CODER HOUSE

TRABAJO FINAL

COMISIÓN 29730 2022

EQUIPO

ANALIA CONSTANZO - ALFREDO DELGADO - ARIEL FELDMAN – EDUARDO DE LA ROSA

ÍNDICE

CONTEXTO

El burnout laboral, también denominado "síndrome del quemado" o "síndrome de estar quemado en el trabajo", es un estado de agotamiento físico, emocional y mental que está vinculado con el ámbito laboral, el estrés causado por el trabajo y el estilo de vida del empleado. Este síndrome puede tener impactos severos en el equilibrio emocional de una persona y, por consecuencia, una disminución en el desempeño laboral. Es un proceso en el que progresivamente el trabajador sufre una pérdida del interés por sus tareas y va desarrollando una reacción psicológica negativa hacia su ocupación laboral.

Los principales síntomas del burnout laboral son: Agotamiento físico y mental generalizado, Despersonalización y cinismo (adopción de una actitud de indiferencia y desapego, reduciendo claramente su compromiso hacia el trabajo) y Descenso en la productividad laboral y desmotivación (disminución de la productividad laboral y desmotivación que genera frustración y evidencia una ausencia de realización personal en el trabajo).

Según una encuesta internacional realizado por la compañía Indeed (1) (https://www.indeed.com/lead/preventing-employee-burnout-report) a 1.500 trabajadores de USA el año 2020, un 52% de los encuestados reconoce presentar burnout en período prepandemia Covid-19 y un 67% consideró que empeoró esta condición en el curso de la pandemia. Esta estadística permite inferir la importancia que tiene la prevención y estudio del burnout en todo tipo de organizaciones, más aún en períodos como el que nos encontramos actualmente (pandemia, inestabilidad política-económica, crisis inflacionaria, entre otros).

PREGUNTAS Y OBJETIVOS GENERALES

A partir de un dataset de 22.750 observaciones, se presentan datos de trabajadores que permitirán predecir el grado de burnout según ciertas condiciones estandarizadas.

El objetivo de esta investigación es crear un modelo predictivo, a través de técnicas de Machine Learning, que permita identificar patrones y lograr pronosticar el nivel de burnout de un empleado, sobre la base de características individuales determinadas en el dataset.

La presente investigación busca responder las siguientes preguntas:

- *¿Cuáles son las características relevantes que pueden determinar el grado de burnout de un trabajador?*

- *¿Qué se puede concluir de los análisis estadísticos sobre el nivel de burnout?*

- *¿Qué modelo y algoritmo de Machine Learning permite obtener mejores resultados en la predicción del nivel de burnout de un empleado?*

PRESENTACIÓN DEL DATASET

OBTENCIÓN Y COMPOSICIÓN

Estos dataset fueron extraídos de la página de kaggle.
Link: https://www.kaggle.com/datasets/blurredmachine/are-your-employees-burning-out?select=train.csv.

A su vez, estos 2 dataset se diferencian en la cantidad de registros de uno u otro y que el "TEST" no cuenta con la variable "Burn Rate".

"TRAIN"

- ***Employee ID:*** *El ID único asignado a cada empleado (ejemplo: fffe390032003000)*

- ***Date of Joining:*** *La fecha y hora en que el empleado se unió a la organización (ejemplo: 2008-12-30)*

- ***Gender:*** *El género del empleado (Hombre/Mujer)*

- ***Company Type:*** *El tipo de empresa donde trabaja el empleado (Servicio/Producto)*

- ***WFH Setup Available:*** *¿Está disponible el trabajo desde casa para el empleado? (Sí/No)*

- ***Designation:*** *La designación del empleado de trabajo en la organización. En el rango de [0.0, 5.0] mayor es la designación mayor.*

- ***Resource Allocation:*** *La cantidad de recursos asignados al empleado para trabajar, es decir. número de horas de trabajo. En el rango de [1.0, 10.0] (más alto significa más recursos)*

- ***Mental Fatigue Score:*** *El nivel de fatiga mental al que se enfrenta el empleado. En el rango de [0.0, 10.0] donde 0.0 significa sin fatiga y 10.0 significa fatiga total.*

- ***Burn Rate:*** *El valor que necesitamos predecir para cada empleado indicando la tasa de Burn out mientras trabaja. En el rango de [0.0, 1.0] donde cuanto más alto es el valor, más se cansa.*

"TEST"

- ***Employee ID:*** *El ID único asignado a cada empleado (ejemplo: fffe390032003000)*

- ***Date of Joining:*** *La fecha y hora en que el empleado se unió a la organización (ejemplo: 2008-12-30)*

- ***Gender:*** *El género del empleado (Hombre/Mujer)*

- ***Company Type:*** *El tipo de empresa donde trabaja el empleado (Servicio/Producto)*

- ***WFH Setup Available:*** *¿Está disponible el trabajo desde casa para el empleado? (Sí/No)*

- ***Designation****: La designación del empleado de trabajo en la organización. En el rango de [0.0, 5.0] mayor es la designación mayor.*

- ***Resource Allocation:*** *La cantidad de recursos asignados al empleado para trabajar, es decir. número de horas de trabajo. En el rango de [1.0, 10.0] (más alto significa más recursos)*

- ***Mental Fatigue Score:*** *El nivel de fatiga mental al que se enfrenta el empleado. En el rango de [0.0, 10.0] donde 0.0 significa sin fatiga y 10.0 significa fatiga total.*

ANÁLISIS DEL DATASET

EDA

	Cantidad	Tipo	Missing	Unicos	Numeric	top	freq	mean	std	min	25%	50%	75%	max	sesgo	kurt
Employee ID	22750	object	0	22750	False	fffe31003200320035003100	1	-	-	-	-	-	-	-	-	-
Date of Joining	22750	object	0	366	False	2008-01-06	86	-	-	-	-	-	-	-	-	-
Gender	22750	object	0	2	False	Female	11908	-	-	-	-	-	-	-	-	-
Company Type	22750	object	0	2	False	Service	14833	-	-	-	-	-	-	-	-	-
WFH Setup Available	22750	object	0	2	False	Yes	12290	-	-	-	-	-	-	-	-	-
Designation	22750	float64	0	6	True	-	-	2.178725	1.135145	0.0	1.0	2.0	3.0	5.0	0.092421	-0.414916
Resource Allocation	21369	float64	1381	10	True	-	-	4.481398	2.047211	1.0	3.0	4.0	6.0	10.0	0.204573	-0.479884
Mental Fatigue Score	20633	float64	2117	101	True	-	-	5.728188	1.920839	0.0	4.6	5.9	7.1	10.0	-0.430895	0.174277
Burn Rate	21626	float64	1124	101	True	-	-	0.452005	0.198226	0.0	0.31	0.45	0.59	1.0	0.045737	-0.261579

- *El dataset cuenta con 22.750 registros y 9 columnas.*

- *Contamos con 9 columnas, de las cuales 5 son de tipo Object y 4 Float, aunque "Date of Joining" podemos transformarla en fecha.*

- ***Employee ID**: no tenemos id repetidos son todos únicos de tipo Object, al no tener repetidos suponemos que son todos empleados distintos.*

- ***Date of Joining**: es de tipo object, pero son fechas, de las cuales tenemos una variación de 366 fechas en las cuales la mayor frecuencia se visualiza para el 01-06-2008.*

- ***Gender:** Viendo los valores de frecuencia y top podemos suponer que las dos categorías de genero están bastante parecidas en porcentaje, no hay un balanceo significativo para una u otro género.*

- ***Company Type:** Son dos tipos de compañía viendo su valor de frecuencia para la categoría top de Service podemos deducir que tiene un 65 % de los datos.*

- ***WFH Setup Available:*** *La mayor frecuencia de los registros nos indica que el 55% a simple vista tiene la disponibilidad de realizar home office.*

- ***Resource Allocation:*** *Tiene 1.381 registros nulos, trataremos los nulos de 3 maneras, para análisis univariado y bivariado rellenando con mediana, para el multivariado eliminándolo, y finalmente para el modelo ya hemos visto que podremos hacerlo con KNNImputer, esto luego de realizar los 3 análisis mencionados.*

- ***Mental Fatigue Score:*** *Tiene 2.117 registros nulos, trataremos los nulos de 3 maneras, para análisis univariado y bivariado rellenando con mediana, para el multivariado eliminándolo, y finalmente para el modelo ya hemos visto que podremos hacerlo con KNNImputer, esto luego de realizar los 3 análisis mencionados.*

- ***Burn Rate:*** *Esta es nuestra variable "target", tiene 1.124 registros nulos, los cuales eliminaremos.*

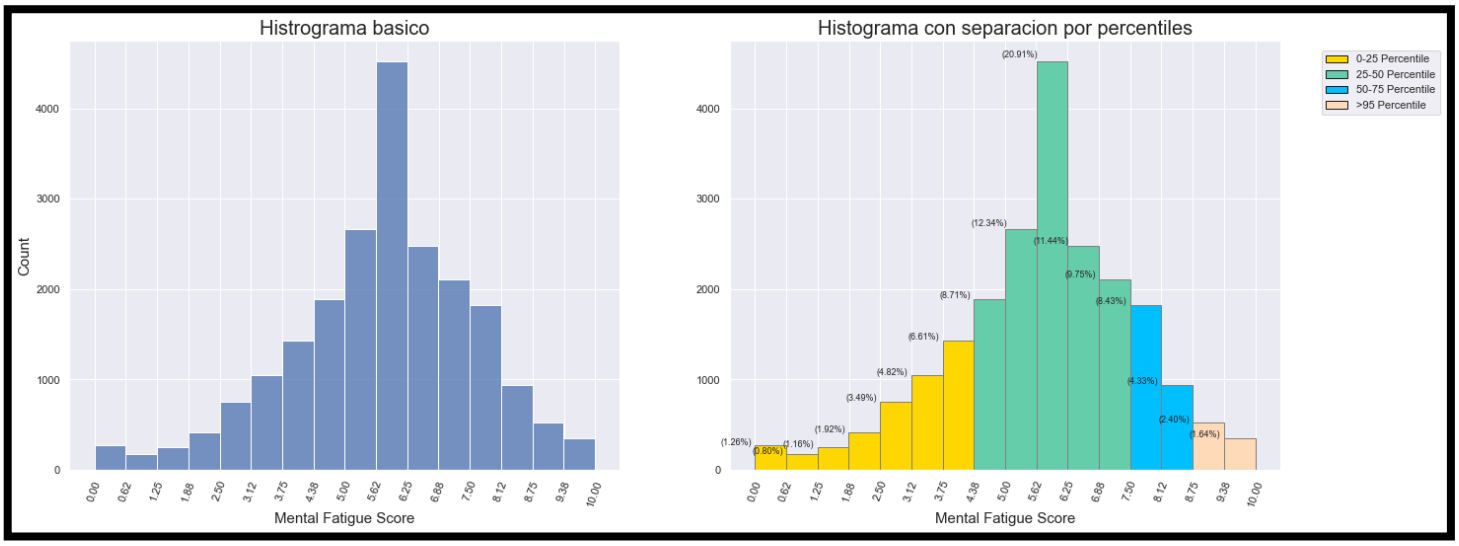
- ***Designation, Resource Allocation, Mental Fatigue Score, Burn Rate:*** *tienen el promedio y la mediana con valores similares.*

ANÁLISIS DEL DATASET

UNIVARIADO

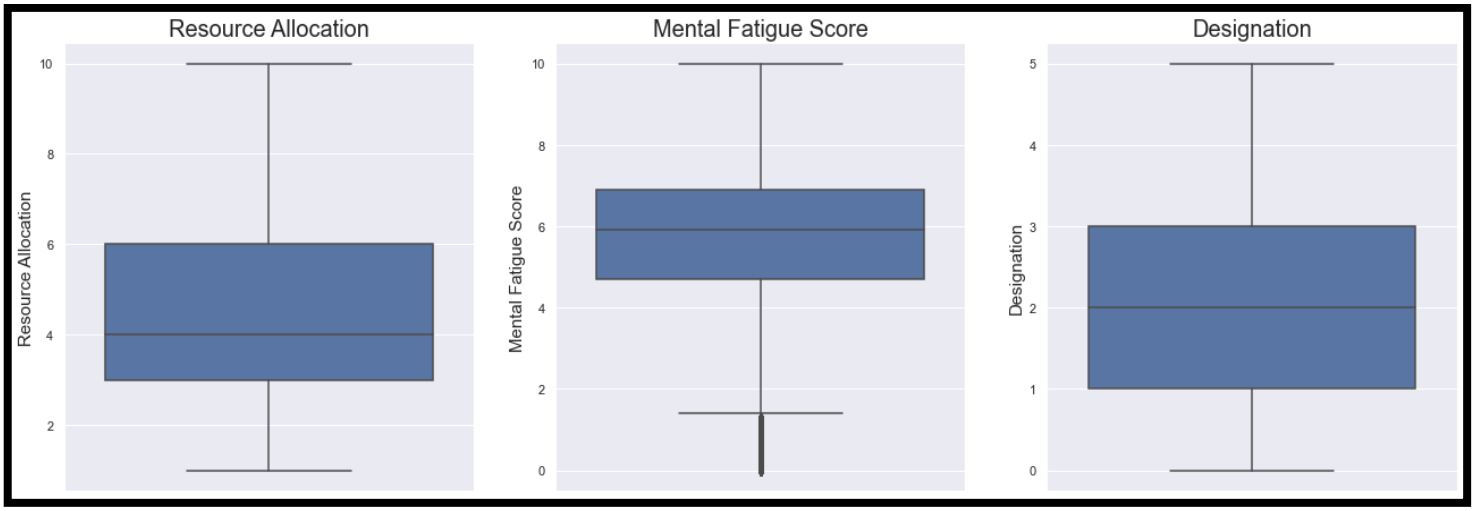
Para este análisis se eliminaron las filas que contienen valores nulos para nuestra variable Target "Burn Rate" y para las siguientes variables "Mental Fatigue Score" y "Resource Allocation" se rellenaron con la mediana.

En las siguientes dos imágenes, se realizaron histogramas según la frecuencia de los niveles de fatiga mental dentro de 16 rangos de valores ubicados en cuartiles.



Se puede apreciar que en el segundo cuartil es donde se concentra la mayor cantidad de observaciones, específicamente, entre 4,38 y 7,50 de nivel de fatiga mental. En este rango se encuentra el 63,15% de la muestra.

A continuación, se presentan tres boxplots, en donde el primero representa la variable Resource Allocation (cantidad de horas de trabajo), el segundo la Mental Fatigue Score (nivel de fatiga mental) y el tercero Designation (rango del trabajador dentro de la empresa).



De estos gráficos se puede concluir lo siguiente:

- *La mayoría de las personas trabajan entre 3,5 y 6 horas diarias.*
- *La mayoría de los trabajadores se encuentran concentrados entre los niveles 4,3 y 7 de fatiga mental.*
- *La mayoría de los trabajadores se encuentran en la designación de su organización entre los niveles 1 y 3.*

ANÁLISIS DEL DATASET

BIVARIADO

Para este análisis se eliminaron las filas que contienen valores nulos para nuestra variable Target "Burn Rate" y para las siguientes variables "Mental Fatigue Score" y "Resource Allocation" se rellenaron con la mediana.

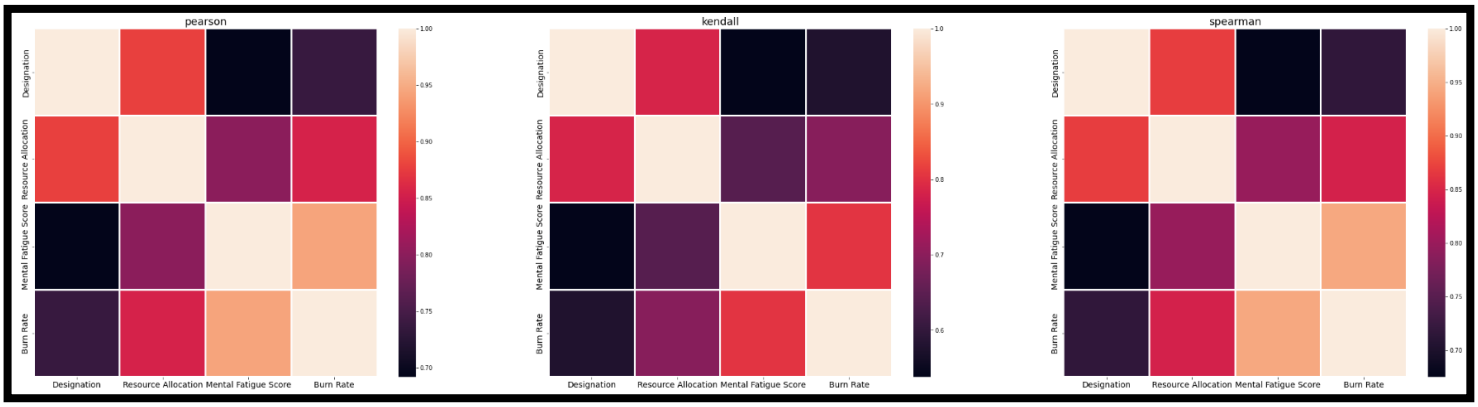
Análisis de variables numéricas

- *Se realiza un DataFrame con una correlación de Pearson con las variables numéricas.*

	Designation	Resource Allocation	Mental Fatigue Score	Burn Rate
Designation	1.000000	0.851383	0.657882	0.737556
Resource Allocation	0.851383	1.000000	0.740061	0.829632
Mental Fatigue Score	0.657882	0.740061	1.000000	0.898926
Burn Rate	0.737556	0.829632	0.898926	1.000000

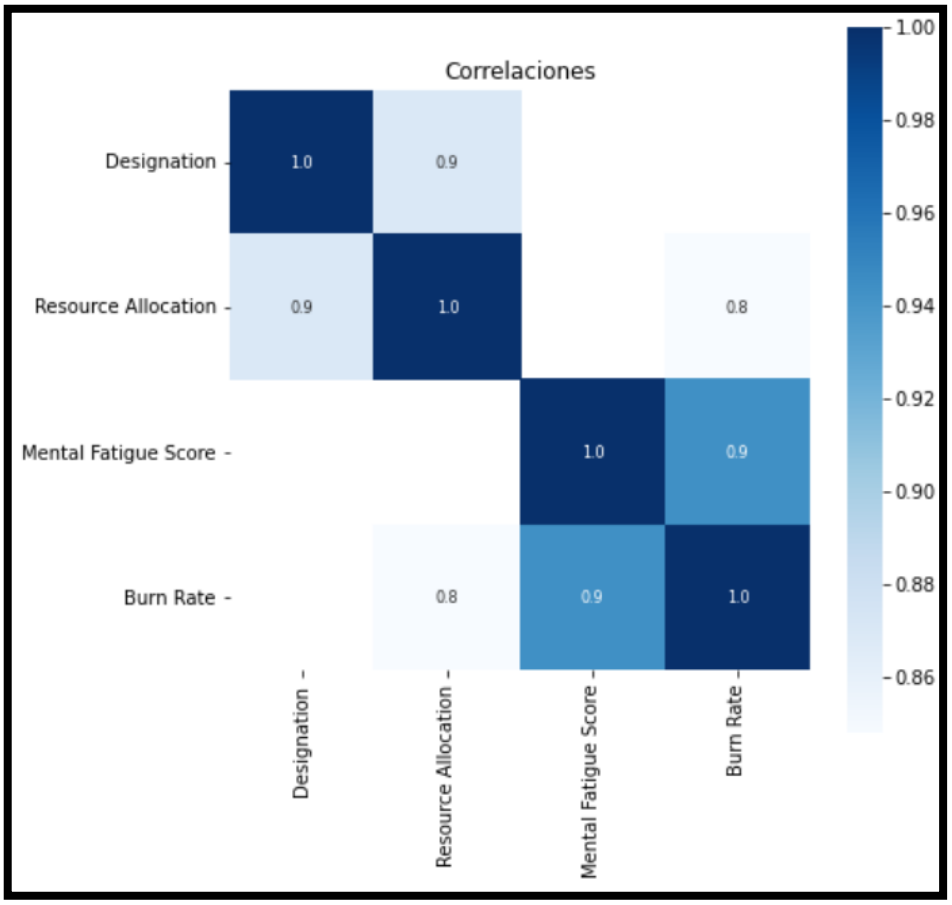
Aquí observamos que las variables más correlacionadas con Burn Rate son Mental Fatigue Score, luego Resource Allocation y por último Designation.

➢ *Se comparan los 3 métodos de correlación*



Se observa que en métodos Pearson y Spearman tienen los mismos valores de correlación.

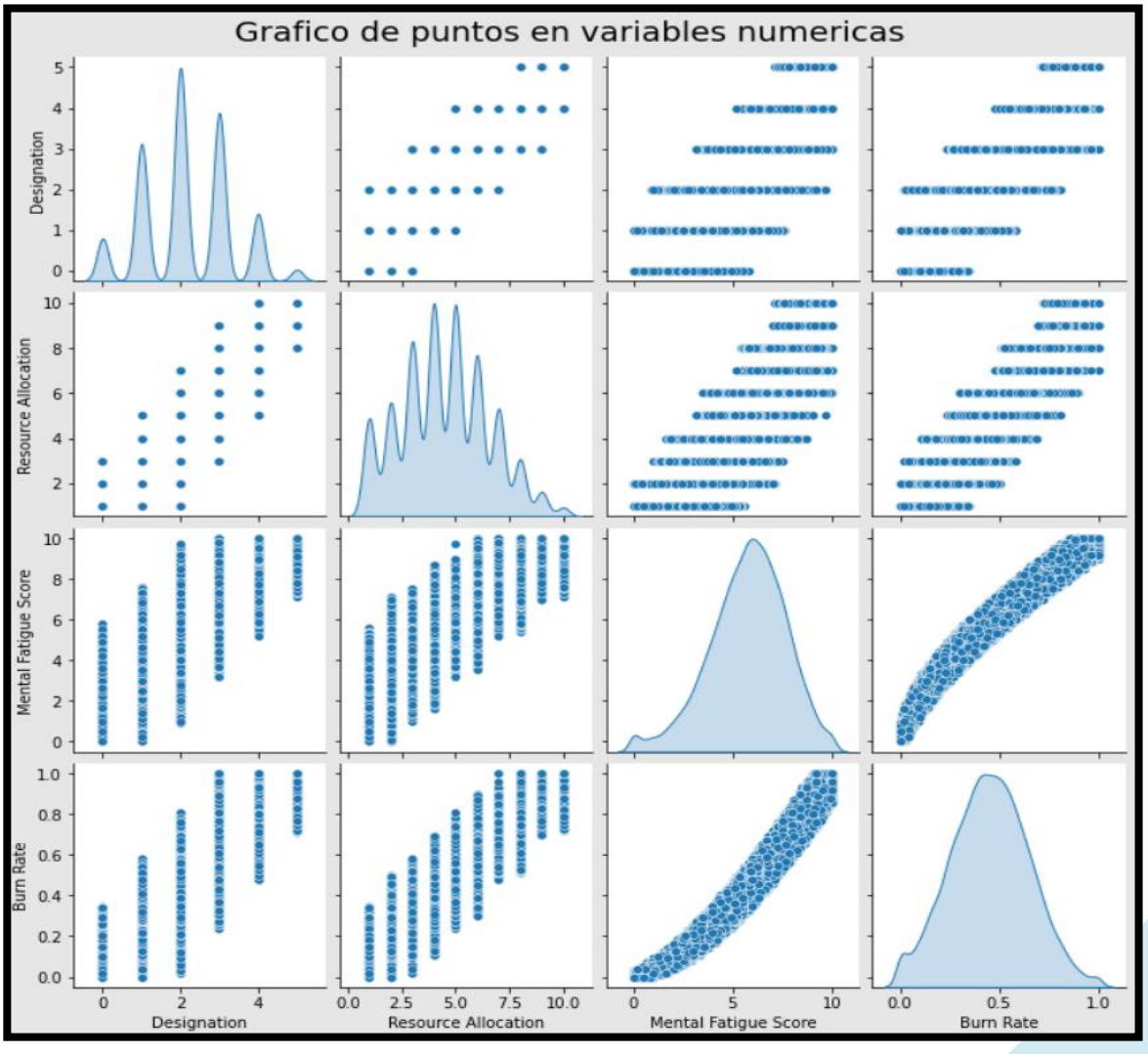
➢ *Se realiza una correlación de Spearman y se pone los datos numéricos*



➢ *Se realiza un DataFrame y se ordena la correlación de las variables de mayor a menor*

	level_0	level_1	0
5	Mental Fatigue Score	Burn Rate	0.895622
1	Designation	Resource Allocation	0.841758
4	Resource Allocation	Burn Rate	0.819000
3	Resource Allocation	Mental Fatigue Score	0.731734
2	Designation	Burn Rate	0.718205

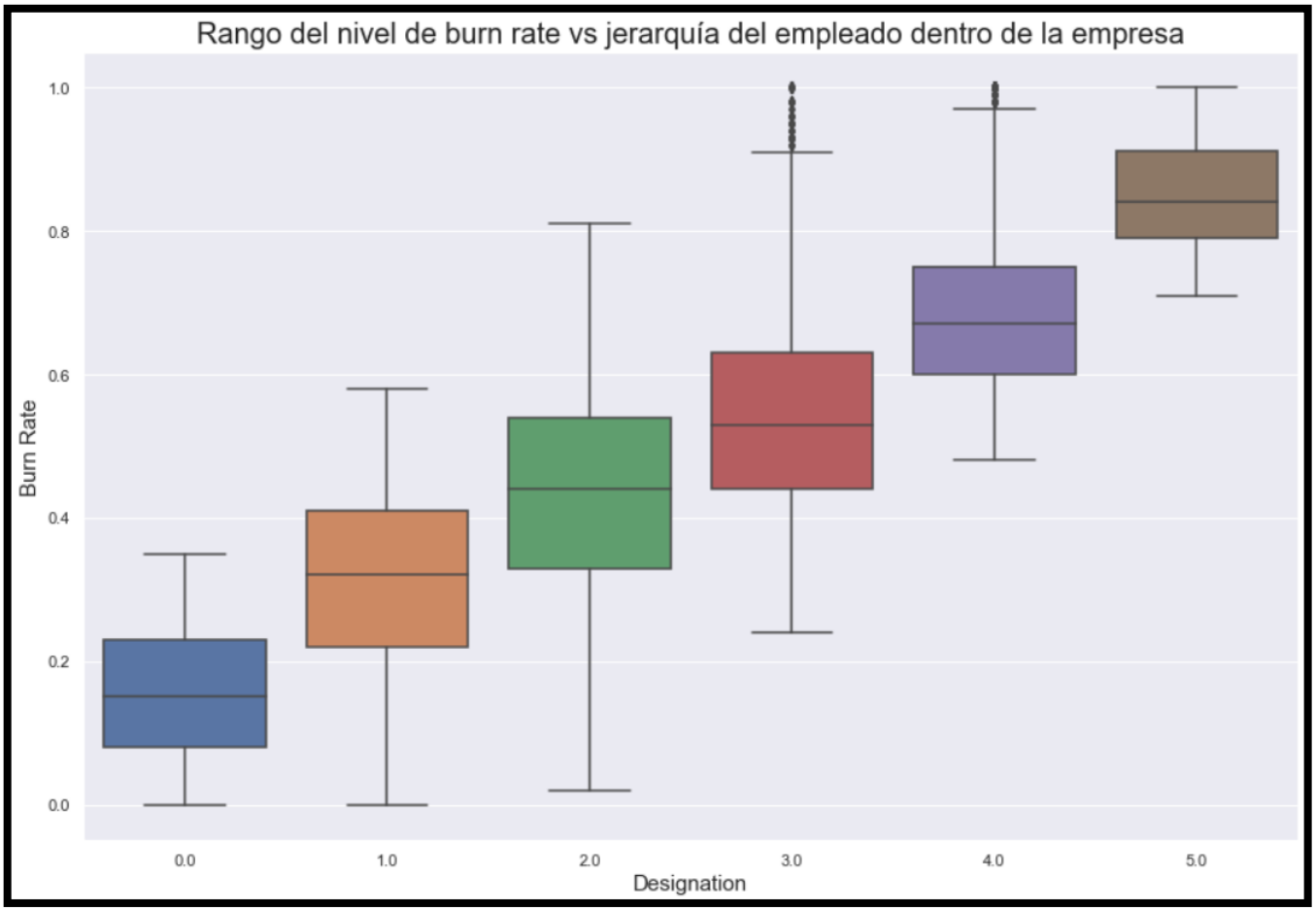
➢ *Se realiza un gráfico de puntos comparativo de las variables cuantitativas*



- *Se observa la gran correlación entre Burn Rate vs Mental Fatigue Score y en segundo lugar con Resource Allocation y con el resto de las variables en menor nivel*

- *Los gráficos de densidad muestran una distribución normal*

- *Gráfico de Boxsplots*



Aquí observamos outlier del nivel de Burn Rate dentro de la categoría 3 y 4 de nivel jerárquico del empleado.

Análisis Categórico vs Numérico

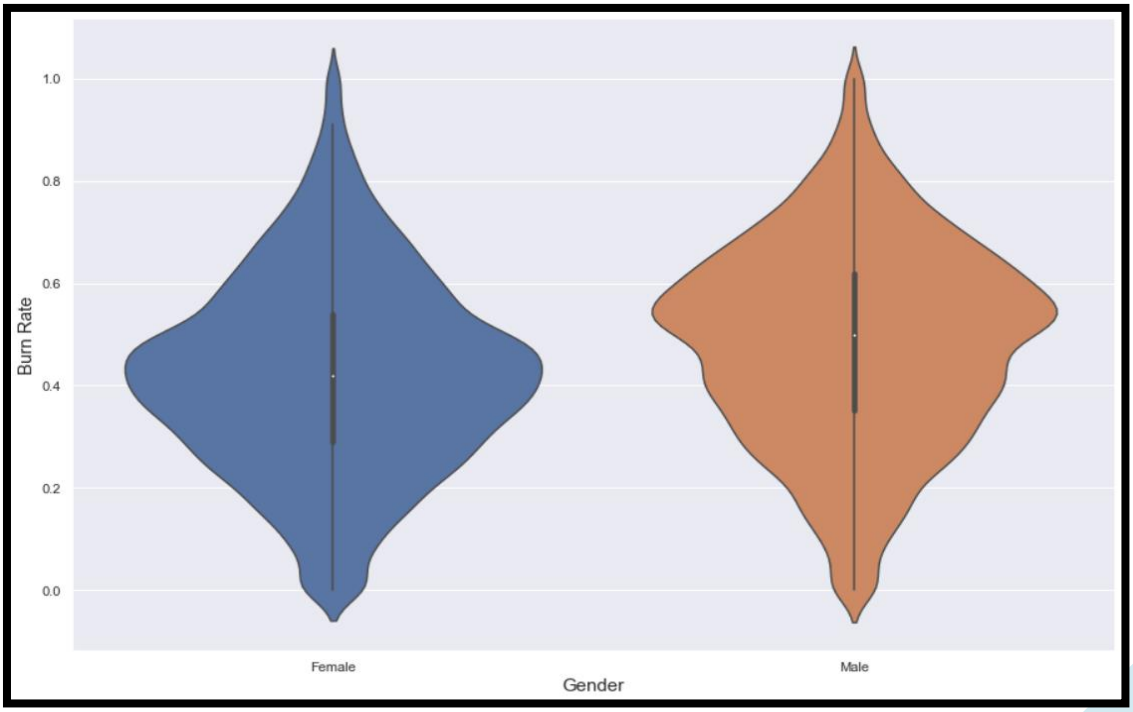
- *Se realiza un Datframe comparativo donde podemos ver que las designaciones altas de los cargos están asignadas a hombres.*

Designation	0.0	1.0	2.0	3.0	4.0	5.0
Gender						
Female	881	2757	3801	2816	921	173
Male	558	1875	3405	2882	1354	203

Aquí podemos ver que los hombres tienen asignadas mayor cantidad de horas que las mujeres.

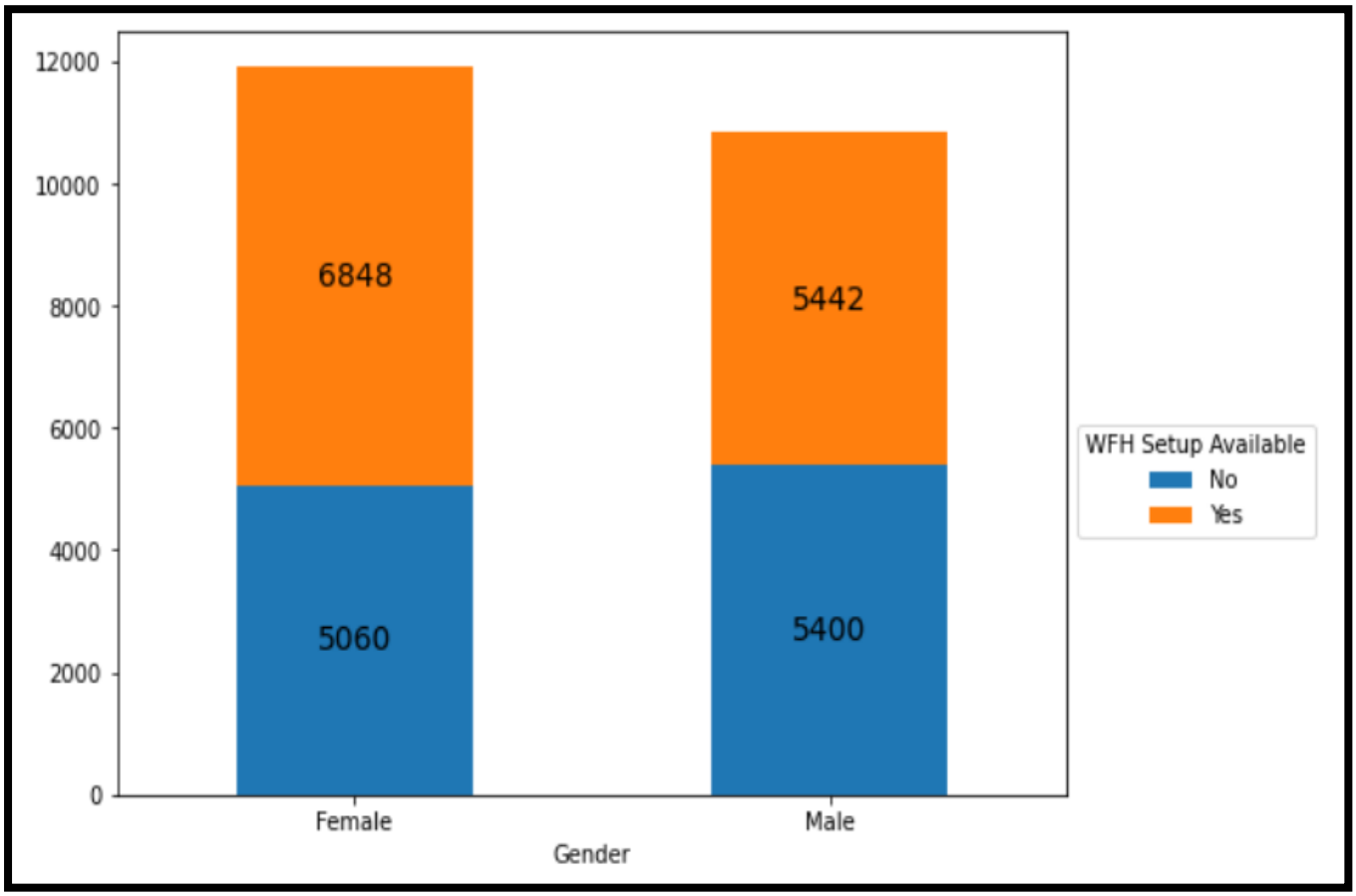
Resource Allocation	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
Gender										
Female	1051	1194	1792	2773	1890	1251	739	405	185	69
Male	650	794	1236	2199	1780	1566	1141	587	243	81

- *Gráfico de Violín*



Aquí vemos que los hombres tienen un nivel de estrés mayor que las mujeres

- *Análisis de Variables Categóricas o Categóricas*



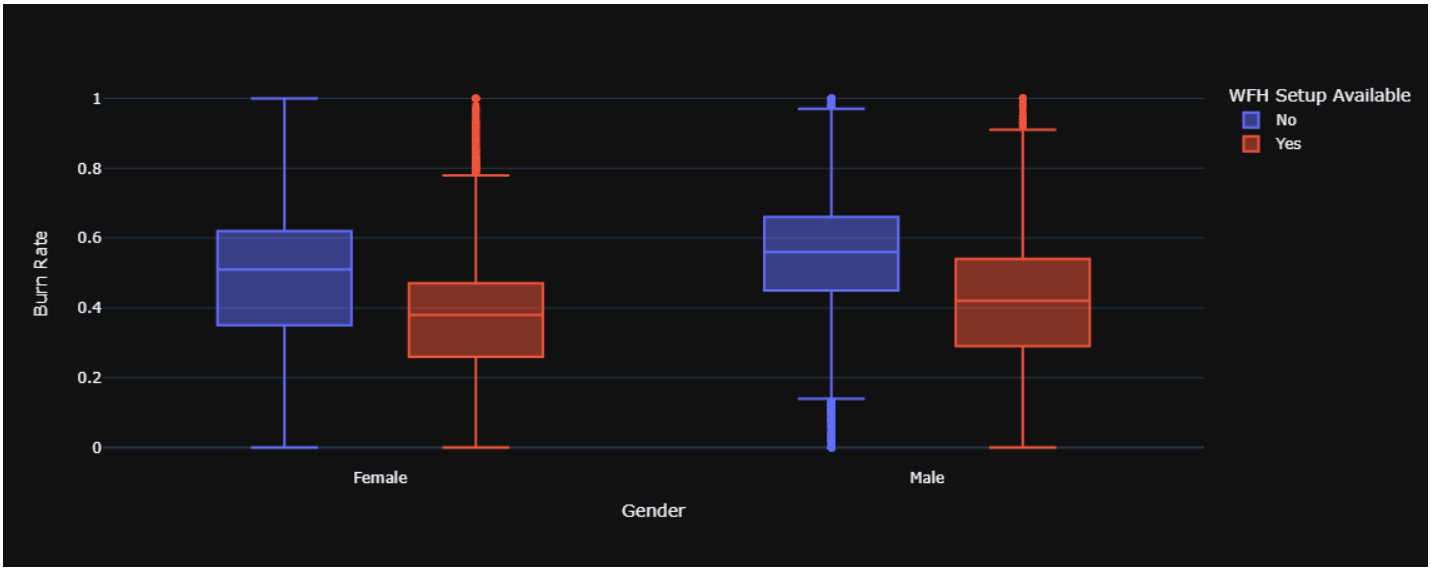
Observamos que las mujeres hacen más home office que los hombres.

ANÁLISIS DEL DATASET

MULTIVARIADO

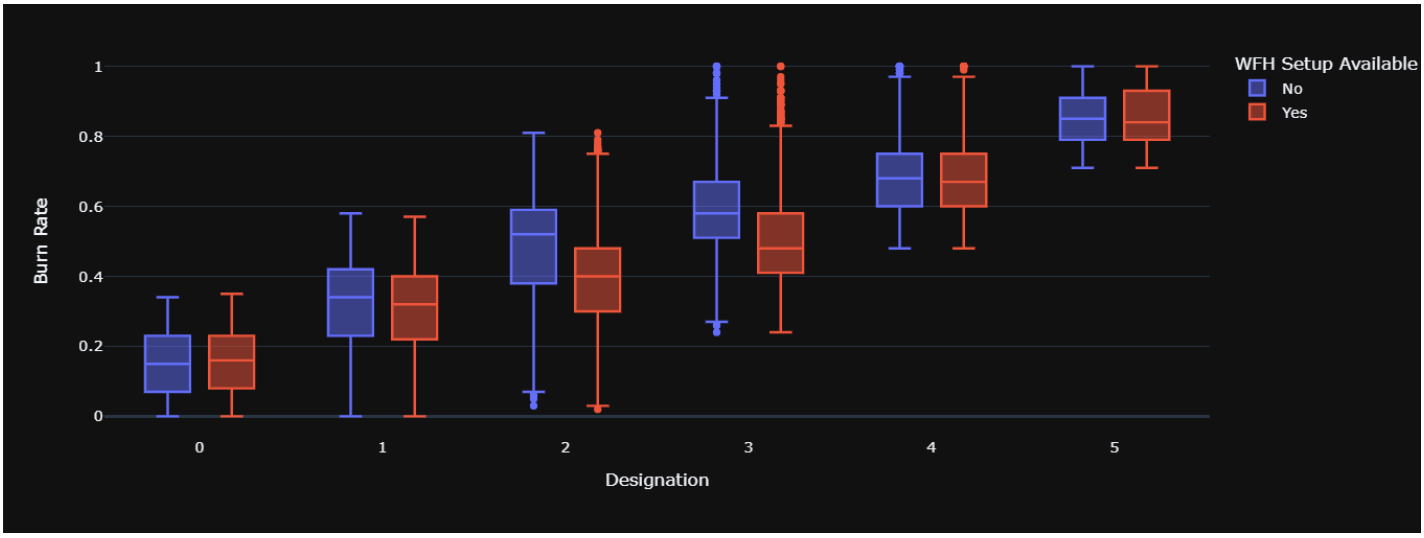
Para este análisis se eliminaron las filas que contienen valores nulos.

Se realiza una comparación por género, posibilidad de home office y nivel de cansancio.



- *Tanto los hombres como mujeres que no realizan home office tienen un mayor nivel de cansancio.*
- *En comparación entre los hombres y mujeres que realizan home office, vemos un rango más amplio en los hombres.*
- *Las mujeres que no realizan home office tienen una mayor dispersión en contra posición que los hombres.*

A continuación, se visualiza un grafica el cual compara el nivel de cansancio, según la jerarquía laboral, la cantidad de horas trabajadas y la fatiga mental del empleado.

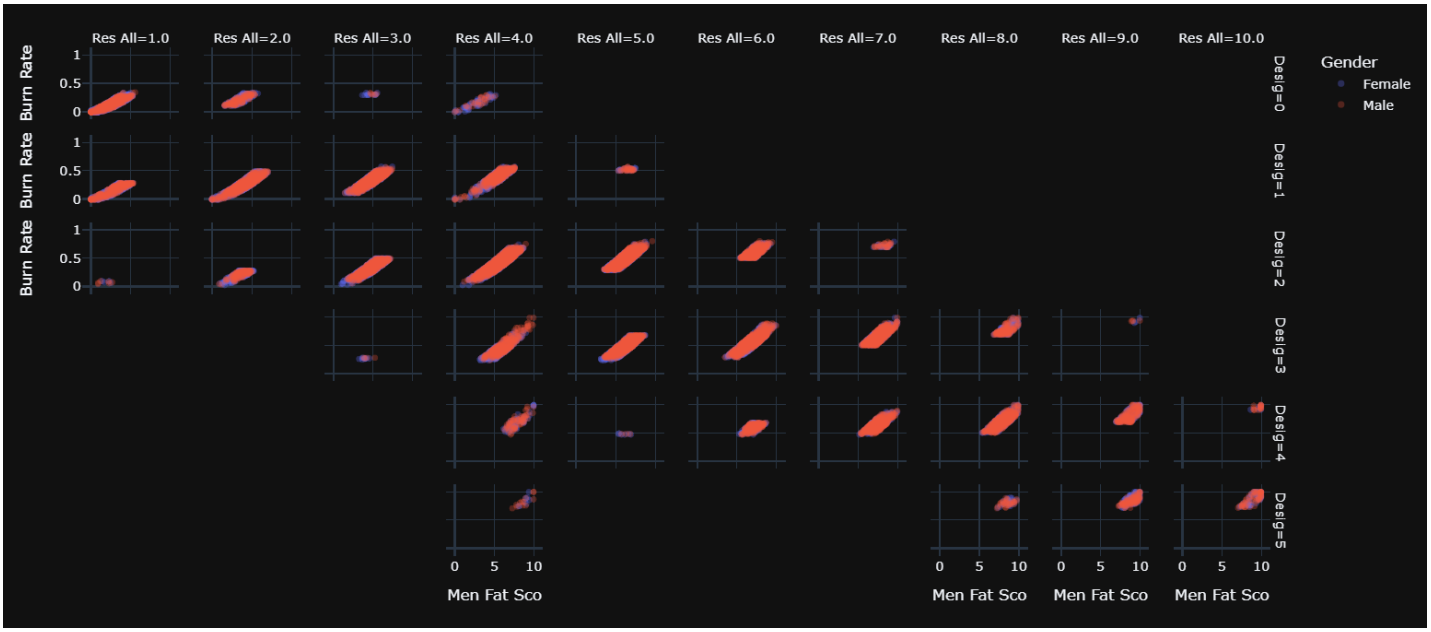


- *Podemos ver que los rangos medios (2 y 3), tienen una notoria diferencia de cansancio entre los que realizan home office y los que no.*

- *Los rangos más altos (4 y 5) tienen minimas diferencias del nivel de cansancio.*

- *Los rangos más bajos (0 y 1) tienen mínimas diferencias del nivel de cansancio.*

Se realiza el siguiente grafico para entender el comportamiento de las jerarquías laborales del dataset con su nivel de estrés laboral y la posibilidad de realizar home office.



- *Se nota un marcado aumento de nivel de estrés laboral a medida que aumenta la jerarquía, esto esta inversamente relacionado.*

- *La cantidad de horas trabajadas según aumento de la jerarquía genera un mayor aumento de nivel de estrés laboral.*

- *La menor jerarquía no muestra un nivel alto de horas trabajadas y tampoco un nivel de fatiga menta y estrés laboral alto.*

- *Hay niveles jerárquicos que no tienen tanta demanda de horas trabajadas, pero sin embargo tienen altos niveles de cansancio, ejemplo: Res All = 4.0 y Desig= 5.0*

- *Cuantas más horas trabajadas genera mayor nivel de fatiga mental.*