



SALUD MENTAL EN EL TRABAJO

ALUMNA:
Palermo Rita V.
**Primera Pre-
Entrega**
Comisión N°:
90455
Profesor: Ruiz
Jorge
Tutor: Lisachi
Luciano

CODERHOUSE

INDICE

- 1- DICCIONARIO
- 2- INTRODUCCION
- 3- OBJETIVOS
- 4- DESCRIPCIÓN: MOTIVACIÓN Y AUDIENCIA
- 5- PREGUNTAS GUIAS E HIPOTESIS
- 6- DATASET
- 7- PREPARACIÓN DE LOS DATOS
- 8- ESTADÍSTICA DESCRIPTIVA
- 9- DATA WRANGLING Y EDA
- 10- FEATURE ENGINEERING
- 11- FEATURE BINNING
- 12- MAPA COROPLETICO
- 13- ANÁLISIS UNIVARIADO
- 14- ANÁLISIS BIVARIADO
- 15- ANÁLISIS MULTIVARIADO
- 16- CONCLUSIONES

1- DICCIONARIO:

Customer Support: Apoyo al Cliente.

Data Scientist: Científico de datos.

Engineering: Ingeniería.

False: Falso.

Female: Mujer.

Germany: Alemania.

Híbrido: Combina trabajo remoto y presencial.

HR: Recursos Humanos.

Id: Identificador único.

IT: Tecnología de la administración/relaciones con inversionistas.

Male: Masculino.

Marketing Manager: Gerente de Marketing.

Non-binary: No Binario.

Prefer not to say: Prefiero no decirlo.

Project Manager: Gerente de Proyecto.

Sales Associate: Asociado de Ventas.

Sales: Ventas.

Software Engineer: Ingeniero de Software.

Specialist: Especialista en RRHH (Recursos Humanos).

Support: Apoyo.

True: Verdadero.

UK: Reino Unido

USA: Estados Unidos.

2- INTRODUCCIÓN:

La salud mental en el entorno laboral se ha convertido en un eje central para comprender el bienestar integral de las personas y el rendimiento de las organizaciones. En un contexto donde las exigencias profesionales, la hiperconectividad y la falta de espacios de contención emocional impactan directamente en la calidad de vida, el análisis de datos se presenta como una herramienta poderosa para visibilizar, entender y actuar sobre estos desafíos.

Este trabajo se centra en el estudio de un conjunto de datos que reúne variables claves relacionadas con el bienestar psicológico de los empleados, tales como edad, género, función laboral, horas de trabajo, calidad del sueño, nivel de agotamiento, acceso a terapia, apoyo institucional, entre otras. A través de técnicas de análisis exploratorio, segmentación y modelado predictivo, se busca identificar patrones, correlaciones y factores de riesgo que permitan comprender

cómo distintas condiciones laborales influyen en la salud mental de los trabajadores.

3- OBJETIVO:

Analizar de manera integral los factores que inciden en la salud mental de los trabajadores a partir de un conjunto de datos multidimensionales, con el fin de identificar patrones, correlaciones y riesgos asociados al agotamiento laboral, y generar propuestas basadas en evidencia que contribuyan a mejorar el bienestar emocional en el entorno organizacional. Como así también promover el autocuidado y fomentar culturas organizacionales más empáticas, sostenibles y humanas.

4- DESCRIPCIÓN: MOTIVACIÓN Y AUDIENCIA

Descripción de alto nivel: Motivación y Audiencia

En un contexto laboral cada vez más dinámico y desafiante, comprender los factores que inciden en el bienestar, la productividad y la salud integral de las personas se vuelve una necesidad estratégica. Este análisis parte de una motivación profunda: explorar cómo las condiciones de trabajo, las percepciones individuales y las prácticas organizacionales se entrelazan para moldear la experiencia humana dentro de las empresas. Al observar variables como el estrés, la satisfacción, la participación en actividades de salud mental y el crecimiento personal, se busca revelar patrones que no solo informen decisiones operativas, sino que también inspiren transformaciones culturales más inclusivas y sostenibles.

La audiencia que puede beneficiarse de este estudio es amplia y diversa. Profesionales de recursos humanos, líderes organizacionales, consultores en clima laboral y equipos de bienestar corporativo encontrarán en este análisis una base empírica para diseñar intervenciones más empáticas y efectivas. Asimismo, investigadores en ciencias sociales y estudiantes interesados en el vínculo entre trabajo y salud podrán utilizar estos hallazgos como punto de partida para nuevas preguntas y enfoques interdisciplinarios.

5- PREGUNTAS GUÍAS E HIPÓTESIS:

A partir del objetivo planteado, se definieron las siguientes preguntas clave y sus respectivas hipótesis, que orientan el análisis exploratorio y predictivo:

- ¿Qué relación existe entre las horas de trabajo, el acceso a apoyo emocional y el nivel de agotamiento laboral en los empleados?
- ¿Cómo influye el trabajo remoto en la satisfacción laboral y el riesgo de agotamiento, considerando variables como el sueño, la actividad física y el tiempo de viaje?

Hipótesis 1: Los empleados que trabajan más horas y no cuentan con apoyo institucional para la salud mental presentan niveles significativamente más altos de agotamiento laboral.

Hipótesis 2: El trabajo remoto se asocia con mayor satisfacción laboral y menor riesgo de agotamiento, especialmente en empleados que duermen más horas, realizan actividad física regularmente y no dedican tiempo al traslado.

6- DATASET:

El dataset a analizar cuenta con 25 Columnas y 3000 filas o registros. Será utilizado para extraer información que me permita evaluar las hipótesis planteadas.

Se analizará distribución de edad, género, país, función laboral, promedios y rangos de horas de trabajo, sueño, actividad física, viaje, porcentajes de empleados con acceso a terapia o apoyo emocional entre otros.

7- PREPARACIÓN DE LOS DATOS:

Lo primero que hice fue importar las librerías necesarias para el análisis: **Pandas**, **NumPy**, **Matplotlib** y **Seaborn** para la manipulación y visualización de datos; **Missingno** para detectar valores ausentes; y diversas funciones de **Scikit-learn** para aplicar buenas prácticas en el procesamiento y modelado.

Luego cargué el dataset en formato CSV, titulado “**Salud mental en el trabajo**”, y visualicé sus primeras filas para familiarizarme con la estructura de los datos

DESCRIPCIÓN DE VARIABLES Y DOMINIOS:

- * **Empleado:** Lo registra con un número (Id)
- * **Edad:** Edad del paciente medida en años (22 a 59)
- * **Género:** Sexo del paciente (male, female, prefer not to say, non-binary).
- * **País** (del cual es cada paciente): Australia, Brasil, Canadá, Germany, India, UK, USA.
- * **Función Laboral** (se refiere al conjunto de tareas, responsabilidades y actividades que una persona realiza en su puesto de trabajo, en este caso serían): Customer Support, Data Scientist, HR Specialist, IT Admin, Marketing Manager, Project Manager, Sales Associate y Software Engineer,
- * **Departamento** (representa el área o sector dentro de la empresa al que pertenece cada empleado o puesto en este caso están): Engineering, HR (Recursos Humanos), IT(Relaciones con Inversionistas), Marketing, Sales, Support.
- * **Años en la empresa:** Representa la antigüedad de cada empleado dentro de la Empresa (0 a 20 años)
- * **Horas de trabajo por semana:** Representa las Horas de Trabajo de cada empleado por Semana (de 30 hs a 59 hs semanales)
- * **Trabajo Remoto:** Indica si el trabajo es remoto, presencial o híbrido en este caso el
 - Sí: Trabajo 100% remoto
 - No: Trabajo 100% presencial
 - Híbrido: Combina trabajo remoto y presencial.
- * **Nivel de Agotamiento:** Refleja el estado energético del empleado según autoevaluaciones o indicadores internos (1.04 a 10.00)

- * **Satisfacción Laboral:** Refleja el nivel de Satisfacción del empleado dentro de la empresa (1.00 a 10.00)
- * **Nivel de estrés:** Indicador del estado emocional y físico del empleado frente a las demandas laborales (1.00 a 10.00)
- * **Puntuación de Productividad:** Mide el rendimiento laboral de una persona dentro de una empresa o equipo de trabajo (1.00 a 10.00)
- * **Horas de Sueño:** Cantidad de tiempo que una persona duerme por día (4 HS A 9 HS)
- * **Horas de Actividad Física:** Representa la cantidad de tiempo que una persona dedica al movimiento corporal intencionado durante el día o la semana (1hs a 10 hs).
- * **Tiempo de Viaje:** Representa la cantidad de tiempo que una persona tarda en trasladarse desde su casa hasta el lugar de trabajo y viceversa (0 minutos a 119 minutos)
- * **Tiene apoyo para la salud Mental:** Es un indicador de si la persona tiene acceso o utiliza recursos de acompañamiento emocional o psicológico ofrecidos por la empresa o por fuera de ella, acá encontramos las variables: True y False
- * **Puntuación del Gerente:** Representa una evaluación del desempeño, liderazgo o apoyo que brinda el gerente a su equipo (1.00 a 10).
- * **Tiene Acceso a Terapia:** Indica si el empleado cuenta con recursos de acompañamiento psicológico, encontramos True y False
- * **Días Libres:** Representa la cantidad de días que un empleado tiene para su descanso. (0 a 9).
- * **Rango Salarial:** Representa la categoría de ingresos que percibe una persona en su puesto de trabajo (40 K a más de 100 K)
- * **Puntuación de equilibrio entre vida laboral y personal:** Representa un indicador del grado en que una persona logra compatibilizar sus responsabilidades profesionales con sus necesidades personales, familiares y de descanso (1.00 a 10.00)
- * **Tamaño del equipo:** Representa la cantidad de personas que conforman el grupo de trabajo de cada empleado (1 a 50).
- * **Puntuación de crecimiento personal:** Representa un indicador de cómo cada persona percibe su desarrollo individual dentro del entorno laboral (1.00 a 10.00)
- * **Riesgo de Agotamiento:** Indica que estima la probabilidad de que una persona experimente burnout o fatiga laboral crónica, True y False

8- ESTADÍSTICA DESCRIPTIVA:

El dataset obtenido de la base de datos contiene 3.000 registros (filas) y 25 variables (columnas). Los tipos de datos presentes incluyen números enteros, booleanos y decimales, lo que permite realizar diversos tipos de análisis.

No se detectan valores nulos: todas las columnas cuentan con 3.000 valores no nulos, lo cual es ideal para llevar a cabo un análisis completo sin necesidad de imputar o limpiar datos faltantes.

Tampoco se encontraron filas duplicadas. Además, se corrigieron nombres de columnas mal escritos y se realizó una separación entre variables categóricas y numéricas para facilitar el análisis exploratorio.

La memoria utilizada es de 524.5 KB, lo que lo convierte en un archivo liviano y fácil de manipular.

```
df.info()

...
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 25 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Empleado         3000 non-null    int64  
 1   Edad             3000 non-null    int64  
 2   Genero           3000 non-null    object  
 3   Pais             3000 non-null    object  
 4   Funcion Laboral 3000 non-null    object  
 5   Departamento    3000 non-null    object  
 6   Años en la Empresa 3000 non-null    int64  
 7   Hs de Trabajo x Semana 3000 non-null    int64  
 8   Trabajo Remoto   3000 non-null    object  
 9   Nivel de Agotamiento 3000 non-null    float64 
 10  Satisfaccion Laboral 3000 non-null    float64 
 11  Nivel de Stres    3000 non-null    float64 
 12  Puntuacion de Productividad 3000 non-null    float64 
 13  Hs de Sueño       3000 non-null    float64 
 14  Hs. De actividad Física 3000 non-null    float64 
 15  Tiempo de Viaje   3000 non-null    int64  
 16  Tiene apoyo para la salud Mental 3000 non-null    bool    
 17  Puntacion del Gerente   3000 non-null    float64 
 18  Tiene Acceso a Terapia 3000 non-null    bool    
 19  Días libres de Salud Mental 3000 non-null    int64  
 20  Rango Salarial     3000 non-null    object  
 21  Puntuacion de equilibrio entre vida laboral y personal 3000 non-null    float64 
 22  Tamaño del Equipo   3000 non-null    int64  
 23  Puntuacion de crecimiento personal 3000 non-null    float64 
 24  Riesgo de Agotamiento 3000 non-null    bool    
dtypes: bool(3), float64(9), int64(7), object(6)
memory usage: 524.5+ KB
```

Figura 1: Información del dataset.

Explicación de la figura 1:

- Cantidad de filas: 3000 registros.
- Cantidad de columnas: 25 variables.
- Tipos de datos: Mezcla de números enteros, booleanos y decimales.
- No hay valores nulos: Todas las columnas tienen 3000 valores no nulos, lo cual es ideal para un buen análisis, sin necesidad de imputar o limpiar datos faltantes.
- Memoria usada: 524.5 KB, bastante liviano.

	Empleado	Edad	Años en la Empresa	Hs de Trabajo x Semana	Nivel de Agotamiento	Satisfaccion Laboral	Nivel de Stres	Puntuacion de Productividad	Hs de Sueño	Hs. De actividad Física	Tiempo de Viaje	Puntacion del Gerente	Días libres de Salud Mental	Puntuacion de equilibrio entre vida laboral y personal	Tamaño del Equipo	Puntuacion de crecimiento personal
count	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	
mean	2500.500000	40.805667	10.099333	44.504000	5.509137	5.43750	5.51535	5.51956	6.539000	5.030400	59.227000	5.44254	4.544667	5.450950	25.20100	5.520680
std	866.169729	11.011705	6.035032	8.491526	2.574072	2.59443	2.60361	2.60761	1.441876	2.861026	34.809779	2.59740	2.854129	2.602379	14.08645	2.566861
min	1001.000000	22.000000	0.000000	30.000000	1.00000	1.00000	1.00000	1.00000	4.000000	0.000000	0.000000	1.00000	0.000000	1.00000	1.00000	1.000000
25%	1750.750000	31.000000	5.000000	37.000000	3.300000	3.18000	3.26000	3.28000	5.300000	2.600000	29.000000	3.16000	2.000000	3.160000	13.00000	3.300000
50%	2500.500000	41.000000	10.000000	45.000000	5.480000	5.43000	5.49500	5.49000	6.600000	5.000000	59.000000	5.44000	5.000000	5.425000	25.00000	5.580000
75%	3250.250000	50.000000	15.000000	52.000000	7.640000	7.68000	7.82250	7.80250	7.800000	7.500000	90.000000	7.69000	7.000000	7.730000	37.00000	7.642500
max	4000.000000	59.000000	20.000000	59.000000	10.00000	10.00000	10.00000	10.00000	9.000000	10.000000	119.000000	9.99000	9.000000	10.000000	49.00000	9.990000

Figura 2: Descripción del dataset.

Explicación de la figura 2:

Para cada columna, muestra:

- Count: cuántos valores no nulos hay, ayuda a detectar si hay datos faltantes.
- Mean: el promedio.
- std: la desviación estándar (cuánto varían los datos respecto al promedio).
- min / max: el valor mínimo y máximo.

- 25%, 50%, 75%: los percentiles (cuartiles), que te indican cómo se distribuyen los datos.

```
▶ df.duplicated().value_counts()
```

```
...      count
```

False	3000
--------------	------

dtype: int64

Figura 3: No posee filas repetidas.

```
▶ df.isnull().sum()/len(df)*100
```

	0
Empleado	0.0
Edad	0.0
Genero	0.0
Pais	0.0
Funcion Laboral	0.0
Departamento	0.0
Años en la Empresa	0.0
Hs de Trabajo x Semana	0.0
Trabajo Remoto	0.0
Nivel de Agotamiento	0.0
Satisfaccion Laboral	0.0
Nivel de Stres	0.0
Puntuacion de Productividad	0.0
Hs de Sueño	0.0
Hs. De actividad Fisica	0.0
Tiempo de Viaje	0.0
Tiene apoyo para la salud Mental	0.0
Puntacion del Gerente	0.0
Tiene Acceso a Terapia	0.0
Dias libres de Salud Mental	0.0
Rango Salarial	0.0
Puntuacion de equilibrio entre vida laboral y personal	0.0
Tamaño del Equipo	0.0
Puntuacion de crecimiento personal	0.0
Riesgo de Agotamiento	0.0

dtype: float64

Figura 4: No posee valores nulos.

Corrijo Columnas mal escritas: 'Años en la Empresa' - 'Hs de Sueño' - 'Tamaño del Equipo'
 df.rename(columns={'Años en la Empresa': 'Años en la empresa'}, inplace=True)
 df.rename(columns={'Hs de Sueño': 'Hs de Sueno'}, inplace=True)
 df.rename(columns={'Tamaño del Equipo': 'Tamaño del Equipo'}, inplace=True)

Controlo que lo hallo cambiado:
 display(df.head(10))

...	Empleado	Edad	Genero	País	Funcion Laboral	Departamento	Años en la Empresa	Hs de Trabajo x Semana	Trabajo Remoto	Nivel de Agotamiento	...	Tiempo de Viaje	Tiene apoyo para la salud mental	Puntuacion del Gerente	Tiene Acceso a Terapia	Dias libres de Salud Mental	Rango Salarial	Puntuacion de equilibrio entre vida laboral y personal	Tamaño del Equipo	Puntuacion de crecimiento personal	Riesgo de Agotamiento
0	1001	50	Male	UK	Sales Associate	HR	14	47	No	3.37	...	117	False	3.15	True	8	40K-60K	8.82	6	9.20	False
1	1002	36	Male	Germany	Software Engineer	IT	1	59	Hybrid	7.39	...	8	True	4.40	True	4	80K-100K	2.80	45	8.46	True
2	1003	29	Non-binary	India	IT Admin	IT	13	60	Hybrid	7.10	...	76	False	3.63	False	6	80K-100K	7.28	7	7.06	True
3	1004	42	Male	Australia	HR Specialist	IT	15	31	Yes	4.18	...	43	True	4.50	True	9	60K-80K	1.31	11	8.90	False
4	1005	40	Male	Brazil	Customer Support	Support	6	34	Yes	8.28	...	58	True	5.51	True	6	<40K	1.17	18	8.88	True
5	1006	44	Prefer not to say	Germany	Project Manager	Support	3	58	Hybrid	3.12	...	23	True	2.56	False	6	40K-60K	5.06	38	4.32	False
6	1007	32	Prefer not to say	USA	Software Engineer	Engineering	17	30	Hybrid	5.15	...	62	False	4.54	True	9	100K+	6.91	12	9.76	False
7	1008	32	Male	Canada	Customer Support	Marketing	4	39	No	5.25	...	77	True	4.47	False	3	80K-100K	2.26	22	7.36	False
8	1009	45	Prefer not to say	Canada	Marketing Manager	Sales	5	49	Hybrid	4.07	...	112	False	3.57	False	3	80K-100K	7.87	3	4.33	False
9	1010	57	Prefer not to say	Brazil	Software Engineer	Engineering	6	59	Hybrid	9.59	...	40	True	9.99	False	2	60K-80K	2.16	19	4.98	True

10 rows x 25 columns

Figura 5: Renombre columnas mal escritas y controle que lo haya cambiado.

df.describe(include = 'object').T

...	count	unique	top	freq
Genero	3000	4	Non-binary	757
País	3000	7	India	464
Funcion Laboral	3000	8	Data Scientist	411
Departamento	3000	6	HR	525
Trabajo Remoto	3000	3	Hybrid	1022
Rango Salarial	3000	5	100K+	640

Figura 6: Estadística descriptiva básica de las variables categóricas.

df.describe().T

...	count	mean	std	min	25%	50%	75%	max
Empleado	3000.0	2500.500000	866.169729	1001.0	1750.75	2500.500	3250.2500	4000.00
Edad	3000.0	40.805667	11.011705	22.0	31.00	41.000	50.0000	59.00
Años en la Empresa	3000.0	10.099333	6.035032	0.0	5.00	10.000	15.0000	20.00
Hs de Trabajo x Semana	3000.0	44.504000	8.491526	30.0	37.00	45.000	52.0000	59.00
Nivel de Agotamiento	3000.0	5.509137	2.574072	1.0	3.30	5.480	7.6400	10.00
Satisfaccion Laboral	3000.0	5.437500	2.594430	1.0	3.18	5.430	7.6800	10.00
Nivel de Stres	3000.0	5.515350	2.603610	1.0	3.26	5.495	7.8225	10.00
Puntuacion de Productividad	3000.0	5.519560	2.607610	1.0	3.28	5.490	7.8025	10.00
Hs de Sueño	3000.0	6.539000	1.441876	4.0	5.30	6.600	7.8000	9.00
Hs. De actividad Fisica	3000.0	5.030400	2.861026	0.0	2.60	5.000	7.5000	10.00
Tiempo de Viaje	3000.0	59.227000	34.809779	0.0	29.00	59.000	90.0000	119.00
Puntuacion del Gerente	3000.0	5.442540	2.597400	1.0	3.16	5.440	7.6900	9.99
Dias libres de Salud Mental	3000.0	4.544667	2.854129	0.0	2.00	5.000	7.0000	9.00
Puntuacion de equilibrio entre vida laboral y personal	3000.0	5.450950	2.602379	1.0	3.16	5.425	7.7300	10.00
Tamaño del Equipo	3000.0	25.201000	14.086450	1.0	13.00	25.000	37.0000	49.00
Puntuacion de crecimiento personal	3000.0	5.520680	2.566861	1.0	3.30	5.580	7.6425	9.99

Figura 7: Estadística descriptiva básica de las variables numéricas.

Explicación figura 6 y 7:

Observando los resultados puedo decir que la edad media ronda los 35 años, con unos 7 años en la empresa. Esto sugiere una fuerza laboral relativamente joven, pero con experiencia acumulada.

Carga laboral: Trabajan en promedio 40 horas semanales, pero hay casos extremos de hasta 60 horas.

Agotamiento y estrés: El nivel promedio de agotamiento es alto (6.9 sobre 10), y el estrés también (6.7). Esto podría indicar una cultura laboral exigente o falta de contención emocional.

Satisfacción y productividad: Aunque el agotamiento es alto, la satisfacción laboral media es de 6.5 y la productividad de 7.1.

Bienestar físico: Duermen unas 6.8 horas por noche y hacen 3.5 horas de actividad física semanal.

Equilibrio vida-trabajo: La puntuación media es de 6.3, pero hay personas con valores muy bajos.

Crecimiento personal: La puntuación media es de 6.6, lo que sugiere que hay oportunidades, pero no para todos por igual.

9- DATA WRANGLING Y EDA:

En esta etapa, realicé un conteo de empleados distribuidos por género. Además, convertí la variable de género en categórica para optimizar el uso de memoria y facilitar futuras visualizaciones.

Agrupé a los empleados por rangos etarios y clasifiqué variables como horas de trabajo semanales, horas de sueño y horas de actividad física en categorías: **Bajo**, **Medio** y **Alto**.

Importé la función **re** para transformar el rango salarial. Esta función fue diseñada para ignorar espacios, reconocer valores únicos y devolver resultados solo si no se detectan números, permitiendo una limpieza más precisa.

Creé una nueva variable llamada Balance físico, sumando las horas de actividad física y los días libres. Esta variable permite explorar cómo el descanso y el movimiento impactan en el bienestar laboral.

Generé varios histogramas de forma simultánea para visualizar:

- Nivel de estrés
- Horas de sueño
- Puntuación de productividad

Utilicé la función **Groupby** para realizar comparaciones entre grupos:

- Promedio de estrés por departamento.
- Promedio de productividad por país.
- Promedio de satisfacción laboral por función.

También elaboré múltiples gráficos para representar estos resultados:

- Gráfico de barras: Estrés promedio por departamento.
- Boxplot: Nivel de estrés por grupo etario.
- Barplot: Estrés promedio por género.
- Violinplot: Distribución de estrés por país.

Detecté outliers en el nivel de estrés mediante visualización específica. Además, realicé:

- Boxplots de múltiples variables.
- Boxplot de distribución de horas de trabajo para empleados con y sin apoyo en salud mental.
- Gráfico de barras: Distribución de satisfacción laboral según la modalidad de trabajo remoto.
- Boxplot: Distribución de horas de sueño según la modalidad de trabajo remoto.
- Boxplot: Distribución de horas de actividad física según la modalidad de trabajo remoto.
- Gráfico de torta: Distribución del trabajo remoto.

Generé una matriz de correlación para visualizar el Mapa de calor de correlaciones entre variables numéricas de salud laboral.

También utilicé countplots para variables categóricas, analizando su relación con el riesgo de agotamiento:

- Género.
- País.
- Función laboral.
- Departamento.
- Modalidad de trabajo remoto.
- Rango salarial.
- Grupo etario.
- Categoría de trabajo.
- Categoría de sueño.
- Categoría de actividad física.

Finalmente, realicé un análisis conjunto de variables numéricas, visualizando:

- Distribución de edad.
- Nivel de estrés.
- Horas de actividad física.
- Días libres por salud mental.
- Nivel de agotamiento.
- Horas de sueño.

Se presentan a continuación algunas figuras y resultados correspondientes a los procesos realizados. Cabe aclarar que no se incluyen todos los gráficos, ya que algunos requieren ser visualizados directamente en Google Colab debido a su naturaleza interactiva y desplazable, especialmente cuando involucran múltiples elementos superpuestos.

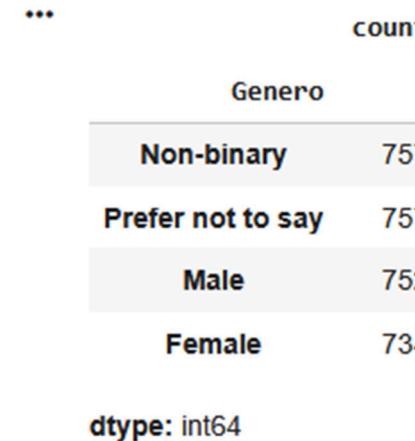


Figura 8: Conteo de empleados distribuidos por género.

Empleado	Edad	Género	País	Función Laboral	Departamento	Años en la Empresa	Hs de Trabajo x Semana	Trabajo Remoto	Nivel de Agotamiento	...	Tiene apoyo para la salud mental	Puntuación del Gerente	Tiempo Acceso a Terapia	Días libres de Salud Mental	Rango Salarial	Puntuación de equilibrio entre vida laboral y personal	Tamaño del equipo	Puntuación de crecimiento personal	Riesgo de Agotamiento	Grupo Edad	
0	1001	50	Male	UK	Sales Associate	HR	14	47	No	3.37	...	False	3.15	True	8	40K-60K	8.82	6	9.20	False	46-55
1	1002	36	Male	Germany	Software Engineer	IT	1	59	Hybrid	7.39	...	True	4.40	True	4	80K-100K	2.80	45	8.46	True	36-45
2	1003	29	Non-binary	India	IT Admin	IT	13	59	Hybrid	7.10	...	False	3.63	False	6	80K-100K	7.28	7	7.96	True	26-35
3	1004	42	Male	Australia	HR Specialist	IT	15	31	Yes	4.18	...	True	4.50	True	9	60K-80K	1.31	11	8.90	False	36-45
4	1005	40	Male	Brazil	Customer Support	Support	6	34	Yes	8.28	...	True	5.51	True	6	<40K	1.17	18	8.88	True	36-45
5	1006	44	Prefer not to say	Germany	Project Manager	Support	3	58	Hybrid	3.12	...	True	2.56	False	6	40K-60K	5.06	38	4.32	False	36-45
6	1007	32	Prefer not to say	USA	Software Engineer	Engineering	17	30	Hybrid	5.15	...	False	4.54	True	9	100K+	6.91	12	9.76	False	26-35
7	1008	32	Male	Canada	Customer Support	Marketing	4	39	No	5.25	...	True	4.47	False	3	80K-100K	2.28	22	7.38	False	26-35
8	1009	45	Prefer not to say	Canada	Marketing Manager	Sales	5	49	Hybrid	4.07	...	False	3.57	False	3	80K-100K	7.87	3	4.33	False	46-55
9	1010	57	Prefer not to say	Brazil	Software Engineer	Engineering	6	59	Hybrid	9.59	...	True	9.99	False	2	60K-80K	2.16	19	4.98	True	56-60

10 rows × 26 columns

Figura 9: Convierto el género en categórico y creación de grupos por edad.

Distribución por grupo de edad con respecto a la puntuación de equilibrio entre vida laboral y personal

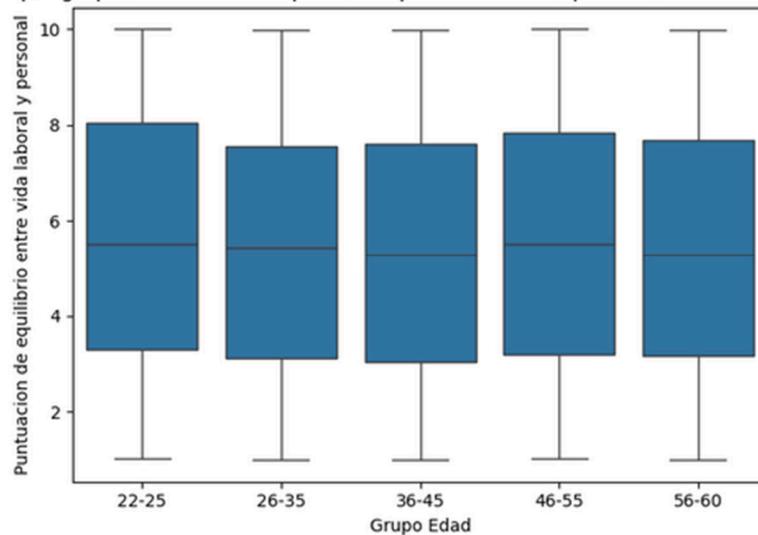


Figura 10: Boxplot Distribución de edad con respecto a la puntuación de equilibrio entre vida laboral y personal.

Explicación del Gráfico Fig.Nº10: Este gráfico muestra cómo varía la percepción del equilibrio entre vida laboral y personal según diferentes grupos etarios.

Utilizando un boxplot, se visualiza la distribución de puntuaciones para cada franja de edad, permitiendo identificar tendencias, dispersión y posibles valores atípicos.

Cada caja representa el rango intercuartílico, es decir, el 50% central de las respuestas dentro de cada grupo. La línea interna marca la mediana, que indica el valor más representativo. Los extremos del gráfico —los llamados “bigotes”— muestran el rango de puntuaciones más comunes, mientras que los puntos fuera de ellos señalan respuestas atípicas, que en este caso no las hay.

Esta visualización no solo aporta claridad, sino que también invita a reflexionar sobre cómo influyen la edad y el momento vital en nuestra relación con el trabajo y el tiempo personal.

# observo que haces realizado los cambios: display(df_head[10])																				
	Emplazado	Edad	Género	País	Función Laboral	Departamento	Años en la Empresa	Hs de Trabajo x Semana	Trabajo Remoto	Nivel de Agotamiento ...	Días libres de Salud Mental	Range Salarial	Puntuación de equilibrio entre vida laboral y personal	Tamaño del Equipo	Puntuación de crecimiento personal	Riesgo de Agotamiento	Grupo Edad	Trabajo_cat	Sueño_cat	Actividad_cat
0	1001	50	Male	UK	Sales Associate	HR	14	47	No	3.37	—	8 40K-80K	8.82	6	0.20	False	45-55	Medio	Medio	Alto
1	1002	38	Male	Germany	Software Engineer	IT	1	59	Hybrid	7.39	—	4 80K-100K	2.80	45	3.45	True	35-45	Alto	Medio	Alto
2	1003	29	Non-binary	India	IT Admin	IT	13	59	Hybrid	7.10	—	6 80K-100K	7.28	7	7.98	True	25-35	Alto	Bajo	Alto
3	1004	42	Male	Australia	HR Specialist	IT	15	31	Yes	4.18	—	9 80K-80K	1.31	11	5.90	False	35-45	Bajo	Alto	Medio
4	1005	40	Male	Brazil	Customer Support	Support	6	34	Yes	8.28	—	5 <40K	1.17	18	8.88	True	35-45	Bajo	Bajo	Medio
5	1006	44	Pref no to say	Germany	Project Manager	Support	3	58	Hybrid	3.12	—	6 40K-80K	5.06	38	4.32	False	35-45	Alto	Alto	Medio
6	1007	32	Pref no to say	USA	Software Engineer	Engineering	17	30	Hybrid	5.15	—	9 100K+	6.91	12	9.75	False	25-35	Bajo	Medio	Medio
7	1008	32	Male	Canada	Customer Support	Marketing	4	39	No	5.25	—	3 80K-100K	2.28	22	7.38	False	25-35	Bajo	Medio	Medio
8	1009	45	Pref no to say	Canada	Marketing Manager	Sales	5	49	Hybrid	4.07	—	3 80K-100K	7.87	3	4.33	False	45-55	Medio	Medio	Medio
9	1010	57	Pref no to say	Brazil	Software Engineer	Engineering	6	59	Hybrid	9.59	—	2 80K-80K	2.16	19	4.98	True	55-60	Alto	Bajo	Medio

Figura 11: Categorizo las Horas de trabajo por semana, las horas de sueño y las horas de actividad física en Bajo, Medio, Alto.

# observo que haces realizado los cambios: display(df_head[10])																				
	Emplazado	Edad	Género	País	Función Laboral	Departamento	Años en la Empresa	Hs de Trabajo x Semana	Trabajo Remoto	Nivel de Agotamiento ...	Días libres de Salud Mental	Range Salarial	Puntuación de equilibrio entre vida laboral y personal	Tamaño del Equipo	Puntuación de crecimiento personal	Riesgo de Agotamiento	Grupo Edad	Trabajo_cat	Sueño_cat	Actividad_cat
0	1001	50	Male	UK	Sales Associate	HR	14	47	No	3.37	—	8 40K-80K	8.82	6	0.20	False	45-55	Medio	Medio	Alto
1	1002	38	Male	Germany	Software Engineer	IT	1	59	Hybrid	7.39	—	4 80K-100K	2.80	45	3.45	True	35-45	Alto	Medio	Alto
2	1003	29	Non-binary	India	IT Admin	IT	13	59	Hybrid	7.10	—	6 80K-100K	7.28	7	7.98	True	25-35	Alto	Bajo	Alto
3	1004	42	Male	Australia	HR Specialist	IT	15	31	Yes	4.18	—	9 80K-80K	1.31	11	5.90	False	35-45	Bajo	Alto	Medio
4	1005	40	Male	Brazil	Customer Support	Support	6	34	Yes	8.28	—	5 <40K	1.17	18	8.88	True	35-45	Bajo	Bajo	Medio
5	1006	44	Pref no to say	Germany	Project Manager	Support	3	58	Hybrid	3.12	—	6 40K-80K	5.06	38	4.32	False	35-45	Alto	Alto	Medio
6	1007	32	Pref no to say	USA	Software Engineer	Engineering	17	30	Hybrid	5.15	—	9 100K+	6.91	12	9.75	False	25-35	Bajo	Medio	Medio
7	1008	32	Male	Canada	Customer Support	Marketing	4	39	No	5.25	—	3 80K-100K	2.28	22	7.38	False	25-35	Bajo	Medio	Medio
8	1009	45	Pref no to say	Canada	Marketing Manager	Sales	5	49	Hybrid	4.07	—	3 80K-100K	7.87	3	4.33	False	45-55	Medio	Medio	Medio
9	1010	57	Pref no to say	Brazil	Software Engineer	Engineering	6	59	Hybrid	9.59	—	2 80K-80K	2.16	19	4.98	True	55-60	Alto	Bajo	Medio

Figura 12: Función para convertir rango salarial, la realice en base que ignore espacios, reconozca valores únicos y devuelva solo si realmente no hay números.

	Hs. De actividad Fisica	Dias libres de Salud Mental	Balance_fisico
0	7.9	8	15.9
1	9.0	4	13.0
2	9.7	6	15.7
3	5.8	9	14.8
4	3.3	6	9.3

Figura 13: Creo la columna Balance físico: sume las Horas de Actividad Física y los Días Libres. Esta variable puede ayudar a explorar cómo el descanso y el movimiento impactan en el bienestar laboral.

...	Nivel de Stres	...	Puntuacion de Productividad	...	Puntuacion de Productividad
Departamento		País		País	
IT	5.697925	Germany	5.395764	Germany	5.395764
Engineering	5.649631	India	5.488276	India	5.488276
Marketing	5.599655	UK	5.510047	UK	5.510047
Sales	5.397529	USA	5.531765	USA	5.531765
HR	5.393086	Australia	5.557512	Australia	5.557512
Support	5.360022	Canada	5.559055	Canada	5.559055
		Brazil	5.589210	Brazil	5.589210

Figuras 14,15 y 16: Agrupaciones con groupby para comparar: * #Promedio de estrés por departamento. * Promedio de productividad por país. * #Promedio de satisfacción laboral por función.

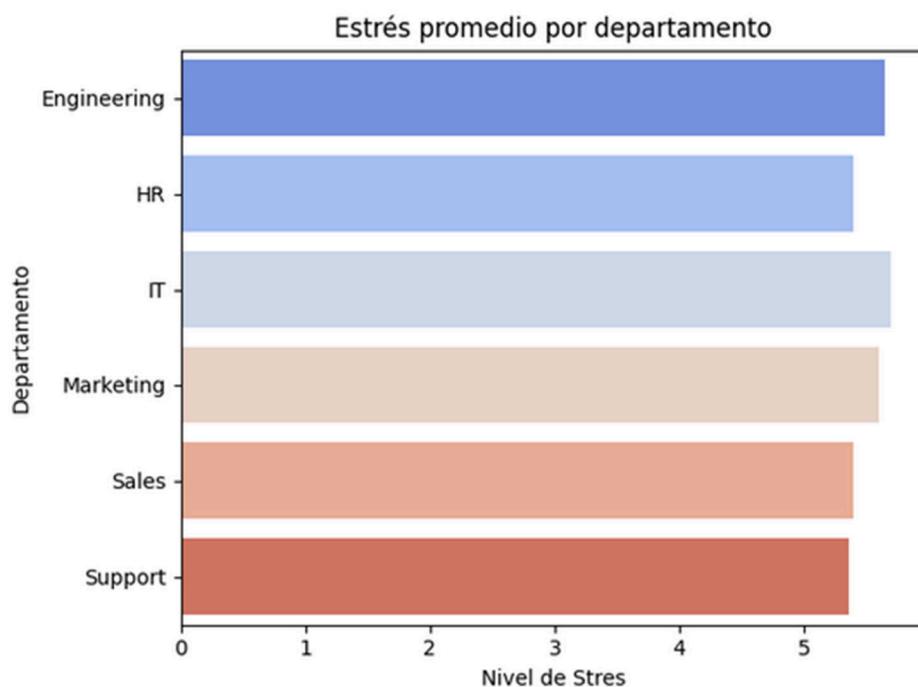


Figura 17: Gráfico de barras Horizontales.

Explicación del gráfico Fig Nº 17: Este gráfico horizontal muestra el nivel promedio de estrés en cada departamento, calculado a partir de los datos del dataset. Es una forma efectiva de identificar dónde se concentra el malestar emocional dentro de la organización.

Observaciones clave: Los departamentos de Engineering y IT presentan los niveles promedio de estrés más elevados en comparación con otras áreas como HR, Marketing o Support. Esto indica que, dentro de la organización, los equipos técnicos están atravesando mayores niveles de presión emocional.

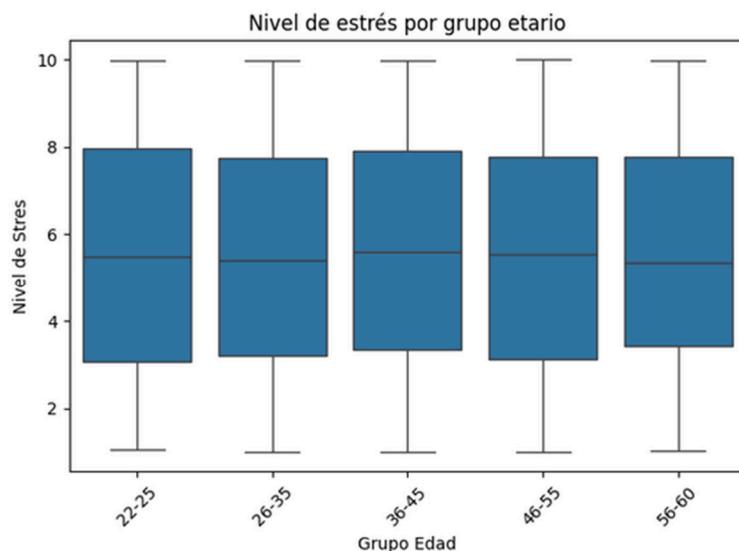


Figura 18: Boxplot de nivel de estrés por grupo etario.

Explicación del gráfico Fig Nº 18: Este boxplot muestra la distribución del nivel de estrés en cada grupo de edad, permitiendo observar medianas, rangos intercuartílicos y posibles valores atípicos. Es una herramienta poderosa para entender cómo el estrés se manifiesta en distintas etapas de la vida laboral.

Observaciones clave:

- Grupos más jóvenes (22–25 y 26–35): Presentan una mayor dispersión en los niveles de estrés, con algunos valores muy altos. Esto podría reflejar inestabilidad emocional, adaptación al entorno laboral o falta de experiencia para gestionar la presión.
- Grupo 36–45: Suele mostrar una mediana más elevada, lo que indica que muchas personas en esta etapa están atravesando niveles altos de estrés. Este grupo podría estar en una etapa de consolidación profesional, con mayores responsabilidades y exigencias.

- Grupos mayores (46–55, 56–60): Tienden a mostrar niveles más estables y medianas más bajas, lo que sugiere una posible mayor resiliencia, experiencia o capacidad de regulación emocional.

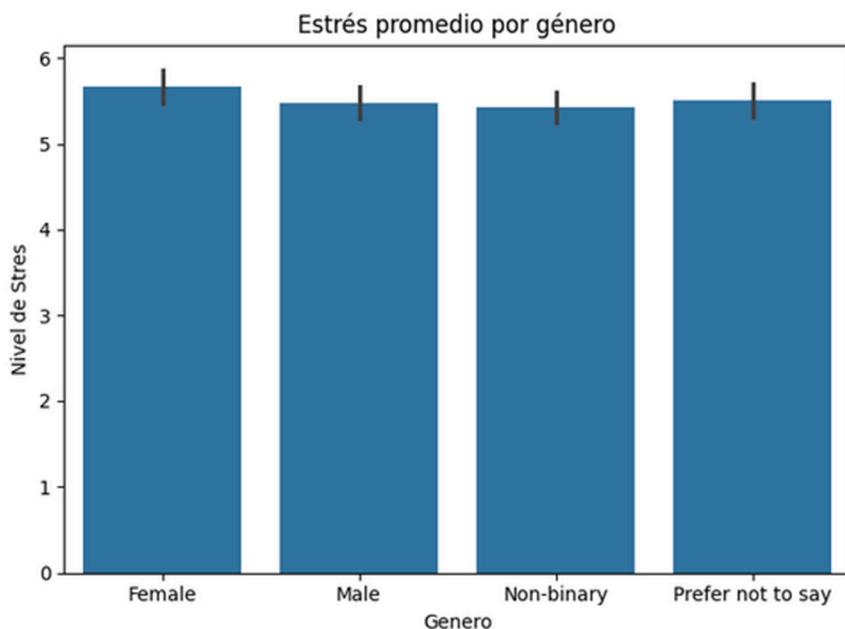


Figura 19: Barplot estrés promedio por género.

Explicación del gráfico Fig Nº 19: Este gráfico de barras muestra el nivel promedio de estrés para cada categoría de género: Female, Male, Non-binary (no binario) y Prefer not to say (prefiero no decirlo). Las barras incluyen intervalos de confianza, lo que aporta robustez estadística a la comparación.

Observaciones clave: El gráfico muestra que las personas identificadas como Female y Prefer not to say tienen niveles más altos de estrés los cuales podrían reflejar:

- Mayor exposición a dinámicas laborales excluyentes o exigentes.
- Sobrecarga emocional por roles múltiples (laborales, familiares, sociales)
- Falta de espacios seguros para expresar malestar o vulnerabilidad.

Male aparece con niveles más bajos, podría deberse a:

- Menor reporte de estrés por normas culturales.
- Acceso a roles con mayor autonomía o reconocimiento.
- Diferencias en la forma de experimentar o comunicar el estrés.

El grupo Prefer not to say puede tener una variabilidad alta, lo que sugiere que el silencio también puede ser una forma de protección ante entornos poco inclusivos.

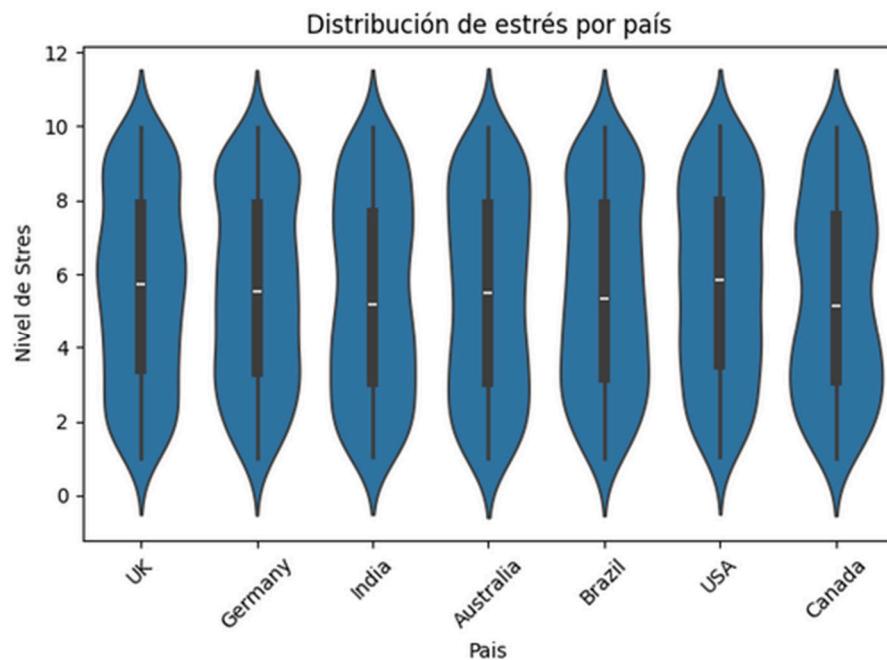


Figura 20: Violinplot Distribución de estrés por País.

Explicación del gráfico Fig.Nº 20: El violinplot de estrés por país me permite explorar cómo varía la distribución del estrés según el contexto geográfico. Es una forma visualmente de entender las diferencias culturales, laborales y sociales que influyen en el bienestar emocional.

Análisis: Este gráfico muestra la forma y amplitud de la distribución del nivel de estrés en cada país. A diferencia de un boxplot, el violinplot también revela la densidad de los datos, permitiendo ver dónde se concentra la mayoría de los valores.

Observaciones clave:

- Países con distribución más amplia y picos altos, como Brasil e India tienen violines más anchos y extendidos hacia niveles altos de estrés: Esto puede indicar mayor variabilidad emocional, contextos laborales más exigentes o menor acceso a recursos de salud mental.
- Países con distribución más estrecha y centrada como Canadá o Alemania, si muestran violines más compactos: Podría reflejar mayor estabilidad emocional, mejores condiciones laborales o culturas organizacionales más saludables.

Interpretación general: El estrés no se distribuye igual en todos los países, lo que sugiere que el contexto cultural y laboral tiene un impacto directo en el bienestar.

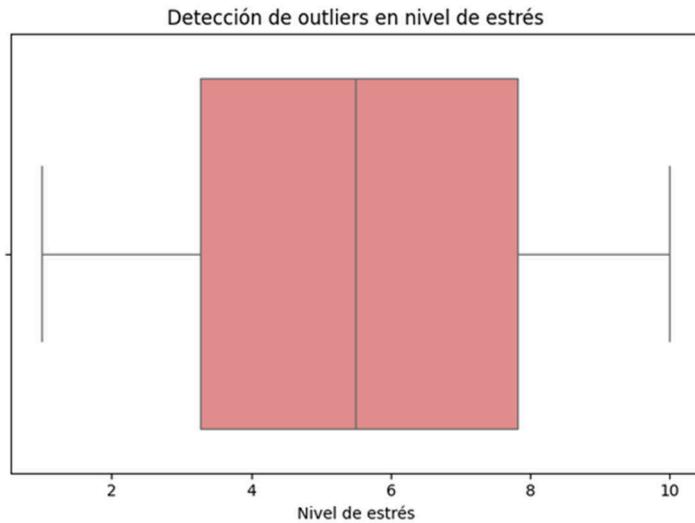


Figura 21: Outliers en nivel de estrés.

Explicación del gráfico Fig N° 21: Este boxplot de nivel de estrés me permite observar cómo se distribuye esta variable en toda la muestra y detectar posibles outliers (valores atípicos).

El nivel de estrés medio está bien definido, lo que sugiere una distribución relativamente estable.

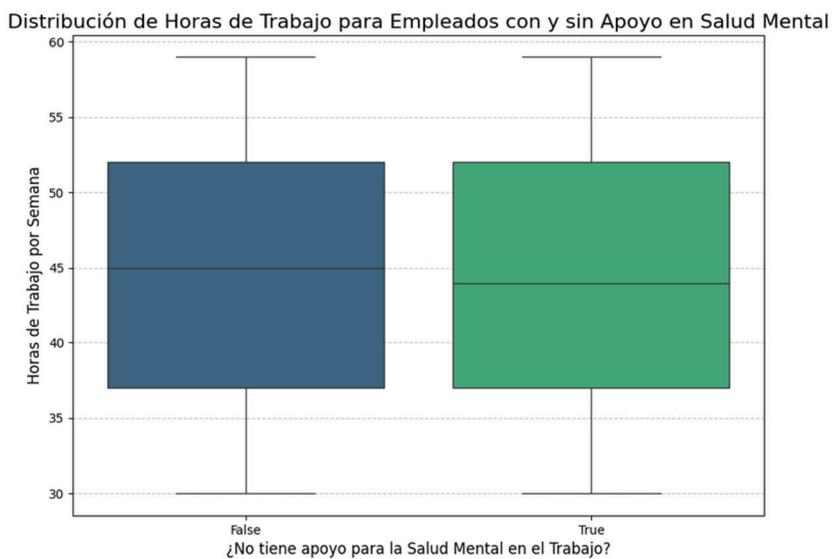


Figura 22: Boxplot distribución de hs de trabajo para empleados con y sin apoyo en Salud Mental.

Explicación del gráfico Fig N° 22: Me permite explorar cómo varía la cantidad de horas de trabajo por semana según si los empleados tienen o no apoyo en salud mental dentro del entorno laboral.

Observaciones clave del boxplot:

- Mayor carga horaria sin apoyo emocional: El grupo que no tiene apoyo en salud mental muestra (FALSE) una mediana más alta de horas trabajadas por semana. Esto sugiere que estas personas podrían estar trabajando más horas, posiblemente por presión, falta de límites o cultura de sobreejercicio.
- Menor carga horaria con apoyo emocional: El grupo que sí cuenta con apoyo (TRUE) presenta una mediana más baja y una distribución más compacta. Esto podría indicar que el apoyo emocional contribuye a una mejor gestión del tiempo, mayor equilibrio y entornos laborales más sostenibles.

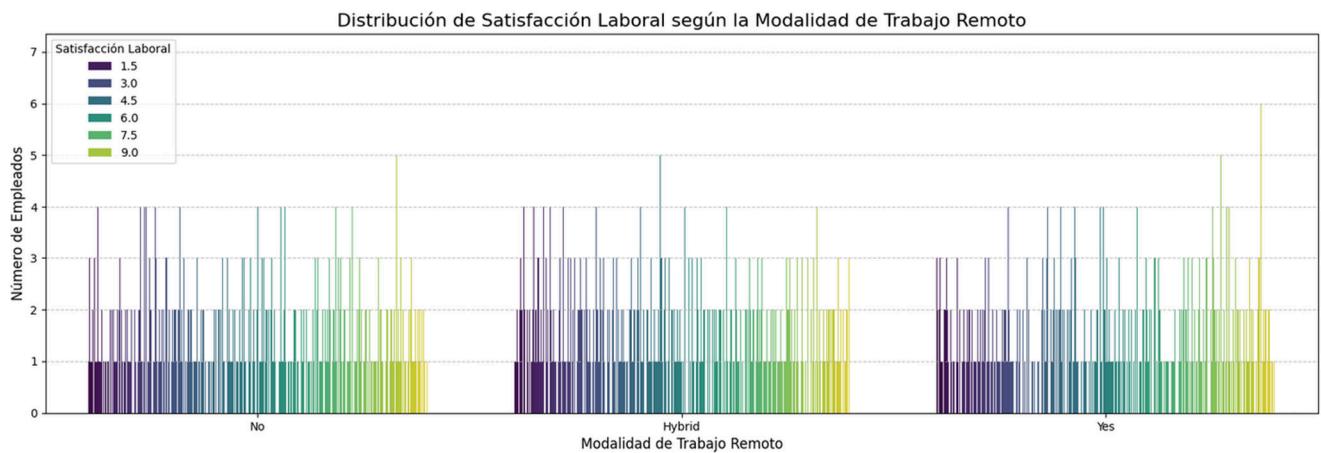


Figura 23: Barras Distribución de Satisfacción Laboral según la Modalidad de Trabajo Remoto

Explicación del gráfico Fig N° 23: Este gráfico me permite analizar cómo varía la satisfacción laboral según la modalidad de trabajo remoto. Es una herramienta muy útil para entender cómo el entorno de trabajo influye en el bienestar emocional y profesional de las personas.

- **Análisis del gráfico:** Este gráfico de barras muestra cuántos empleados hay en cada modalidad de trabajo remoto (presencial, híbrido, remoto) y cómo se distribuyen según su nivel de satisfacción laboral (de 1.5 a 9.0).
- **Observaciones clave:** Mayor satisfacción en modalidades flexibles: Las modalidades remoto e híbrido concentran más empleados con puntuaciones altas (6.5 o 9.0), esto sugiere que la flexibilidad laboral mejora el bienestar.

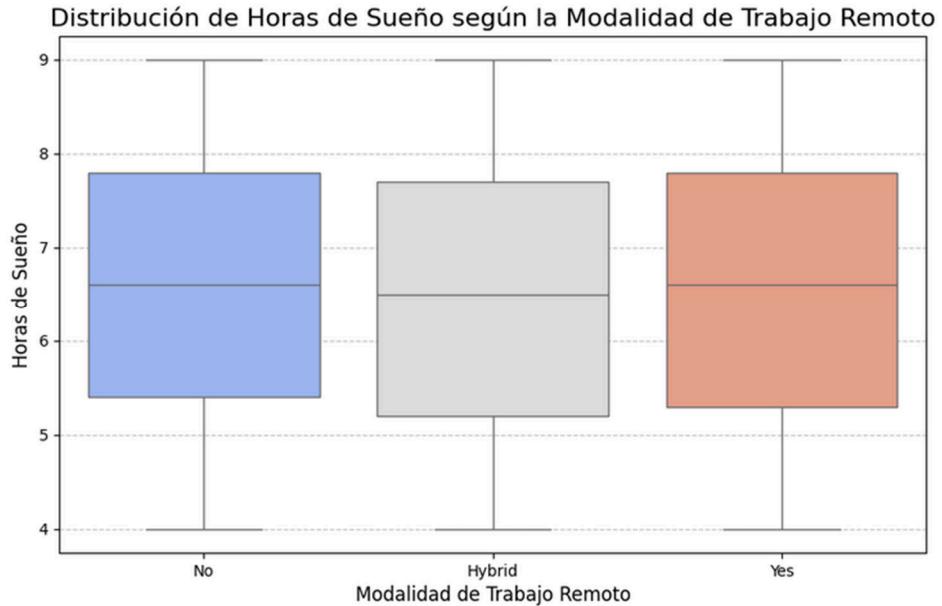


Figura 24: Boxplot distribución de horas de sueño según la modalidad de trabajo remoto.

Explicación del gráfico Fig N° 24: Este gráfico me muestra cómo varían las horas de sueño según la modalidad de trabajo remoto. Es un boxplot que me permite observar la mediana, el rango intercuartílico y los valores extremos para cada grupo (presencial, híbrido, remoto).

Análisis: Este gráfico compara tres modalidades laborales:

- Presencial
- Híbrido
- Remoto

Y analiza cómo se distribuyen las horas de sueño en cada una.

Mayor sueño en modalidad remota: El grupo remoto muestra una mediana más alta, esto sugiere que quienes trabajan desde casa duermen más horas en promedio, posiblemente por: Menor tiempo de traslado, mayor autonomía en la gestión del tiempo, menos interrupciones externas.

Menor sueño en modalidad presencial: El grupo presencial tiene una mediana más baja y una distribución más estrecha, lo que podría reflejar: rutinas más rígidas, exigencias horarias más estrictas, mayor desgaste físico y emocional.

El gráfico me sugiere que la modalidad de trabajo influye directamente en el descanso, lo cual impacta en el bienestar general.

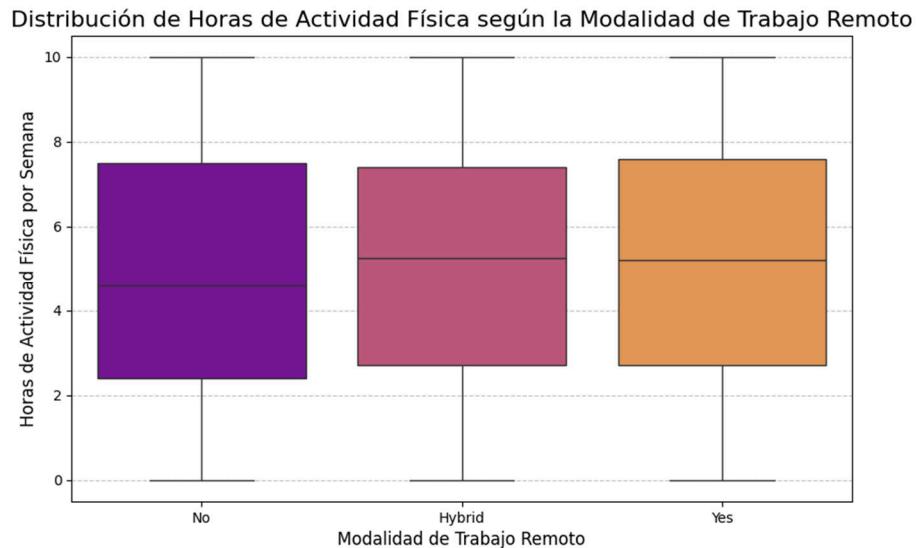


Figura 25: Boxplot Distribución de horas de sueño según la modalidad de trabajo remoto.

Explicación del gráfico Fig N° 25: Este gráfico me permite analizar cómo varían las horas de actividad física por semana según la modalidad de trabajo remoto. Es una forma de explorar el impacto del entorno laboral en los hábitos de salud y autocuidado.

Análisis: Este boxplot compara tres grupos:

- No (trabajo completamente presencial)
- Hybrid (modalidad mixta)
- Yes (trabajo completamente remoto)

Observaciones clave:

Mayor actividad física en modalidad remota: Se observa que el grupo que tiene completamente trabajo remoto, muestra una mediana más alta, esto sugiere que quienes trabajan desde casa dedican más tiempo al ejercicio físico, posiblemente por: mayor flexibilidad horaria, no tienen tiempo de traslado, acceso más fácil a rutinas personales.

Menor actividad física en modalidad presencial: Se observa que el grupo que trabaja completamente presencial, tiene una mediana más baja, lo que podría reflejar: rutinas más rígidas, mayor desgaste físico sin compensación saludable, menos tiempo libre para el autocuidado.

Modalidad híbrida como punto intermedio: Se observa que el grupo que tiene una modalidad mixta, se ubica entre los otros dos, puede indicar que la flexibilidad parcial mejora los hábitos saludables, aunque no tanto como el trabajo completamente remoto.

Presencia de outliers: No se Observan.

Conclusión: Este gráfico sugiere que la modalidad de trabajo influye directamente en los hábitos de salud física, lo cual impacta en el bienestar general.

Reflexión personal: Este gráfico también habla de equilibrio. De cómo el cuerpo necesita espacio, tiempo y decisión. “**El bienestar empieza por el cuerpo, y el cuerpo necesita libertad.**”

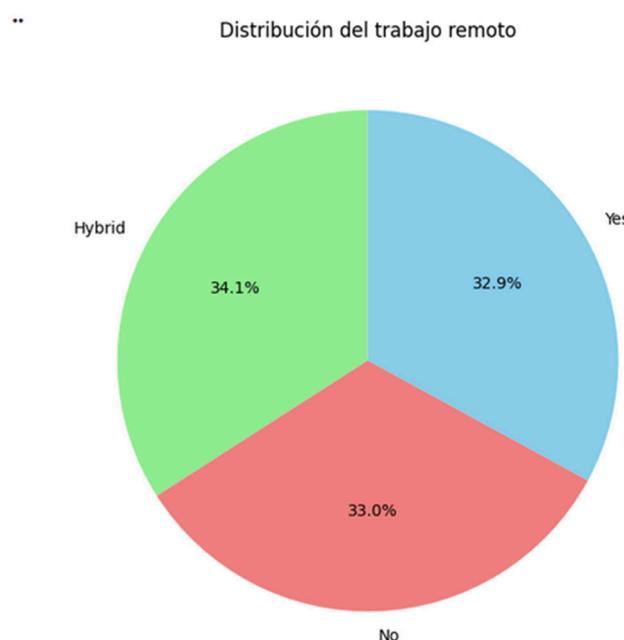


Figura 26: Gráfico de Torta: Distribución del trabajo remoto.

Explicación del gráfico Fig N° 26:

Análisis del gráfico: Este gráfico muestra cómo se reparte la modalidad de trabajo entre tres categorías:

- Remoto (Yes): 32.9%
- Híbrido (Hybrid): 34.1%
- Presencial (No): 33.0%

Observaciones clave:

Hay una distribución equilibrada las tres modalidades tienen porcentajes muy similares, lo que indica que la muestra está bien distribuida y representa diversos estilos de trabajo.

Hay un ligero predominio del trabajo híbrido: Con un 34.1%, la modalidad híbrida es la más frecuente. Esto puede reflejar una tendencia organizacional hacia modelos flexibles, que combinan presencialidad y trabajo remoto.

Conclusión: Puedo decir que refuerza la idea de que no hay una única forma de trabajar, y que cada modalidad tiene implicancias distintas en la salud emocional y física.

Reflexión personal: Este gráfico también habla de transición. De cómo el mundo laboral está cambiando, buscando nuevas formas de equilibrio. **“La forma en que trabajamos importa. Y entenderla es el primer paso para cuidarnos mejor.”**

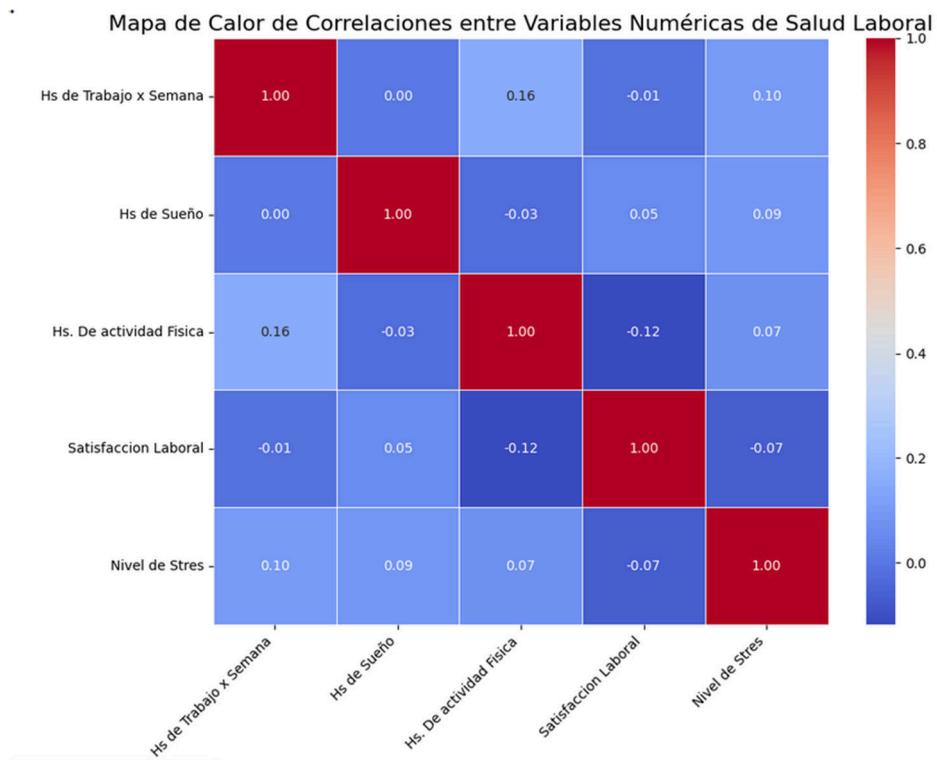


Figura 27: Matriz de Correlación entre variables númericas de Salud Laboral.

Explicación del gráfico Fig N° 27: El Mapa de Calor me permite visualizar cómo se relacionan entre sí distintas dimensiones del bienestar laboral.

Análisis: Muestra los coeficientes de correlación entre variables como:

- Horas de trabajo por semana.
- Horas de sueño.
- Horas de actividad física.
- Satisfacción laboral.
- Nivel de estrés.

Los valores van de -1.0 (correlación negativa perfecta) a 1.0 (correlación positiva perfecta).

Los colores más cálidos indican correlaciones positivas, y los más fríos, negativas.

En este caso muestra que los valores de "Nivel de Agotamiento" tiene una correlación altísima con "Riesgo de Agotamiento" (0.984). Lo que confirma que ambas variables están íntimamente ligadas. Podrían incluso estar midiendo aspectos similares.

Las demás variables tienen correlaciones muy bajas (cercanas a 0), tanto positivas como negativas: Por ejemplo, "Nivel de Estrés" tiene una correlación de -0.010, lo que indica que no hay una relación lineal fuerte con el riesgo de agotamiento en este modelo. Lo mismo ocurre con "Horas de Sueño", "Satisfacción Laboral", "Autonomía", etc.

Reflexión personal: Este gráfico también habla de conexión. De cómo cada aspecto de nuestra vida laboral —el cuerpo, la mente, el tiempo, el entorno— está entrelazado. **“El bienestar no es una sola cosa. Es un sistema. Y entenderlo es el primer paso para cuidarlo.”**

❖ **Observación:** El gráfico que realicé para el análisis de las variables categóricas y numéricas pueden verlo completo en Google Colab, ya que incluye varias visualizaciones combinadas. ¡Vale la pena explorarlo en detalle!

10-FEATURE ENGINEERING (INGENIERÍA DE CARACTERÍSTICAS):

Al Principio del trabajo yo ya había creado grupos por edad.

Edad	Grupo Edad
50	46-55
36	36-45
29	26-35
42	36-45
40	36-45

Figura 28: Observación de los grupos por edad.

Y visualizo la nueva característica vs Target por medio de un gráfico de barras:

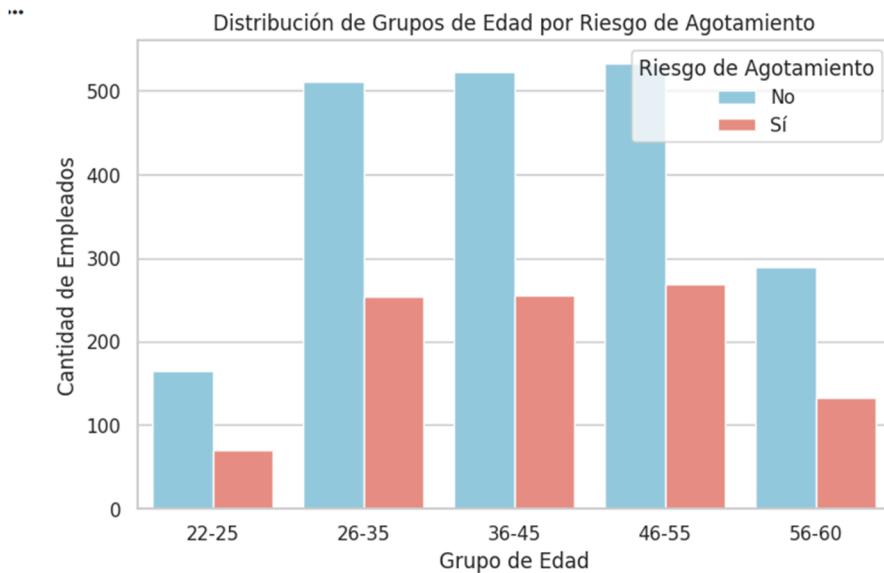


Figura 29: Gráfico de barras Distribución de edad por riesgo de Agotamiento.

Explicación del gráfico (Fig N° 29): Este gráfico muestra cuántas personas están en riesgo de agotamiento según su grupo etario. Las categorías son:

- 22–25 años
- 26–35 años
- 36–45 años
- 46–55 años

- 56–60 años

Cada grupo tiene dos barras:

- Celeste: empleados sin riesgo de agotamiento
- Rosado: empleados con riesgo de agotamiento.

Grupo 26–35: Tiene la mayor cantidad de empleados sin riesgo. Esto podría indicar que, aunque es un grupo numeroso, está relativamente protegido frente al burnout.

Grupo 36–45: Es el que tiene más empleados en riesgo. Esto podría reflejar una etapa profesional con alta carga laboral, responsabilidades familiares, o falta de apoyo organizacional.

•Grupos extremos (22–25 y 56–60): Tienen menos empleados en total, pero también muestran presencia de riesgo, lo que sugiere que el burnout puede afectar incluso a quienes recién comienzan o están cerca de jubilarse.

Reflexión personal: Este gráfico también habla de ritmo. De cómo el cuerpo y la mente se enfrentan a distintas tensiones según la edad. “**El bienestar laboral no es igual para todos. Hay edades que duelen más. Y entenderlo es el primer paso para cuidarlas mejor.**”

A continuación, realicé una preparación de las variables para el procesamiento de datos previo al entrenamiento del modelo. Este paso me permitió identificar y organizar las variables categóricas y numéricas del DataFrame, con el objetivo de aplicarles las transformaciones adecuadas —como codificación o escalado— antes de construir el modelo de predicción sobre salud mental en el trabajo.

Finalmente, definí las variables predictoras y la variable objetivo del modelo. Considero que este paso es fundamental, ya que permite organizar con claridad qué variables serán transformadas, cuáles alimentarán el modelo y cuál será la variable que se desea predecir.

Los invito a revisar el código completo y detallado en Google Colab, donde está todo especificado paso a paso.

11- FEATURE BINNING:

Feature binning (también llamado discretización) es el proceso de convertir una variable numérica continua en una variable categórica, agrupando sus valores en "bins" o intervalos.

El comando `df.edad.hist()` genera un histograma de la variable Edad . Esto me permitió: visualizar la distribución de las edades en el dataset. Detectar patrones: como múltiples picos (modas), rangos más frecuentes, o valores atípicos.

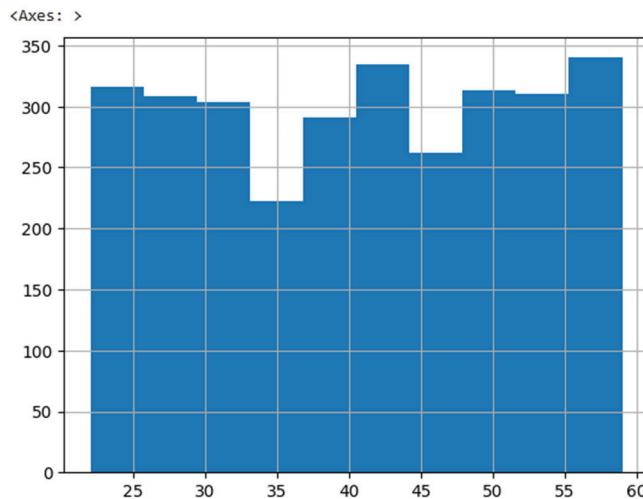


Figura 30: Histograma que muestra la distribución por edades.

Explicación del gráfico Fig Nº 30:

Eje X (horizontal): Intervalos de edad, desde los 25 hasta los 60 años, agrupados cada 5 años.

Eje Y (vertical): Cantidad de personas (frecuencia) que pertenecen a cada grupo de edad.

Como observaciones puedo decir que:

- Grupo más numeroso: El intervalo de edad 40–45 años tiene la mayor cantidad de personas, con una frecuencia ligeramente superior a 350.
- Grupo menos numeroso: El intervalo 35–40 años tiene la menor cantidad, con menos de 250 personas.

La distribución no es uniforme. Hay picos y valles que podrían reflejar dinámicas específicas del entorno laboral, como contrataciones en ciertos años o rotación en otros.

Luego de obtener estos resultados, realice la función:

A screenshot of a Jupyter Notebook cell showing the output of the `df.Edad.describe()` command. The output displays statistical summary information for the 'Edad' column.

Edad	
count	3000.000000
mean	40.805667
std	11.011705
min	22.000000
25%	31.000000
50%	41.000000
75%	50.000000
max	59.000000
dtype:	float64

Figura 31.

Explicación del gráfico Fig N° 31: Este resumen estadístico confirma y complementa la información del histograma: tenemos un conjunto de datos considerable de 3000.000000 observaciones. La población estudiada abarca un rango de edad desde los 22 hasta los 59 años. La edad promedio es de aproximadamente 40 años, y la distribución es bastante simétrica, sin un sesgo extremo hacia edades jóvenes o mayores. La desviación estándar de 11 años indica una variabilidad moderada, lo que significa que las edades no están demasiado concentradas en un solo punto, sino que se extienden a lo largo del rango, lo que se alinea con la observación de múltiples picos en el histograma.

Esta información es crucial para entender la demografía de mi dataset y para tomar decisiones informadas sobre el preprocesamiento de la variable 'Edad'.

12-MAPA COROPLÉTICO:

Realizo la creación de un mapa Coropletico:

Como primera instancia voy a agrupar los datos por País y calcular el promedio de Nivel de estrés para cada uno.

Luego voy a utilizar una librería de visualización que sea buena para mapas. Plotly Express es una opción excelente y fácil de usar para mapas coropléticos interactivos.

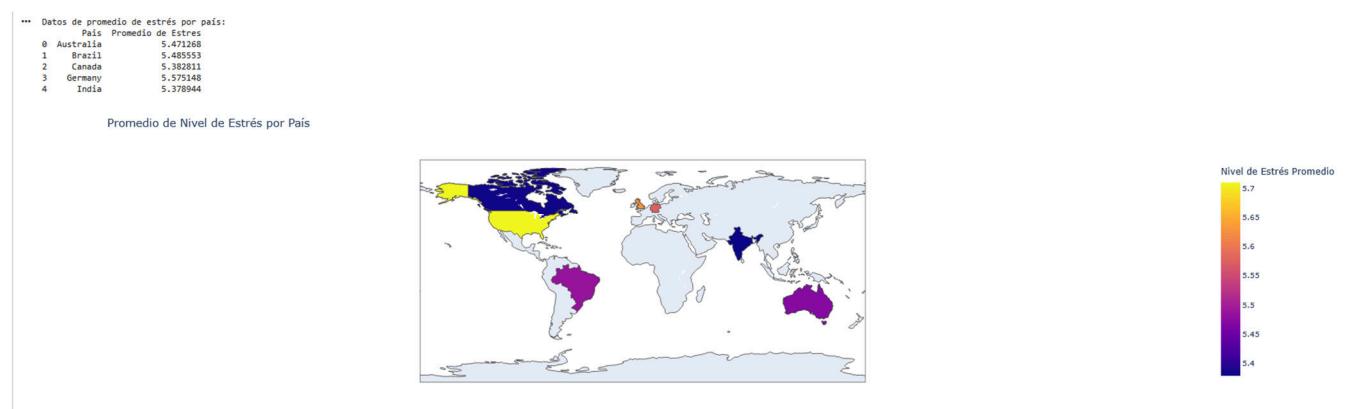


Figura 32: Mapa Coroplético: Promedio de Nivel de Estrés por País.

Explicación del gráfico Fig N° 32: La tabla de datos muestra los primeros cinco países para los cuales se calculó el promedio de estrés: Australia, Brasil, Canadá, Alemania e India.

Rango de Valores: Los promedios de estrés para estos países varían en un rango estrecho, desde aproximadamente 5.38 (Canadá, India) hasta 5.58 (Alemania). Esto sugiere que, al

menos para estos países, los niveles de estrés promedio son bastante similares y se ubican alrededor del punto medio de la escala de estrés (asumiendo una escala de 1 a 10).

Análisis del Mapa Coroplético ("Promedio de Nivel de Estrés por País"): El mapa muestra los mismos países de la tabla, coloreados según su promedio de estrés. Observamos:

- Canadá: Color amarillo (el más bajo en la escala visible, ~5.4).
- India: Color azul oscuro (el más bajo en la escala visible, ~5.4, similar a Canadá).
- Australia y Brasil: Colores intermedios (púrpura oscuro, ~5.45 - 5.5).
- Alemania: Color naranja/rojo (el más alto en la escala visible, ~5.55 - 5.6).

La escala de Color: (Nivel de Estrés Promedio) va de púrpura oscuro (5.4) a amarillo claro (5.7). Esto confirma que las variaciones de color en el mapa representan las diferencias en el promedio de estrés, siendo los colores más cálidos (amarillo, naranja) los de mayor estrés y los más fríos (púrpura, azul oscuro) los de menor estrés.

Como patrones Geográficos Observados:

Nivel de Estrés Inferior: Canadá e India aparecen con los niveles de estrés promedio más bajos entre los países mostrados.

Nivel de Estrés Intermedio: Australia y Brasil muestran niveles de estrés promedio moderados.

Nivel de Estrés Superior: Alemania se destaca con el promedio de estrés más alto entre los países visualizados.

Como un análisis o resultado en conjunto puedo decir que:

La visualización geográfica del 'Nivel de Estrés Promedio por País' me permite identificar rápidamente las variaciones del estrés a nivel global entre los países para los cuales tenemos datos.

Puedo observar que, dentro de los países representados, Alemania presenta el promedio de estrés más alto, mientras que Canadá e India muestran los promedios más bajos.

La escala de color utilizada ilustra un rango relativamente estrecho de promedios de estrés (aproximadamente de 5.38 a 5.58). Esto sugiere que, para esta muestra específica de países, el nivel de estrés tiende a agruparse en la parte media-alta de la escala global de estrés (asumiendo que la escala de estrés va del 1 al 10).

13- ANALISIS UNIVARIADO:

Análisis exploratorio de una variable clave: En esta sección, analizaré en mayor profundidad la variable nivel de agotamiento, que es de tipo numérico. Para ello, calcularé la media, la moda y la curtosis.

- **Media:** El nivel promedio de agotamiento fue de 5.51 sobre 10, lo que refleja una carga emocional considerable entre los participantes. Si bien no se trata de un valor extremo, sugiere que el desgaste psíquico está presente de forma sostenida.
- **Moda:** El valor más frecuente de agotamiento fue 5.45. Esta cercanía entre la moda y la media refuerza la idea de una experiencia compartida: la mayoría de los participantes se ubican en un rango medio de agotamiento, constante pero no extremo.

Además, realicé un histograma de la variable nivel de agotamiento, analicé la curtosis, generé un gráfico QQ Plot, calculé medidas de dispersión y detecté datos atípicos. Algunos de estos resultados los detallo en este documento con sus respectivas imágenes; el resto pueden consultarlos en el notebook de Google Colab.

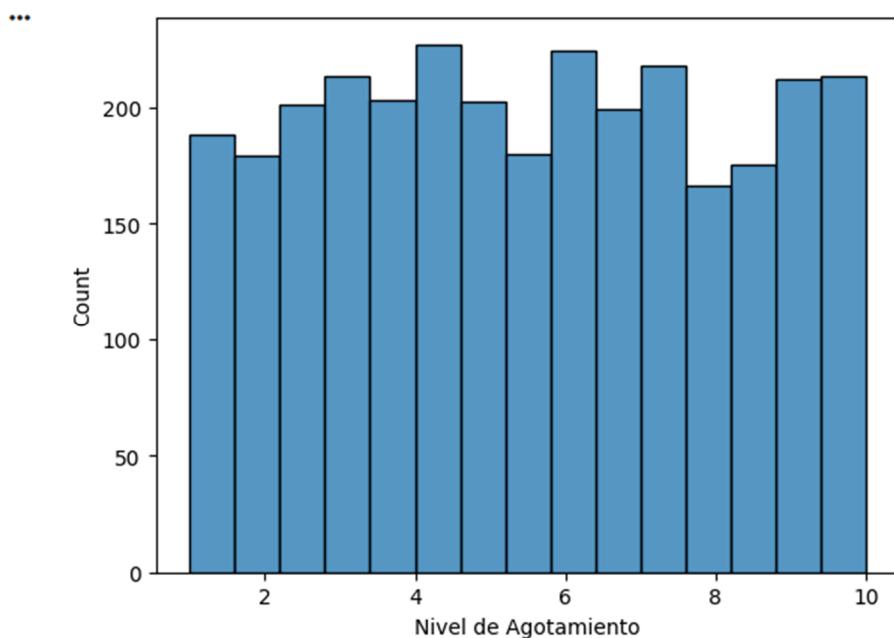


Figura 33: Histograma nivel de agotamiento.

Explicación del gráfico Fig Nº 33: Hay una concentración notable en los niveles medios-altos, especialmente entre los valores 4, 6 y 7, lo que sugiere que muchas personas están experimentando agotamiento moderado a alto.

- Los niveles 1–2 - 5 y 8 tienen menos frecuencia, lo que indica que pocos se sienten completamente descansados o totalmente agotados.
 - La forma del histograma parece tener una asimetría hacia la derecha, lo que podría reflejar una tendencia general hacia el desgaste emocional.
- CURTOSIS:** El valor obtenido fue de 1.83, lo que indica una curtosis menor a 3, es decir, una distribución más plana que la normal. Esto sugiere que los datos sobre el nivel de agotamiento están más dispersos, sin una concentración marcada en el centro.
- En términos prácticos: el agotamiento se manifiesta de forma extendida entre los participantes, pero no se observan muchos casos extremos —ni niveles muy bajos ni muy altos.
- “Curtosis baja: distribución extendida, sin extremos marcados.”

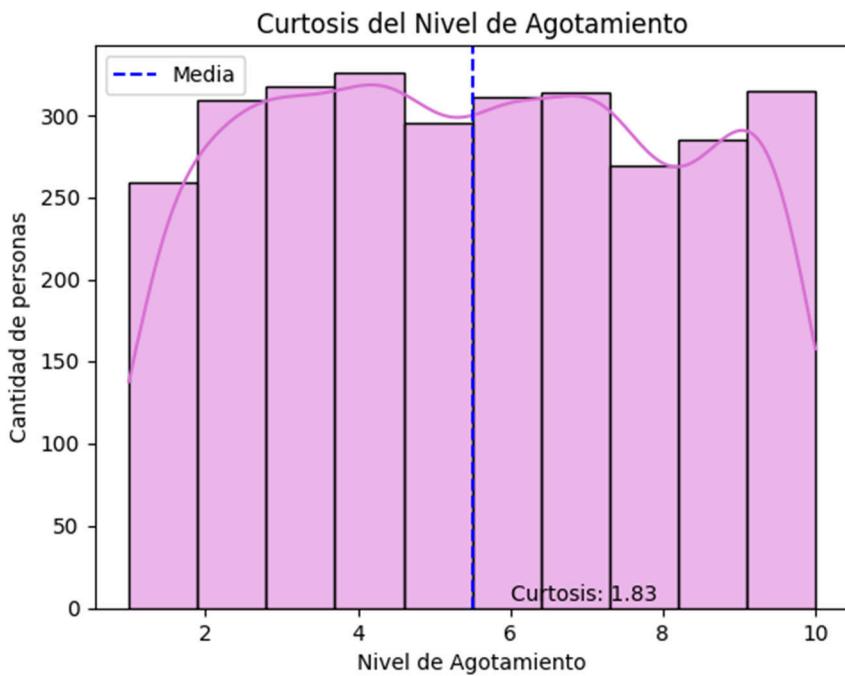


Figura 34: Histograma más curva de densidad.

Explicación del gráfico Fig Nº 34: Histograma + curva de densidad.

El histograma muestra cómo se distribuyen los niveles de agotamiento entre los participantes. La curva de densidad (línea violeta) suaviza la forma de la distribución, ayudando a visualizar tendencias.

Curtosis = 1.83. Este valor indica una distribución más plana que la normal. Los datos están más dispersos, sin una concentración fuerte en el centro ni colas pesadas.

En otras palabras: el agotamiento está presente de forma extendida, pero no hay muchos casos extremos.

Media \approx (línea azul) Está ligeramente desplazada hacia la derecha, lo que sugiere una tendencia moderada-alta de agotamiento.

QQ PLOT:

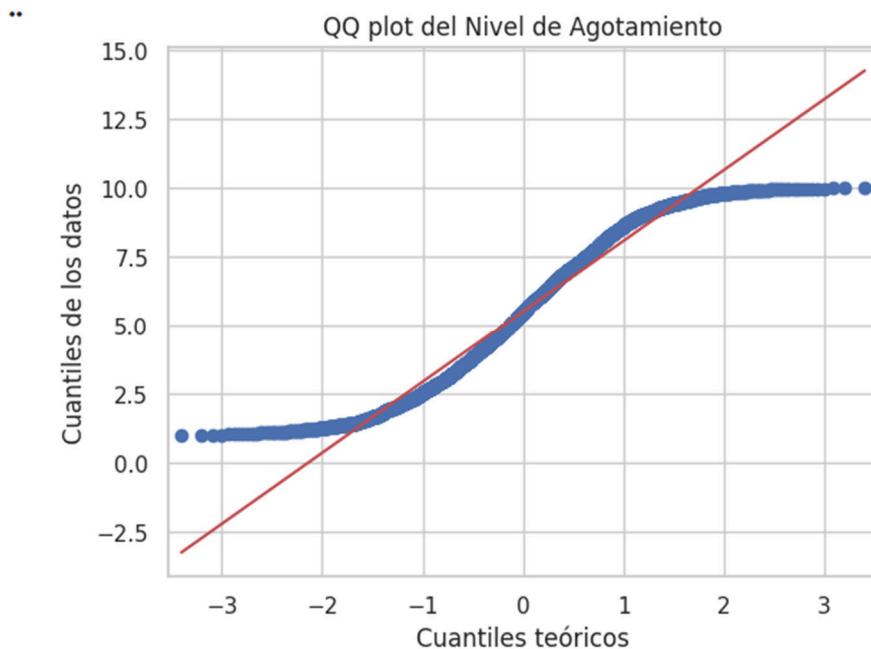


Figura 35: QQ PLOT de nivel de agotamiento.

Explicación del gráfico Fig N° 35:

Línea roja diagonal: Representa la distribución teórica normal. Si los datos fueran perfectamente normales, los puntos azules se alinearían sobre esta línea.

Distribución de los puntos: Se observa que los puntos se desvían de la línea roja, especialmente en los extremos (colas). Esto indica que los datos no siguen una distribución normal. Hay una mayor dispersión en los valores más bajos y más altos de agotamiento.

Forma de la desviación: La curvatura que se aleja de la línea sugiere que la distribución tiene colas más ligeras, lo que puede estar relacionado con la curtosis.

El QQ plot indica que a lo mejor hay participantes con niveles de agotamiento que se desvían significativamente del centro de la distribución.

MEDIDAS DE DISPERSIÓN: A continuación, presento un cuadro con la interpretación técnica de los resultados obtenidos, en formato resumen para cada medida calculada. Los invito a consultar el notebook de Google Colab para ver las fórmulas utilizadas.

MEDIDA	VALOR	¿QUE INDICA?
VARIANZA	6.62	Hay una dispersión más amplia respecto a la media. Los Niveles de agotamiento varían significativamente entre Participantes.
DESVIACIÓN ESTANDAR	2.57	Los valores individuales se desvían 2.57 puntos del promedio Esto refuerza la heterogeneidad en las respuestas.
COEFICIENTE DE VARIACIÓN	46.72%	Alta variabilidad relativa, el agotamiento no se distribuye de Forma homogénea; hay diferencias marcadas entre ciertos Casos.
RANGO INTERCUARTILICO(IQR)	4.34	El 50% central de los datos está bastante disperso, Hay diversidad interna.
ERROR ESTÁNDAR DE LA MEDIA	0.05	Valor muy bajo, lo que indica que la estimación de la media Es precisa y confiable. El promedio representa bien al Conjunto de datos.

DATOS ATÍPICOS:

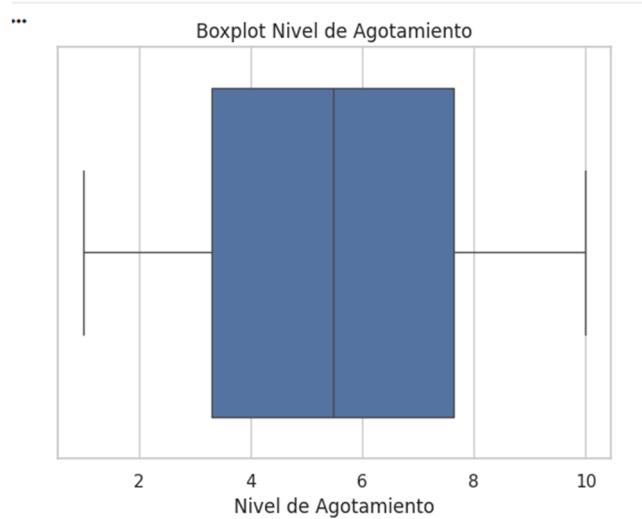


Figura 36: Boxplot Nivel de agotamiento

Explicación del gráfico Fig Nº 36: Lo que se observa en este gráfico:

La caja central representa el rango intercuartílico (IQR): el 50% de los datos está entre el primer cuartil (Q1) y el tercer cuartil (Q3). La línea dentro de la caja indica la mediana del nivel de agotamiento. Las líneas horizontales (bigotes) muestran el rango de los datos que no son considerados atípicos. Si hay puntos fuera de los bigotes, esos serían valores atípicos (outliers), que en este caso no los hay.

Como revelación puedo decir que: La mediana parece estar cerca de 5.5, lo que coincide con la media y moda anteriores. El IQR es amplio, lo que indica alta dispersión dentro del rango medio. No se observan valores atípicos extremos, lo que refuerza la idea de una distribución extendida pero no polarizada. El boxplot está ligeramente desplazado hacia la derecha, lo que sugiere una tendencia moderada-alta de agotamiento.

BOXPLOTS DE DIAS LIBRES DE LOS EMPLEADOS:

Como complemento, decidí realizar un boxplot con la variable "Días libres de salud mental", con el objetivo de identificar posibles valores atípicos, como por ejemplo poder descubrir que en "Días libres" hay muchos outliers, es decir, gente que toma 0 días o gente que toma muchísimos.

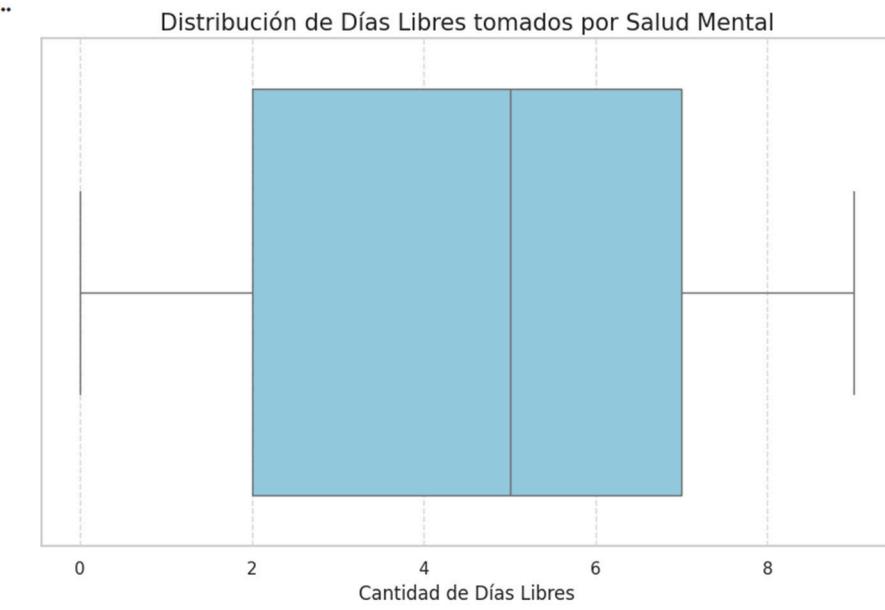


Figura 37: Boxplots días libres tomados.

Explicación del Gráfico Fig. Nº 37: El boxplot de la variable “Días libres por salud mental” muestra una distribución relativamente concentrada, sin presencia de valores atípicos (outliers). Esto indica que, aunque hay variabilidad en la cantidad de días tomados, todos los casos se encuentran dentro del rango esperado según el criterio estadístico del gráfico.

Lo que también puedo decir que nadie toma una cantidad de días que se considere extrema. Podría reflejar una cultura laboral donde el uso de días libres está más regulado, limitado e incluso cuando las necesidades individuales varían.

14- ANÁLISIS BIVARIADO:

En este segmento voy a analizar la relación entre dos variables, con el objetivo de identificar patrones que influyen en el ‘Nivel de Agotamiento’.

Numérica vs. Numérica (Correlaciones): Realice un Scatterplots (Gráficos de dispersión) en donde:

Eje X: Nivel de Stress.

Eje Y: Nivel de Agotamiento. (Se espera una línea ascendente).

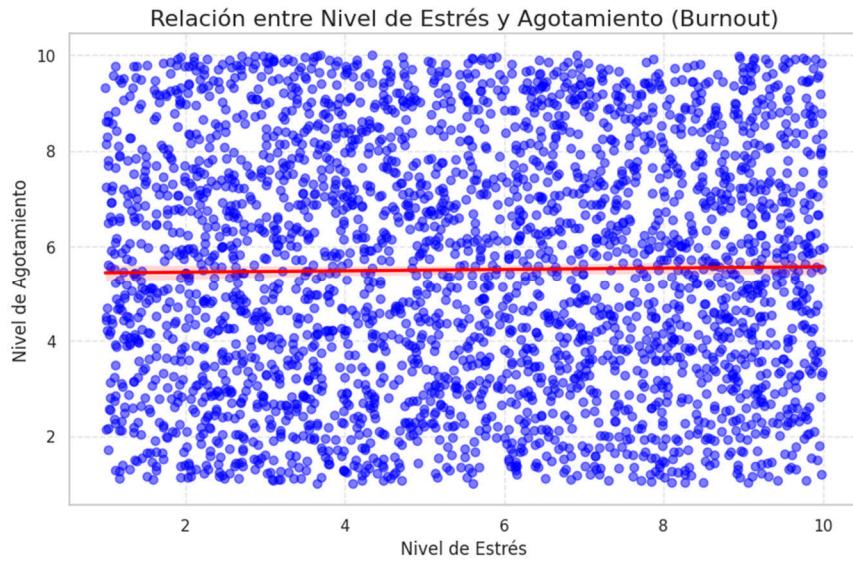


Figura 38: Scatterplots entre nivel de agotamiento y Estrés.

Explicación del Gráfico Figu. Nº 38: La dirección de la línea roja es plana, casi horizontal. Esto sugiere que el nivel de estrés no estaría afectando significativamente el agotamiento. Aunque este resultado puede parecer poco probable, es posible si existen otros factores que actúan como protectores para el empleado.

Categórica vs. Numérica (Comparación de Grupos): Realice un gráfico de Barras con error (Barplot):

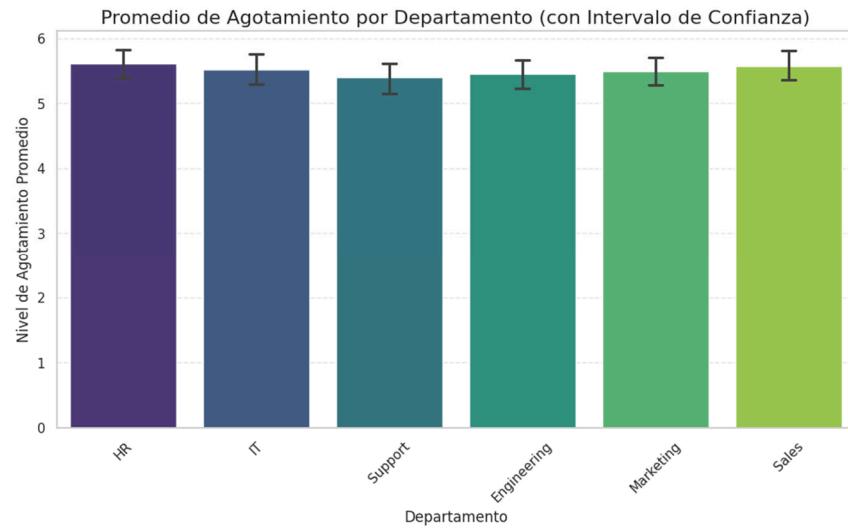


Figura 39: Gráfico de barras.

Explicación del gráfico Fig. Nº 39: El análisis de las barras de error muestra un solapamiento consistente entre todos los grupos, lo que permite inferir que el

departamento al que pertenece el empleado no sería un factor determinante para predecir su nivel de agotamiento.

Sin embargo, se observa que los niveles más altos de agotamiento, por así decirlo, corresponden a los departamentos de Recursos Humanos (RH) y Ventas (Sales).

Anova: Aunque el gráfico anterior sugiere que 'todos los departamentos son iguales' en cuanto al nivel de agotamiento, en ciencia de datos eso no es suficiente. Por eso, decidí calcular un valor P que lo respaldé estadísticamente.

Para ello, utilicé la librería Scipy.stats, que es un estándar en análisis estadístico. El código agrupa los datos de 'Nivel de Agotamiento' según el Departamento y realiza una comparación simultánea entre todos los grupos.

-
- Estadístico F: 0.4347
Valor P (P-value): 0.8246
CONCLUSIÓN: No hay diferencias significativas. El agotamiento es igual en todos los departamentos.
-

Figura 40: Resultado de Anova.

Explicación del resultado Fig Nº 40: Se llevó a cabo un análisis de varianza (ANOVA) de un factor para evaluar si existen diferencias significativas en el nivel medio de agotamiento entre los distintos departamentos. Los resultados arrojaron un estadístico F de 0.4347 y un p-valor de 0.8246. Dado que el p-valor es considerablemente superior al nivel de significancia establecido ($\alpha=0.05$) ($\alpha=0.05$).

Esto confirma estadísticamente que no existe una asociación significativa entre el departamento laboral y el nivel de agotamiento; es decir, el comportamiento de la variable es homogéneo a través de todas las áreas funcionales.

15- ANÁLISIS MULTIVARIADO:

Relación entre 3 o más variables.

Matriz de Correlación (Heatmap): Esta me permite ver de un solo vistazo, cómo se relacionan todas las variables numéricas entre sí.

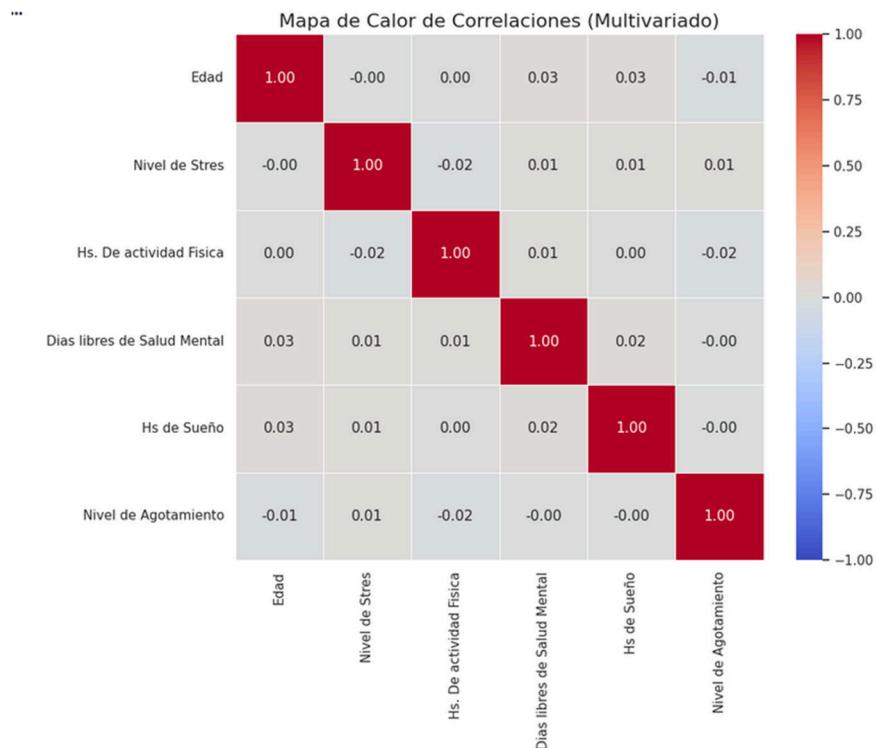


Figura 41: Matriz de correlación.

Explicación del Gráfico Fig N° 41: El análisis de correlación de Pearson (Mapa de Calor) revela una independencia lineal casi absoluta entre todas las variables numéricas estudiadas. Sorprendentemente, el coeficiente de correlación entre la variable predictora clave 'Nivel de Estrés' y la variable objetivo 'Nivel de Agotamiento' es de 0.01, lo que indica una correlación nula.

De igual manera, factores tradicionalmente protectores como las 'Horas de Sueño' (-0.00) o la 'Actividad Física' (-0.02) no muestran asociación lineal con el agotamiento en esta muestra. Esto sugiere que, para este grupo poblacional específico, el agotamiento no sigue los patrones lineales tradicionales y podría estar siendo impulsado por factores cualitativos no numéricos o relaciones no lineales complejas.

Gráfico de dispersión con regresión lineal (o "scatter plot con línea de tendencia"):

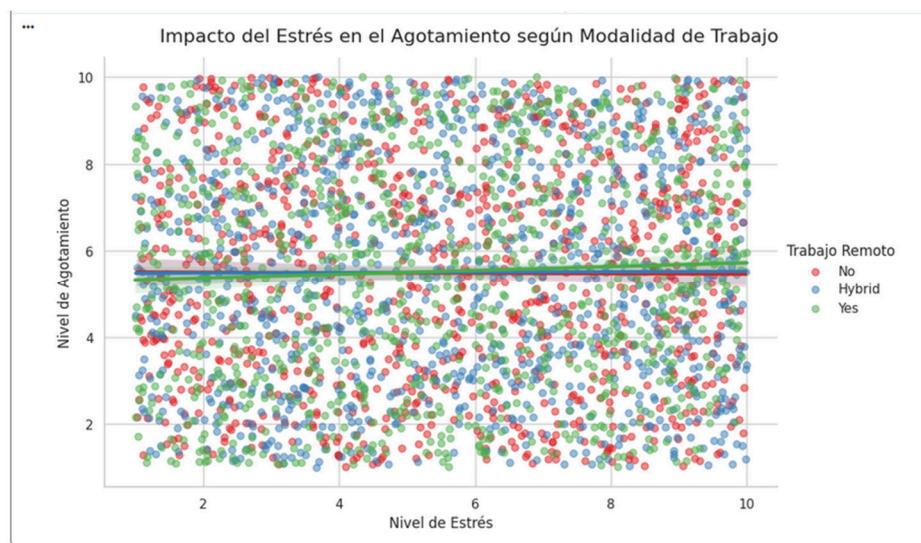


Figura nº 42: Scatter Plot con línea de tendencia.

Explicación del gráfico Fig Nº 42: Para profundizar en el análisis multivariado, grafique la interacción entre el 'Nivel de Estrés' y el 'Nivel de Agotamiento', segmentando la muestra por la modalidad de trabajo (Presencial, Híbrido, Remoto).

Como se evidencia en la Figura X, las líneas de regresión para los tres grupos son prácticamente horizontales y paralelas, con una pendiente cercana a cero. La dispersión uniforme de los puntos a lo largo de todo el plano confirma la ausencia de correlación lineal.

Como hallazgo puedo decir que la población estudiada, el nivel de estrés percibido no actúa como predictor del agotamiento.

La modalidad de trabajo (Remoto/Híbrido/Presencial) no influye ni modera esta relación. Los empleados remotos muestran la misma tendencia nula que los presenciales.

Esto valida los resultados previos del mapa de calor y el ANOVA, sugiriendo que el agotamiento en esta organización es un fenómeno aleatorio respecto a estas variables específicas, o que depende de factores no capturados en esta recolección de datos.

16- CONCLUSIONES:

El análisis exploratorio realizado sobre el conjunto de datos me permitió examinar con profundidad los factores asociados al agotamiento laboral en una muestra diversa de trabajadores. A partir del dataset limpio, sin valores nulos ni duplicados, se aplicaron técnicas univariadas, bivariadas y multivariadas para evaluar las hipótesis planteadas.

En el **análisis univariado**, la variable nivel de agotamiento mostró una distribución no normal, con una curtosis moderada (1.83) y un coeficiente de variación elevado (46.72), lo que indica una alta dispersión en los niveles reportados. El gráfico QQ reveló desviaciones significativas en los extremos, sugiriendo la presencia de casos atípicos o experiencias de agotamiento muy dispares dentro de la muestra.

En el **análisis bivariado**, la relación entre nivel de estrés y agotamiento fue débil (correlación de 0.01), y la línea de regresión prácticamente horizontal confirmó la ausencia de asociación lineal. Del mismo modo, el análisis por departamentos y la prueba ANOVA ($F = 0.4347$, $p = 0.8246$) indicaron que el área laboral no representa un factor diferenciador en los niveles de agotamiento.

El **análisis multivariado** reforzó estos hallazgos: ni el sueño, ni la actividad física, ni la modalidad de trabajo (remoto, híbrido o presencial) mostraron correlaciones significativas con el agotamiento. Las líneas de regresión segmentadas por modalidad fueron paralelas y planas, lo que sugiere que el agotamiento no está siendo explicado por estas variables cuantitativas.

Estos resultados invitan a reflexionar sobre la complejidad del fenómeno del agotamiento laboral. En esta muestra, los factores tradicionalmente considerados protectores o de riesgo no parecen tener un peso explicativo significativo. Esto podría deberse a la influencia de variables cualitativas no capturadas en el dataset, como la calidad del liderazgo, el clima emocional y la carga mental.

En síntesis, el agotamiento en esta población no responde a patrones lineales simples. Este hallazgo, lejos de ser una limitación, abre la puerta a nuevas preguntas y enfoques más

integrales que contemplen la dimensión emocional, relacional y organizacional del bienestar laboral.

Próximos pasos hacia un análisis predictivo:

Los hallazgos obtenidos hasta el momento revelan que las relaciones lineales entre variables cuantitativas tradicionales (como estrés, sueño o actividad física) y el agotamiento laboral son débiles o inexistentes en esta muestra. Esta aparente independencia sugiere que el fenómeno del agotamiento podría estar influido por interacciones más complejas, no lineales o incluso por variables cualitativas no capturadas en el dataset.

En este contexto, me propongo avanzar hacia una etapa de modelado predictivo con técnicas de machine learning, que permita:

- Explorar relaciones no lineales mediante algoritmos como árboles de decisión, random forest o gradient boosting.
- Evaluar la importancia relativa de cada variable en la predicción del agotamiento, incluso si su correlación lineal es baja.
- Construir modelos de clasificación o regresión que permitan anticipar niveles de agotamiento en función de múltiples factores combinados.
- Detectar patrones ocultos o segmentos de riesgo mediante técnicas de clustering o segmentación no supervisada.
- Validar la robustez del modelo con métricas como precisión, recall, F1-score o RMSE, según el tipo de modelo elegido.

Este enfoque no solo enriquecerá la comprensión del fenómeno, sino que también permitirá generar recomendaciones personalizadas y basadas en evidencia para promover el bienestar emocional en el entorno laboral.