PRACTICA 2

De laboratorio 'Programación en python'

Práctica de Laboratorio - Unidad 2: Introducción a la Limpieza de

Datos Materia: Introducción a la Ciencia de Datos

Unidad 2: Procesamiento y Limpieza de Datos

Título de la Práctica: Limpieza de una Base de Datos Ensuciada

Nombre: Saldaña Morales Alexa Esmeralda

Día y horario de la materia: lunes 9:00 a 10:00

martes y jueves 8:00 a 10:00

Objetivo:

Desarrollar habilidades en el preprocesamiento de datos, incluyendo la identificación y tratamiento de valores faltantes, datos duplicados, y formatos inconsistentes en una base de datos.

Instrucciones:

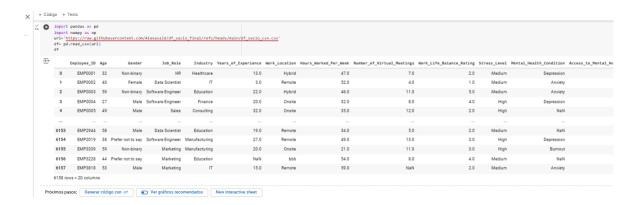
1. Recepción de la Base de Datos

Descargue el archivo de la base de datos "ensuciada" proporcionado por el profesor. Este archivo ha sido alterado con errores comunes, como valores duplicados, valores faltantes (NaNs), y errores en el formato de algunas columnas.

2. Análisis Inicial de la Base de Datos

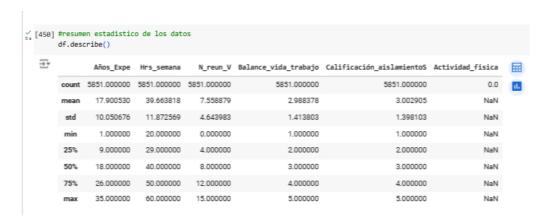
Antes de comenzar a limpiar la base de datos, debe realizar un análisis preliminar para comprender la naturaleza y distribución de los errores.

Visualización de la base de datos inicial.
 Podemos notar que tenemos un total de 6158 datos, pero existen columnas que mas adelante no serán tan necesarias.



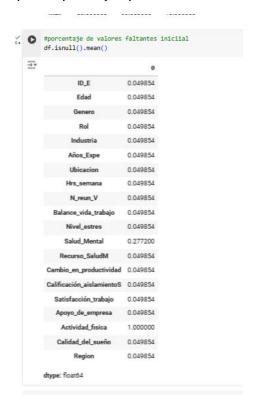
Resumen estadístico de los datos.

con el comando de describe(), podemos visualizar la tabla con el conteo completo en casi todas las columnas excepto en la de actividad física. También se muestran la falta de columnas ya que el tipo de dato no concuerda con ser dato numérico por ello más adelante tendremos que modificar estos datos.



• Porcentaje de valores faltantes por columna

El siguiente comando. isnull().mean(), muestra la tabla de porcentajes de cada columna, aquí podemos interpretar que en su mayoría los porcentajes de datos faltantes son del 4.9854%, la de salud mental es de 27.72% y la de actividad física de un 100% lo cual es preocupante ya que sus datos son nulos.



Existencia duplicados

Con el comando. duplicated().sum Se muestra el conteo total de filas duplicadas

```
os [452] #Contar filas duplicadas
df.duplicated().sum()

145
```

• Tipo de dato

En la siguiente tabla observamos lo mencionado anteriormente, los datos están definidos de manera incorrecta lo cual nos provocara algunos conflictos para el análisis

```
√
0s [333] #tipo de datos
                  df1.info()
        <class 'pandas.core.frame.DataFrame'>
                   Index: 6013 entries, 0 to 6157
                  Data columns (total 20 columns):
# Column
                                                                                               Non-Null Count Dtype
                                                                                           5706 non-null object
5707 non-null object
5708 non-null object
                    0 ID_E
1 Edad
                    2 Genero

        2
        Genero
        5708 non-null

        3
        Rol
        5706 non-null

        4
        Industria
        5706 non-null

        5
        Años_Expe
        5706 non-null

        6
        Ubicacion
        5707 non-null

        7
        Hrs_semana
        5706 non-null

        8
        N_reun_V
        5707 non-null

        9
        Balance_vida_trabajo
        5707 non-null

        10
        Nivel_estres
        5706 non-null

        11
        Salud_Mental
        4340 non-null

                                                                                                                                   object
                                                                                                                                   object
                                                                                                                                   float64
                                                                                                                                  float64
object
                dtypes: float64(6), object(14)
                   memory usage: 986.5+ KB
```

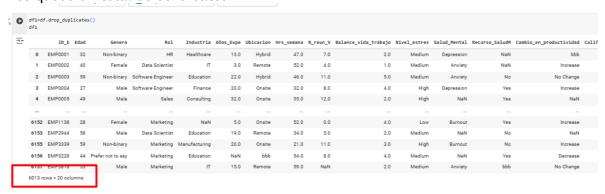
Ahora al conocer un poco mas los datos con los que se trabajara y los detalles que se tendrán que tomar en cuenta, iniciare con el proceso de limpieza teniendo en cuenta el nuevo etiquetado de datos y la resolución de los datos faltantes.

Uno de los problemas es que la cantidad de datos faltantes es importante ya que en la siguiente fase de limpieza suelen perderse muchos de los datos, haciendo así la base de datos muy pequeña.

Limpieza de datos.

Eliminación de duplicados

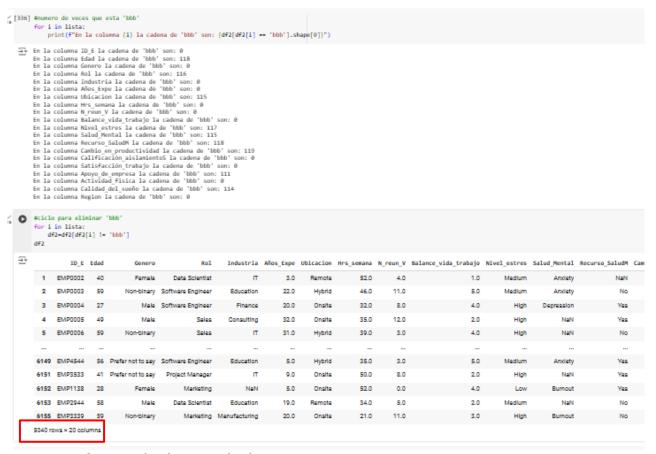
Agregue un nuevo DF1 para no trabajar sobre el mismo, quite los duplicados de todo el DF, anteriormente teníamos un total de 6158 y ahora eliminando duplicados nos quedo un total de 6013 datos



Corrección de datos 'bbb'

Coloque un ciclo para saber cuántos datos había escritos de esa manera, después realice otro ciclo for para eliminar aquellos datos inválidos. Dejando así solo datos NaN y los datos normales para después rellenar los datos vacíos.

Nota (profesor intente convertirlos a datos NaN pero me las cadenas de caracteres no me dejaba colocarlos a datos nulos 22)



Corrección de tipos de datos:

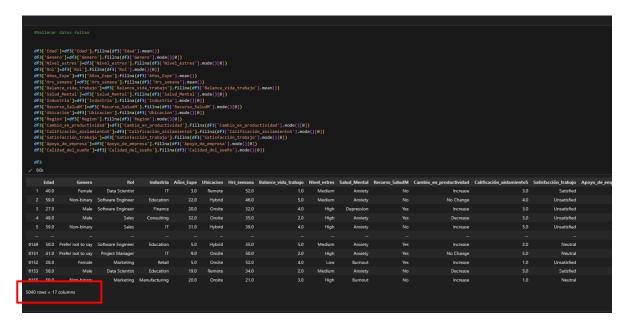
Convertir solamente a datos numéricos aquellos que estaban etiquetados como object para poder usarlos como números y se reconocieran de esa manera.

Como lo vimos anteriormente, la fila de actividad física tenia el 100% de datos faltantes por lo que decidí eliminarla, junto con el ID del empleado lo cual no se hace relevante en el análisis futuro

											r Coolgo T Markdown			
	<pre>####################################</pre>													
	0.0s													
	Eda	d Genero	Rol	Industria	Años_Expe	Ubicacion	Hrs_semana	Balance_vida_trabajo	Nivel_estres	Salud_Mental	Recurso_SaludM	Cambio_en_productividad	Cali	
	1 40.) Female	Data Scientist		3.0	Remote	52.0	1.0	Medium	Anxiety	NaN	Increase		
	2 59.	Non-binary	Software Engineer	Education	22.0	Hybrid	46.0		Medium	Anxiety	No	No Change		
	3 27.) Male	Software Engineer	Finance	20.0	Onsite	32.0	4.0	High	Depression	Yes	Increase		
	4 49.) Male	Sales	Consulting	32.0	Onsite	35.0	2.0	High	NaN	Yes	NaN		
	5 59.	Non-binary	Sales		31.0	Hybrid	39.0	4.0	High	NaN	No	Increase		
614	9 56.	Prefer not to say	Software Engineer	Education	5.0	Hybrid	35.0	5.0	Medium	Anxiety	Yes	Increase		
615	1 41.	Prefer not to say	Project Manager		9.0	Onsite	50.0	2.0	High	NaN	Yes	No Change		
615	2 28.) Female	Marketing	NaN	5.0	Onsite	52.0	4.0	Low	Burnout	Yes	Increase		
615	3 58.) Male	Data Scientist	Education	19.0	Remote	34.0	2.0	Medium	NaN	No	NaN		
615	5 59.	Non-binary	Marketing	Manufacturing	20.0	Onsite	21.0	3.0	High	Burnout	No	Increase		
5040	5040 rows × 17 columns													

Eliminación o imputación de valores faltantes:

Rellene los valores nulos por moda y promedio, así no serían muchos datos eliminados o perdidos. Ya que como se vio en la tabla de porcentajes, era una cantidad considerable de datos faltantes lo que haría que el análisis fuera complicado y no muy preciso.



Resultados.

En conclusión, la base de datos ha sido trabajada en un proceso de limpieza, pudimos ver los conflictos que había en ella. Los datos ahora nos permitirán hacer el análisis correcto, sin embargo el análisis pudiera no llegar a ser 100% confiable ya que en el transcurso de limpieza siempre llegan a perderse algunos datos.

Tipo de dato final

Tabla de datos faltantes, a comparación de la tabla que se presentó al inicio con porcentajes un tanto altos en cada columna, en la actual podemos observar que no existen datos faltantes.

```
#%de datos faltantes
df3.isnull().mean()

007

Edad 0.0

Genero 0.0

Rol 0.0

Industria 0.0

Años_Expe 0.0

Ubicacion 0.0

Hrs_semana 0.0

Balance_vida_trabajo 0.0

Nivel_estres 0.0

Salud_Mental 0.0

Recurso_SaludM 0.0

Cambio_en_productividad 0.0

Calificación_trabajo 0.0

Satisfacción_trabajo 0.0

Apoyo_de_empresa 0.0

Apoyo_de_empresa 0.0

Calidad_del_sueño 0.0

Region 0.0

Region 0.0

dtype: float64
```

Datos inválidos

De igual manera los datos 'bbb' ya se encuentran eliminados de la base de datos