

Importing Libraries

In [1]:

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

Importing Datasets

In [2]:

```
df=pd.read_csv("rainfall_bihar.csv")
df
```

Out[2]:

	index	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV
0	897	BIHAR	1901	51.8	19.6	11.9	1.1	65.6	66.3	245.9	319.4	155.1	8.3	7.3
1	898	BIHAR	1902	4.6	0.7	24.3	17.3	66.3	118.2	361.0	225.5	358.7	28.5	1.1
2	899	BIHAR	1903	5.3	4.7	2.0	4.7	28.2	192.9	115.0	342.6	173.9	147.0	0.1
3	900	BIHAR	1904	6.3	1.7	3.5	5.3	118.7	191.6	394.4	351.3	84.4	98.1	10.6
4	901	BIHAR	1905	16.0	30.1	32.6	21.4	77.5	50.5	409.1	495.3	353.9	11.6	0.0
...
110	1007	BIHAR	2011	4.2	7.7	9.2	23.9	74.5	211.0	241.1	278.7	234.1	10.0	2.0
111	1008	BIHAR	2012	18.1	2.7	7.3	20.4	18.8	96.2	354.0	240.4	233.8	34.3	6.4
112	1009	BIHAR	2013	5.1	22.6	0.6	32.3	89.5	183.3	182.0	213.6	143.3	197.1	0.4
113	1010	BIHAR	2014	17.0	33.5	8.4	0.7	103.9	115.2	265.4	307.6	160.3	47.8	0.0
114	1011	BIHAR	2015	12.8	1.8	27.2	38.7	39.5	122.1	231.5	287.0	101.7	10.4	0.0

115 rows × 20 columns



head

In [3]:

```
df.head(5)
df
```

Out[3]:

	index	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV
0	897	BIHAR	1901	51.8	19.6	11.9	1.1	65.6	66.3	245.9	319.4	155.1	8.3	7.3
1	898	BIHAR	1902	4.6	0.7	24.3	17.3	66.3	118.2	361.0	225.5	358.7	28.5	1.1

	index	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV
2	899	BIHAR	1903	5.3	4.7	2.0	4.7	28.2	192.9	115.0	342.6	173.9	147.0	0.1
3	900	BIHAR	1904	6.3	1.7	3.5	5.3	118.7	191.6	394.4	351.3	84.4	98.1	10.6
4	901	BIHAR	1905	16.0	30.1	32.6	21.4	77.5	50.5	409.1	495.3	353.9	11.6	0.0
...
110	1007	BIHAR	2011	4.2	7.7	9.2	23.9	74.5	211.0	241.1	278.7	234.1	10.0	2.0
111	1008	BIHAR	2012	18.1	2.7	7.3	20.4	18.8	96.2	354.0	240.4	233.8	34.3	6.4
112	1009	BIHAR	2013	5.1	22.6	0.6	32.3	89.5	183.3	182.0	213.6	143.3	197.1	0.4
113	1010	BIHAR	2014	17.0	33.5	8.4	0.7	103.9	115.2	265.4	307.6	160.3	47.8	0.0
114	1011	BIHAR	2015	12.8	1.8	27.2	38.7	39.5	122.1	231.5	287.0	101.7	10.4	0.0

115 rows × 20 columns

tail

In [4]:

```
df.tail(5)
df
```

Out[4]:

	index	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV
0	897	BIHAR	1901	51.8	19.6	11.9	1.1	65.6	66.3	245.9	319.4	155.1	8.3	7.3
1	898	BIHAR	1902	4.6	0.7	24.3	17.3	66.3	118.2	361.0	225.5	358.7	28.5	1.1
2	899	BIHAR	1903	5.3	4.7	2.0	4.7	28.2	192.9	115.0	342.6	173.9	147.0	0.1
3	900	BIHAR	1904	6.3	1.7	3.5	5.3	118.7	191.6	394.4	351.3	84.4	98.1	10.6
4	901	BIHAR	1905	16.0	30.1	32.6	21.4	77.5	50.5	409.1	495.3	353.9	11.6	0.0
...
110	1007	BIHAR	2011	4.2	7.7	9.2	23.9	74.5	211.0	241.1	278.7	234.1	10.0	2.0
111	1008	BIHAR	2012	18.1	2.7	7.3	20.4	18.8	96.2	354.0	240.4	233.8	34.3	6.4
112	1009	BIHAR	2013	5.1	22.6	0.6	32.3	89.5	183.3	182.0	213.6	143.3	197.1	0.4
113	1010	BIHAR	2014	17.0	33.5	8.4	0.7	103.9	115.2	265.4	307.6	160.3	47.8	0.0
114	1011	BIHAR	2015	12.8	1.8	27.2	38.7	39.5	122.1	231.5	287.0	101.7	10.4	0.0

115 rows × 20 columns

Data Cleaning and Data Preprocessing

describe()

In [5]:

```
df.describe()
```

Out[5]:

	index	YEAR	JAN	FEB	MAR	APR	MAY	JUN
count	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000
mean	954.000000	1958.000000	13.386087	14.393913	10.124348	16.918261	53.081739	174.315652
std	33.341666	33.341666	14.791960	15.075036	11.695340	15.978278	27.941714	76.167930
min	897.000000	1901.000000	0.000000	0.000000	0.000000	0.100000	1.300000	48.100000
25%	925.500000	1929.500000	2.350000	2.750000	1.800000	5.250000	31.550000	117.100000
50%	954.000000	1958.000000	9.400000	8.400000	6.500000	12.600000	46.200000	165.500000
75%	982.500000	1986.500000	18.700000	21.400000	12.850000	24.500000	76.200000	211.000000
max	1011.000000	2015.000000	81.200000	66.300000	65.500000	91.400000	118.700000	446.000000

shape

In [6]:

```
np.shape(df)
```

Out[6]: (115, 20)

size

In [7]:

```
np.size(df)
```

Out[7]: 2300

dropna

In [8]:

```
df=df.dropna()
```

columns

In [9]:

```
df.columns
```

Out[9]: Index(['index', 'SUBDIVISION', 'YEAR', 'JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL', 'AUG', 'SEP', 'OCT', 'NOV', 'DEC', 'ANNUAL', 'Jan-Feb',

```
'Mar-May', 'Jun-Sep', 'Oct-Dec'],
dtype='object')
```

info()

In [10]:

```
df.info()
```

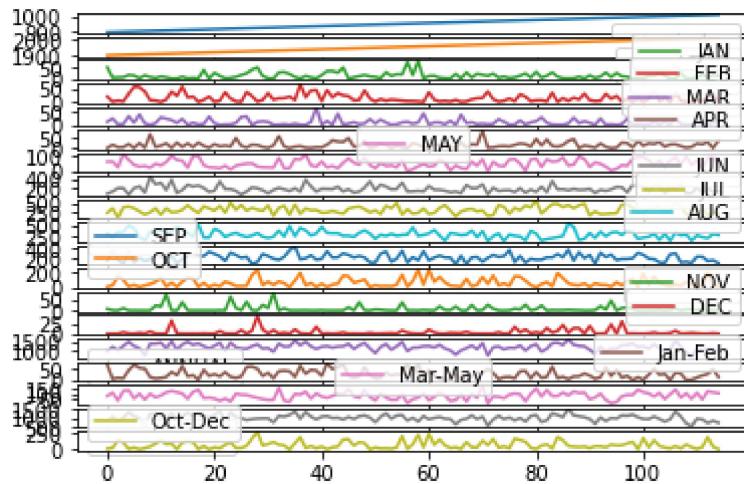
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 115 entries, 0 to 114
Data columns (total 20 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   index       115 non-null    int64  
 1   SUBDIVISION 115 non-null    object  
 2   YEAR        115 non-null    int64  
 3   JAN         115 non-null    float64 
 4   FEB         115 non-null    float64 
 5   MAR         115 non-null    float64 
 6   APR         115 non-null    float64 
 7   MAY         115 non-null    float64 
 8   JUN         115 non-null    float64 
 9   JUL         115 non-null    float64 
 10  AUG         115 non-null    float64 
 11  SEP         115 non-null    float64 
 12  OCT         115 non-null    float64 
 13  NOV         115 non-null    float64 
 14  DEC         115 non-null    float64 
 15  ANNUAL      115 non-null    float64 
 16  Jan-Feb     115 non-null    float64 
 17  Mar-May     115 non-null    float64 
 18  Jun-Sep     115 non-null    float64 
 19  Oct-Dec     115 non-null    float64 
dtypes: float64(17), int64(2), object(1)
memory usage: 18.9+ KB
```

Line chart

In [11]:

```
df.plot.line(subplots=True)
```

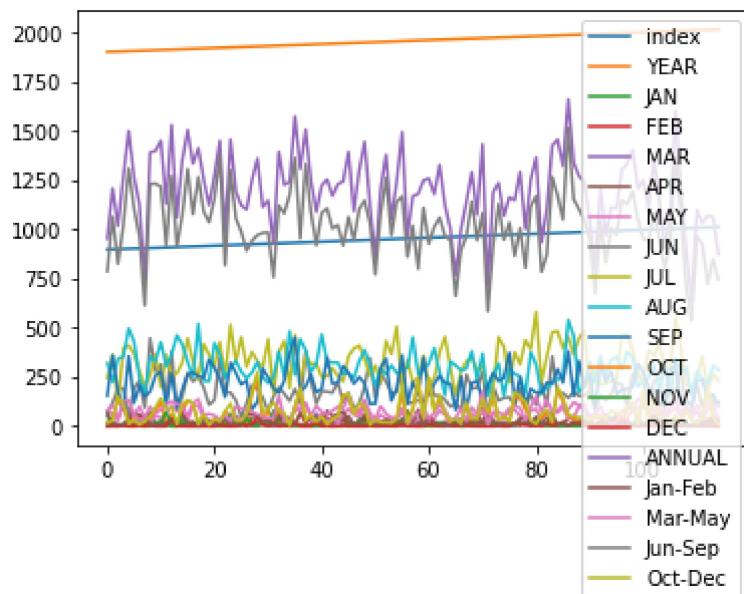
```
Out[11]: array([<AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>,
   <AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>,
   <AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>,
   <AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>,
   <AxesSubplot:>, <AxesSubplot:>, <AxesSubplot:>], dtype=object)
```



Line chart

In [12]: `df.plot.line()`

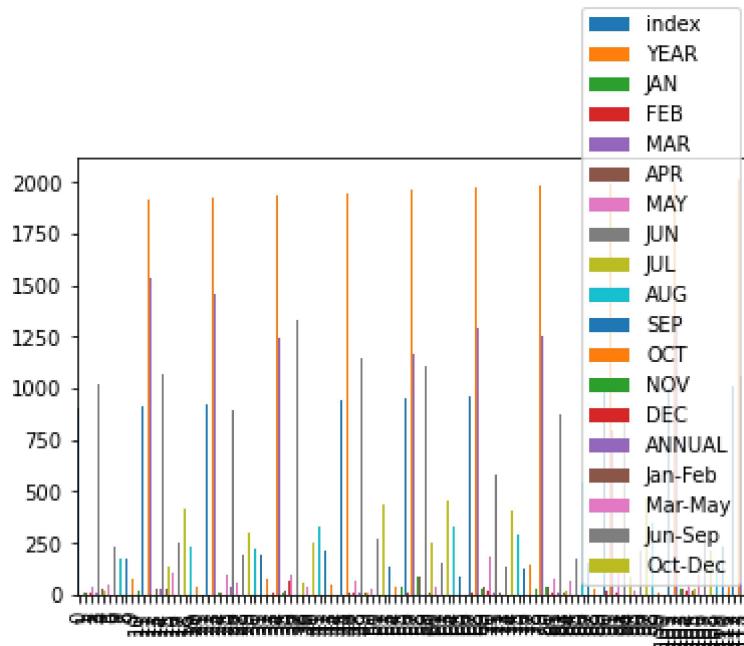
Out[12]: <AxesSubplot:>



Bar chart

In [13]: `df.plot.bar()`

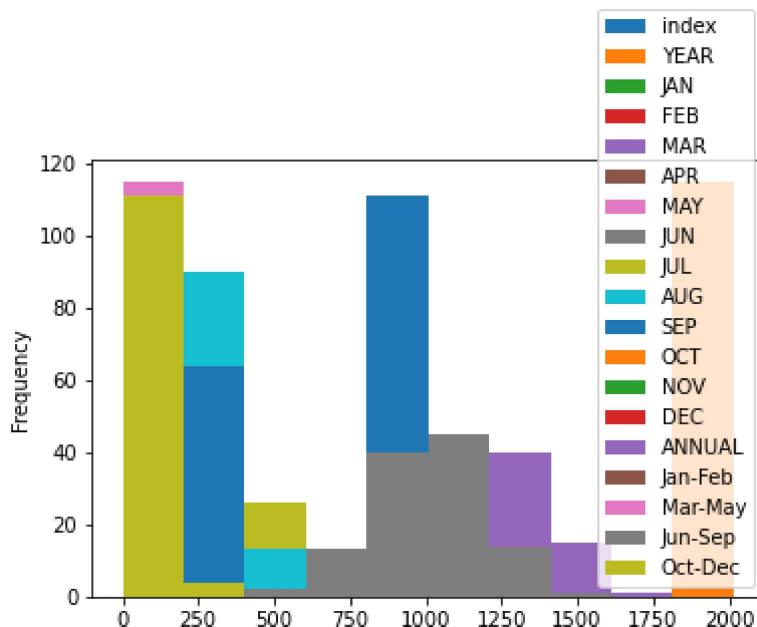
Out[13]: <AxesSubplot:>



Histogram

In [14]: `df.plot.hist()`

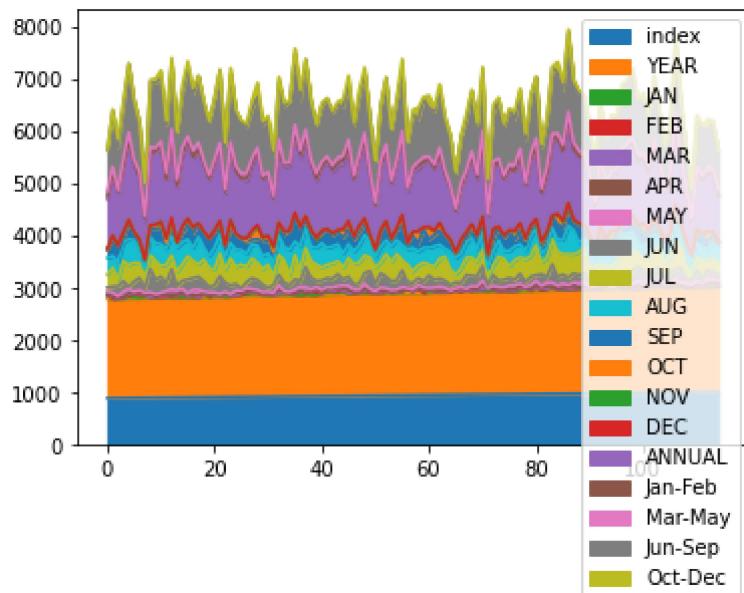
Out[14]: <AxesSubplot:ylabel='Frequency'>



Area chart

In [15]: `df.plot.area()`

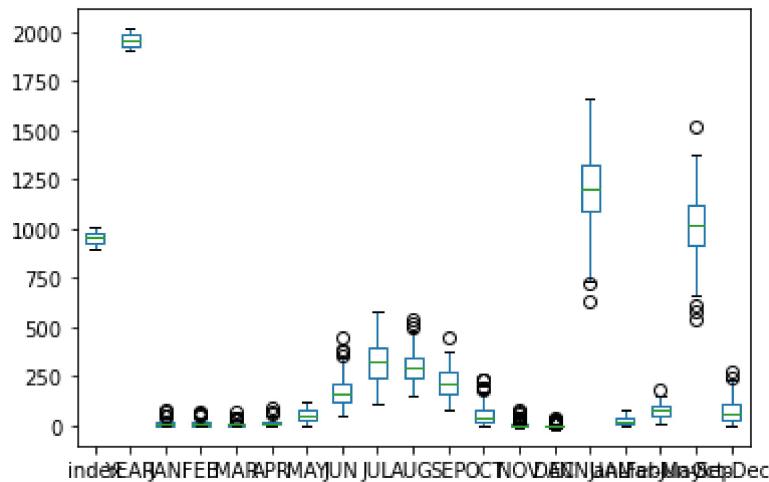
Out[15]: <AxesSubplot:>



Box chart

```
In [16]: df.plot.box()
```

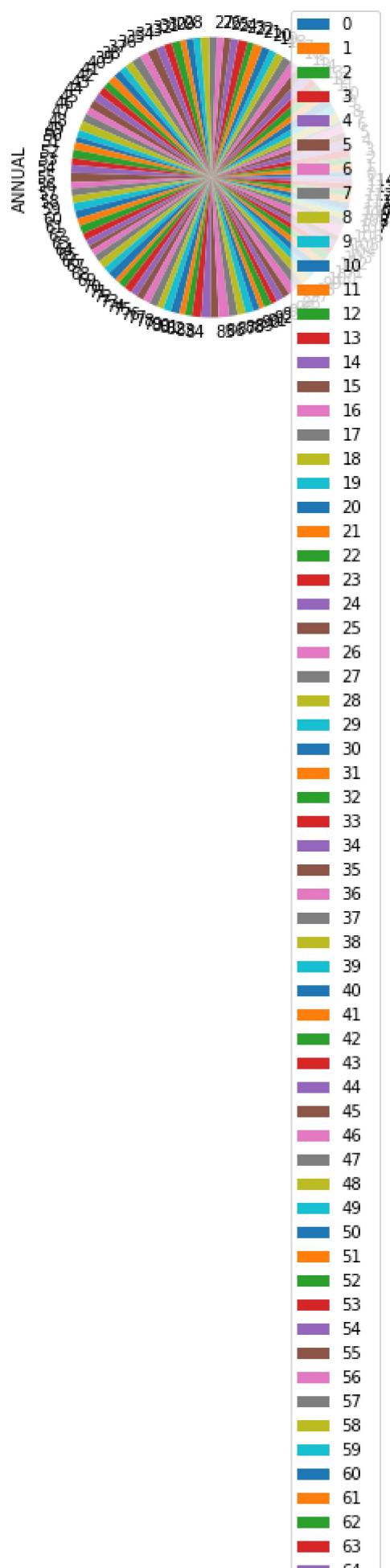
```
Out[16]: <AxesSubplot:>
```

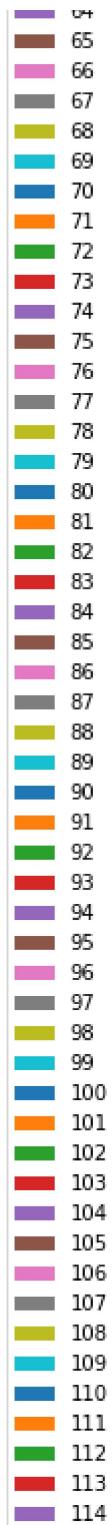


Pie chart

```
In [17]: df.plot.pie(y='ANNUAL')
```

```
Out[17]: <AxesSubplot:ylabel='ANNUAL'>
```

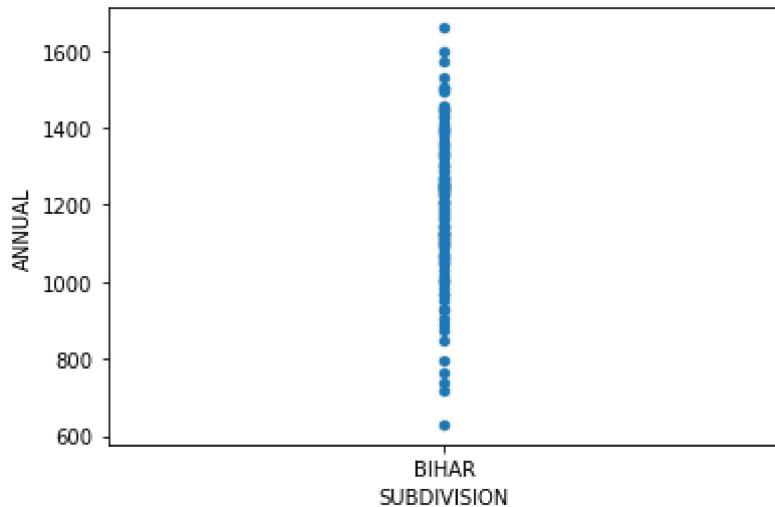




Scatter chart

In [18]: `df.plot.scatter(x='SUBDIVISION' ,y='ANNUAL')`

Out[18]: <AxesSubplot:xlabel='SUBDIVISION', ylabel='ANNUAL'>



In [19]:

`df.info()`

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 115 entries, 0 to 114
Data columns (total 20 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   index       115 non-null    int64  
 1   SUBDIVISION 115 non-null    object  
 2   YEAR        115 non-null    int64  
 3   JAN         115 non-null    float64 
 4   FEB         115 non-null    float64 
 5   MAR         115 non-null    float64 
 6   APR         115 non-null    float64 
 7   MAY         115 non-null    float64 
 8   JUN         115 non-null    float64 
 9   JUL         115 non-null    float64 
 10  AUG         115 non-null    float64 
 11  SEP         115 non-null    float64 
 12  OCT         115 non-null    float64 
 13  NOV         115 non-null    float64 
 14  DEC         115 non-null    float64 
 15  ANNUAL      115 non-null    float64 
 16  Jan-Feb     115 non-null    float64 
 17  Mar-May     115 non-null    float64 
 18  Jun-Sep     115 non-null    float64 
 19  Oct-Dec     115 non-null    float64 
dtypes: float64(17), int64(2), object(1)
memory usage: 18.9+ KB
```

In [20]:

`df.describe()`

Out[20]:

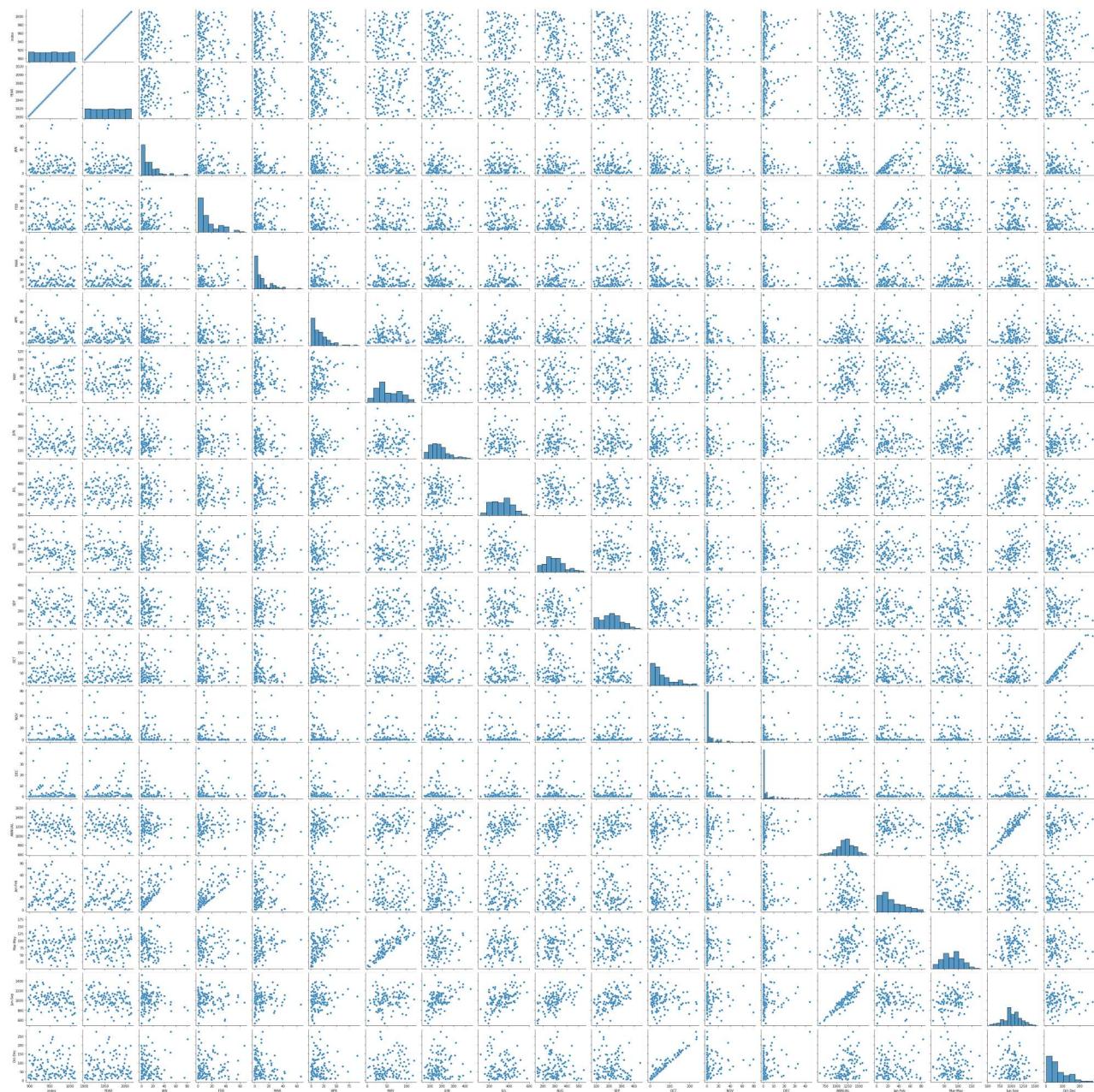
	index	YEAR	JAN	FEB	MAR	APR	MAY	JUN
count	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000	115.000000
mean	954.000000	1958.000000	13.386087	14.393913	10.124348	16.918261	53.081739	174.315652
std	33.341666	33.341666	14.791960	15.075036	11.695340	15.978278	27.941714	76.167930
min	897.000000	1901.000000	0.000000	0.000000	0.000000	0.100000	1.300000	48.100000
25%	925.500000	1929.500000	2.350000	2.750000	1.800000	5.250000	31.550000	117.100000

	index	YEAR	JAN	FEB	MAR	APR	MAY	JUN
50%	954.000000	1958.000000	9.400000	8.400000	6.500000	12.600000	46.200000	165.500000
75%	982.500000	1986.500000	18.700000	21.400000	12.850000	24.500000	76.200000	211.000000
max	1011.000000	2015.000000	81.200000	66.300000	65.500000	91.400000	118.700000	446.000000

EDA AND VISUALIZATION

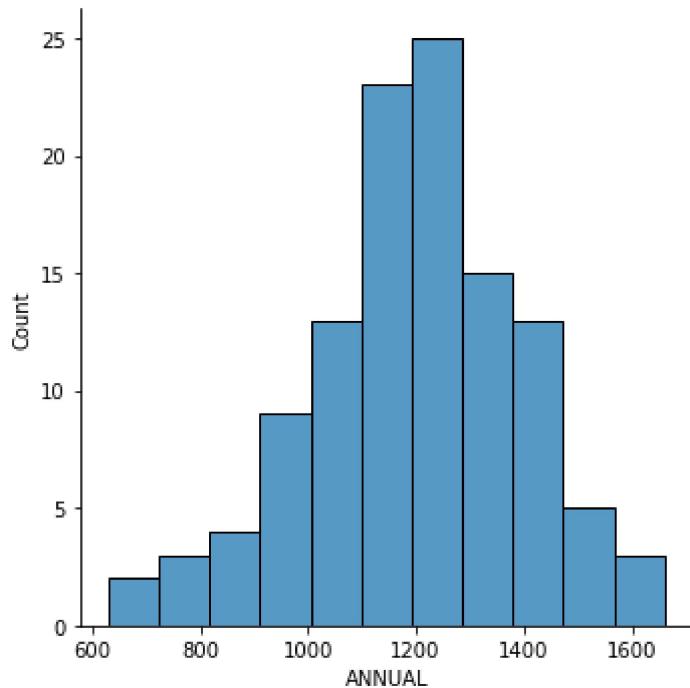
In [21]: `sns.pairplot(df)`

Out[21]: <seaborn.axisgrid.PairGrid at 0x175c3243a30>



In [22]: `sns.displot(df['ANNUAL'])`

Out[22]: <seaborn.axisgrid.FacetGrid at 0x175ce12bca0>

In [23]:
sns.heatmap(df.corr())

Out[23]: <AxesSubplot:>

