Pandas

In [1]:

```
import pandas as pd
```

In [2]:

```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:

```
data=pd.read_csv("Greenhouse-gas-emissions-industry-and-household-March-2022-quarter.csv
df=pd.DataFrame(data)
df
```

Out[5]:

	Anzsic	Anzsic_description	Period	Data_value	Variable	Units	Gas	Sta
0	ZPZ	Primary industries	2014Q1	11316.0	Seasonally adjusted	Kilotonnes	Carbon dioxide equivalents	
1	ZPZ	Primary industries	2014Q2	NaN	Seasonally adjusted	Kilotonnes	Carbon dioxide equivalents	
2	ZPZ	Primary industries	2014Q3	11126.0	Seasonally adjusted	Kilotonnes	Carbon dioxide equivalents	
3	ZPZ	Primary industries	2014Q4	11190.0	Seasonally adjusted	Kilotonnes	Carbon dioxide equivalents	
4	ZPZ	Primary industries	2015Q1	NaN	Seasonally adjusted	Kilotonnes	Carbon dioxide equivalents	
...	...	...	...	...	...	...	...	
1051	ZZZ	Total	2021Q1	22211.0	Actual	Kilotonnes	Carbon dioxide equivalents	
1052	ZZZ	Total	2021Q2	19019.0	Actual	Kilotonnes	Carbon dioxide equivalents	
1053	ZZZ	Total	2021Q3	17788.0	Actual	Kilotonnes	Carbon dioxide equivalents	
1054	ZZZ	Total	2021Q4	20312.0	Actual	Kilotonnes	Carbon dioxide equivalents	
1055	ZZZ	Total	2022Q1	21462.0	Actual	Kilotonnes	Carbon dioxide equivalents	

1056 rows × 8 columns

Finding and replacing missing values

In [3]:

```
df.isnull()
```

Out[6]:

	Anzsic	Anzsic_description	Period	Data_value	Variable	Units	Gas	Status
0	False	False	False	False	False	False	False	False
1	False	False	False	True	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	True	False	False	False	False
...	...	...	...	...	...	...	...	...
1051	False	False	False	False	False	False	False	False
1052	False	False	False	False	False	False	False	False
1053	False	False	False	False	False	False	False	False
1054	False	False	False	False	False	False	False	False
1055	False	False	False	False	False	False	False	False

1056 rows × 8 columns

In [4]:

```
df['Data_value'] = df['Data_value'].fillna(df['Data_value'].mean())
df
```

Out[7]:

	Anzsic	Anzsic_description	Period	Data_value	Variable	Units	Gas	S
0	ZPZ	Primary industries	2014Q1	11316.000000	Seasonally adjusted	Kilotonnes	Carbon dioxide equivalents	
1	ZPZ	Primary industries	2014Q2	4992.389685	Seasonally adjusted	Kilotonnes	Carbon dioxide equivalents	
2	ZPZ	Primary industries	2014Q3	11126.000000	Seasonally adjusted	Kilotonnes	Carbon dioxide equivalents	
3	ZPZ	Primary industries	2014Q4	11190.000000	Seasonally adjusted	Kilotonnes	Carbon dioxide equivalents	
4	ZPZ	Primary industries	2015Q1	4992.389685	Seasonally adjusted	Kilotonnes	Carbon dioxide equivalents	
...	...	...	...	...	...	...	...	
1051	ZZZ	Total	2021Q1	22211.000000	Actual	Kilotonnes	Carbon dioxide equivalents	
1052	ZZZ	Total	2021Q2	19019.000000	Actual	Kilotonnes	Carbon dioxide equivalents	
1053	ZZZ	Total	2021Q3	17788.000000	Actual	Kilotonnes	Carbon dioxide equivalents	
1054	ZZZ	Total	2021Q4	20312.000000	Actual	Kilotonnes	Carbon dioxide equivalents	
1055	ZZZ	Total	2022Q1	21462.000000	Actual	Kilotonnes	Carbon dioxide equivalents	

1056 rows × 8 columns

Reshaping the Dataset

In [8]:

```
df['Variable']=df['Variable'].map({'Seasonally adjusted':0,'Actual':1}).astype(float)
df
```

Out[8]:

	Anzsic	Anzsic_description	Period	Data_value	Variable	Units	Gas	Sta
0	ZPZ	Primary industries	2014Q1	11316.000000	0.0	Kilotonnes	Carbon dioxide equivalents	
1	ZPZ	Primary industries	2014Q2	4992.389685	0.0	Kilotonnes	Carbon dioxide equivalents	
2	ZPZ	Primary industries	2014Q3	11126.000000	0.0	Kilotonnes	Carbon dioxide equivalents	
3	ZPZ	Primary industries	2014Q4	11190.000000	0.0	Kilotonnes	Carbon dioxide equivalents	
4	ZPZ	Primary industries	2015Q1	4992.389685	0.0	Kilotonnes	Carbon dioxide equivalents	
...	...	...	...	...	...	...	...	
1051	ZZZ	Total	2021Q1	22211.000000	1.0	Kilotonnes	Carbon dioxide equivalents	
1052	ZZZ	Total	2021Q2	19019.000000	1.0	Kilotonnes	Carbon dioxide equivalents	
1053	ZZZ	Total	2021Q3	17788.000000	1.0	Kilotonnes	Carbon dioxide equivalents	
1054	ZZZ	Total	2021Q4	20312.000000	1.0	Kilotonnes	Carbon dioxide equivalents	
1055	ZZZ	Total	2022Q1	21462.000000	1.0	Kilotonnes	Carbon dioxide equivalents	

1056 rows × 8 columns

Filtering in Pandas

In [9]:

```
df=df[df['Data_value']>=12567]
df
```

Out[9]:

	Anzsic	Anzsic_description	Period	Data_value	Variable	Units	Gas	Statu
330	ZZ9	Total industry	2014Q1	18508.0	0.0	Kilotonnes	Carbon dioxide equivalents	
331	ZZ9	Total industry	2014Q2	18644.0	0.0	Kilotonnes	Carbon dioxide equivalents	
332	ZZ9	Total industry	2014Q3	18747.0	0.0	Kilotonnes	Carbon dioxide equivalents	
333	ZZ9	Total industry	2014Q4	18889.0	0.0	Kilotonnes	Carbon dioxide equivalents	
334	ZZ9	Total industry	2015Q1	18789.0	0.0	Kilotonnes	Carbon dioxide equivalents	
...	...	...	...	...	...	...	...	.
1051	ZZZ	Total	2021Q1	22211.0	1.0	Kilotonnes	Carbon dioxide equivalents	
1052	ZZZ	Total	2021Q2	19019.0	1.0	Kilotonnes	Carbon dioxide equivalents	
1053	ZZZ	Total	2021Q3	17788.0	1.0	Kilotonnes	Carbon dioxide equivalents	
1054	ZZZ	Total	2021Q4	20312.0	1.0	Kilotonnes	Carbon dioxide equivalents	
1055	ZZZ	Total	2022Q1	21462.0	1.0	Kilotonnes	Carbon dioxide equivalents	

140 rows × 8 columns

In [10]:

```
df=df.drop(['Units'],axis=1)
df
```

Out[10]:

	Anzsic	Anzsic_description	Period	Data_value	Variable	Gas	Status
330	ZZ9	Total industry	2014Q1	18508.0	0.0	Carbon dioxide equivalents	P
331	ZZ9	Total industry	2014Q2	18644.0	0.0	Carbon dioxide equivalents	P
332	ZZ9	Total industry	2014Q3	18747.0	0.0	Carbon dioxide equivalents	P
333	ZZ9	Total industry	2014Q4	18889.0	0.0	Carbon dioxide equivalents	P
334	ZZ9	Total industry	2015Q1	18789.0	0.0	Carbon dioxide equivalents	P
...	...	...	...	...	...	...	...
1051	ZZZ	Total	2021Q1	22211.0	1.0	Carbon dioxide equivalents	P
1052	ZZZ	Total	2021Q2	19019.0	1.0	Carbon dioxide equivalents	P
1053	ZZZ	Total	2021Q3	17788.0	1.0	Carbon dioxide equivalents	P
1054	ZZZ	Total	2021Q4	20312.0	1.0	Carbon dioxide equivalents	P
1055	ZZZ	Total	2022Q1	21462.0	1.0	Carbon dioxide equivalents	P

140 rows × 7 columns

Remove Duplicate

In [11]:

```
non_duplicate = df[~df.duplicated('Anzsic_description')]
non_duplicate
```

Out[11]:

	Anzsic	Anzsic_description	Period	Data_value	Variable	Gas	Status
330	ZZ9	Total industry	2014Q1	18508.0	0.0	Carbon dioxide equivalents	P
495	ZZZ	Total	2014Q1	20480.0	0.0	Carbon dioxide equivalents	P
528	ZPZ	Primary industries	2014Q1	12881.0	1.0	Carbon dioxide equivalents	P
581	AAZ	Agriculture, forestry, fishing	2019Q1	12685.0	1.0	Carbon dioxide equivalents	P

Factorize

In [12]:

```
df['Period'] = pd.factorize(df['Period'])[0]
df
```

Out[12]:

	Anzsic	Anzsic_description	Period	Data_value	Variable	Gas	Status
330	ZZ9	Total industry	0	18508.0	0.0	Carbon dioxide equivalents	P
331	ZZ9	Total industry	1	18644.0	0.0	Carbon dioxide equivalents	P
332	ZZ9	Total industry	2	18747.0	0.0	Carbon dioxide equivalents	P
333	ZZ9	Total industry	3	18889.0	0.0	Carbon dioxide equivalents	P
334	ZZ9	Total industry	4	18789.0	0.0	Carbon dioxide equivalents	P
...	...	...	...	...	...	...	...
1051	ZZZ	Total	28	22211.0	1.0	Carbon dioxide equivalents	P
1052	ZZZ	Total	29	19019.0	1.0	Carbon dioxide equivalents	P
1053	ZZZ	Total	30	17788.0	1.0	Carbon dioxide equivalents	P
1054	ZZZ	Total	31	20312.0	1.0	Carbon dioxide equivalents	P
1055	ZZZ	Total	32	21462.0	1.0	Carbon dioxide equivalents	P

140 rows × 7 columns

Normalization:

max abs scaling

In [13]:

```
df.head()
```

Out[13]:

	Anzsic	Anzsic_description	Period	Data_value	Variable	Gas	Status
330	ZZ9	Total industry	0	18508.0	0.0	Carbon dioxide equivalents	P
331	ZZ9	Total industry	1	18644.0	0.0	Carbon dioxide equivalents	P
332	ZZ9	Total industry	2	18747.0	0.0	Carbon dioxide equivalents	P
333	ZZ9	Total industry	3	18889.0	0.0	Carbon dioxide equivalents	P
334	ZZ9	Total industry	4	18789.0	0.0	Carbon dioxide equivalents	P

In [14]:

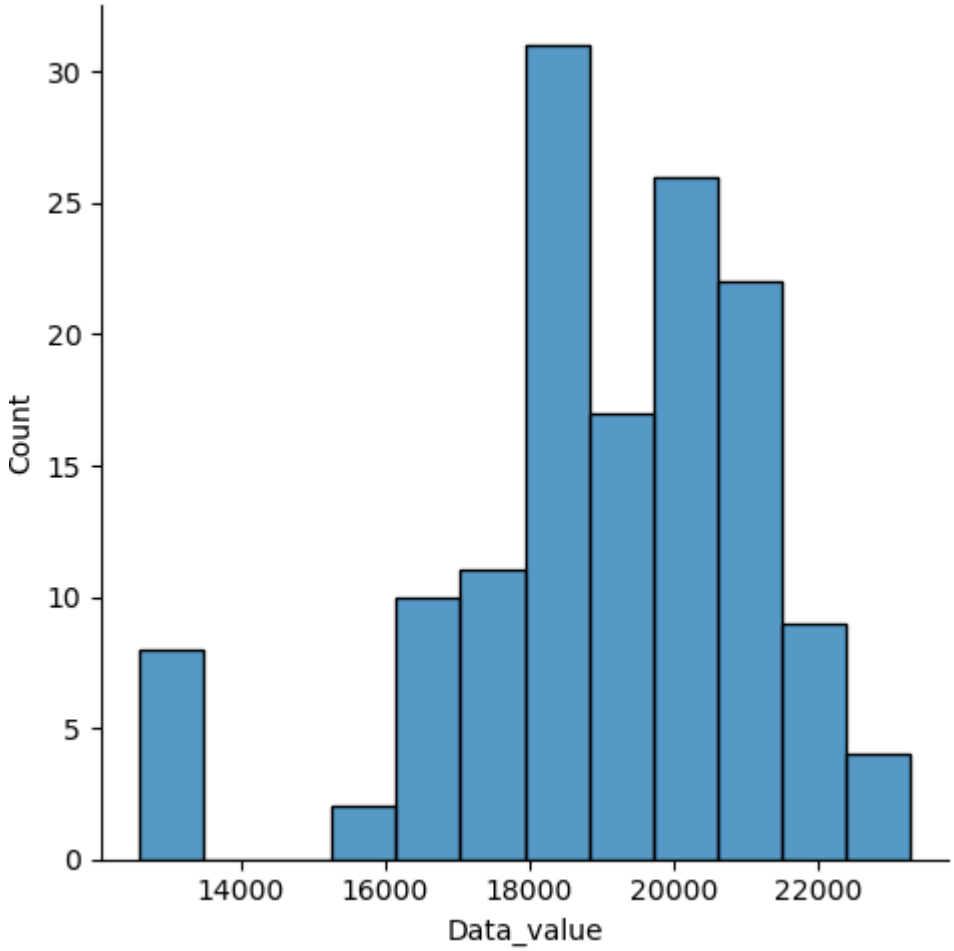
```
df.describe()
```

Out[14]:

	Period	Data_value	Variable
count	140.000000	140.000000	140.000000
mean	16.171429	19055.778571	0.528571
std	9.642489	2216.007130	0.500975
min	0.000000	12598.000000	0.000000
25%	8.000000	18141.250000	0.000000
50%	16.500000	19117.500000	1.000000
75%	24.000000	20587.000000	1.000000
max	32.000000	23277.000000	1.000000

In [15]:

```
sns.displot(df['Data_value'])
plt.show()
```



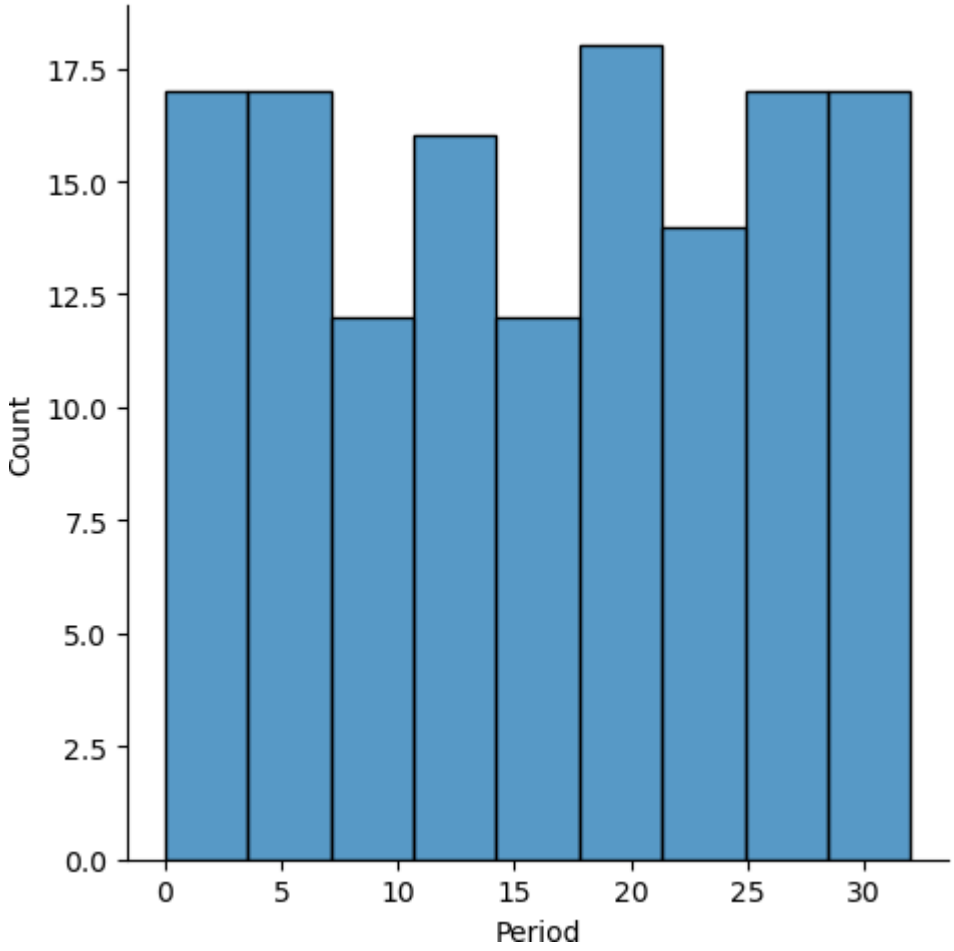
In [16]:

```
df_copy=df.copy()
df_copy['Data_value']=df_copy['Data_value']/df_copy['Data_value'].abs().max()
```

Min-max Scaling

In [17]:

```
sns.displot(df_copy['Period'])
plt.show()
```

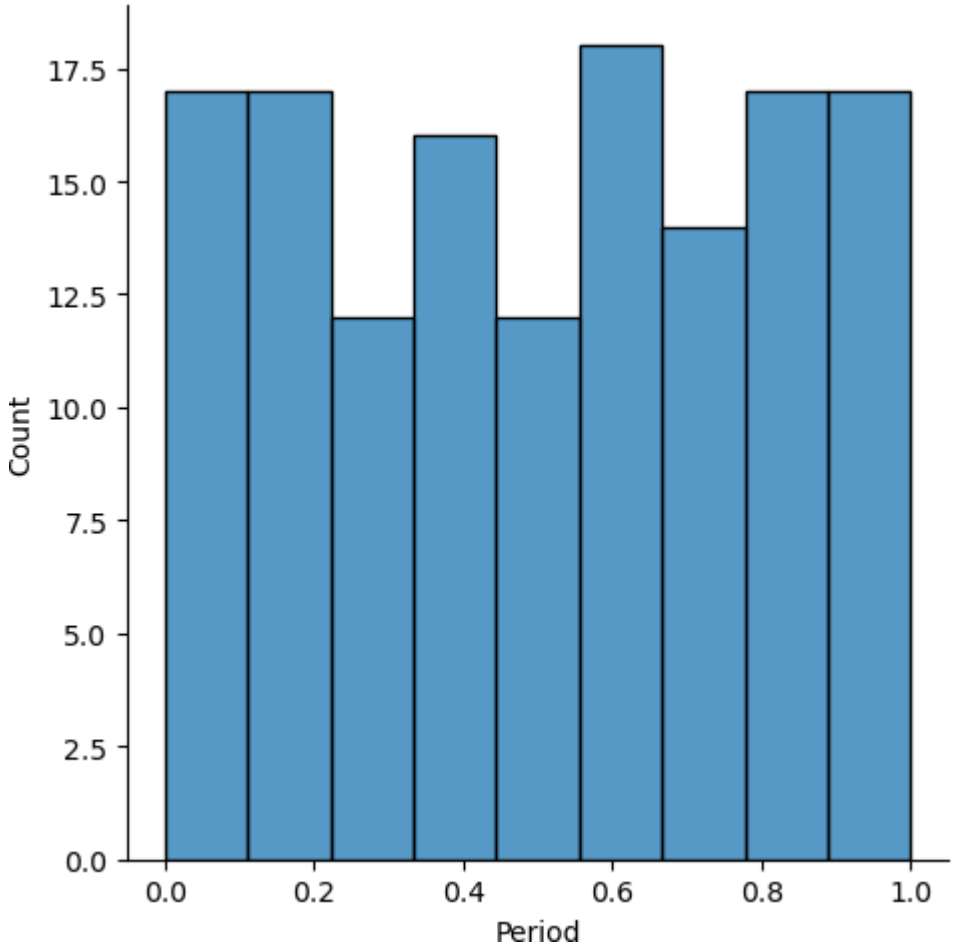


In [18]:

```
df_copy=df.copy()
df_copy['Period']=(df_copy['Period'] - df_copy['Period'].min())/(df_copy['Period'].max()
```

In [19]:

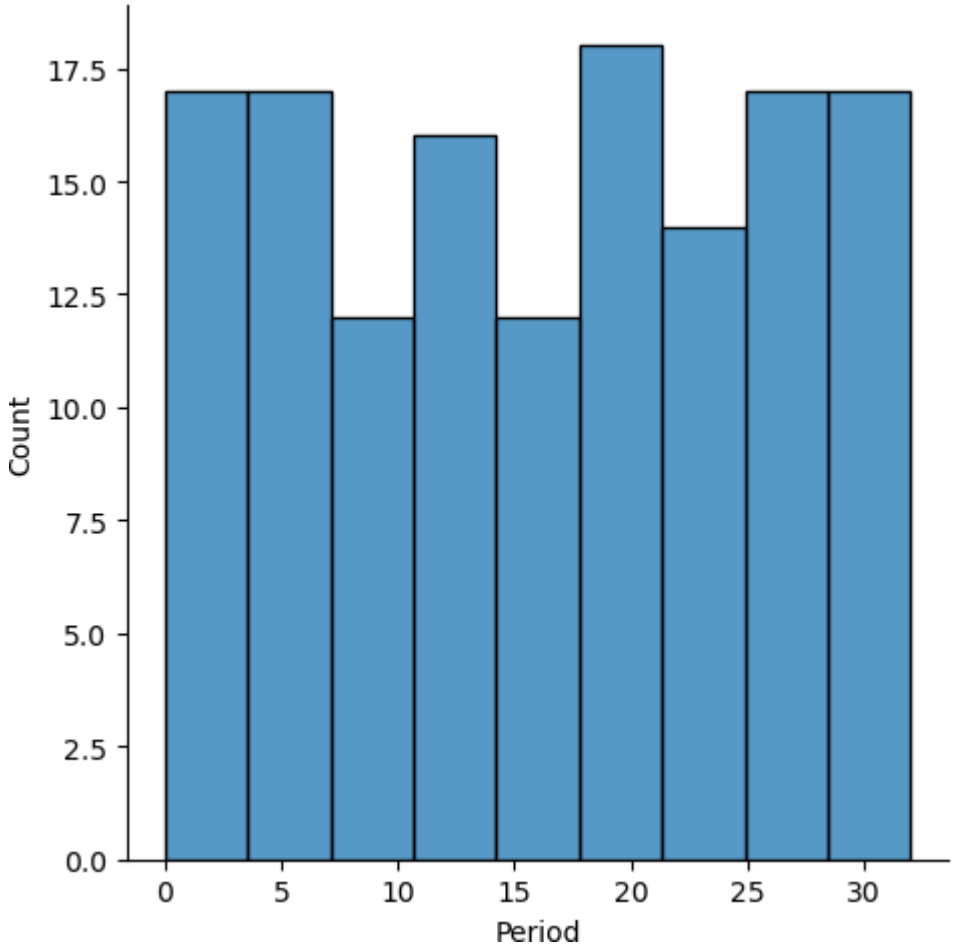
```
sns.displot(df_copy['Period'])
plt.show()
```



Z-index Scaling

In [20]:

```
sns.displot(df['Period'])
plt.show()
```



In [21]:

```
z_copy=df.copy()
for col in ['Period']:
    z_copy[col]=(z_copy[col]-z_copy[col].mean())/z_copy[col].std()
```

In [22]:

```
sns.displot(z_copy['Period'])
plt.show()
```