCTRICA con 1308 registros (3.3%).

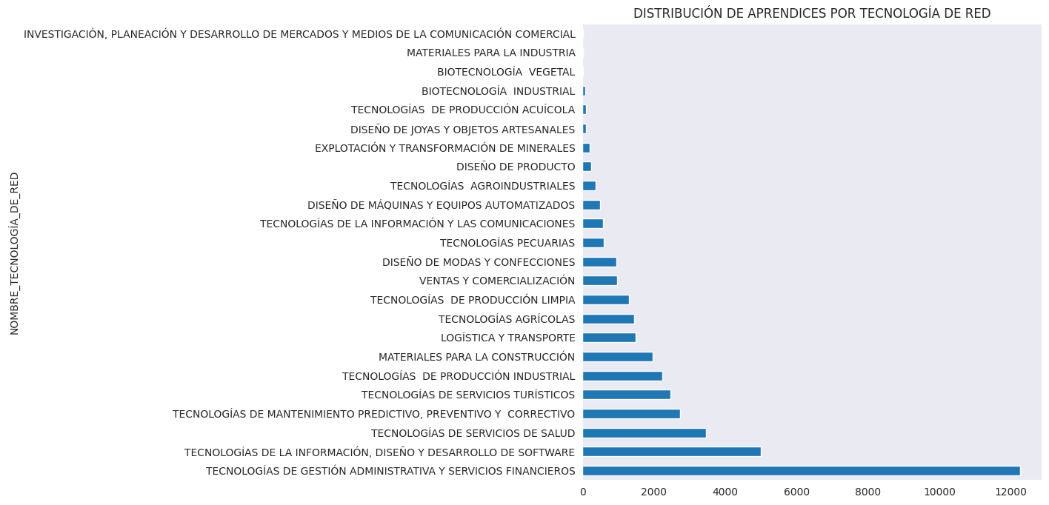
COD\_LINEA\_TECNOLÓGICA: Se encontraron 5 valores diferentes. Código 5 20.643 registros (52.7%), código 3 9.012 registros (23%), código 1 5.568 registros (14.2%), código 4 2.105 registros (5.4%) y código 2 1.774 registros (4.5%)



NOM\_LINEA\_TECNOLÓGICA: Se encontraron 5 valores diferentes. CLIENTE 20.643 registros (52.7%), PRODUCCIÓN Y TRANSFORMACIÓN 9.012 registros (23%), TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES GESTIÓN DE INFORMACIÓN 5.568 registros (14.2%), MATERIALES HERRAMIENTAS 2.105 registros (5.4%) y DISEÑO 1.774 registros (4.5%)

CÓDIGO\_TECNOLOGÍA\_DE\_RED: Se encontraron 24 valores diferentes. Cerca del 53% de los registros se encontraron en los códigos 25 12.248 registros (31.3%), 1 4.992 registros (12.7%), 24 3.467 registros (8.8%).

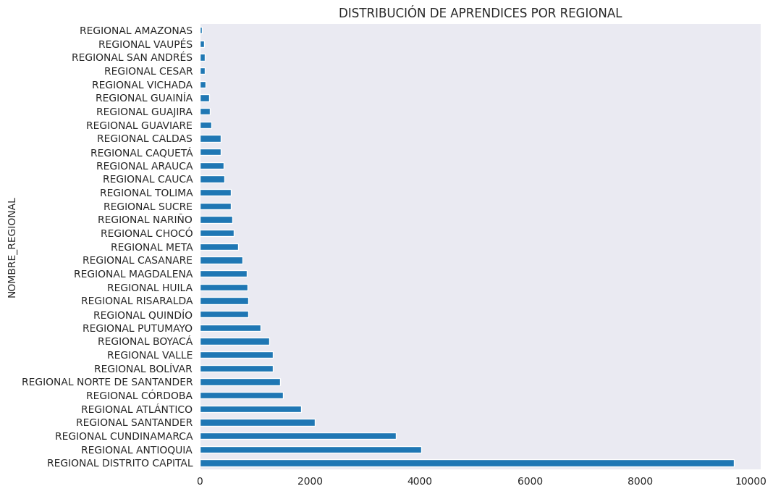




NOMBRE\_TECNOLOGÍA\_DE\_RED: Se encontraron 24 valores diferentes. Cerca del 53% de los registros se encontraron en las TECNOLOGÍAS DE GESTIÓN ADMINISTRATIVA Y SERVICIOS FINANCIEROS 12.248 registros (31.3%), TECNOLOGÍAS DE LA INFORMACIÓN, DISEÑO Y DESARROLLO DE SOFTWARE 4.992 registros (12.7%), TECNOLOGÍAS DE SERVICIOS DE SALUD 3.467 registros (8.8%).

FICHA\_NUM\_CURS: Se observaron 1381 números de curso diferentes, la ficha con más apariciones es la 1581575 con 216 registros, seguida de la 1581582 con 211 registros. Hay 4 fichas con el menor número de registros (3 en cada una).

CÓDIGO\_REGIONAL: Se observaron 33 códigos diferentes. Cuatro códigos de regional concentran casi el 50% de los registros, el código 11 con 9699 (24.8%), código 5 con 1024 (10.3%), código 25 con 3557 registros (9%) y código 68 con 2088 registros (5.3%).

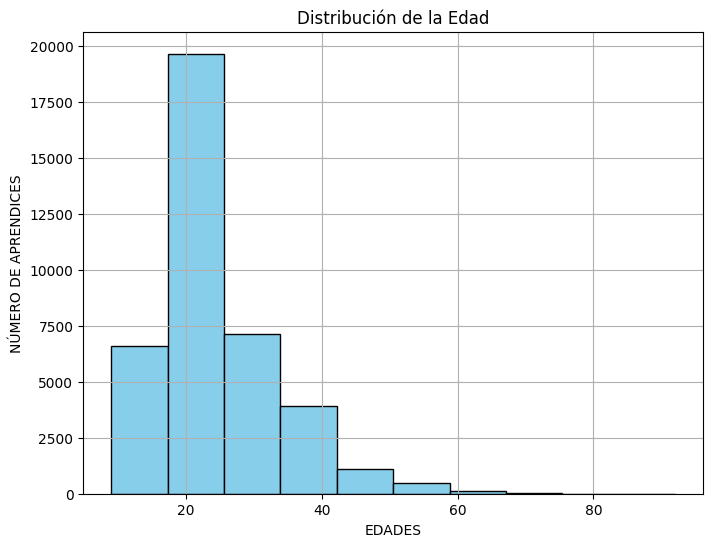


NOMBRE\_REGIONAL: Se observaron 33 regionales. Cuatro de las regionales concentran casi el 50% de los registros, REGIONAL DISTRITO CAPITA con 9699 (24.8%), REGIONAL ANTIOQUIA con 1024 (10.3%), REGIONAL CUNDINAMARCA con 3557 registros (9%) y REGIONAL SANTANDER con 2088 registros (5.3%).

La regional con menor cantidad de apariciones es REGIONAL AMAZONAS, con 36 registros.

CÓDIGO\_CENTRO: Se encontraron 110 códigos. 5 códigos concentran el 20% de los registros, código 9404 con 2235 registros (5.7%), código 9405 con 2006 registros (5.1%), código 9210 con 1442 registros (3.7%), código 9518 con 1103 registros (2.8%) y código 9512 con 984 registros (2.5%)

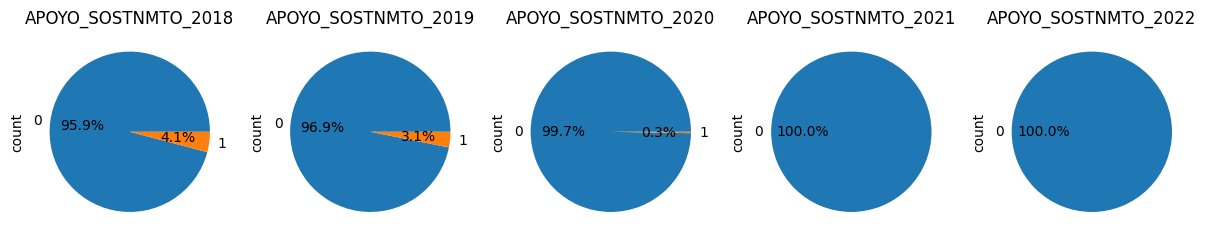
SEDE: Se encontraron 103 sedes, 5 de ellas concentran el 23% de los registros, CENTRO DE GESTION ADMINISTRATIVA con 2235 registros (5.7%), CENTRO DE COMERCIO Y SERVICIOS con 2211 registros (5.65%), código CENTRO DE SERVICIOS FINANCIEROS con 2006 registros (5.1%), CENTRO DE ELECTRICIDAD, ELECTRÓNICA Y TELECOMUNICACIONES con 1442 registros (3.7%) y CENTRO DE LA INDUSTRIA, LA EMPRESA Y LOS SERVICIOS con 1127 registros (2.9%).



EDAD: Se encuentran 76 valores diferentes, todos los registros tienen un valor. Cerca del 40% de los registros tienen edades entre los 17 y 21 años, así, 17 4909 (12.5%), 18 4447 (11.3%), 19 3370 (8.6%), 20 2770 (7.1%) y 21 2297 (5.9%).

CÓDIGO\_TIPO\_POBLACIÓN: Se encontraron 41 diferentes valores, siendo el código 34 con 22518 registros (57.6%) el que más se repite, seguido del código 3 con 4291 registros (11%) y código 8 con 3760 registros (9.6%)

NOMBRE\_TIPO\_POBLACIÓN: Se encontraron 41 diferentes valores, siendo NINGUNA con 22518 registros (57.6%) el que más se repite, seguido de DESPLAZADOS POR LA VIOLENCIA con 4291 registros (11%) y JOVENES VULNERABLES con 3760 registros (9.6%).



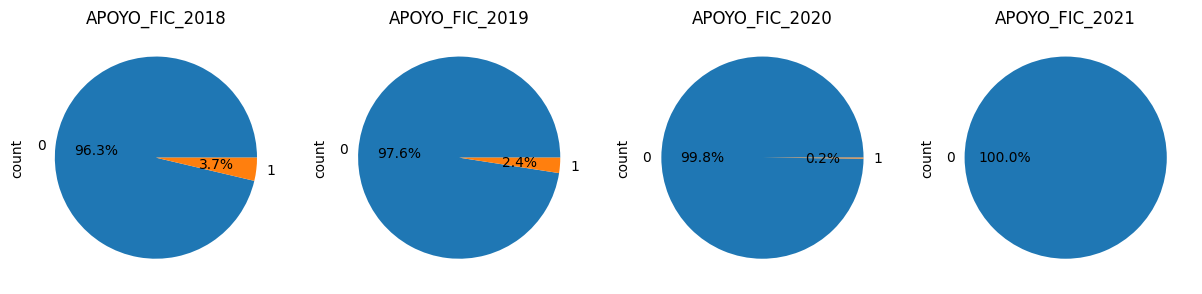
APOYO\_SOSTNMTO\_2018: Se encontraron 37502 registros (95.9%) con valor 0 y 1600 registros con valor 1 (4.1%)

APOYO\_SOSTNMTO\_2019: Se encontraron 37878 registros (96.9%) con valor 0 y 1224 registros con valor 1 (3.1%)

APOYO\_SOSTNMTO\_2020: Se encontraron 38993 registros (99.7%) con valor 0 y 109 registros con valor 1 (0.3%)

APOYO\_SOSTNMTO\_2021: Se encontraron 39102 registros (100%) con valor 0.

APOYO\_SOSTNMTO\_2022: Se encontraron 39102 registros (100%) con valor 0.



APOYO\_FIC\_2018: Se encontraron 37659 registros (96.3%) con valor 0 y 1443 registros con valor 1 (3.7%)

APOYO\_FIC\_2019: Se encontraron 38146 registros (97.6%) con valor 0 y 956 registros con valor 1 (2.4%)

APOYO\_FIC\_2020: Se encontraron 39043 registros (99.8%) con valor 0 y 59 registros con valor 1 (0.2%)

APOYO\_FIC\_2021: Se encontraron 39102 registros (100%) con valor 0.

Como conclusión de este tercer acercamiento a los datos, se pudo determinar que el número de registros correspondía a la oferta académica indicada, correspondía a todos los aprendices matriculados y aumentaron las características, desechando algunas observadas en los archivos anteriores.

Hay varios sedes o centros con el mismo nombre, pero diferente código

Referencias:

Carlos Leiva Vigorena. Metodología de tratamiento de datos. https://es.linkedin.com/pulse/metodolog%C3%ADa-de-tratamiento-datos-carlos-leiva-vigorena-io2ue