

✓ Cargar librerías:

```
from google.colab import drive
drive.mount('/content/drive')
```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True)

```
%cd '/content/01_Sleep_health_and_lifestyle_dataset.csv'
!ls
```

[Errno 20] Not a directory: '/content/01_Sleep_health_and_lifestyle_dataset.csv'

```
/content
01_Sleep_health_and_lifestyle_dataset.csv  drive  sample_data
```

✓ Con el código de arriba (!ls) saco un listado de todos los datasets que tengo cargados y subidos para trabajarlos. En este caso, solo tengo 1 dataset.

```
import pandas as pd
#from google.colab import files
#files.upload()
```

```
datos = pd.read_csv(
    "/content/drive/MyDrive/Colab Notebooks/01_Sleep_health_and_lifestyle_dataset.csv",
    index_col="Person ID"
)
datos.head()
```

Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	NaN
2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN

En caso de que nos diera error a la hora de abrir para leer el dataset, tendríamos que añadir: header=1 o header=2, para que Pandas lo pueda leer y abrir sin errores.

COMO EN ESTE EJERCICIO NO VOY A HACER UN MODELO DE MACHINE LEARNING PARA LA PREDICCIÓN, NO NECESITO:

- TOCAR LOS VALORES NULOS o FALTANTES

- MODIFICAR EL TIPO DE DATO Y CONVERTIRLO A NUMÉRICO o CUANTITATIVO.

✓ Preguntas "semilla u objetivo" a responder con el análisis de este Dataset.

1- ¿Quiénes tienen peor calidad del sueño, hombres o mujeres?

2- ¿Existe alguna relación entre la calidad del sueño de las personas y su profesión?

3- ¿La actividad física influye en el sueño o la calidad del mismo?

4- ¿Qué profesión presenta la peor calidad de sueño?

5- ¿Qué profesión tienen las personas con el mayor nivel de estrés y el mayor índice de masa corporal?

6- ¿En qué rango de edades se encuentran la mayor cantidad de trastornos del sueño?

7- La cantidad de pasos al día...¿Afecta a la calidad del sueño? ¿Y al índice de masa corporal?

```
datos.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 374 entries, 1 to 374
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Gender                                374 non-null    object
1   Age                                   374 non-null    int64
2   Occupation                            374 non-null    object
3   Sleep Duration                        374 non-null    float64
4   Quality of Sleep                      374 non-null    int64
5   Physical Activity Level               374 non-null    int64
6   Stress Level                          374 non-null    int64
7   BMI Category                          374 non-null    object
8   Blood Pressure                       374 non-null    object
9   Heart Rate                           374 non-null    int64
10  Daily Steps                           374 non-null    int64
11  Sleep Disorder                        155 non-null    object
dtypes: float64(1), int64(6), object(5)
memory usage: 38.0+ KB
```

✓ Tengo un dataset con 374 filas y 12 columnas distribuidas de la siguiente manera:

- 5 columnas de tipo "object",
- 6 columnas de tipo "int64" y
- 1 columna de tipo "float"

Con el método (.shape) usado a continuación consigo lo mismo, es decir el nº de filas y columnas.

```
datos.shape
```

```
(374, 12)
```

```
datos["Gender"].value_counts()
```

```
Gender
Male      189
Female    185
Name: count, dtype: int64
```

✓ Con esto he comprobado que la muestra de DATOS está BALANCEADA, ya que tan solo hay 4 datos de diferencia entre el nº de hombres y mujeres.

```
datos.describe()
```

```

      Age      Sleep      Quality      Physical      Stress      Heart      Daily
      Duration of Sleep Activity Level Level Rate
count 374.000000 374.000000 374.000000 374.000000 374.000000 374.000000 374.000000
mean  42.184492  7.132086  7.312834  59.171123  5.385027  70.165775  6816.1
std    8.673133  0.795657  1.196956  20.830804  1.774526  4.135676  1617.1
min   27.000000  5.800000  4.000000  30.000000  3.000000  65.000000  3000.0
25%   35.250000  6.400000  6.000000  45.000000  4.000000  68.000000  5600.0
50%   43.000000  7.200000  7.000000  60.000000  5.000000  70.000000  7000.0
75%   50.000000  7.800000  8.000000  75.000000  7.000000  72.000000  8000.0

```

Conclusiones:

Al observar la tabla de edades usando el método describe, he descubierto que:

- la edad media (mean) es de aprox 42 años,
- la desviación estándar es de 8.6 años,
- la edad mín es de 27 años y
- la máx de 59.

Con esto se puede deducir que tanto la desviación estándar (muy alta) como el valor promedio (está muy cerca del valor máximo), son inusualmente altos y puede ser debido a que:

- por un lado, el dataset tiene muy pocas filas de datos y,
- por otro lado puede haber una "deficiente distribución" en los rangos de edades.

Esto afectará al análisis de nuestros datos, ya que la información correspondiente a las edades es muy dispersa o amplia, lo que nos puede generar conclusiones erróneas.

✓ RESPUESTA a Pregunta "semilla u objetivo":

1- ¿Quiénes tienen peor calidad del sueño, hombres o mujeres?

```
datos.groupby("Gender")[["Sleep Duration", "Quality of Sleep"]].mean()
```

	Sleep Duration	Quality of Sleep
Gender		
Female	7.229730	7.664865
Male	7.036508	6.968254

Conclusión:

Se observa que las mujeres duermen durante más tiempo y con mayor calidad.

✓ RESPUESTA a Pregunta "semilla u objetivo":

2- ¿Existe alguna relación entre la calidad del sueño de las personas y su profesión?

Análisis de las profesiones.

```
datos.groupby("Occupation")["Quality of Sleep"].value_counts()
```

Occupation	Quality of Sleep
Accountant	8 7 9
Doctor	7 6 9
Engineer	9 8 7 5 6
Lawyer	8 7
Manager	7
Nurse	6 9 5 8 7
Sales Representative	4
Salesperson	6
Scientist	6 4
Software Engineer	8 4 6
Teacher	7 8 6


```
Name: count, dtype: int64
```

✓ Ahora voy a ver lo mismo, la calidad y la duración del sueño, pero en cada una de las profesiones usando el siguiente código.



Además, lo voy a ordenar en función de la Calidad del sueño, de menor a mayor.

Si quiero verlo al contrario, es decir de mayor a menor, tan solo tengo que añadir `ascending = False`.

```
calidad_profesiones=datos.groupby("Occupation")[["Sleep Duration", "Quality of Sleep"]].mean()
calidad_profesiones.sort_values(by="Quality of Sleep")
```



	Sleep Duration	Quality of Sleep
Occupation		
Sales Representative	5.900000	4.000000
Scientist	6.000000	5.000000
Salesperson	6.403125	6.000000
Software Engineer	6.750000	6.500000
Doctor	6.970423	6.647887
Teacher	6.690000	6.975000
Manager	6.900000	7.000000
Nurse	7.063014	7.369863
Accountant	7.113514	7.891892
Lawyer	7.410638	7.893617
Engineer	7.987302	8.412698

Conclusión:

- El Representante de ventas es la profesión que peor calidad de sueño tiene.
- Los Ingenieros por su parte, son los que mejor calidad de sueño tienen.
- También parece existir una relación directa entre la duración y la calidad el sueño, ya que en líneas generales, a menor duración menor calidad y viceversa.

✓ RESPUESTA a Preguntas "semilla u objetivo":

3- ¿La actividad física influye en el sueño o la calidad del mismo?.

7 y ÚLTIMA- La cantidad de pasos al día...¿Afecta a la calidad del sueño? ¿Y al índice de masa corporal?

Resumen estadístico (hallando el promedio) en base a las columnas numéricas del conjunto de datos, ordenando o teniendo como referencia la "Calidad del sueño":

Lo haré mediante el siguiente código, usando para ello las columnas de:

- pasos diarios acumulados,
- el nivel de actividad física,
- la calidad del sueño y
- la duración.

```
pasos=datos.groupby("Occupation")[["Daily Steps", "Physical Activity Level", "Sleep Duration", "Stress Level", "Quality of Sleep"]]
pasos.sort_values(by="Quality of Sleep")
```



	Daily Steps	Physical Activity Level	Sleep Duration	Stress Level	Quality of Sleep
Occupation					
Sales Representative	3000.000000	30.000000	5.900000	8.000000	4.000000
Scientist	5350.000000	41.000000	6.000000	7.000000	5.000000
Salesperson	6000.000000	45.000000	6.403125	7.000000	6.000000
Software Engineer	5800.000000	48.000000	6.750000	6.000000	6.500000
Doctor	6808.450704	55.352113	6.970423	6.732394	6.647887
Teacher	5957.500000	45.625000	6.690000	4.525000	6.975000
Manager	5500.000000	55.000000	6.900000	5.000000	7.000000
Nurse	8057.534247	78.589041	7.063014	5.547945	7.369863
Accountant	6881.081081	58.108108	7.113514	4.594595	7.891892

Conclusiones:

- Parece existir una relación proporcional y directa (aunque no estricta) entre los "Pasos diarios" y la "Calidad del sueño", ya que como podemos apreciar en la tabla, el "Representante de Ventas" tiene el menor nº de "Pasos al día" (3.000 pasos diarios) y también, la peor "Calidad del sueño" (un 4).

En el lado opuesto, encontramos a las "Enfermeras" con el mayor nº de "Pasos diarios" (un poco más de 8.000 pasos) y, con una de las tasas más altas de la "Calidad del sueño" (por encima del 7), aunque no la más alta, que corresponde a los "Ingenieros" (con casi 8 y medio), seguidos de "Abogados" y "Contables" (con casi 8).

- Asimismo, también parece existir una relación inversamente proporcional entre la "Calidad del sueño", el "Nivel de Stress" y el "Nivel de Actividad física", ya que a mayor "Nivel de Stress", menor "Calidad de sueño" y menor "Nivel de Actividad física".

RESPUESTA a Pregunta "semilla u objetivo":

4- ¿Qué profesión presenta la peor calidad de sueño?

```
datos.groupby("Occupation")["Gender"].value_counts()
```



Occupation	Gender	
Accountant	Female	36
	Male	1
Doctor	Male	69
	Female	2
Engineer	Female	32
	Male	31
Lawyer	Male	45
	Female	2
Manager	Female	1
Nurse	Female	73
Sales Representative	Male	2
Salesperson	Male	32
Scientist	Female	4
Software Engineer	Male	4
Teacher	Female	35
	Male	5
Name: count, dtype: int64		

```
datos.groupby("Sleep Disorder")["Occupation"].value_counts()
```



Sleep Disorder	Occupation	
Insomnia	Salesperson	29
	Teacher	27
	Accountant	7
	Engineer	5
	Doctor	3

```

Nurse 3
Lawyer 2
Software Engineer 1
Sleep Apnea Nurse 61
Doctor 4
Teacher 4
Lawyer 3
Sales Representative 2
Scientist 2
Engineer 1
Salesperson 1
Name: count, dtype: int64

```

Conclusión:

Se observa que hay 3 profesiones claramente diferenciadas con respecto al resto, en cuanto a los desordenes producidos por la falta de sueño o la mala calidad del mismo, que son:

- 1- Enfermeras,
- 2- Vendedores y
- 3- Profesores.

✓ RESPUESTA a Pregunta "semilla u objetivo":

5- ¿Qué profesión tienen las personas con el mayor nivel de estrés y el mayor índice de masa corporal?

Calculando el "Nivel de Stress" por profesión.

```

nivel_stress=datos.groupby("Gender")["Occupation", "Stress Level"].value_counts()
nivel_stress.sort_values()

```

```

Gender Occupation Stress Level
Female Lawyer 5 1
Male Software Engineer 6 1
Engineer 6 1
Software Engineer 8 1
Female Lawyer 6 1
Manager 5 1
Male Accountant 6 1
Doctor 5 2
3 2
Lawyer 6 2
Teacher 5 2
Sales Representative 8 2
Software Engineer 5 2
Engineer 3 2
Female Accountant 6 2
Nurse 6 2
Scientist 6 2
8 2
Teacher 6 2
7 2
Accountant 3 2
Doctor 3 2
Male Teacher 7 3
Engineer 7 3
Female Nurse 7 4
4 4
Accountant 7 6
Male Engineer 4 9
5 16
Female Accountant 4 26
Teacher 4 31
Nurse 3 31
Engineer 3 32
Male Doctor 6 32
Salesperson 7 32
Female Nurse 8 32
Male Doctor 8 33
Lawyer 5 43
Name: count, dtype: int64

```

✓ Conclusión:

Se observa que hay varias profesiones muy estresantes que son:

- 1- Doctores (Hombres),
- 2- Enfermeras (Mujeres),
- 3- Vendedores (Hombres),
- 4- Ingenieros y Profesores (Hombres) y
- 5- Contables, Científicas y Profesoras (Mujeres)

✓ Calculando el "Índice de masa corporal" por profesión.

```
masa_corporal=datos.groupby("Occupation")["BMI Category"].value_counts()
masa_corporal.sort_values()
```

Occupation	BMI Category	
Lawyer	Normal Weight	1
Software Engineer	Overweight	1
	Obese	1
	Overweight	1
Manager	Overweight	1
Teacher	Obese	1
Doctor	Normal Weight	2
Lawyer	Obese	2
	Overweight	2
Software Engineer	Normal Weight	2
Sales Representative	Obese	2
Engineer	Overweight	3
	Normal Weight	4
Doctor	Obese	4
Scientist	Overweight	4
Accountant	Normal Weight	5
	Overweight	6
Teacher	Normal	6
Nurse	Normal Weight	7
Accountant	Normal	26
Salesperson	Overweight	32
Teacher	Overweight	33
Lawyer	Normal	42
Engineer	Normal	56
Doctor	Normal	65
Nurse	Overweight	66

Name: count, dtype: int64

Conclusión:

Las profesiones con mayor tasa de sobrepeso de forma muy destacada con respecto al resto, son:

- 1- Enfermeras,
- 2- Profesores y
- 3- Vendedores

✓ RESPUESTA a Pregunta "semilla u objetivo":

6- ¿En qué rango de edades se encuentran la mayor cantidad de trastornos del sueño?

Análisis de edades

- ✓ Agregando una nueva columna al dataset, para hacer rangos de edades y tratar de minimizar la deficiente distribución primaria.

```
datos.loc[datos["Age"] <= 18, "Grupo_edades"] = "Adolescente"
datos.loc[datos["Age"] >= 19, "Grupo_edades"] = "Joven"
datos.loc[datos["Age"] >= 30, "Grupo_edades"] = "Adulto"
datos.loc[datos["Age"] >= 46, "Grupo_edades"] = "Maduro"
datos.loc[datos["Age"] >= 55, "Grupo_edades"] = "Maduro viejo"
datos.head(25)
```



	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category
Person ID								
1	Male	27	Software Engineer	6.1	6	42	6	Overweight
2	Male	28	Doctor	6.2	6	60	8	Normal
3	Male	28	Doctor	6.2	6	60	8	Normal
4	Male	28	Sales Representative	5.9	4	30	8	Obese
5	Male	28	Sales Representative	5.9	4	30	8	Obese
6	Male	28	Software Engineer	5.9	4	30	8	Obese
7	Male	29	Teacher	6.3	6	40	7	Obese
8	Male	29	Doctor	7.8	7	75	6	Normal
9	Male	29	Doctor	7.8	7	75	6	Normal
10	Male	29	Doctor	7.8	7	75	6	Normal
11	Male	29	Doctor	6.1	6	30	8	Normal
12	Male	29	Doctor	7.8	7	75	6	Normal
13	Male	29	Doctor	6.1	6	30	8	Normal
14	Male	29	Doctor	6.0	6	30	8	Normal
15	Male	29	Doctor	6.0	6	30	8	Normal
16	Male	29	Doctor	6.0	6	30	8	Normal
17	Female	29	Nurse	6.5	5	40	7	Normal Weight
18	Male	29	Doctor	6.0	6	30	8	Normal
19	Female	29	Nurse	6.5	5	40	7	Normal Weight

Pasos siguientes:

Generar código con datos

Ver gráficos recomendados

```
datos["Grupo_edades"].value_counts()
```



```
Grupo_edades
Adulto      245
Maduro       75
Maduro viejo  35
Joven        19
Name: count, dtype: int64
```

Conclusión:

Se observa que la mayor parte de los datos se centran en las edades comprendidas entre los 19 y los 30 años (calificada como adulto) y entre los 31 y 46 años (maduro).

```
datos.groupby("Gender")["Age"].agg(["mean", "std"])
```



	mean	std
Gender		
Female	47.405405	8.093407
Male	37.074074	5.662006

Conclusión:

- Las mujeres tienen una media de edad mucho más alta que la de los hombres (10 años de diferencia).
- La desviación estándar es muy alta en ambos casos, pero sobre todo en las mujeres, lo que nos puede conllevar a errores en nuestro análisis final.

✓ ANEXO

Para Imprimir el resultado del análisis como un PDF, desde este IDLE (Google Colab) debo escribir el siguiente código:

```
#!apt-get install texlive texlive-xetex texlive-latex-extra pandoc
#!pip install pypandoc

#!apt-get install texlive-xetex texlive-fonts-recommended texlive-generic-recommended

#!pip install pandas-profiling

#!jupyter nbconvert --to PDF /content/drive/MyDrive/Colab Notebooks/01_Sleep_health_and_lifestyle_dataset.csv

#from pandas_profiling import ProfileReport
```