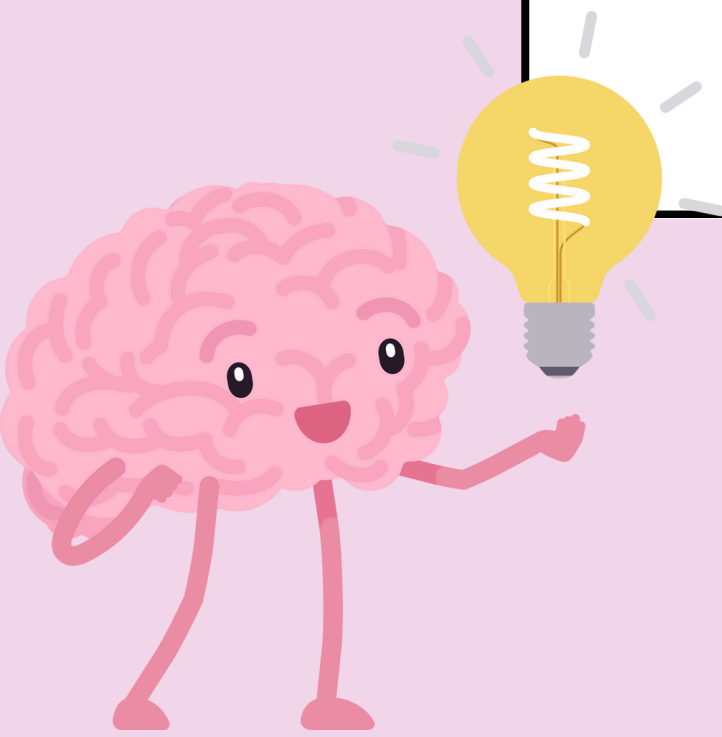


# **DATASET**

## **IMPACT OF REMOTE WORK ON MENTAL HEALTH**

**VIERO HEDFAM PUTRI**  
**RIZKI NUR LAILI**

- SIB GNFI BATCH 7 KELAS A -



# penjelasan data pada tabel

**EMPLOYEE\_ID**

menunjukkan bahawa  
ada 5000 karyawan  
yang di data

**AGE**

menunjukkan umur  
pegawai yang bervariasi  
mulai dari 27 hingga 49

**GENDER**

menunjukkan gender  
pekerja yang  
dikategorikan dalam  
Male, Female, dan Non-  
binary

**JOB\_ROLE**

menunjukkan jenis  
pekerjaan pekerja yang  
dikategorikan dalam HR,  
Data Scientist, Software  
Engineer dan Sales

# penjelasan data pada tabel

## INDUSTRY

menunjukkan jenis bidang pekerjaan yang dikategorikan dalam healthcare (kesehatan), IT, Education (pendidikan), finance (keuangan), dan consulting

## YEAR OF EXPERIENCE

menunjukkan seberapa lama pekerja bekerja dalam perusahaan. dikategorikan dalam data numeric bervariasi mulai dari 3 hingga 32 tahun

## WORK LOCATION

menunjukkan jenis tempat bekerja yang dikategorikan dalam hybrid, remote, dan onsite

## HOURS WORKED OF WEEK

menunjukkan seberapa para pekerja berkerja dalam seminggu. dikategorikan dalam data numeric bervariasi mulai 24 hingga 58 jam setiap minggunya

# penjelasan data pada tabel

## NUMBER OF VIRTUAL MEETING

merepresentasikan jumlah pertemuan virtual yang dilakukan oleh setiap individu dalam dataset selama periode tertentu.

## WORK LIFE BALANCE RATING

menunjukkan seberapa seimbang kehidupan bekerja para pekerja yang dikategorikan dalam data numeric rentang 1 sampai 5

## STRESS LEVEL

menunjukkan tingkat stress pekerja dikategorikan dalam low, medium, high

## MENTAL HEALTH CONDITION

menunjukkan kondisi mental pekerja yang dikategorikan dalam depression, anxiety, dan burnout

# penjelasan data pada tabel

## ACCESS TO MENTAL HEALTH RESOURCES

menunjukkan apakah karyawan memiliki akses ke sumber daya kesehatan mental yang disediakan oleh perusahaan atau tidak. "Yes" menunjukkan adanya akses, sedangkan "No" berarti tidak ada akses.

## PRODUCTIVITY CHANGE

mengukur perubahan produktivitas karyawan sejak bekerja remote. Nilai-nilai yang muncul adalah "Increase" (meningkat), "Decrease" (menurun), "No Change" (tidak berubah).

## SOCIAL ISOLATION RATING

mengukur isolasi sosial yang dirasakan karyawan akibat bekerja secara remote. Semakin tinggi nilainya, maka semakin tinggi tingkat isolasi sosial yang dirasakan.

## SATISFACTION WITH REMOTE WORK

mengukur tingkat kepuasan karyawan terhadap pekerjaan jarak jauh. "Satisfied" menunjukkan kepuasan, "Unsatisfied" menunjukkan ketidakpuasan, dan "Neutral" menunjukkan perasaan netral

# penjelasan data pada tabel

## **COMPANY\_SUPPORT\_FOR\_REMOTE\_WORK**

menunjukkan apakah perusahaan memberikan dukungan terhadap adanya kerja jarak jauh atau tidak dan dikategorikan dalam data numeric dengan rentang 1 sampai 5.

## **PHYSICAL\_ACTIVITY**

mengukur aktivitas fisik yang dilakukan oleh karyawan, apakah karyawan tsb rutin melakukan aktivitas fisik dalam daily (harian) atau (mingguan)

## **SLEEP\_QUALITY**

mengukur tingkat kualitas tidur atau istirahat karyawan dengan nilai yang muncul adalah good (baik), average (cukup), atau poor (kurang)

## **REGION**

menunjukkan wilayah geografis di mana karyawan bekerja secara remote.

# DATA CLEANING

## MASUKAN DATASET

UNTUK MENAMPILKAN DATA SET YANG AKAN DIGUNAKAN

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

data=pd.read_csv('/content/drive/MyDrive/GNFI/Impact_of_Remote_Work_on_Mental_Health.csv')
data
```

# OUTPUT

```
data.isnull().sum()
```

	0
Employee_ID	0
Age	0
Gender	0
Job_Role	0
Industry	0
Years_of_Experience	0
Work_Location	0
Hours_Worked_Per_Week	0
Number_of_Virtual_Meetings	0
Work_Life_Balance_Rating	0
Stress_Level	0
Mental_Health_Condition	1196
Access_to_Mental_Health_Resources	0
Productivity_Change	0
Social_Isolation_Rating	0
Satisfaction_with_Remote_Work	0
Company_Support_for_Remote_Work	0
Physical_Activity	1629
Sleep_Quality	0
Region	0

dtype: int64

- terdapat data yang lebih mencolok daripada lainnya yang berarti terdapat missing value di Mental\_Health\_Condition dan Physical\_Activity



# DATA CLEANING

## DATA ISNULL

GUNAKAN ISNULL UNTUK PERIKSA DATA APAKAH ADA MV DAN  
VALUE\_COUNT UNTUK MELIHAT DISTRIBUSI NILAI

```
# memeriksa data  
data.isnull().sum()  
for x in data.columns:  
    print(data[x].value_counts())
```

## CARI MISSING VALUE

```
data.isnull().sum()
```

# OUTPUT

Mental_Health_Condition
Depression
Anxiety
Anxiety
Depression
NaN
...
Burnout
Depression
Burnout
NaN
Depression

Physical_Activity
Weekly
Weekly
NaN
NaN
Weekly
...
Weekly
NaN
Daily
Daily
NaN

- interpolate yang dilakukan tidak menghasilkan perubahan sehingga tetap terdapat missing value di kolom Mental\_Health\_Condition dan Physical\_Activity

# DATA CLEANING

## INTERPOLATE

digunakan untuk mengganti mv atau nilai yang hilang

```
[ ] data.interpolate()  
data
```

## DATA DROP

buang data yang tidak diperlukan

```
[23] # membuang yang tidak penting  
data=data.dropna()  
data=data.drop("Employee_ID", axis=1)  
data=data.drop("Mental_Health_Condition", axis=1)  
data=data.drop("Physical_Activity", axis=1)  
data
```

# OUTPUT

## SEBELUM

	Employee_ID	Age	Gender	Job_Role	Industry	Years_of_Experience	Work_Location	Hours_Worked_Per_Week	Number_of_Virtual_Meetings	Work_Life_Balance_Rating	Stress_Level
0	EMP0001	0.263158	Non-binary	HR	Healthcare	0.352941	Hybrid	0.675	0.466667	0.25	Medium
1	EMP0002	0.473684	Female	Data Scientist	IT	0.058824	Remote	0.800	0.266667	0.00	Medium
2	EMP0003	0.973684	Non-binary	Software Engineer	Education	0.617647	Hybrid	0.650	0.733333	1.00	Medium
3	EMP0004	0.131579	Male	Software Engineer	Finance	0.558824	Onsite	0.300	0.533333	0.75	High
4	EMP0005	0.710526	Male	Sales	Consulting	0.911765	Onsite	0.375	0.800000	0.25	High
...	...	...	...	...	...	...	...	...	...	...	...
4995	EMP4996	0.263158	Male	Sales	Consulting	0.088235	Onsite	0.100	0.133333	1.00	High
4996	EMP4997	0.447368	Female	Sales	Healthcare	0.764706	Onsite	0.700	1.000000	0.00	Low
4997	EMP4998	0.526316	Female	Sales	Healthcare	0.588235	Hybrid	0.350	0.066667	0.75	High
4998	EMP4999	0.131579	Female	Sales	Healthcare	0.735294	Remote	0.950	0.000000	1.00	Low
4999	EMP5000	0.184211	Male	HR	IT	0.852941	Onsite	0.000	1.000000	0.00	Low

5000 rows × 20 columns

## SESUDAH

	Age	Gender	Job_Role	Industry	Years_of_Experience	Work_Location	Hours_Worked_Per_Week	Number_of_Virtual_Meetings	Work_Life_Balance_Rating	Stress_Level	Ac
0	32	Non-binary	HR	Healthcare	13	Hybrid	47	7	2	Medium	
1	40	Female	Data Scientist	IT	3	Remote	52	4	1	Medium	
2	59	Non-binary	Software Engineer	Education	22	Hybrid	46	11	5	Medium	
3	27	Male	Software Engineer	Finance	20	Onsite	32	8	4	High	
4	49	Male	Sales	Consulting	32	Onsite	35	12	2	High	
...	...	...	...	...	...	...	...	...	...	...	...
4995	32	Male	Sales	Consulting	4	Onsite	24	2	5	High	
4996	39	Female	Sales	Healthcare	27	Onsite	48	15	1	Low	
4997	42	Female	Sales	Healthcare	21	Hybrid	34	1	4	High	
4998	27	Female	Sales	Healthcare	26	Remote	58	0	5	Low	
4999	29	Male	HR	IT	30	Onsite	20	15	1	Low	

5000 rows × 17 columns

- dilakukan data drop untuk menghapus mv dan data yang dirasa tidak penting
- tersisa 17 dari 20 kolom diawal yang menunjukkan bahwa kolom yang tidak diperlukan dan kolom yang memiliki mv telah di hapus

- data set yang sudah dibersihkan kemudian bisa di ubah ke data numeric dan di analisis lebih lanjut

# REPLACE

mengubah semua data  
menjadi data numerik

```
0d
wl = {"Hybrid":1, "Remote":2, "Onsite":3}
sl = {"Low":1, "Medium":2, "High":3}
sq = {"Poor":0, "Average":1, "Good":2}
gdr = {"Male":1, "Female":2, "Non-binary":3}
jr = {"HR":1, "Data Scientist":2, "Software Engineer":3, "Sales":4}
i = {"Healthcare":1, "IT":2, "Education":3, "Finance":4, "Consulting":5}
mhr = {"No":1, "Yes":2}
pc = {"Decrease":1, "Increase":2, "No Change":3}
swrw = {"Unsatisfied":1, "Satisfied":2, "Neutral":3}
r = {"Europe":1, "Asia":2, "North America":3, "Africa":4, "Oceania":5}
data["Work_Location"] = data["Work_Location"].replace(wl)
data["Stress_Level"] = data["Stress_Level"].replace(sl)
data["Sleep_Quality"] = data["Sleep_Quality"].replace(sq)
data["Gender"] = data["Gender"].replace(gdr)
data["Job_Role"] = data["Job_Role"].replace(jr)
data["Industry"] = data["Industry"].replace(i)
data["Access_to_Mental_Health_Resources"] = data["Access_to_Mental_Health_Resources"].replace(mhr)
data["Productivity_Change"] = data["Productivity_Change"].replace(pc)
data["Satisfaction_with_Remote_Work"] = data["Satisfaction_with_Remote_Work"].replace(swrw)
data["Region"] = data["Region"].replace(r)
data_one_hot_encoded = pd.get_dummies(data, columns=[
    "Work_Location", "Stress_Level", "Sleep_Quality", "Gender",
    "Job_Role", "Industry", "Access_to_Mental_Health_Resources", |
    "Productivity_Change", "Satisfaction_with_Remote_Work", "Region"
])

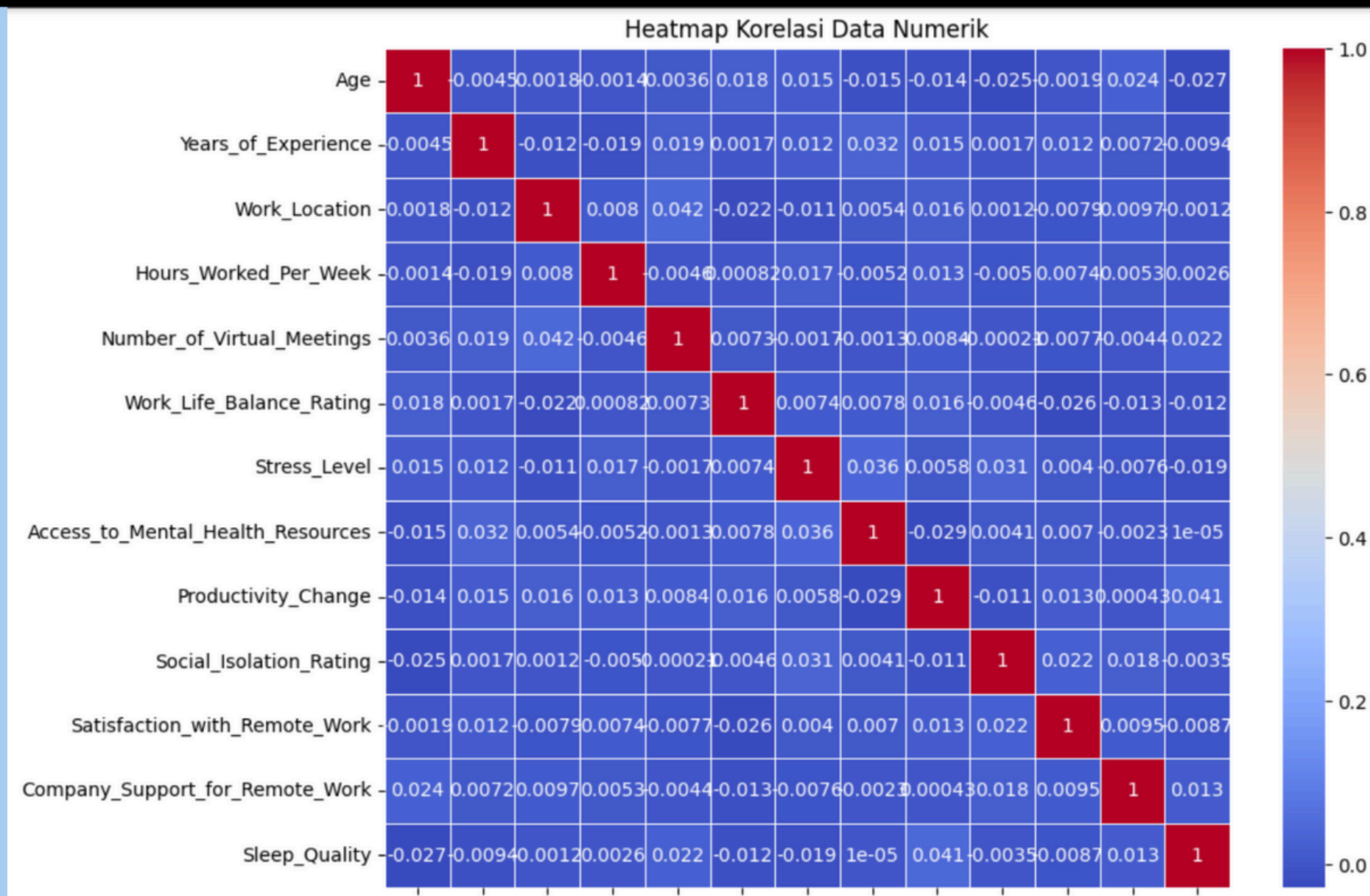
data_one_hot_encoded
```

# HEATMAP

visualisasi data yang  
menggunakan warna untuk  
menggambarkan distribusi  
data di suatu area

```
1 # Memfilter hanya kolom numerik
2 numerical_data = data.select_dtypes(include=['float64', 'int64'])
3
4 # Membuat heatmap korelasi
5 plt.figure(figsize=(10, 8))
6 sns.heatmap(numerical_data.corr(), annot=True, cmap="coolwarm", linewidths=0.5)
7 plt.title("Heatmap Korelasi Data Numerik")
8 plt.show()
```

# OUTPUT



Beberapa variabel menunjukkan korelasi yang cukup kuat, baik positif maupun negatif.

Misalnya, variabel "Work\_Life\_Balance\_Rating" memiliki korelasi negatif yang kuat dengan "Stress\_Level", yang berarti semakin tinggi penilaian keseimbangan kerja-hidup, semakin rendah tingkat stres.



# BOXPLOT

meringkas dan menampilkan distribusi data secara visual berupa grafik

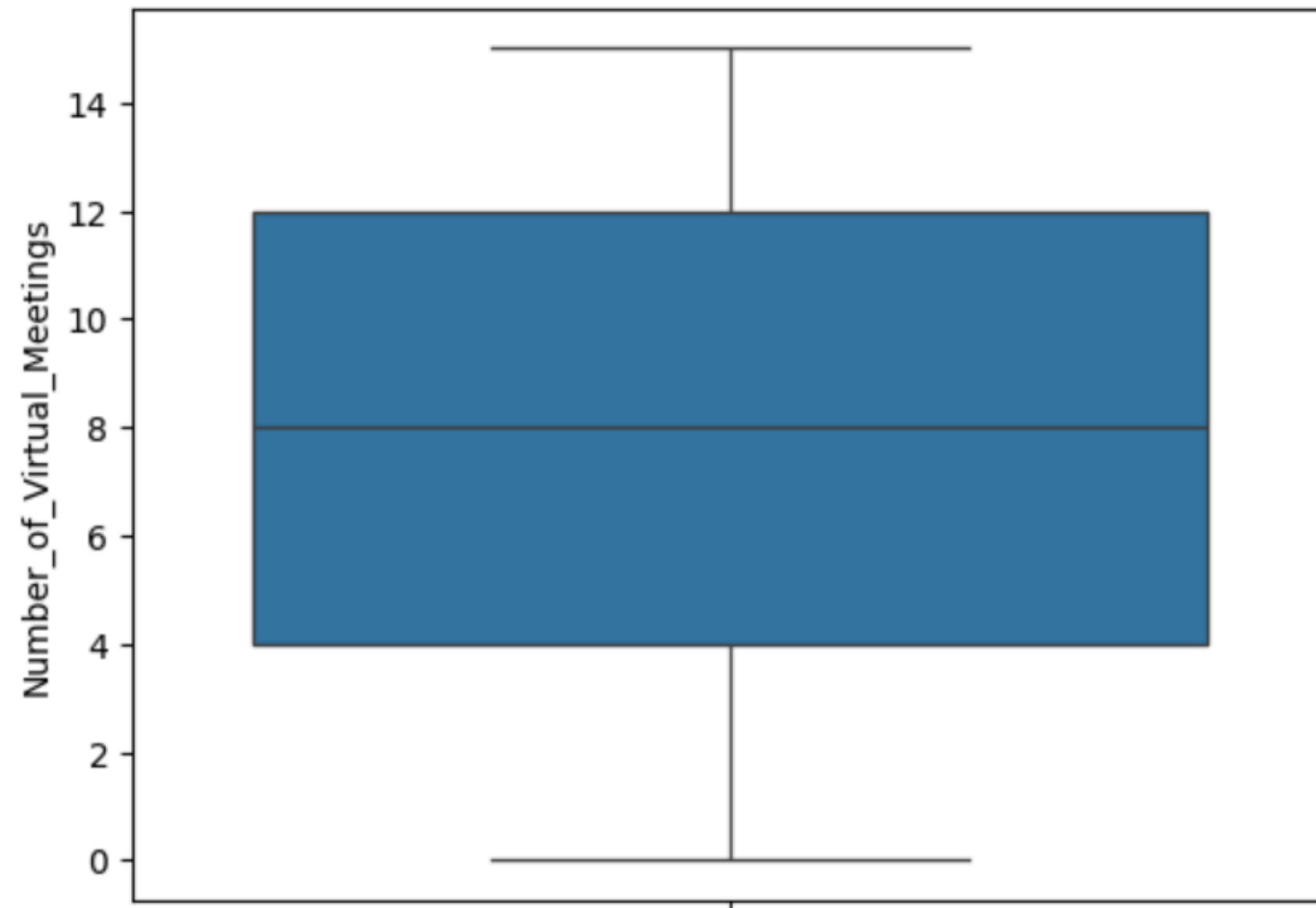
```
1 # mencari outlier
2 sns.boxplot(data["Number_of_Virtual_Meetings"])
```

# KUARTIL

membagi data yang telah disusun secara berurutan menjadi empat bagian yang sama besar

```
1 # Menghitung kuartil dan IQR
2 desc = data.describe()
3 q1 = desc.loc["25%"]
4 q3 = desc.loc["75%"]
5 iqr = q3 - q1
6
7 # Menghitung batas bawah dan atas untuk mendeteksi outliers
8 lower = q1 - 1.5 * iqr
9 upper = q3 + 1.5 * iqr
10
11 # Menghilangkan outliers di setiap kolom numerik
12 for col in data.select_dtypes(include='number').columns:
13     data = data[(data[col] > lower[col]) & (data[col] < upper[col])]
14
15 # Menampilkan data yang sudah dibersihkan dari outliers
16 data
```

# OUTPUT



- Sebagian besar responden memiliki sekitar 8 pertemuan virtual.
- Data cukup menyebar dengan beberapa responden memiliki jauh lebih sedikit atau lebih banyak pertemuan dibandingkan rata-rata.
- Ada beberapa responden yang memiliki jumlah pertemuan virtual yang sangat sedikit (dekat dengan 0) atau sangat banyak (mendekati 14).



# OUTPUT

	Age	Gender	Job_Role	Industry	Years_of_Experience	Work_Location	Hours_Worked_Per_Week	Number_of_Virtual_Meetings	Work_Life_Balance_Rating	Stress_Level	Access_to_Mental_Health_Resources	Productivity_Change	Social_Isolation_Rating	Satisfaction_with_Remote_Work
0	32	3	1	1	13	1	47	7	2	2	1	1	1	1
1	40	2	2	2	3	2	52	4	1	2	1	2	3	2
2	59	3	3	3	22	1	46	11	5	2	1	3	4	1
3	27	1	3	4	20	3	32	8	4	3	2	2	3	1
4	49	1	4	5	32	3	35	12	2	3	2	1	3	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4995	32	1	4	5	4	3	24	2	5	3	2	1	4	3
4996	39	2	4	1	27	3	48	15	1	1	2	1	1	2
4997	42	2	4	1	21	1	34	1	4	3	1	2	3	2
4998	27	2	4	1	26	2	58	0	5	1	2	2	3	1
4999	29	1	1	2	30	3	20	15	1	1	2	3	3	1

5000 rows × 17 columns

Company_Support_for_Remote_Work	Sleep_Quality	Region
1	2	1
2	2	2
5	0	3
3	0	1
3	1	3
...	...	...
1	1	2
1	1	4
1	0	5
4	1	2
5	0	2

- Kolom "Hours\_Worked\_Per\_Week" merepresentasikan distribusi jumlah jam kerja per minggu dari seluruh individu dalam dataset.
- Kuartil Pertama: Menunjukkan 25% individu bekerja kurang dari atau sama dengan jumlah jam yang tertera Q1.
- Median (Q2): Menunjukkan 50% individu bekerja kurang dari atau sama dengan jumlah jam yang tertera pada median.
- Kuartil Ketiga (Q3): Menunjukkan bahwa 75% individu bekerja kurang dari atau sama dengan jumlah jam yang tertera pada Q3.

# OUTPUT

	Age	Gender	Job_Role	Industry	Years_of_Experience	Work_Location	Hours_Worked_Per_Week	Number_of_Virtual_Meetings	Work_Life_Balance_Rating	Stress_Level	Access_to_Mental_Health_Resources	Productivity_Change	Social_Isolation_Rating	Satisfaction_with_Remote_Work
0	32	3	1	1	13	1	47	7	2	2	1	1	1	1
1	40	2	2	2	3	2	52	4	1	2	1	2	3	2
2	59	3	3	3	22	1	46	11	5	2	1	3	4	1
3	27	1	3	4	20	3	32	8	4	3	2	2	3	1
4	49	1	4	5	32	3	35	12	2	3	2	1	3	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4995	32	1	4	5	4	3	24	2	5	3	2	1	4	3
4996	39	2	4	1	27	3	48	15	1	1	2	1	1	2
4997	42	2	4	1	21	1	34	1	4	3	1	2	3	2
4998	27	2	4	1	26	2	58	0	5	1	2	2	3	1
4999	29	1	1	2	30	3	20	15	1	1	2	3	3	1

5000 rows × 17 columns

Company_Support_for_Remote_Work	Sleep_Quality	Region
1	2	1
2	2	2
5	0	3
3	0	1
3	1	3
...	...	...
1	1	2
1	1	4
1	0	5
4	1	2
5	0	2


- Nilai Minimum: Menunjukkan jumlah jam kerja paling sedikit yang dilakukan oleh seorang individu dalam seminggu.
- Nilai Maksimum: Menunjukkan jumlah jam kerja paling banyak yang dilakukan oleh seorang individu dalam seminggu.

# MINMAX SCALER

mengubah data numeric menjadi 0  
dan 1

```
from sklearn.preprocessing import MinMaxScaler  
scaler = MinMaxScaler()  
scaled_data = scaler.fit_transform(data_one_hot_encoded)  
scaled_data_df = pd.DataFrame(scaled_data, columns=data_one_hot_encoded.columns)  
scaled_data_df
```

# OUTPUT



	Age	Years_of_Experience	Hours_Worked_Per_Week	Number_of_Virtual_Meetings	Work_Life_Balance_Rating	Social_Isolation_Rating	Company_Support_Rating
0	0.263158	0.352941	0.675	0.466667	0.25	0.00	
1	0.473684	0.058824	0.800	0.266667	0.00	0.50	
2	0.236842	0.676471	0.775	0.466667	0.50	1.00	
3	0.210526	0.794118	0.925	0.400000	0.00	0.25	
4	0.473684	0.000000	0.025	0.466667	0.25	0.25	
...	...	...	...	...	...	...	...
2572	1.000000	0.235294	0.925	0.266667	0.25	0.75	
2573	0.868421	0.588235	0.625	0.466667	0.25	0.25	
2574	0.473684	0.470588	0.800	0.066667	0.25	0.50	
2575	0.263158	0.088235	0.100	0.133333	1.00	0.75	
2576	0.526316	0.588235	0.350	0.066667	0.75	0.50	

2577 rows × 48 columns