

CÓDIGO DEL PROCESO DE LIMPIEZA DE DATOS

```
pip install pandas

[1] ✓ 3.7s Python

... Requirement already satisfied: pandas in c:\users\gabriela\appdata\local\programs\python\python313\lib\site-packages (2.2.3)
Requirement already satisfied: numpy>=1.26.0 in c:\users\gabriela\appdata\local\programs\python\python313\lib\site-packages (from pandas) (2.1.2)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\gabriela\appdata\local\programs\python\python313\lib\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\gabriela\appdata\local\programs\python\python313\lib\site-packages (from pandas) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in c:\users\gabriela\appdata\local\programs\python\python313\lib\site-packages (from pandas) (2024.2)
Requirement already satisfied: six>=1.5 in c:\users\gabriela\appdata\local\programs\python\python313\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
Note: you may need to restart the kernel to use updated packages.

+ Code + Markdown

Cargar base de datos

import pandas as pd
df = pd.read_csv('https://raw.githubusercontent.com/Gaby14-nt/PROYECTO_CDD/refs/heads/main/df_sucio.csv')
df

[2] ✓ 1.8s Python

... User_ID Age Gender Technology_Usage_Hours Social_Media_Usage_Hours Gaming_Hours Screen_Time_Hours Mental_Health_Status Stress_Level Sleep_Hours Physical_Activity_Hours
0 USER-00001 23.0 Female 6.57 6.0 0.68 12.36 Good Low 8.01
1 USER-00002 21.0 Male 3.01 2.57 3.74 7.61 Poor High 7.28
2 USER-00003 51.0 Male 3.04 6.14 1.26 3.16 Fair High 8.04
3 USER-00004 25.0 Female 3.84 4.48 2.59 13.08 Excellent Medium 5.62
4 USER-00005 53.0 Male 1.2 0.56 0.29 12.63 Good Low 5.55
... ..
12524 USER-06709 24.0 Female 5.08 1.58 1.13 6.58 Poor Low 7.74
12525 USER-09996 42.0 Male 7.05 0.41 0.53 13.90 invalid Medium 7.37
12526 USER-02814 36.0 Male 4.42 0.08 2.20 10.48 Good Medium 6.91
12527 USER-06331 35.0 Female 4.28 0.9 4.90 NaN Fair High 8.51
12528 USER-07333 50.0 Female 11.85 0.4 1.70 6.21 Good High 4.82

12529 rows × 14 columns

df.describe()

[3] ✓ 0.0s Python

... Gaming_Hours Screen_Time_Hours Online_Support_Usage
count 12164.000000 12164.000000 0.0
mean 2.520382 7.976828 NaN
std 1.450498 4.050130 NaN
min 0.000000 1.000000 NaN
25% 1.260000 4.520000 NaN
50% 2.520000 7.920000 NaN
75% 3.810000 11.510000 NaN
max 5.000000 15.000000 NaN

df.columns

[4] ✓ 0.0s Python

... Index(['User_ID', 'Age', 'Gender', 'Technology_Usage_Hours',
'Social_Media_Usage_Hours', 'Gaming_Hours', 'Screen_Time_Hours',
'Mental_Health_Status', 'Stress_Level', 'Sleep_Hours',
'Physical_Activity_Hours', 'Support_Systems_Access',
'Work_Environment_Impact', 'Online_Support_Usage'],
dtype='object')
```

df.info()

[5] ✓ 0.0s Python

...

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12529 entries, 0 to 12528
Data columns (total 14 columns):
Column Non-Null Count Dtype
--- --- --- ---
0 User_ID 12417 non-null object
1 Age 12403 non-null object
2 Gender 12421 non-null object
3 Technology_Usage_Hours 12405 non-null object
4 Social_Media_Usage_Hours 12406 non-null object
5 Gaming_Hours 12164 non-null float64
6 Screen_Time_Hours 12164 non-null float64
7 Mental_Health_Status 12412 non-null object
8 Stress_Level 12397 non-null object
9 Sleep_Hours 12388 non-null object
10 Physical_Activity_Hours 12405 non-null object
11 Support_Systems_Access 12393 non-null object
12 Work_Environment_Impact 12378 non-null object
13 Online_Support_Usage 0 non-null float64
dtypes: float64(3), object(11)
memory usage: 1.3+ MB

df.isnull().sum()

[6] ✓ 0.0s Python

...

User_ID 112
Age 126
Gender 108
Technology_Usage_Hours 124
Social_Media_Usage_Hours 123
Gaming_Hours 365
Screen_Time_Hours 365
Mental_Health_Status 117
Stress_Level 132
Sleep_Hours 141
Physical_Activity_Hours 124
Support_Systems_Access 136
Work_Environment_Impact 151
Online_Support_Usage 12529
dtype: int64

#Elimino columna Online_Support_Usage, todos sus datos son NaN
df=df.drop(columns=['Online_Support_Usage'])
df

[7] ✓ 0.0s Python

...

	User_ID	Age	Gender	Technology_Usage_Hours	Social_Media_Usage_Hours	Gaming_Hours	Screen_Time_Hours	Mental_Health_Status	Stress_Level	Sleep_Hours	Physi
0	USER-00001	23.0	Female	6.57		6.0	0.68	12.36	Good	Low	8.01
1	USER-00002	21.0	Male	3.01	2.57	3.74	7.61	Poor	High	7.28	
2	USER-00003	51.0	Male	3.04	6.14	1.26	3.16	Fair	High	8.04	
3	USER-00004	25.0	Female	3.84	4.48	2.59	13.08	Excellent	Medium	5.62	
4	USER-00005	53.0	Male	1.2	0.56	0.29	12.63	Good	Low	5.55	
...
12524	USER-06709	24.0	Female	5.08	1.58	1.13	6.58	Poor	Low	7.74	
12525	USER-09996	42.0	Male	7.05	0.41	0.53	13.90	invalid	Medium	7.37	
12526	USER-02814	36.0	Male	4.42	0.08	2.20	10.48	Good	Medium	6.91	
12527	USER-06331	35.0	Female	4.28	0.9	4.90	NaN	Fair	High	8.51	
12528	USER-07333	50.0	Female	11.85	0.4	1.70	6.21	Good	High	4.82	

12529 rows × 13 columns

```
df.duplicated()
[38] ✓ 0.0s Python

...
0      False
1      False
2      False
3      False
4      False
...
12524   True
12525   False
12526   True
12527   False
12528   True
Length: 12529, dtype: bool

df=df.drop_duplicates()
df
[39] ✓ 0.0s Python
```

	User_ID	Age	Gender	Technology_Usage_Hours	Social_Media_Usage_Hours	Gaming_Hours	Screen_Time_Hours	Mental_Health_Status	Stress_Level	Sleep_Hours	Physi
0	USER-00001	23.0	Female	6.57	6.0	0.68	12.36	Good	Low	8.01	
1	USER-00002	21.0	Male	3.01	2.57	3.74	7.61	Poor	High	7.28	
2	USER-00003	51.0	Male	3.04	6.14	1.26	3.16	Fair	High	8.04	
3	USER-00004	25.0	Female	3.84	4.48	2.59	13.08	Excellent	Medium	5.62	
4	USER-00005	53.0	Male	1.2	0.56	0.29	12.63	Good	Low	5.55	
...
12519	USER-01052	36.0	Other	8.97	3.02	4.05	6.52	Good	Medium	5.4	
12520	USER-05760	46.0	Male	invalid	2.2	4.71	13.31	Excellent	Medium	6.51	
12523	USER-02010	25.0	Other	3.82	3.71	3.61	6.57	Poor	Low	4.78	
12525	USER-09996	42.0	Male	7.05	0.41	0.53	13.90	invalid	Medium	7.37	
12527	USER-06331	35.0	Female	4.28	0.9	4.90	NaN	Fair	High	8.51	

11386 rows × 13 columns

Renombrar nombres de columnas.

```
df2 = df.rename(columns= {'User_ID': 'Usuario', 'Age': 'Edad', 'Gender': 'Género', 'Technology_Usage_Hours': 'Horas_de_uso_de_la_tecnología', 'Social_Media_Us
df2
[8] ✓ 0.0s Python
```

	Usuario	Edad	Género	Horas_de_uso_de_la_tecnología	Horas_de_uso_de_las_redes_sociales	Horas_de_juego	Horas_de_tiempo_en_pantalla	Estado_de_salud_mental	Nivel_de_estrés	H
0	USER-00001	23.0	Female	6.57	6.0	0.68	12.36	Good	Low	
1	USER-00002	21.0	Male	3.01	2.57	3.74	7.61	Poor	High	
2	USER-00003	51.0	Male	3.04	6.14	1.26	3.16	Fair	High	
3	USER-00004	25.0	Female	3.84	4.48	2.59	13.08	Excellent	Medium	
4	USER-00005	53.0	Male	1.2	0.56	0.29	12.63	Good	Low	
...
12524	USER-06709	24.0	Female	5.08	1.58	1.13	6.58	Poor	Low	
12525	USER-09996	42.0	Male	7.05	0.41	0.53	13.90	invalid	Medium	
12526	USER-02814	36.0	Male	4.42	0.08	2.20	10.48	Good	Medium	
12527	USER-06331	35.0	Female	4.28	0.9	4.90	NaN	Fair	High	
12528	USER-07333	50.0	Female	11.85	0.4	1.70	6.21	Good	High	

12529 rows × 13 columns

Renombrar datos de columnas.

```
#Género
gen={
    'Female': 'Mujer',
    'Male': 'Hombre',
    'Other': 'Otro'
}

#Salud mental
sm={
    'Excellent': 'Excelente',
    'Good': 'Buena',
    'Fair': 'Neutral',
    'Poor': 'Mala'
}

#Nivel de estrés
ne={
    'Low': 'Bajo',
    'Medium': 'Medio',
    'High': 'Alto'
}

#Acceso a sistemas de soporte
ss={
    'Yes': 'Si'
}

#Impacto en el entorno laboral
el={
    'Negative': 'Negativo',
    'Positive': 'Positivo'
}
```

[9]

✓ 0.0s

Python

```
df2['Género']=df2['Género'].replace(gen)
df2['Estado de salud mental']=df2['Estado de salud mental'].replace(sm)
df2['Nivel de estrés']=df2['Nivel de estrés'].replace(ne)
df2['Acceso a sistemas de soporte']=df2['Acceso a sistemas de soporte'].replace(ss)
df2['Impacto en el entorno laboral']=df2['Impacto en el entorno laboral'].replace(el)
df2
```

[10]

✓ 0.0s

Python

...	Usuario	Edad	Género	Horas de uso de la tecnología	Horas de uso de las redes sociales	Horas de juego	Horas de tiempo en pantalla	Estado de salud mental	Nivel de estrés	...
0	USER-00001	23.0	Mujer	6.57	6.0	0.68	12.36	Buena	Bajo	
1	USER-00002	21.0	Hombre	3.01	2.57	3.74	7.61	Mala	Alto	
2	USER-00003	51.0	Hombre	3.04	6.14	1.26	3.16	Neutral	Alto	
3	USER-00004	25.0	Mujer	3.84	4.48	2.59	13.08	Excelente	Medio	
4	USER-00005	53.0	Hombre	1.2	0.56	0.29	12.63	Buena	Bajo	
...
12524	USER-06709	24.0	Mujer	5.08	1.58	1.13	6.58	Mala	Bajo	
12525	USER-09996	42.0	Hombre	7.05	0.41	0.53	13.90	invalid	Medio	
12526	USER-02814	36.0	Hombre	4.42	0.08	2.20	10.48	Buena	Medio	
12527	USER-06331	35.0	Mujer	4.28	0.9	4.90	NaN	Neutral	Alto	
12528	USER-07333	50.0	Mujer	11.85	0.4	1.70	6.21	Buena	Alto	

12529 rows × 13 columns

Eliminamos los invalid.

```
lista_col=df2.columns
for i in lista_col:
    print(f"En la columna {i} los invalid son: {df2[df2[i] == 'invalid'].shape[0]}")
```

[11] ✓ 0.0s Python

... En la columna Usuario los invalid son: 232
En la columna Edad los invalid son: 249
En la columna Género los invalid son: 265
En la columna Horas_de_uso_de_la_tecnología los invalid son: 249
En la columna Horas_de_uso_de_las_redes_sociales los invalid son: 253
En la columna Horas_de_juego los invalid son: 0
En la columna Horas_de_tiempo_en_pantalla los invalid son: 0
En la columna Estado_de_salud_mental los invalid son: 246
En la columna Nivel_de_estrés los invalid son: 257
En la columna Horas_de_sueño los invalid son: 252
En la columna Horas_de_actividad_física los invalid son: 273
En la columna Acceso_a_sistemas_de_soporte los invalid son: 237
En la columna Impacto_en_el_entorno_laboral los invalid son: 234

```
for i in lista_col:
    df2=df2[df2[i] != 'invalid']
df2
```

[12] ✓ 0.1s Python

...

	Usuario	Edad	Género	Horas_de_uso_de_la_tecnología	Horas_de_uso_de_las_redes_sociales	Horas_de_juego	Horas_de_tiempo_en_pantalla	Estado_de_salud_mental	Nivel_de_estrés	
0	USER-00001	23.0	Mujer	6.57	6.0	0.68	12.36	Buena	Bajo	
1	USER-00002	21.0	Hombre	3.01	2.57	3.74	7.61	Mala	Alto	
2	USER-00003	51.0	Hombre	3.04	6.14	1.26	3.16	Neutral	Alto	
3	USER-00004	25.0	Mujer	3.84	4.48	2.59	13.08	Excelente	Medio	
4	USER-00005	53.0	Hombre	1.2	0.56	0.29	12.63	Buena	Bajo	
...	
12522	USER-05686	62.0	Otro	9.69	3.28	4.98	2.93	Neutral	Bajo	
12523	USER-02010	25.0	Otro	3.82	3.71	3.61	6.57	Mala	Bajo	
12524	USER-06709	24.0	Mujer	5.08	1.58	1.13	6.58	Mala	Bajo	
12527	USER-06331	35.0	Mujer	4.28	0.9	4.90	NaN	Neutral	Alto	
12528	USER-07333	50.0	Mujer	11.85	0.4	1.70	6.21	Buena	Alto	

10043 rows × 13 columns

```
lista_col=df2.columns
for i in lista_col:
    print(f"En la columna {i} los invalid son: {df2[df2[i] == 'invalid'].shape[0]}")
```

[13] ✓ 0.1s Python

... En la columna Usuario los invalid son: 0
En la columna Edad los invalid son: 0
En la columna Género los invalid son: 0
En la columna Horas_de_uso_de_la_tecnología los invalid son: 0
En la columna Horas_de_uso_de_las_redes_sociales los invalid son: 0
En la columna Horas_de_juego los invalid son: 0
En la columna Horas_de_tiempo_en_pantalla los invalid son: 0
En la columna Estado_de_salud_mental los invalid son: 0
En la columna Nivel_de_estrés los invalid son: 0
En la columna Horas_de_sueño los invalid son: 0
En la columna Horas_de_actividad_física los invalid son: 0
En la columna Acceso_a_sistemas_de_soporte los invalid son: 0
En la columna Impacto_en_el_entorno_laboral los invalid son: 0

Reemplazamos los espacios vacíos.

```
df2['Edad'] = pd.to_numeric(df2['Edad'], errors='coerce')
df2['Horas_de_uso_de_la_tecnología'] = pd.to_numeric(df2['Horas_de_uso_de_la_tecnología'], errors='coerce')
df2['Horas_de_uso_de_las_redes_sociales'] = pd.to_numeric(df2['Horas_de_uso_de_las_redes_sociales'], errors='coerce')
df2['Horas_de_sueño'] = pd.to_numeric(df2['Horas_de_sueño'], errors='coerce')
df2['Horas_de_actividad_física'] = pd.to_numeric(df2['Horas_de_actividad_física'], errors='coerce')
df2
```

[14] ✓ 0.0s Python

...

	Usuario	Edad	Género	Horas de uso de la tecnología	Horas de uso de las redes sociales	Horas de juego	Horas de tiempo en pantalla	Estado de salud mental	Nivel de estrés	
0	USER-00001	23.0	Mujer	6.57	6.00	0.68	12.36	Buena	Bajo	
1	USER-00002	21.0	Hombre	3.01	2.57	3.74	7.61	Mala	Alto	
2	USER-00003	51.0	Hombre	3.04	6.14	1.26	3.16	Neutral	Alto	
3	USER-00004	25.0	Mujer	3.84	4.48	2.59	13.08	Excelente	Medio	
4	USER-00005	53.0	Hombre	1.20	0.56	0.29	12.63	Buena	Bajo	
...	
12522	USER-05686	62.0	Otro	9.69	3.28	4.98	2.93	Neutral	Bajo	
12523	USER-02010	25.0	Otro	3.82	3.71	3.61	6.57	Mala	Bajo	
12524	USER-06709	24.0	Mujer	5.08	1.58	1.13	6.58	Mala	Bajo	
12527	USER-06331	35.0	Mujer	4.28	0.90	4.90	NaN	Neutral	Alto	
12528	USER-07333	50.0	Mujer	11.85	0.40	1.70	6.21	Buena	Alto	

10043 rows × 13 columns

```
df2['Edad'].fillna(df2['Edad'].mean(), inplace=True)
df2['Horas de uso de la tecnología'].fillna(df2['Horas de uso de la tecnología'].mean(), inplace=True)
df2['Horas de uso de las redes sociales'].fillna(df2['Horas de uso de las redes sociales'].mean(), inplace=True)
df2['Horas de sueño'].fillna(df2['Horas de sueño'].mean(), inplace=True)
df2['Horas de actividad física'].fillna(df2['Horas de actividad física'].mean(), inplace=True)
df2['Horas de juego'].fillna(df2['Horas de juego'].mean(), inplace=True)
df2['Horas de tiempo en pantalla'].fillna(df2['Horas de tiempo en pantalla'].mean(), inplace=True)
df2
```

[17] ✓ 0.0s Python

...

	Usuario	Edad	Género	Horas de uso de la tecnología	Horas de uso de las redes sociales	Horas de juego	Horas de tiempo en pantalla	Estado de salud mental	Nivel de estrés	
0	USER-00001	23.0	Mujer	6.57	6.00	0.68	12.360000	Buena	Bajo	
1	USER-00002	21.0	Hombre	3.01	2.57	3.74	7.610000	Mala	Alto	
2	USER-00003	51.0	Hombre	3.04	6.14	1.26	3.160000	Neutral	Alto	
3	USER-00004	25.0	Mujer	3.84	4.48	2.59	13.080000	Excelente	Medio	
4	USER-00005	53.0	Hombre	1.20	0.56	0.29	12.630000	Buena	Bajo	
...	
12522	USER-05686	62.0	Otro	9.69	3.28	4.98	2.930000	Neutral	Bajo	
12523	USER-02010	25.0	Otro	3.82	3.71	3.61	6.570000	Mala	Bajo	
12524	USER-06709	24.0	Mujer	5.08	1.58	1.13	6.580000	Mala	Bajo	
12527	USER-06331	35.0	Mujer	4.28	0.90	4.90	7.965271	Neutral	Alto	
12528	USER-07333	50.0	Mujer	11.85	0.40	1.70	6.210000	Buena	Alto	

10043 rows × 13 columns

```
#Elimino esta columna Usuario porque no aporta nada a la investigación.
df3=df2.drop(columns=['Usuario'])
df3
```

[18] ✓ 0.0s Python

...

	Edad	Género	Horas de uso de la tecnología	Horas de uso de las redes sociales	Horas de juego	Horas de tiempo en pantalla	Estado de salud mental	Nivel de estrés	Horas de s
0	23.0	Mujer	6.57	6.00	0.68	12.360000	Buena	Bajo	
1	21.0	Hombre	3.01	2.57	3.74	7.610000	Mala	Alto	
2	51.0	Hombre	3.04	6.14	1.26	3.160000	Neutral	Alto	
3	25.0	Mujer	3.84	4.48	2.59	13.080000	Excelente	Medio	
4	53.0	Hombre	1.20	0.56	0.29	12.630000	Buena	Bajo	
...	
12522	62.0	Otro	9.69	3.28	4.98	2.930000	Neutral	Bajo	
12523	25.0	Otro	3.82	3.71	3.61	6.570000	Mala	Bajo	
12524	24.0	Mujer	5.08	1.58	1.13	6.580000	Mala	Bajo	
12527	35.0	Mujer	4.28	0.90	4.90	7.965271	Neutral	Alto	
12528	50.0	Mujer	11.85	0.40	1.70	6.210000	Buena	Alto	

10043 rows × 12 columns

```
df3['Género'].fillna("Otro", inplace=True)
df3['Estado_de_salud_mental'].fillna("Neutral", inplace=True)
df3['Nivel_de_estrés'].fillna("Medio", inplace=True)
df3['Impacto_en_el_entorno_laboral'].fillna("Neutral", inplace=True)

[20] ✓ 0.0s Python

df3.isnull().sum()

[21] ✓ 0.0s Python

...
Edad 0
Género 0
Horas_de_uso_de_la_tecnología 0
Horas_de_uso_de_las_redes_sociales 0
Horas_de_juego 0
Horas_de_tiempo_en_pantalla 0
Estado_de_salud_mental 0
Nivel_de_estrés 0
Horas_de_sueño 0
Horas_de_actividad_física 0
Acceso_a_sistemas_de_soporte 114
Impacto_en_el_entorno_laboral 0
dtype: int64

#Elimino los NaN restantes, son de la columna 'Acceso a sistemas de soporte' y no pueden ser rellenados porque solo hay datos de sí y no.
df4=df3.dropna()

[22] ✓ 0.0s Python

df4.isnull().sum()

[23] ✓ 0.0s Python

...
Edad 0
Género 0
Horas_de_uso_de_la_tecnología 0
Horas_de_uso_de_las_redes_sociales 0
Horas_de_juego 0
Horas_de_tiempo_en_pantalla 0
Estado_de_salud_mental 0
Nivel_de_estrés 0
Horas_de_sueño 0
Horas_de_actividad_física 0
Acceso_a_sistemas_de_soporte 0
Impacto_en_el_entorno_laboral 0
dtype: int64

▷ ~
df4 = df4.reset_index(drop=True)

[24] ✓ 0.0s Python

df4['Edad']=df4['Edad'].astype(int)

[26] ✓ 0.0s Python

▷ ~
df4.info()

[27] ✓ 0.0s Python

...
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9929 entries, 0 to 9928
Data columns (total 12 columns):
# Column Non-Null Count Dtype
---
0 Edad 9929 non-null int64
1 Género 9929 non-null object
2 Horas_de_uso_de_la_tecnología 9929 non-null float64
3 Horas_de_uso_de_las_redes_sociales 9929 non-null float64
4 Horas_de_juego 9929 non-null float64
5 Horas_de_tiempo_en_pantalla 9929 non-null float64
6 Estado_de_salud_mental 9929 non-null object
7 Nivel_de_estrés 9929 non-null object
8 Horas_de_sueño 9929 non-null float64
9 Horas_de_actividad_física 9929 non-null float64
10 Acceso_a_sistemas_de_soporte 9929 non-null object
11 Impacto_en_el_entorno_laboral 9929 non-null object
dtypes: float64(6), int64(1), object(5)
memory usage: 931.0+ KB

#Guardar los resultados en un csv
df4.to_csv("Base limpia proyecto.csv", index=True)

[29] ✓ 0.1s Python
```