Reading and understanding the data

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import folium
import os, re
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import normalize
from IPython.display import IFrame
from sklearn.cluster import AgglomerativeClustering
import scipy.cluster.hierarchy as shc

%matplotlib inline

import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
path_to_data = './crime'
cd = os.path.dirname(os.path.abspath(path_to_data))

i = 0
columns = range(1,100)
dfList = []

for root, dirs, files in os.walk(cd):
    for fname in files:
        if re.match("^.*.csv$", fname):
            frame = pd.read_csv(os.path.join(root, fname))
            frame['key'] = "file{}".format(i)
            dfList.append(frame)
            i += 1

dataset = pd.concat(dfList)
```

In [3]:
```
dataset.head()
```

Out[3]:

	Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	Crime type
0	NaN	2021-06	Essex Police	Essex Police	0.864094	51.971811	On or near Bear Street	E01029906	Babergh 009D	Anti-social behaviour
1	NaN	2021-06	Essex Police	Essex Police	0.436057	51.639952	On or near Coach Mews	E01021237	Basildon 001A	Anti-social behaviour
2	d91fddaaae8b0664cf330fc1a85bfdcddc57256d0bd2b3...	2021-06	Essex Police	Essex Police	0.437217	51.642455	On or near Bridleway	E01021238	Basildon 001B	Vehicle crime
3	f5104dc9cd4aaa31f162b0bed7b7f7714f0bdf266fa388...	2021-06	Essex Police	Essex Police	0.435880	51.643391	On or near Penwood Close	E01021238	Basildon 001B	Violence and sexual offences
4	faa6b0a7146e1e2816512d2f2505d98c384451518f3935...	2021-06	Essex Police	Essex Police	0.435880	51.643391	On or near Penwood Close	E01021238	Basildon 001B	Violence and sexual offences

In [4]:
```
print(dataset.shape)
```

```
(4019944, 13)
```

In [5]:
```
name_number = 'PreciousAdaugoReginald1-2325671.csv'
dataset.to_csv(name_number, index=False)
```

In [6]:
```
data = pd.read_csv(name_number)
```

In [7]:
```
data['Crime type'].value_counts()
```

Out[7]:
```
Violence and sexual offences    1642341
Anti-social behaviour            581039
Public order                     373768
Criminal damage and arson        315248
Other theft                      249128
Vehicle crime                    243219
Shoplifting                      180652
Burglary                         138510
Drugs                            106552
Other crime                       70927
Bicycle theft                     32775
Possession of weapons             29659
Robbery                           29241
Theft from the person             26885
Name: Crime type, dtype: int64
```

In [8]:
```
data['Month'].value_counts()
```

Out[8]:
```
2021-07    349353
2021-06    345914
2022-03    324881
2022-05    321613
2021-08    318269
2021-09    315571
2021-10    310156
2022-06    306299
2021-11    299060
2022-04    295070
2022-01    287375
2021-12    281485
2022-02    264898
Name: Month, dtype: int64
```

In [9]:
```
data['town'] = data['LSOA name'].str.split(' ').str[0]
```

In [10]:
```
data.head()
```

Out[10]:

	Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	Crime type
0	NaN	2021-06	Essex Police	Essex Police	0.864094	51.971811	On or near Bear Street	E01029906	Babergh 009D	Anti-social behaviour
1	NaN	2021-06	Essex Police	Essex Police	0.436057	51.639952	On or near Coach Mews	E01021237	Basildon 001A	Anti-social behaviour
2	d91fddaaae8b0664cf330fc1a85bfdcddc57256d0bd2b3...	2021-06	Essex Police	Essex Police	0.437217	51.642455	On or near Bridleway	E01021238	Basildon 001B	Vehicle crime
3	f5104dc9cd4aaa31f162b0bed7b7f7714f0bdf266fa388...	2021-06	Essex Police	Essex Police	0.435880	51.643391	On or near Penwood Close	E01021238	Basildon 001B	Violence and sexual offences
4	faa6b0a7146e1e2816512d2f2505d98c384451518f3935...	2021-06	Essex Police	Essex Police	0.435880	51.643391	On or near Penwood Close	E01021238	Basildon 001B	Violence and sexual offences

In [11]:
```
towns = ['Chelmsford']
filtered_data = data[data.town.str.contains('|'.join(towns), na=False)]
filtered_data.head()
```

Out[11]:

	Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	
4807	NaN	2021-06	Essex Police	Essex Police	0.497521	51.818432	On or near The Crescent	E01021538	Chelmsford 001A	beh
4808	NaN	2021-06	Essex Police	Essex Police	0.508854	51.832013	On or near Shimbrooks	E01021538	Chelmsford 001A	beh
4809	NaN	2021-06	Essex Police	Essex Police	0.509951	51.824076	On or near Catherines Close	E01021538	Chelmsford 001A	beh
4810	NaN	2021-06	Essex Police	Essex Police	0.509951	51.824076	On or near Catherines Close	E01021538	Chelmsford 001A	beh
4811	4595f85a0c9b5060cddc75414a58e6345b77b6a9b260f1...	2021-06	Essex Police	Essex Police	0.504922	51.828374	On or near Old Moors	E01021538	Chelmsford 001A	

Q2 answer

In [12]:
```
filtered_data['Crime type'].value_counts()
```

Out[12]:
```
Violence and sexual offences    150100
Anti-social behaviour            49666
Public order                     32927
Criminal damage and arson        26144
Other theft                      25992
Vehicle crime                    21071
Shoplifting                      19817
Burglary                         16074
Drugs                             9291
Other crime                       7809
Bicycle theft                     7676
Theft from the person             4560
Possession of weapons             2299
Robbery                           2147
Name: Crime type, dtype: int64
```

Q3 answer

The most common type of crime commited in Chelmsford is violence and sexual offences, this shows a count of 7900. The most commited crime in the Essex area is the same and it shows a count of 86439

In [13]:
```
filtered_data['LSOA code'].value_counts().nlargest(10)
```

Out[13]:
```
E01033141    41097
E01033140    31768
E01021574    14744
E01021542    10336
E01033138     7904
E01021540     7429
E01021573     6650
E01033144     6384
E01021613     6118
E01021631     6023
Name: LSOA code, dtype: int64
```

Q4 answer

The first code selected is E01033141, which is the code that clo tains the areas with the most crime rates. When the map is observed for this code, it has been seen that the active areas are Burgess Sprinngs, Park Road, and Victoria Rd S. The second LS0A code that has been chosen is E01021631, when looking at the map there are only active areas (shown in green), those are Exmoor Close and Sheerwood Dr

Preparing the data for clustering

Columns selection

In [14]:
```
filtered_important_data = filtered_data[['LSOA code','Crime type']]
filtered_important_data = pd.get_dummies(filtered_important_data, columns=['Crime type'])
clustering_data = filtered_important_data.groupby(['LSOA code']).agg(
 {'Crime type_Anti-social behaviour':'sum',
 'Crime type_Bicycle theft':'sum',
 'Crime type_Burglary':'sum',
 'Crime type_Criminal damage and arson':'sum',
 'Crime type_Drugs':'sum',
 'Crime type_Other crime':'sum',
 'Crime type_Other theft':'sum',
 'Crime type_Possession of weapons':'sum',
 'Crime type_Public order':'sum',
 'Crime type_Robbery':'sum',
 'Crime type_Shoplifting':'sum',
 'Crime type_Theft from the person':'sum',
 'Crime type_Vehicle crime':'sum',
 'Crime type_Violence and sexual offences':'sum'
 }
 ).reset_index()
```

In [15]:
```python
clustering_data[:5]
```

Out[15]:

	LSOA code	Crime type_Anti-social behaviour	Crime type_Bicycle theft	Crime type_Burglary	Crime type_Criminal damage and arson	Crime type_Drugs	Crime type_Other crime	Crime type_Other theft	Crime type_Possession of weapons	Crime type_Public order	type_Ro
0	E01021533	228.0	0.0	152.0	133.0	0.0	19.0	76.0	0	171.0	
1	E01021535	247.0	0.0	171.0	190.0	57.0	19.0	133.0	0	133.0	
2	E01021536	76.0	0.0	57.0	285.0	38.0	19.0	133.0	38	152.0	
3	E01021537	855.0	19.0	228.0	266.0	190.0	114.0	361.0	0	513.0	
4	E01021538	646.0	0.0	114.0	323.0	19.0	38.0	285.0	38	418.0	

In [16]:
```python
clustering_data_original = clustering_data.copy()
clustering_data_original.head()
```

Out[16]:

	LSOA code	Crime type_Anti-social behaviour	Crime type_Bicycle theft	Crime type_Burglary	Crime type_Criminal damage and arson	Crime type_Drugs	Crime type_Other crime	Crime type_Other theft	Crime type_Possession of weapons	Crime type_Public order	type_Ro
0	E01021533	228.0	0.0	152.0	133.0	0.0	19.0	76.0	0	171.0	
1	E01021535	247.0	0.0	171.0	190.0	57.0	19.0	133.0	0	133.0	
2	E01021536	76.0	0.0	57.0	285.0	38.0	19.0	133.0	38	152.0	
3	E01021537	855.0	19.0	228.0	266.0	190.0	114.0	361.0	0	513.0	
4	E01021538	646.0	0.0	114.0	323.0	19.0	38.0	285.0	38	418.0	

In [17]:
```python
clustering_data.drop(['LSOA code'], axis = 1, inplace = True, errors = 'ignore')
clustering_data.head()
```

Out[17]:

	Crime type_Anti-social behaviour	Crime type_Bicycle theft	Crime type_Burglary	Crime type_Criminal damage and arson	Crime type_Drugs	Crime type_Other crime	Crime type_Other theft	Crime type_Possession of weapons	Crime type_Public order	Crime type_Robbery	type_
0	228.0	0.0	152.0	133.0	0.0	19.0	76.0	0	171.0	0.0	
1	247.0	0.0	171.0	190.0	57.0	19.0	133.0	0	133.0	38.0	
2	76.0	0.0	57.0	285.0	38.0	19.0	133.0	38	152.0	0.0	
3	855.0	19.0	228.0	266.0	190.0	114.0	361.0	0	513.0	0.0	
4	646.0	0.0	114.0	323.0	19.0	38.0	285.0	38	418.0	0.0	

Normalization

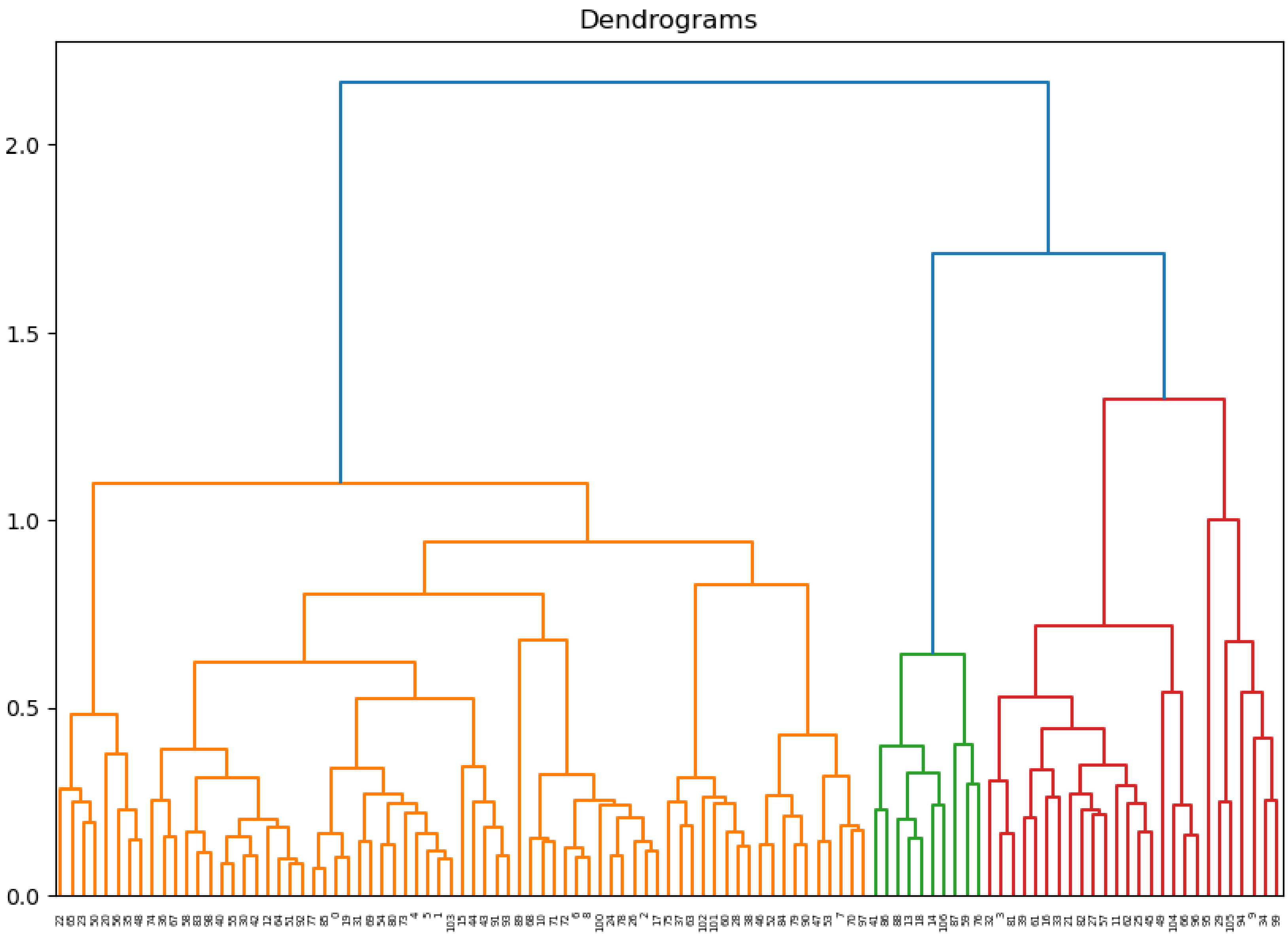
In [18]:
```python
data_scaled = normalize(clustering_data)
data_scaled = pd.DataFrame(data_scaled, columns=clustering_data.columns)
data_scaled.head()
```

Out[18]:

	Crime type_Anti-social behaviour	Crime type_Bicycle theft	Crime type_Burglary	Crime type_Criminal damage and arson	Crime type_Drugs	Crime type_Other crime	Crime type_Other theft	Crime type_Possession of weapons	Crime type_Public order	Crime type_Robbery	type_
0	0.160071	0.000000	0.106714	0.093375	0.000000	0.013339	0.053357	0.000000	0.120053	0.000000	
1	0.230350	0.000000	0.159473	0.177192	0.053158	0.017719	0.124035	0.000000	0.124035	0.035438	
2	0.052559	0.000000	0.039419	0.197096	0.026279	0.013140	0.091978	0.026279	0.105118	0.000000	
3	0.349721	0.007772	0.093259	0.108802	0.077716	0.046629	0.147660	0.000000	0.209833	0.000000	
4	0.278876	0.000000	0.049213	0.139438	0.008202	0.016404	0.123034	0.016404	0.180449	0.000000	

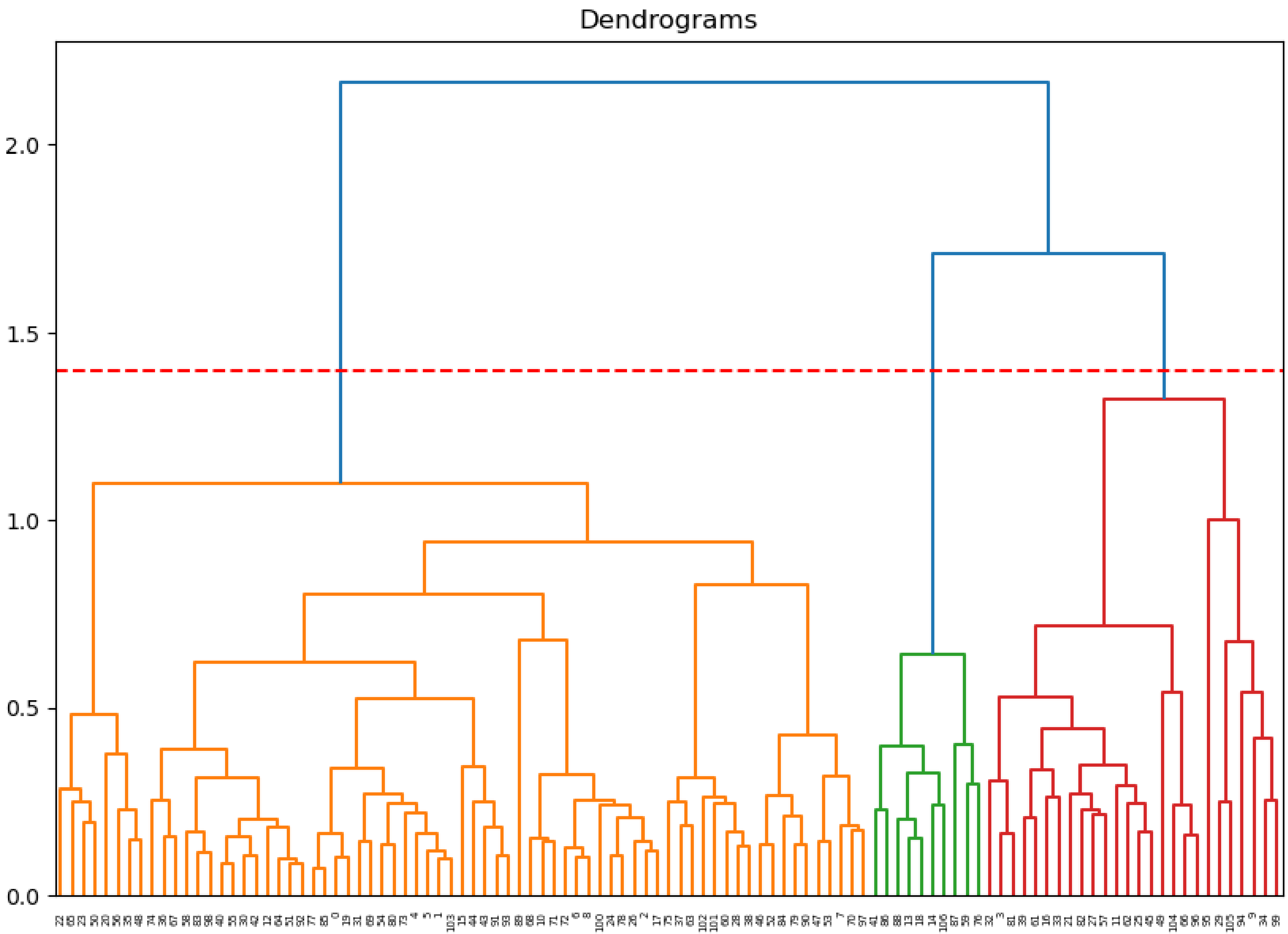
Determining number of clusters using dendograms

In [19]:
```
plt.figure(figsize=(10, 7))
plt.title("Dendrograms")
dend = shc.dendrogram(shc.linkage(data_scaled, method='ward'))
```


Dendrograms
2.0
1.5
1.0
0.5
0.0

In [20]:
```
plt.figure(figsize=(10, 7))
plt.title("Dendrograms")
dend = shc.dendrogram(shc.linkage(data_scaled, method='ward'))
plt.axhline(y=1.40, color='r', linestyle='--')
```

Out[20]:
```
<matplotlib.lines.Line2D at 0x22f61715dc0>
```


Dendrograms
2.0
1.5
1.0
0.5
0.0

Q5 answer

When the dendogram is being cut in a different level, the number of k(klusters) will change, changing then the outcome of the dataset

Agglomerative clustering

In [21]:
```
cluster = AgglomerativeClustering(n_clusters=3, affinity='euclidean', linkage='ward')
cluster_ids = cluster.fit_predict(data_scaled)
```

In [22]:
```
clustering_data['cluster'] = cluster_ids
clustering_data.head()
```

Out[22]:

	Crime type_Anti-social behaviour	Crime type_Bicycle theft	Crime type_Burglary	Crime type_Criminal damage and arson	Crime type_Drugs	Crime type_Other crime	Crime type_Other theft	Crime type_Possession of weapons	Crime type_Public order	Crime type_Robbery	type_
0	228.0	0.0	152.0	133.0	0.0	19.0	76.0	0	171.0	0.0	
1	247.0	0.0	171.0	190.0	57.0	19.0	133.0	0	133.0	38.0	
2	76.0	0.0	57.0	285.0	38.0	19.0	133.0	38	152.0	0.0	
3	855.0	19.0	228.0	266.0	190.0	114.0	361.0	0	513.0	0.0	
4	646.0	0.0	114.0	323.0	19.0	38.0	285.0	38	418.0	0.0	

In [23]:
```
hierarchical_cluster = pd.DataFrame(round(clustering_data.groupby('cluster').mean(),1))
hierarchical_cluster
```

Out[23]:

Crime rglary	Crime type_Criminal damage and arson	Crime type_Drugs	Crime type_Other crime	Crime type_Other theft	Crime type_Possession of weapons	Crime type_Public order	Crime type_Robbery	Crime type_Shoplifting	Crime type_Theft from the person	Crime type_Vehicle crime
176.1	216.3	68.7	51.9	344.2	13.2	229.5	10.2	104.5	45.3	271.1
144.2	263.6	97.1	85.1	208.5	23.8	345.5	22.5	155.2	46.6	177.2
125.4	180.5	60.8	41.8	224.2	26.6	243.2	28.5	608.0	7.6	144.4

Q6 Answer

Based on my dataset a set of conlcusions can be figured out. Cluster ID 1, contains the LSOA codes with the highest crimes, therefore the post codes of those areas are of high risk. Therefore it is not adviced to live in such locations. Cluster ID2 is the one that contains the LSOA codes with the lowest number of crimes. So there are low risk areas. Where as cluster ID 0 is the one that contains LSOSA codes that show moderate risk areas.

Visualising clusters

A

In [24]:
```
clustering_data_original['cluster'] = cluster_ids
clusters = clustering_data_original[['LSOA code', 'cluster']]
```

In [25]:
```
clusters.head()
```

Out[25]:

	LSOA code	cluster
0	E01021533	1
1	E01021535	1
2	E01021536	1
3	E01021537	0
4	E01021538	1

In [26]:
```
clusters.shape
```

Out[26]:
```
(107, 2)
```

In [27]:
```
clustered_full = pd.merge(filtered_data, clusters, on='LSOA code')
clustered_full.head()
```

Out[27]:

	Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	Crir ty
0	NaN	2021-06	Essex Police	Essex Police	0.497521	51.818432	On or near The Crescent	E01021538	Chelmsford 001A	Ar soc behavic
1	NaN	2021-06	Essex Police	Essex Police	0.508854	51.832013	On or near Shimbrooks	E01021538	Chelmsford 001A	Ar soc behavic
2	NaN	2021-06	Essex Police	Essex Police	0.509951	51.824076	On or near Catherines Close	E01021538	Chelmsford 001A	Ar soc behavic
3	NaN	2021-06	Essex Police	Essex Police	0.509951	51.824076	On or near Catherines Close	E01021538	Chelmsford 001A	Ar soc behavic
4	4595f85a0c9b5060cddc75414a58e6345b77b6a9b260f1...	2021-06	Essex Police	Essex Police	0.504922	51.828374	On or near Old Moors	E01021538	Chelmsford 001A	Otl th

In [28]:
```
def get_color(cluster_id):
    if cluster_id == 1:
        return 'darkred'
    if cluster_id == 2:
        return 'green'
    if cluster_id == 0:
        return 'amber'
```

In [29]:
```
#create a map
this_map = folium.Map(location =[clustered_full["Latitude"].mean(), clustered_full["Longitude"].mean()], zoom_start=5)

def plot_dot(point):
    '''input: series that contains a numeric named latitude and a numeric named longitude
    this function creates a CircleMarker and adds it to your this_map'''
    folium.CircleMarker(location=[point.Latitude, point.Longitude],
                        radius=2,
                        color=point.color,
                        weight=1).add_to(this_map)


clustered_full["color"] = clustered_full["cluster"].apply(lambda x: get_color(x))

#use df.apply(,axis=1) to iterate through every row in your dataframe
clustered_full.apply(plot_dot, axis = 1)


#Set the zoom to the maximum possible
this_map.fit_bounds(this_map.get_bounds())

#Save the map to an HTML file
this_map.save(os.path.join('Crime_map.html'))
#IFrame(src='Crime_map.html', width=1000, height=600)
```

In [ ]:
```

```

This website below is used for proof of crime rates in Chelmsford

https://crimerate.co.uk/essex/chelmsford#:~:text=The%20most%20common%20crimes%20in,2021's%20crime%20rate%20of%2046 (https://crimerate.co.uk/essex/chelmsford#:~:text=The%20most%20common%20crimes%20in,2021's%20crime%20rate%20of%2046).

In this first part of the workshop the first map file is labelled as Crime_map.html, where as for the the map file for question 9 it will be named Crime_map2.html for clarification purposes

Q7 answer

The aim of this workshop is to investigate crime rates in a specific location with the use of LSOA codes and types per each code loaction. Therefore the hierachila clustering algorith was appied for this dataset after the step of normalization. For Hierarchical clustering the duration that was considered is June 2021 to June 2022 and precisely Essex police was used to investigate crime rates and assign LSOA codes to each cluster.

Based on the clustering technique it was possible to find the areas that are of very high risk of crimes and areas that are of low risk of crimes. Therefore it is posiible to predict high risk areas and low risk areas. An example is that ClusterID1 contains areas of high risk whrere crime rates are very high.

When it comes to pre processing steps the data (with all the locations) was converted using pandas into a data frame then a aspecific location or town which in this case is Chelmsord was analysed. The Data was prepared for clustering using only crime types and LSOA codes and clusters were created for just that particular town.

Based on the results and the concept of hierarchical clustering it can be seen that even though there are sub clusters therefore many codes belonging to multiple clusters there are locations within the clusters that might not have a high crime rate even though the clusteID in itself might represent LSOA codes where crimes are committed the most.

Q8 answer

In [ ]:

```

```

The answer to question 9 is found in the next notebook in this same folder and it is named named Q9 answer