In [126]:

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("1rec-crime-pfa.csv", parse_dates=["12 months ending"])
df
```

Out[126]:

	12 months ending	PFA	Region	Offence	Rolling year total number of offences
0	2003-03-31	Avon and Somerset	South West	All other theft offences	25959
1	2003-03-31	Avon and Somerset	South West	Bicycle theft	3090
2	2003-03-31	Avon and Somerset	South West	Criminal damage and arson	26202
3	2003-03-31	Avon and Somerset	South West	Death or serious injury caused by illegal driving	2
4	2003-03-31	Avon and Somerset	South West	Domestic burglary	14561
...	...	...	...	...	...
46464	2018-12-31	Wiltshire	South West	Stalking and harassment	2380
46465	2018-12-31	Wiltshire	South West	Theft from the person	347
46466	2018-12-31	Wiltshire	South West	Vehicle offences	2895
46467	2018-12-31	Wiltshire	South West	Violence with injury	5701
46468	2018-12-31	Wiltshire	South West	Violence without injury	5840

46469 rows × 5 columns

In [127]:

```
offences_stat = {"min": df["Rolling year total number of offences"].min(),
            "max": df["Rolling year total number of offences"].max(),
            "mean": df["Rolling year total number of offences"].mean(),
            "median": df["Rolling year total number of offences"].median(),
            "mode": df["Rolling year total number of offences"].mode().to_list(),
            "var": df["Rolling year total number of offences"].var(ddof=0),
            "std": df["Rolling year total number of offences"].std(ddof=0),
            "range": df["Rolling year total number of offences"].max() - df["Rolling year total number of
            "interquartile_range": df["Rolling year total number of offences"].quantile(0.75) - df["Rolli
            "skew": df["Rolling year total number of offences"].skew()
            }
offences_stat
```

Out[127]:

```
{'min': -53,
 'max': 308901,
 'mean': 5266.331705007639,
 'median': 2011.0,
 'mode': [0],
 'var': 166811319.2283264,
 'std': 12915.545641912555,
 'range': 308954,
 'interquartile_range': 5051.0,
 'skew': 11.135052591328055}
```

Видим отрицательные значения, но количество преступлений не может быть отрицательным. Скорее всего, это опечатка, возьмём значения по модулю.

In [128]:

```
df["Rolling year total number of offences"] = df["Rolling year total number of offences"].abs()
```

In [153]:

```
df["Rolling year total number of offences"].min(),
```

Out[153]:

```
(0,)
```

In [130]:

```
df['Region'].unique()
```

Out[130]:

```
array(['South West', 'East', 'British Transport Police', 'North West',
       'London', 'North East', 'East Midlands', 'Wales', 'South East',
       'Yorkshire and The Humber', 'West Midlands', 'Fraud: Action Fraud',
       'Fraud: CIFAS', 'Fraud: UK Finance'], dtype=object)
```

In [131]:

```
cols = list(df.columns)
nom_cols_data = [{name: df[col].to_list().count(name) for name in df[col].unique()}
                 for col in cols
                 if df[col].dtype == "object"]
nom_cols_data
```

Out[131]:


```
[{'Avon and Somerset': 1054,
  'Bedfordshire': 1054,
  'British Transport Police': 1054,
  'Cambridgeshire': 1054,
  'Cheshire': 1054,
  'City of London': 1054,
  'Cleveland': 1054,
  'Cumbria': 1054,
  'Derbyshire': 1054,
  'Devon and Cornwall': 1054,
  'Dorset': 1054,
  'Durham': 1054,
  'Dyfed-Powys': 1054,
  'Essex': 1054,
  'Gloucestershire': 1054,
  'Greater Manchester': 1054,
  'Gwent': 1054,
  'Hampshire': 1054,
  'Hertfordshire': 1054,
  'Humberside': 1054,
  'Kent': 1054,
  'Lancashire': 1054,
  'Leicestershire': 1054,
  'Lincolnshire': 1054,
  'Merseyside': 1054,
  'Metropolitan Police': 1054,
  'Norfolk': 1054,
  'North Wales': 1054,
  'North Yorkshire': 1054,
  'Northamptonshire': 1054,
  'Northumbria': 1054,
  'Nottinghamshire': 1054,
  'South Wales': 1054,
  'South Yorkshire': 1054,
  'Staffordshire': 1054,
  'Suffolk': 1054,
  'Surrey': 1054,
  'Sussex': 1054,
  'Thames Valley': 1054,
  'Warwickshire': 1054,
  'West Mercia': 1054,
  'West Midlands': 1054,
  'West Yorkshire': 1054,
  'Wiltshire': 1054,
  'Action Fraud': 31,
  'CIFAS': 31,
  'UK Finance': 31},
 {'South West': 5270,
  'East': 6324,
  'British Transport Police': 1054,
  'North West': 5270,
  'London': 2108,
  'North East': 3162,
  'East Midlands': 5270,
  'Wales': 4216,
  'South East': 5270,
  'Yorkshire and The Humber': 4216,
  'West Midlands': 4216,
  'Fraud: Action Fraud': 31,
  'Fraud: CIFAS': 31,
  'Fraud: UK Finance': 31},
 {'All other theft offences': 2288,
  'Bicycle theft': 2288,
```

```
 'Criminal damage and arson': 2288,
 'Death or serious injury caused by illegal driving': 2288,
 'Domestic burglary': 2288,
 'Drug offences': 2288,
 'Fraud offences': 1408,
 'Homicide': 2288,
 'Miscellaneous crimes against society': 2288,
 'Non-domestic burglary': 2288,
 'Possession of weapons offences': 2288,
 'Public order offences': 2288,
 'Robbery': 2288,
 'Sexual offences': 2288,
 'Shoplifting': 2288,
 'Stalking and harassment': 2288,
 'Theft from the person': 2288,
 'Vehicle offences': 2288,
 'Violence with injury': 2288,
 'Violence without injury': 2288,
 'Action Fraud': 31,
 'CIFAS': 31,
 'UK Finance': 31,
 'Non-residential burglary': 748,
 'Residential burglary': 748}]
```

Видим странные данные, которые встречаются лишь в 31 строчке, тогда как другие исчисляются тысячами. А так же они записаны во всех колонках одинакого. Посмотрим на них внимательнее

In [132]:

```
same_df = df.loc[df["PFA"] == df["Offence"]]
same_df
```

Out[132]:

	12 months ending	PFA	Region	Offence	Rolling year total number of offences
19360	2011-06-30	Action Fraud	Fraud: Action Fraud	Action Fraud	8140
19361	2011-06-30	CIFAS	Fraud: CIFAS	CIFAS	52334
19362	2011-06-30	UK Finance	Fraud: UK Finance	UK Finance	34266
20243	2011-09-30	Action Fraud	Fraud: Action Fraud	Action Fraud	19613
20244	2011-09-30	CIFAS	Fraud: CIFAS	CIFAS	109192
...	...	...	...	...	...
44721	2018-09-30	CIFAS	Fraud: CIFAS	CIFAS	279613
45436	2018-09-30	UK Finance	Fraud: UK Finance	UK Finance	74889
45542	2018-12-31	Action Fraud	Fraud: Action Fraud	Action Fraud	306126
45648	2018-12-31	CIFAS	Fraud: CIFAS	CIFAS	296896
46363	2018-12-31	UK Finance	Fraud: UK Finance	UK Finance	72930

93 rows × 5 columns

Видим очень странные данные, которые начинаются лишь с 2011 года, хотя датасет ведётся с 2003 г. Во всех трёх столбцах по-сути записано одно и то же. Погуглив, узнаём, что Action Fraud - это национальное бюро по расследованию случаев мошенничества - это полицейское подразделение в Соединенном Королевстве, которое занимается сбором и анализом разведывательных данных, касающихся мошенничества и киберпреступлений, мотивированных в финансовом отношении. CIFAS - это служба предотвращения мошенничества в Великобритании. Это некоммерческая членская ассоциация, представляющая организации из государственного, частного и добровольного секторов. UK Finance - это торговая ассоциация для сектора банковских и финансовых услуг Великобритании, образованная 1 июля 2017 года. Она представляет около 300 фирм в Великобритании, предоставляющих кредитные, банковские, рыночные и платежные услуги. По всей видимости, это организации, составляющие свою статистику по преступлению мошенничество. Очень не удобных данные, особенно не понятна ситуация с UK Finance, потому что основана на в 2017 году, а данные о преступлениях есть с 2011.

In [134]:

```
df = df.loc[(df["PFA"] != "Action Fraud") & (df["PFA"] != "CIFAS") & (df["PFA"] != "UK Finance")]
cols = list(df.columns)
nom_cols_data = [{name: df[col].to_list().count(name) for name in df[col].unique()}
                 for col in cols
                 if df[col].dtype == "object"]
nom_cols_data
```

Out[134]:


```
[{'Avon and Somerset': 1054,
  'Bedfordshire': 1054,
  'British Transport Police': 1054,
  'Cambridgeshire': 1054,
  'Cheshire': 1054,
  'City of London': 1054,
  'Cleveland': 1054,
  'Cumbria': 1054,
  'Derbyshire': 1054,
  'Devon and Cornwall': 1054,
  'Dorset': 1054,
  'Durham': 1054,
  'Dyfed-Powys': 1054,
  'Essex': 1054,
  'Gloucestershire': 1054,
  'Greater Manchester': 1054,
  'Gwent': 1054,
  'Hampshire': 1054,
  'Hertfordshire': 1054,
  'Humberside': 1054,
  'Kent': 1054,
  'Lancashire': 1054,
  'Leicestershire': 1054,
  'Lincolnshire': 1054,
  'Merseyside': 1054,
  'Metropolitan Police': 1054,
  'Norfolk': 1054,
  'North Wales': 1054,
  'North Yorkshire': 1054,
  'Northamptonshire': 1054,
  'Northumbria': 1054,
  'Nottinghamshire': 1054,
  'South Wales': 1054,
  'South Yorkshire': 1054,
  'Staffordshire': 1054,
  'Suffolk': 1054,
  'Surrey': 1054,
  'Sussex': 1054,
  'Thames Valley': 1054,
  'Warwickshire': 1054,
  'West Mercia': 1054,
  'West Midlands': 1054,
  'West Yorkshire': 1054,
  'Wiltshire': 1054},
 {'South West': 5270,
  'East': 6324,
  'British Transport Police': 1054,
  'North West': 5270,
  'London': 2108,
  'North East': 3162,
  'East Midlands': 5270,
  'Wales': 4216,
  'South East': 5270,
  'Yorkshire and The Humber': 4216,
  'West Midlands': 4216},
 {'All other theft offences': 2288,
  'Bicycle theft': 2288,
  'Criminal damage and arson': 2288,
  'Death or serious injury caused by illegal driving': 2288,
  'Domestic burglary': 2288,
  'Drug offences': 2288,
  'Fraud offences': 1408,
  'Homicide': 2288,
  'Miscellaneous crimes against society': 2288,
  'Non-domestic burglary': 2288,
  'Possession of weapons offences': 2288,
  'Public order offences': 2288,
  'Robbery': 2288,
  'Sexual offences': 2288,
  'Shoplifting': 2288,
  'Stalking and harassment': 2288,
  'Theft from the person': 2288,
  'Vehicle offences': 2288,
  'Violence with injury': 2288,
  'Violence without injury': 2288,
```

```
 'Non-residential burglary': 748,
 'Residential burglary': 748}]
```

In [135]:

```
df
```

Out[135]:

	12 months ending	PFA	Region	Offence	Rolling year total number of offences
0	2003-03-31	Avon and Somerset	South West	All other theft offences	25959
1	2003-03-31	Avon and Somerset	South West	Bicycle theft	3090
2	2003-03-31	Avon and Somerset	South West	Criminal damage and arson	26202
3	2003-03-31	Avon and Somerset	South West	Death or serious injury caused by illegal driving	2
4	2003-03-31	Avon and Somerset	South West	Domestic burglary	14561
...	...	...	...	...	...
46464	2018-12-31	Wiltshire	South West	Stalking and harassment	2380
46465	2018-12-31	Wiltshire	South West	Theft from the person	347
46466	2018-12-31	Wiltshire	South West	Vehicle offences	2895
46467	2018-12-31	Wiltshire	South West	Violence with injury	5701
46468	2018-12-31	Wiltshire	South West	Violence without injury	5840

46376 rows × 5 columns

In [137]:

```
dfch = pd.read_excel("population.xlsx", parse_dates=["12 months ending"])
dfch
```

Out[137]:

	12 months ending	Region	Population
0	2003-03-31	South West	4991000.0
1	2004-03-31	South West	5022000.0
2	2005-03-31	South West	5062000.0
3	2006-03-31	South West	5103000.0
4	2007-03-31	South West	5145000.0
...	...	...	...
515	2017-12-31	West Midlands	5836003.0
516	2018-03-31	West Midlands	5873000.0
517	2018-06-30	West Midlands	5873001.0
518	2018-09-30	West Midlands	5873002.0
519	2018-12-31	West Midlands	5873003.0

520 rows × 3 columns

В эту таблицу не попал регион "Британская транспортная полиция", потому что, это, собственно, и не регион. Придётся также избавиться от этих данных.

In [138]:

```
df = pd.merge(df, dfch, on=("Region", "12 months ending"))
df
```

Out[138]:

	12 months ending	PFA	Region	Offence	Rolling year total number of offences	Population
0	2003-03-31	Avon and Somerset	South West	All other theft offences	25959	4991000.0
1	2003-03-31	Avon and Somerset	South West	Bicycle theft	3090	4991000.0
2	2003-03-31	Avon and Somerset	South West	Criminal damage and arson	26202	4991000.0
3	2003-03-31	Avon and Somerset	South West	Death or serious injury caused by illegal driving	2	4991000.0
4	2003-03-31	Avon and Somerset	South West	Domestic burglary	14561	4991000.0
...	...	...	...	...	...	...
44414	2018-12-31	West Midlands	West Midlands	Stalking and harassment	15002	5873003.0
44415	2018-12-31	West Midlands	West Midlands	Theft from the person	3230	5873003.0
44416	2018-12-31	West Midlands	West Midlands	Vehicle offences	37250	5873003.0
44417	2018-12-31	West Midlands	West Midlands	Violence with injury	30561	5873003.0
44418	2018-12-31	West Midlands	West Midlands	Violence without injury	24861	5873003.0

44419 rows × 6 columns

In [139]:

```
df['Region'].unique()
```

Out[139]:

```
array(['South West', 'East', 'North West', 'London', 'North East',
       'East Midlands', 'Wales', 'South East', 'Yorkshire and The Humber',
       'West Midlands'], dtype=object)
```

In [149]:

```
df1 = df.copy()
```

In [150]:

```
df1["Rolling year total number of offences"] = df1["Rolling year total number of offences"]/df1["Populati
df1
```

Out[150]:

	12 months ending	PFA	Region	Offence	Rolling year total number of offences	Population
0	2003-03-31	Avon and Somerset	South West	All other theft offences	5.201162	4991000.0
1	2003-03-31	Avon and Somerset	South West	Bicycle theft	0.619114	4991000.0
2	2003-03-31	Avon and Somerset	South West	Criminal damage and arson	5.249850	4991000.0
3	2003-03-31	Avon and Somerset	South West	Death or serious injury caused by illegal driving	0.000401	4991000.0
4	2003-03-31	Avon and Somerset	South West	Domestic burglary	2.917451	4991000.0
...	...	...	...	...	...	...
44414	2018-12-31	West Midlands	West Midlands	Stalking and harassment	2.554400	5873003.0
44415	2018-12-31	West Midlands	West Midlands	Theft from the person	0.549974	5873003.0
44416	2018-12-31	West Midlands	West Midlands	Vehicle offences	6.342581	5873003.0
44417	2018-12-31	West Midlands	West Midlands	Violence with injury	5.203641	5873003.0
44418	2018-12-31	West Midlands	West Midlands	Violence without injury	4.233098	5873003.0

44419 rows × 6 columns

In [ ]: