In [1]:

```
# importing modules
import pandas as pd #install pandas using pip
import numpy as np #numpy module is used to perform sum on rows/columns
import matplotlib.pyplot as plt #to plot graphs
import seaborn as sns
```

Data cleaning

In [2]:

```
df = pd.read_csv(r"D:\Data anlaysis on forest fires\archive\amazon.csv")   #importing the csv file
```

In [3]:

```
df
```

Out[3]:

	year	state	month	number	date
0	1998	Acre	Janeiro	0.0	1998-01-01
1	1999	Acre	Janeiro	0.0	1999-01-01
2	2000	Acre	Janeiro	0.0	2000-01-01
3	2001	Acre	Janeiro	0.0	2001-01-01
4	2002	Acre	Janeiro	0.0	2002-01-01
...	...	...	...	...	...
6449	2012	Tocantins	Dezembro	128.0	2012-01-01
6450	2013	Tocantins	Dezembro	85.0	2013-01-01
6451	2014	Tocantins	Dezembro	223.0	2014-01-01
6452	2015	Tocantins	Dezembro	373.0	2015-01-01
6453	2016	Tocantins	Dezembro	119.0	2016-01-01

6454 rows × 5 columns

In [4]:

```
df.dtypes
```

Out[4]:

```
year        int64
state      object
month      object
number    float64
date       object
dtype: object
```

In [5]:

```
df.shape
```

Out[5]:

```
(6454, 5)
```

In [6]:

```
#checking for missing/null values
df.isna().sum()
```

Out[6]:

```
year      0
state     0
month     0
number    0
date      0
dtype: int64
```

There are no missing values in this dataset

In [7]:

```
# checking if there are any duplicate values
df.duplicated().sum()
```

Out[7]:

```
32
```

There are 32 duplicate rows

In [8]:

```
# deleting the duplicate rows
df = df.drop_duplicates()
```

In [9]:

```
df.shape
```

Out[9]:

```
(6422, 5)
```

Data analysis

- Total no. of fires by state

In [10]:

```
# goruping the data
pivot1 = pd.pivot_table(df,values = "number",index = ["state"], aggfunc=np.sum)
pivot1 = pivot1.rename(index={'Pari':'Pará'})
pivot1
```

Out[10]:

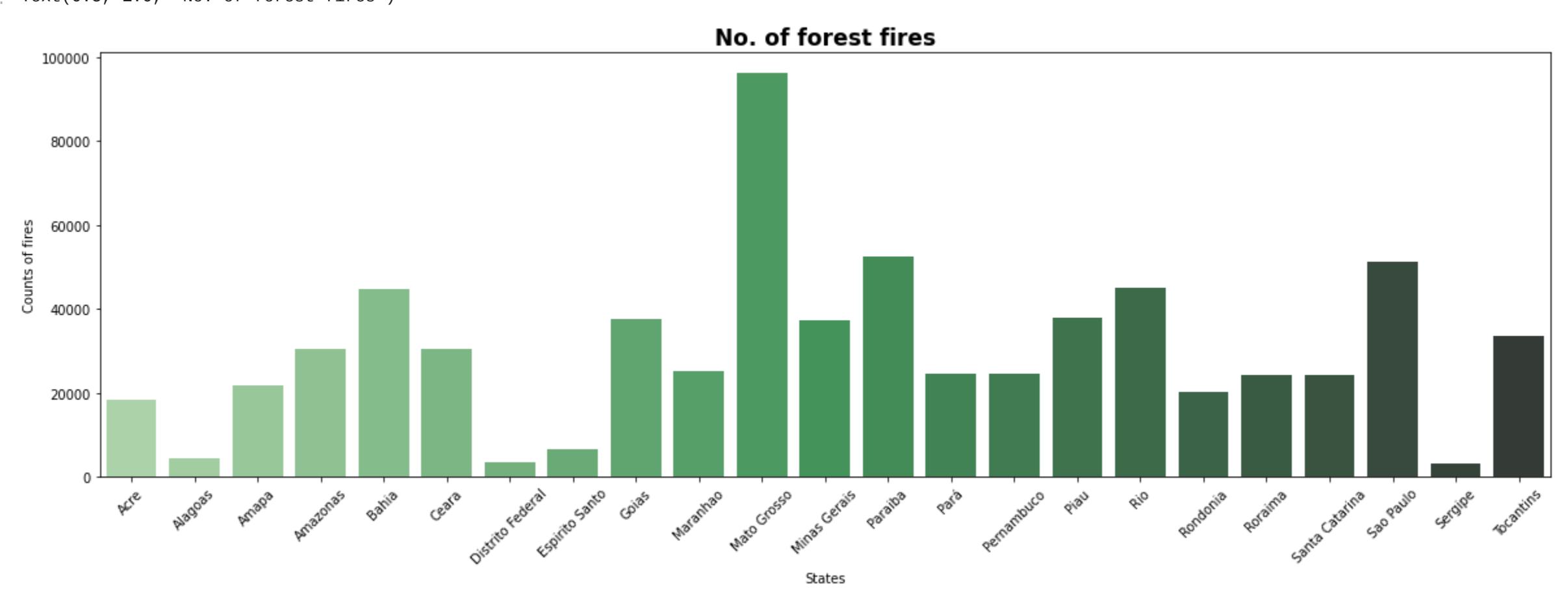
state	number
Acre	18464.030
Alagoas	4606.000
Amapa	21831.576
Amazonas	30650.129
Bahia	44746.226
Ceara	30428.063
Distrito Federal	3561.000
Espirito Santo	6546.000
Goias	37695.520
Maranhao	25129.131
Mato Grosso	96246.028
Minas Gerais	37475.258
Paraiba	52426.918
Pará	24512.144
Pernambuco	24498.000
Piau	37803.747
Rio	45094.865
Rondonia	20285.429
Roraima	24385.074
Santa Catarina	24359.852
Sao Paulo	51121.198
Sergipe	3237.000
Tocantins	33707.885

In [11]:

```
#plotting a graph
plt.figure(figsize=(20,6))
ax = sns.barplot(x=pivot1.index, y='number', data=pivot1,palette='Greens_d')
ax.set_xticklabels(ax.get_xticklabels(), rotation=45)
ax.set_xlabel('States')
ax.set_ylabel('Counts of fires')
ax.set_title('No. of forest fires',fontdict={'fontsize': '17', 'fontweight' : 'bold'})
```

Out[11]:

```
Text(0.5, 1.0, 'No. of forest fires')
```



Conclusions:

1. Large no. of forest fires occur in Manto Grosso
2. Alagaos, Distritio Fedaral,Espririto Santo and Sergipe see a small no. of fires

Study: One of the reasons as to why some states have more forest fires than other can be due to forestation in that area. Brazil is divied into five regions namely North, Northeast, Central west, South, Southeast. Forests are reletively more denser in North and Northeast compared to other regions.

Does this hold true?

- Analysis by region

In [12]:

```
# dividing states by region
states = pivot1.index
state_region = ['North','Northeast','North','North','Northeast','Northeast','Centralwest','Southeast','Centralwest','Northeast',
                'Centralwest','Southeast','Northeast','North','Northeast','Northeast','Southeast','North','North','South',
                'Southeast','Northeast','North']
state_index = 0

for i in states:
    pivot1.loc[(pivot1.index==i),'Region'] = state_region[state_index]
    state_index = state_index + 1
pivot1
```

Out[12]:

state	number	Region
Acre	18464.030	North
Alagoas	4606.000	Northeast
Amapa	21831.576	North
Amazonas	30650.129	North
Bahia	44746.226	Northeast
Ceara	30428.063	Northeast
Distrito Federal	3561.000	Centralwest
Espirito Santo	6546.000	Southeast
Goias	37695.520	Centralwest
Maranhao	25129.131	Northeast
Mato Grosso	96246.028	Centralwest
Minas Gerais	37475.258	Southeast
Paraiba	52426.918	Northeast
Pará	24512.144	North
Pernambuco	24498.000	Northeast
Piau	37803.747	Northeast
Rio	45094.865	Southeast
Rondonia	20285.429	North
Roraima	24385.074	North
Santa Catarina	24359.852	South
Sao Paulo	51121.198	Southeast
Sergipe	3237.000	Northeast
Tocantins	33707.885	North

In [14]:

```
pivot1.Region.value_counts()
```

Out[14]:

```
Northeast      8
North          7
Southeast      4
Centralwest    3
South          1
Name: Region, dtype: int64
```

Yes, there are more Northeast and North rows within the dataset. Does this reflect on the no. of fires?

In [30]:

```
# Number of fires by Region
region = pd.pivot_table(pivot1, values='number',index=['Region'],aggfunc=np.sum)
region.sort_values(by='number',ascending=False)
```

Out[30]:

Region	number
Northeast	222875.085
North	173836.267
Southeast	140237.321
Centralwest	137502.548
South	24359.852

Yes... This shows that there are more no. fires in the North and Northwest. So, the fires are also effected by goegraphical conditions.

- Total no. of fires by month

In [46]:

```
#translating the month names
# If the months are mapped SettingWithCopy warning arises

portuguese_month = df.month.unique()
english_months = ['January','February', 'March', 'April', 'May', 'June', 'July',
                  'August', 'September', 'October', 'November', 'December']

month_index=0

for p_month in portuguese_month:
    df.loc[(df.month == p_month),'month'] = english_months[month_index]
    month_index = month_index + 1
```

In [44]:

```
# checking if the months are in order
df.month.unique()
```

Out[44]:

```
array(['January', 'February', 'March', 'April', 'May', 'June', 'July',
       'August', 'September', 'October', 'November', 'December'],
      dtype=object)
```

In [33]:

```
# aggregating the data
pivot2 = pd.pivot_table(df,values = "number",index = ["month"], aggfunc=np.sum)
#arranging the data(months) in order
pivot2.index = pd.CategoricalIndex(pivot2.index, categories = english_months, ordered = True)
pivot2 = pivot2.sort_index()
```

In [34]:

```
pivot2
```

Out[34]:

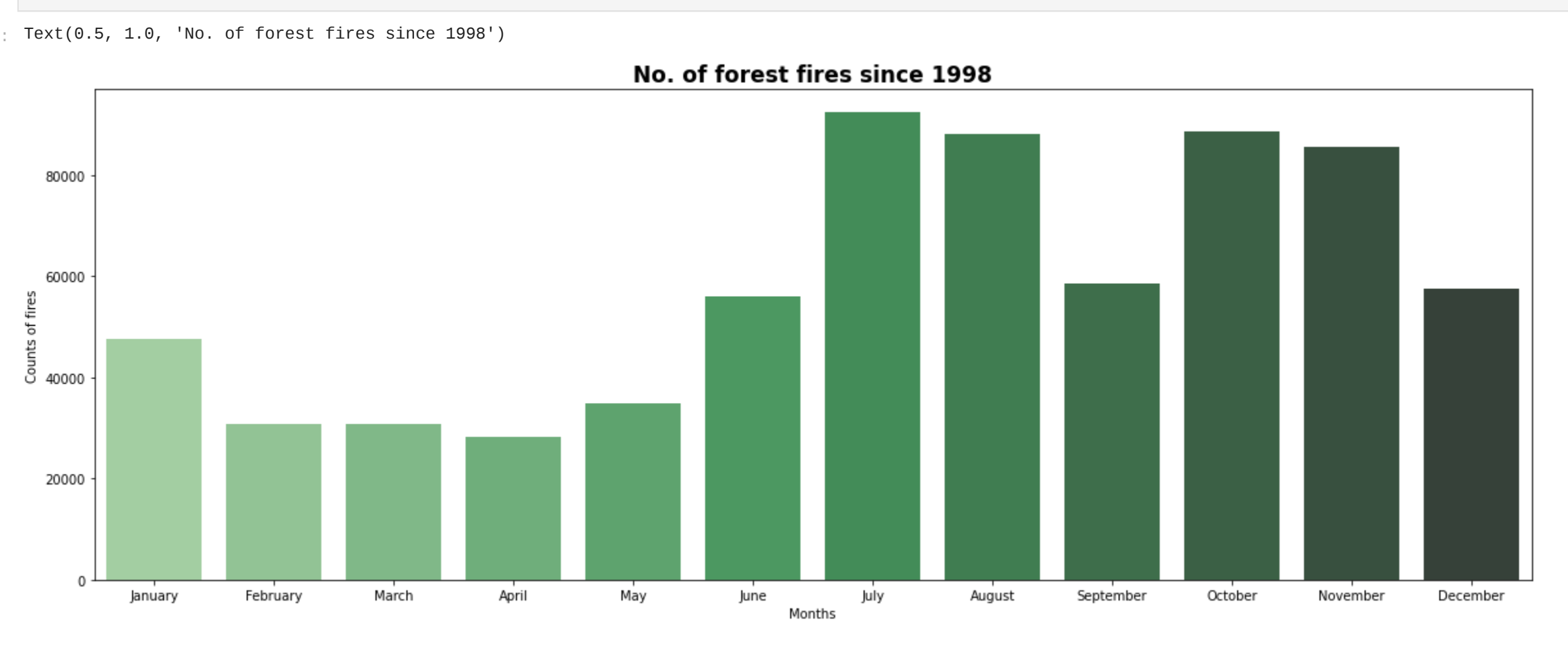
month	number
January	47681.844
February	30839.050
March	30709.405
April	28184.770
May	34725.363
June	55997.675
July	92319.113
August	88050.435
September	58578.305
October	88681.579
November	85508.054
December	57535.480

In [35]:

```
#plotting a graph
plt.figure(figsize=(20,7))
ax = sns.barplot(x=pivot2.index, y='number', data=pivot2,palette='Greens_d')
ax.set_xlabel('Months')
ax.set_ylabel('Counts of fires')
ax.set_title('No. of forest fires since 1998',fontdict={'fontsize': '17', 'fontweight' : 'bold'})
```

Out[35]:

```
Text(0.5, 1.0, 'No. of forest fires since 1998')
```



Conclusions:

1. In February,March,April and May the lowest no. of forest fires occured
2. In the four months(July,august,October and November) the maximum no. forest fires occured
3. There is a sudden increase in July.
4. It is observed that the no. of fires are more in the latter half of the year.

Discussion:

Accoring to a study conducted by the forest dept. of Taiwan, weather is a primary cause for forest fires. We are analysing about forest fires in amazon located in Brazil which lies in the southern hemisphere. So the seasons in brazil is opposite to that of the northern hemisphere. In other words,summer comes at the end of the year where as winter comes at the start of the year.

- Observing temperatures in brazil

In [36]:

```
state1 = pd.read_csv(r"D:\Data anlaysis on forest fires\archive\station_rio.csv")
state2 = pd.read_csv(r"D:\Data anlaysis on forest fires\archive\station_sao_paulo.csv")
```

In [37]:

```
rio_temp = state1.iloc[35:42,1:13].apply(np.mean)
rio_temp = rio_temp.reset_index()
rio_temp = rio_temp.rename(columns={0:'Temp'})

sao_paulo_temp = state2.iloc[61:67,1:13].apply(np.mean)
sao_paulo_temp = sao_paulo_temp.reset_index()
sao_paulo_temp = sao_paulo_temp.rename(columns={0:'Temp'})
```

In [38]:

```
sao_paulo_temp
```

Out[38]:

	index	Temp
0	JAN	23.538333
1	FEB	24.943333
2	MAR	23.813333
3	APR	22.083333
4	MAY	19.056667
5	JUN	17.905000
6	JUL	18.275000
7	AUG	19.420000
8	SEP	20.541667
9	OCT	21.723333
10	NOV	22.258333
11	DEC	23.673333

In [39]:

```
rio_temp
```

Out[39]:

	index	Temp
0	JAN	27.182857
1	FEB	28.188571
2	MAR	26.464286
3	APR	25.005714
4	MAY	22.722857
5	JUN	21.610000
6	JUL	21.257143
7	AUG	22.075714
8	SEP	22.964286
9	OCT	24.095714
10	NOV	24.942857
11	DEC	26.705714

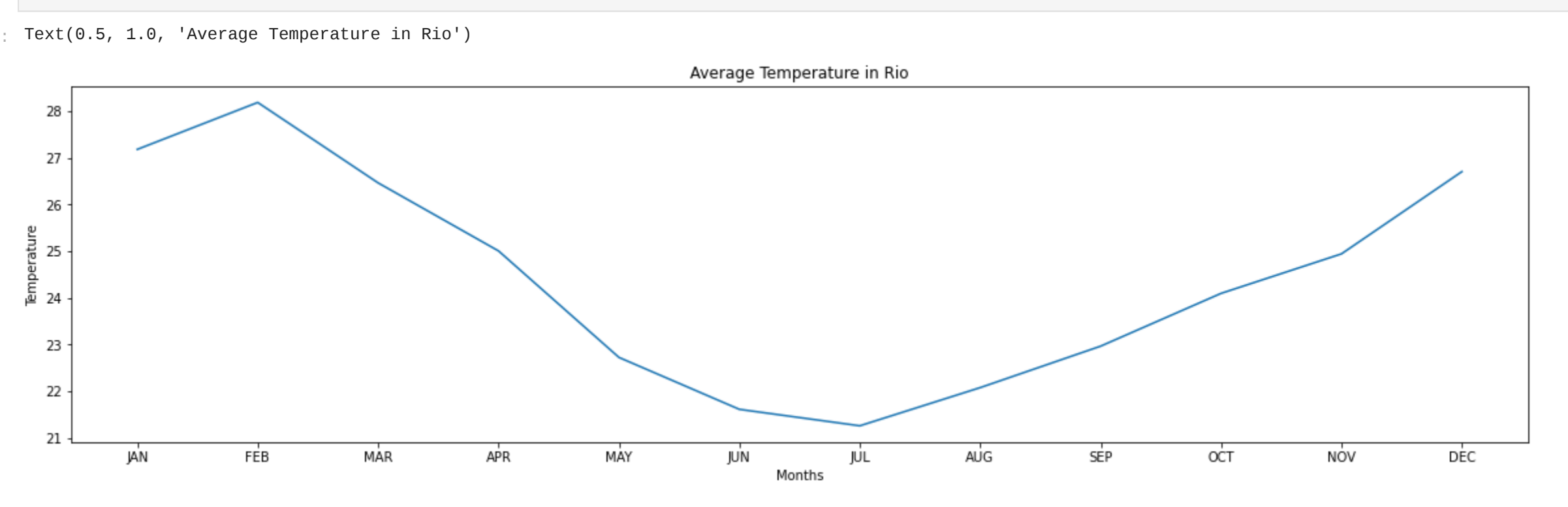
- plotting graphs

In [40]:

```
plt.figure(figsize=(20,5))
ax = sns.lineplot(x = 'index', y = 'Temp' , data = rio_temp)
ax.set_xlabel('Months')
ax.set_ylabel('Temperature')
ax.set_title('Average Temperature in Rio')
```

Out[40]:

```
Text(0.5, 1.0, 'Average Temperature in Rio')
```

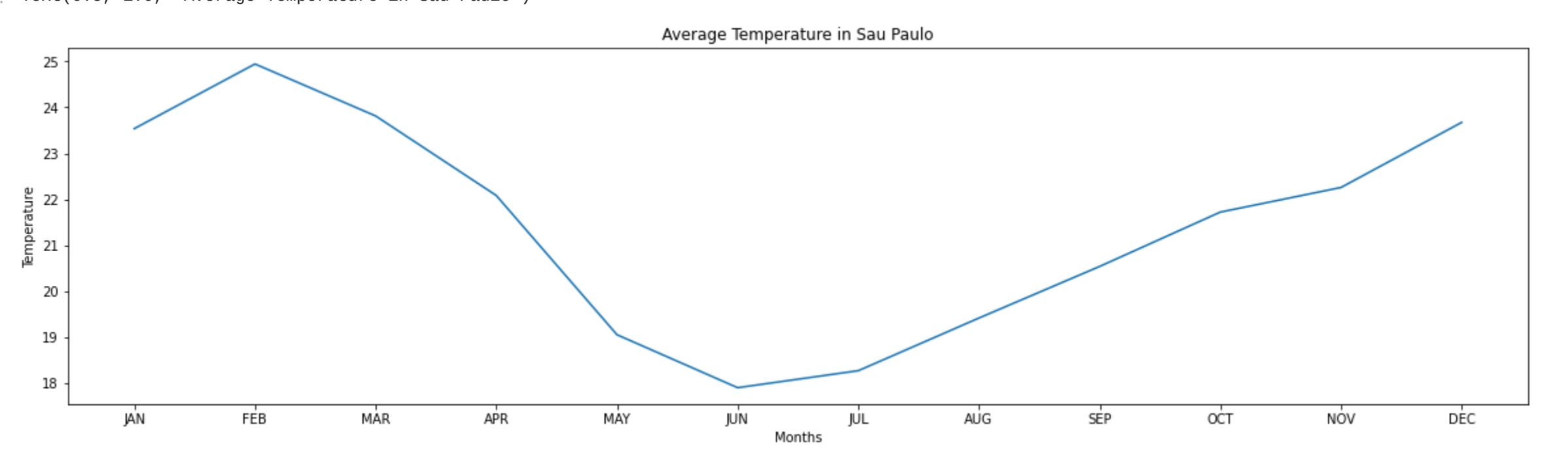


In [41]:

```
plt.figure(figsize=(20,5))
ax = sns.lineplot(x = 'index', y = 'Temp' , data = sao_paulo_temp)
ax.set_xlabel('Months')
ax.set_ylabel('Temperature')
ax.set_title('Average Temperature in Sau Paulo')
```

Out[41]:

```
Text(0.5, 1.0, 'Average Temperature in Sau Paulo')
```



- Total no. of fires by year

In [42]:

```
pivot3 = pd.pivot_table(df,values = "number",index = ["year"], aggfunc=np.sum)
pivot3
```

Out[42]:

year	number
1998	20013.971
1999	26882.821
2000	27351.251
2001	29054.612
2002	37390.600
2003	42760.674
2004	38450.163
2005	35004.965
2006	33824.161
2007	33028.413
2008	29378.964
2009	39116.178
2010	37037.449
2011	34633.545
2012	40084.860
2013	35137.118
2014	39621.183
2015	41208.292
2016	42212.229
2017	36619.624

In [43]:

```
#plotting a graph
plt.figure(figsize=(20,7))
ax = sns.barplot(x=pivot3.index, y='number', data=pivot3,palette='Greens_d')
ax.set_xlabel('Years')
ax.set_ylabel('Counts of fires')
ax.set_title('No. of forest fires since 1998',fontdict={'fontsize': '17', 'fontweight' : 'bold'})
```

Out[43]:

```
Text(0.5, 1.0, 'No. of forest fires since 1998')
```