UK Police Crime Data Analysis

Phase 1 – EDA

Step 1: Data cleaning and preprocessing

- **Dataset 1: all_crimes18_hdr.csv – (df_crime)**
  1. Removing unnecessary and NaN columns/features.
     a. Column 0: Index column: Unneeded
     b. Column 1: Computer generated ID: Unneeded.
     c. Column 3: Falls Within: Duplicate column
     d. Column 11: Result: Unneeded
     e. Column 12: NaN Column
  2. Renaming columns appropriately

```
'_c1':'Month', '_c3':'FallsWithin', '_c4':'Longitude', '_c5':'Latitude',
'_c6':'Location', '_c7':'LsoaCode', '_c8':'LsoaName', '_c9':'Crime',}, inplace = True)
```

  3. Keeping LSOA rows that are not NaN: as without LSOA information we cannot continue.
  4. Converting columns to appropriate data types.
  5. Final dataframe:

	Month	FallsWithin	Longitude	Latitude	Location	LsoaCode	LsoaName	Crime
3054704	2015-07	West Midlands Police	-1.471840	52.437358	Roseberry Avenue	E01009607	Coventry 004C	Anti-social behaviour
426334	2018-03	Sussex Police	-0.460277	50.922312	Steyning Crescent	E01031627	Horsham 012C	Anti-social behaviour
3443881	2015-01	Bedfordshire Police	-0.261424	52.086462	Back Street	E01017385	Central Bedfordshire 005C	Violence and sexual offences
1635333	2016-12	Hertfordshire Constabulary	-0.202379	51.902935	Shopping Area	E01023758	Stevenage 008D	Violence and sexual offences
1519264	2017-02	South Yorkshire Police	-1.353814	53.433115	Shopping Area	E01007714	Rotherham 017B	Anti-social behaviour

- **Dataset 2: postcodes.csv – (df_postcodes)**
  1. Keeping necessary and not NaN columns/features.
     a. Postcodes
     b. InUse
     c. Latitude
     d. Longitude
     e. District

f. **LSOA Name (We use this column for merging with Dataset1 – df_crime)**

2. Converting columns to appropriate data types.
3. Keeping only rows where InUse column is 'Yes' – in order to remove expired postcodes
4. Final dataframe:

	Postcode	InUse	Latitude	Longitude	District	LsoaName
2222238	SY20 9PA	Yes	52.656606	-3.703466	Powys	Powys 004A
1184980	LE1 3EJ	Yes	52.643123	-1.136511	Leicester	Leicester 008B
1381726	MK2 2EE	Yes	51.995973	-0.725524	Milton Keynes	Milton Keynes 030B
856894	G51 4DL	Yes	55.854512	-4.332342	Glasgow City	Drumoyne and Shieldhall - 03
902907	GL56 0LR	Yes	51.987405	-1.699088	Cotswold	Cotswold 002E

- **1st Merge on df_crime and df_postcodes on LSOA to gain District level information.**
  1. Resultant df_crime

	Month	FallsWithin	Longitude	Latitude	Location	LsoaCode	LsoaName	Crime	District
3182504	2015-05	Metropolitan Police Service	-0.087597	51.388785	Willow Wood Crescent	E01001109	Croydon 013B	Anti-social behaviour	Croydon
423950	2018-03	Surrey Police	-0.419013	51.387655	Parking Area	E01030359	Elmbridge 007C	Anti-social behaviour	Elmbridge
2735852	2015-10	Cleveland Police	-1.054011	54.535084	Parking Area	E01012117	Redcar and Cleveland 018A	Anti-social behaviour	Redcar and Cleveland
3057969	2015-07	West Yorkshire Police	-1.611850	53.691559	Petrol Station	E01011123	Kirklees 018C	Anti-social behaviour	Kirklees
2092491	2016-07	South Wales Police	-3.177560	51.477741	St Mary Street	W01001941	Cardiff 032G	Violence and sexual offences	Cardiff

- **Dataset 4: postrans.csv – (df_population)**

```
columns = ["Date","Lsoa","LsoaCode","Rural_Urban","TotalPopulation","Males",
           "Females","Household","Communal","Child_Student","Area_Hectares","Density"]
```

1. No cleaning needed; we just renamed the columns.
2. Final Dataframe:

	Date	Lsoa	LsoaCode	Rural_Urban	TotalPopulation	Males	Females	Household	Communal	Child_Student	Area_Hectares	Density
16493	2011	Cambridge 012F	E01032795	Total	1568	758	810	1479	89	25	46.62	33.6
9069	2011	Leeds 070C	E01011618	Total	1581	742	839	1581	0	8	12.76	123.9
26759	2011	New Forest 007C	E01023045	Total	1580	768	812	1538	42	14	716.75	2.2
23291	2011	Hounslow 004E	E01002631	Total	1628	844	784	1628	0	21	20.73	78.5
10594	2011	Charnwood 006A	E01025740	Total	1499	760	739	1454	45	23	564.49	2.7

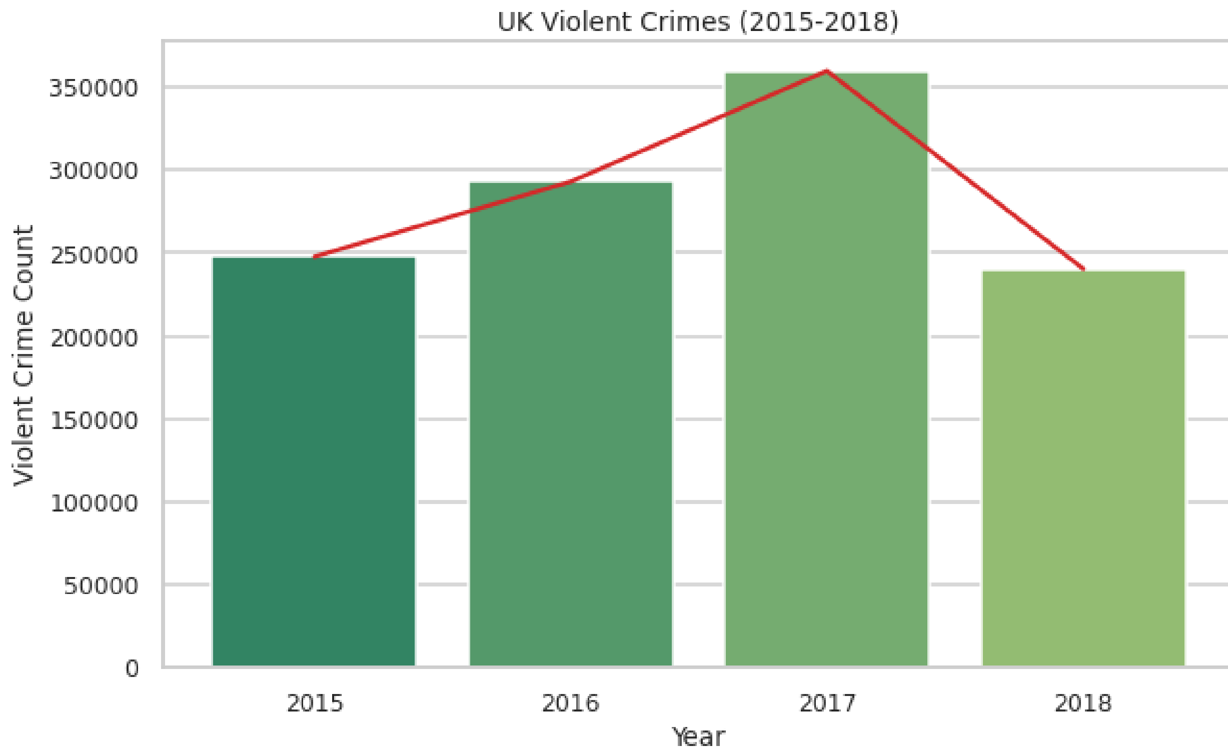
- **2nd merge on updated df_crime and df_population on LSOA to gain population information.**
    1. Resultant df_crime

	Month	FallsWithin	Longitude	Latitude	Location	LsoaCode	LsoaName	Crime	District	Population
222219	2018-05	Merseyside Police	-2.876089	53.435485	Baron'S Hey	E01006644	Liverpool 017D	Violence and sexual offences	Liverpool	2040
3155147	2015-05	Devon & Cornwall Police	-3.603080	50.535978	Nightclub	E01020217	Teignbridge 014B	Violence and sexual offences	Teignbridge	1657
714166	2017-11	Metropolitan Police Service	-0.155055	51.462581	Forthbridge Road	E01004589	Wandsworth 009E	Violence and sexual offences	Wandsworth	1700
3212943	2015-05	Thames Valley Police	-1.327300	52.063314	Victoria Place	E01028440	Cherwell 004F	Anti-social behaviour	Cherwell	1634
1910614	2016-09	Northumbria Police	-1.526912	55.116328	Keats Avenue	E01027422	Northumberland 024B	Anti-social behaviour	Northumberland	1845

Step 2: Exploratory Data Analysis

- **Exploring Claim 1: "The violent crime is increasing with time."**
    - **Technique:**
        - We make 4 dataframes from df_crime for each year. (2015, 2016, 2017, 2018)
        - Each dataframe is then filtered to only contain rows for "Violence and Sexual Offences."
        - We then find the length of each dataframe and store it. The length of each dataframe corresponds to the number of instances of violent crimes for that year.

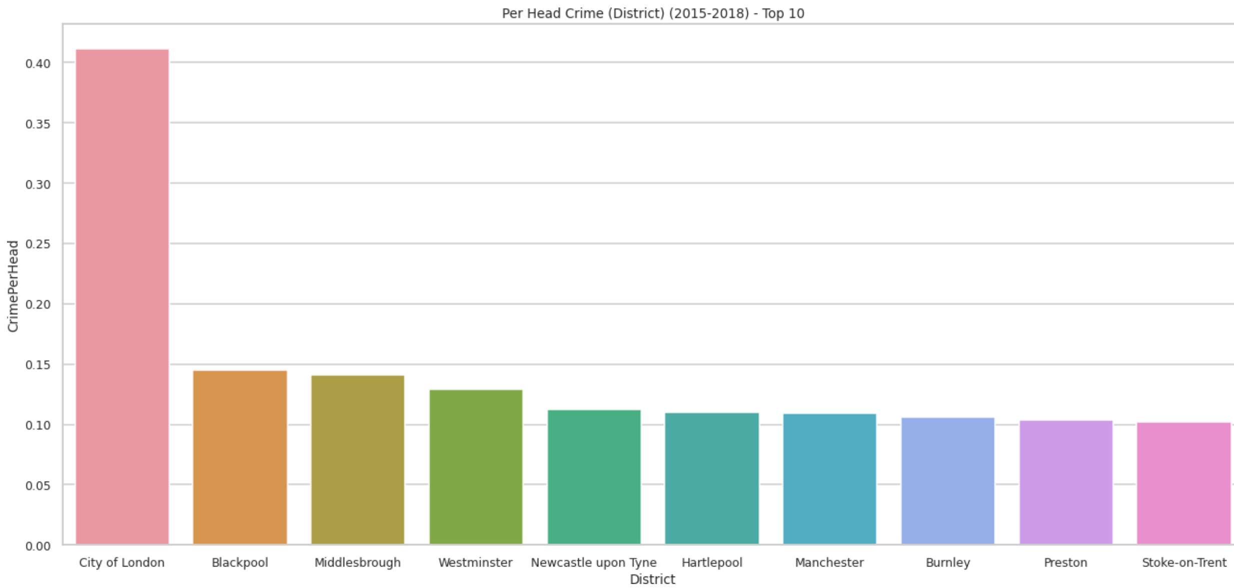
- **Results**



- **Analysis**
  - The claim can be confidently refuted. The violent crime did in fact increased from 2015 to 2017, however violent crime decreased drastically to levels even lower than to 2015.

- **Exploring Claim 2: “In Birmingham per head crime rate is higher than anywhere else in UK.”**
  - **Technique:**
    - We group the dataframe by district, then aggregate on the sum of population count and sum of crime instances.
    - Next, we create a new column where we store per head crime rate for each district (total crime/total population).

- **Results:**



- **Analysis**
  - Although it seems like the results for the city of London are not accurate as there might be discrepancies in recording the true population for this district. However, we can still confidently refute the claim as Birmingham is not even the in the top 10 of the results.

- **Exploring correlation between District and Most Prevalent Crime Type**
  - **Technique:**
    - We filter df_crime to only include District and Crime columns
    - We then group the resulting dataframe by district and aggregate by the mode (most common occurring) crime type.

  - **Results**
    - The resultant dataframe contains only the district and the most common crime. The first five rows are such:

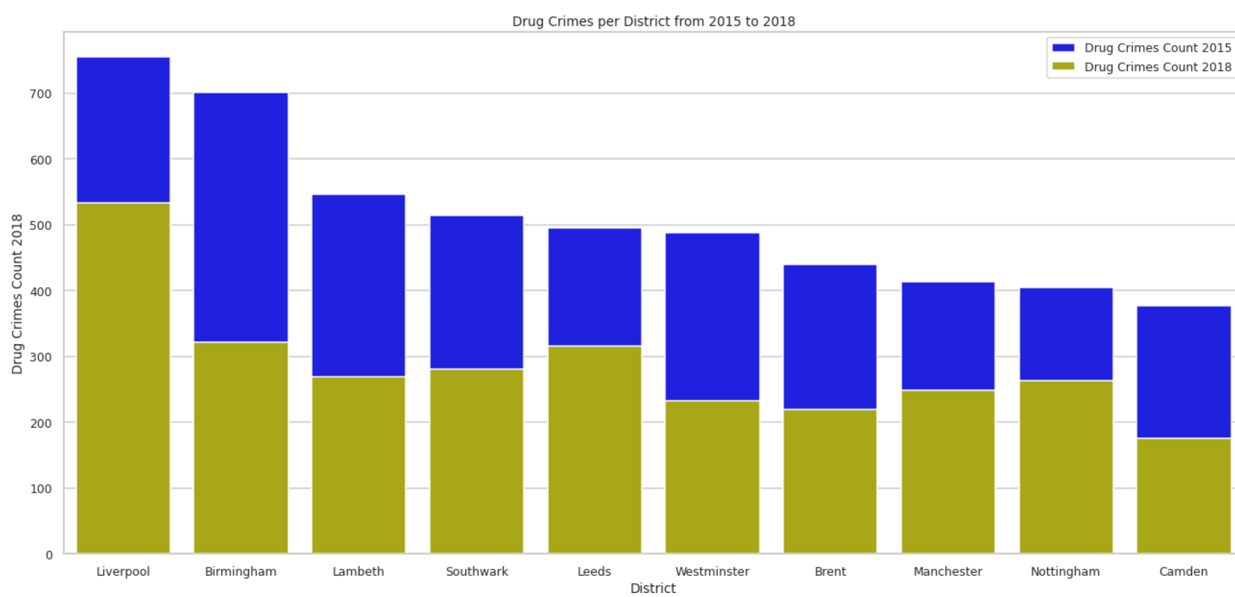
	District	Most Common Crime
0	Adur	Anti-social behaviour
1	Allerdale	Anti-social behaviour
2	Amber Valley	Anti-social behaviour
3	Arun	Anti-social behaviour
4	Ashfield	Anti-social behaviour

- **Other interesting insights**

1. **Exploring the change in trend in top 10 districts for drug crimes from 2015 to 2016.**
    a. **Technique:**
        i. Extract all drug crimes in 2015.
        ii. Group by district and aggregate on count of drug crimes
        iii. Select Top 10 Districts for 2015.
        iv. Extract all drug crimes in 2018.
        v. Group by district and aggregate on count of drug crimes
        vi. Select Top 10 Districts for 2018
        vii. Merge both dataframes into one

	District	Drug Crimes Count 2015	Drug Crimes Count 2018
0	Liverpool	756	533
1	Birmingham	702	322
2	Lambeth	547	269
3	Southwark	515	281
4	Leeds	496	316
5	Westminster	488	232
6	Brent	440	220
7	Manchester	413	249
8	Nottingham	405	264
9	Camden	377	176

b. Results



c. Analysis

i. The graph shows a significant decrease in number of drug related crimes in 2018 in the top 10 districts.

Phase 2 – Cluster Analysis

Step 1: Data cleaning and preprocessing

- **Reading Dataset made in previous phase – df_crime**

	Month	FallsWithin	Longitude	Latitude	Location	LsoaCode	LsoaName	Crime	District	Population
0	2018-07	Avon and Somerset Constabulary	-2.511761	51.409966	Caernarvon Close	E01014399	Bath and North East Somerset 001A	Anti-social behaviour	Bath and North East Somerset	1624
1	2018-07	Avon and Somerset Constabulary	-2.494870	51.422276	Conference/Exhibition Centre	E01014399	Bath and North East Somerset 001A	Violence and sexual offences	Bath and North East Somerset	1624
2	2018-07	Avon and Somerset Constabulary	-2.512773	51.411751	Westfield Close	E01014399	Bath and North East Somerset 001A	Violence and sexual offences	Bath and North East Somerset	1624
3	2018-07	Avon and Somerset Constabulary	-2.496204	51.417982	Abbey Park	E01014400	Bath and North East Somerset 001B	Anti-social behaviour	Bath and North East Somerset	1944
4	2018-07	Avon and Somerset Constabulary	-2.502805	51.414033	St Keyna Road	E01014400	Bath and North East Somerset 001B	Criminal damage and arson	Bath and North East Somerset	1944

- **Preprocessing for Cluster Analysis**

1. Group dataframe by district and crime type – df_cluster.

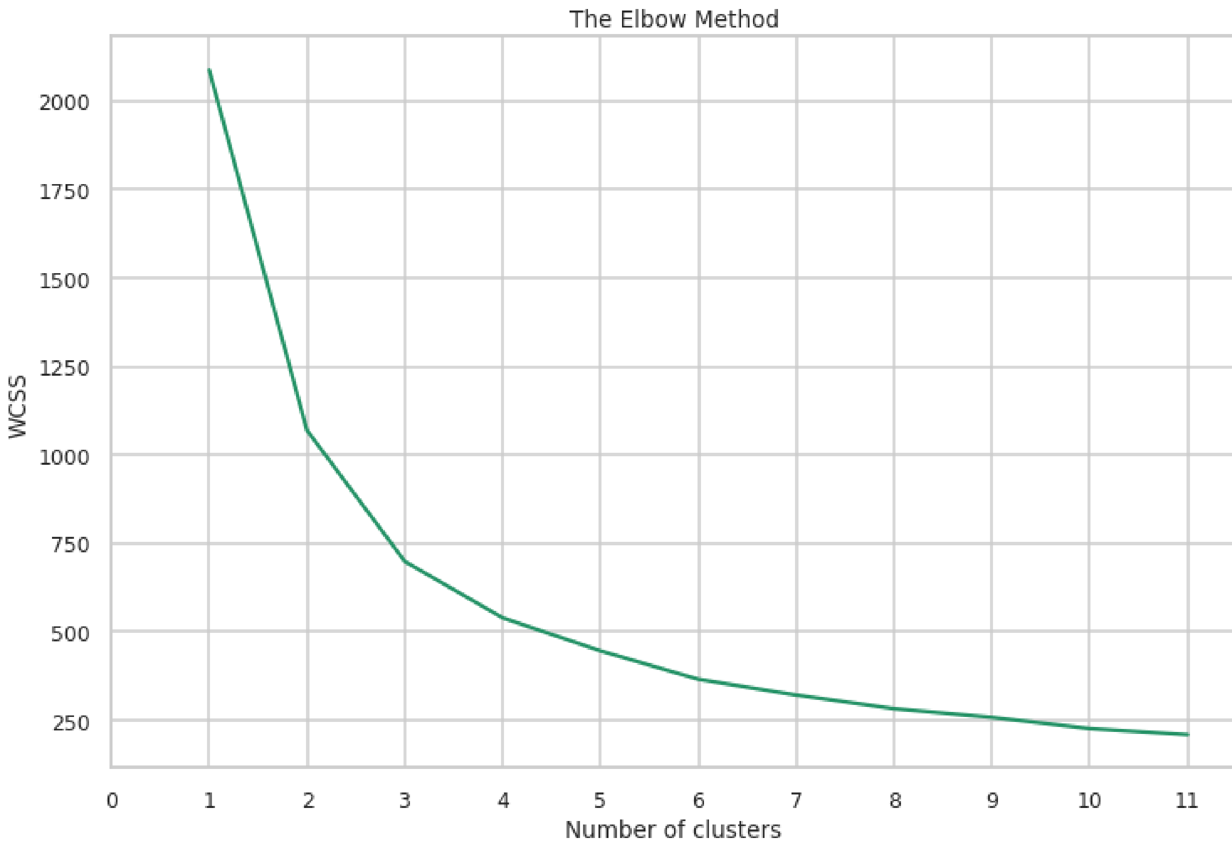
Crime	District	Anti-social behaviour	Criminal damage and arson	Drugs	Possession of weapons	Robbery	Violence and sexual offences
0	Adur	1204.0	471.0	133.0	26.0	22.0	1033.0
1	Allerdale	1970.0	1016.0	156.0	34.0	17.0	1563.0
2	Amber Valley	3957.0	852.0	199.0	36.0	31.0	1300.0
3	Arun	3403.0	1112.0	255.0	99.0	56.0	2651.0
4	Ashfield	3188.0	1263.0	182.0	62.0	69.0	2473.0

2. Checking if some districts don't have records for a specific crime type and replacing NaN with Zero. We can see that for all 348 districts we now have instances for all crime types. df_cluster.describe():

Crime	Anti-social behaviour	Criminal damage and arson	Drugs	Possession of weapons	Robbery	Violence and sexual offences
count	348.000000	348.000000	348.000000	348.000000	348.000000	348.000000
mean	4345.267241	1412.454023	341.459770	83.540230	157.385057	3279.077586
std	3789.438223	1242.531306	365.269899	91.669473	313.978681	2945.814276
min	17.000000	6.000000	3.000000	0.000000	0.000000	18.000000
25%	1955.000000	666.750000	125.500000	28.000000	22.000000	1417.000000
50%	3108.500000	1036.000000	225.000000	53.000000	50.500000	2325.500000
75%	5556.250000	1761.750000	383.250000	111.000000	128.750000	4451.000000
max	29611.000000	9438.000000	2980.000000	764.000000	2972.000000	22344.000000

Step 2: K-Means Clustering

1. **Standardizing the data for K-means using Z-score normalization.**
2. **Find optimal number of clusters using the Elbow-Curve Method. We observe that the optimal number of clusters is 3.**



3. **Predicting clusters using K-Means Algorithm**
4. **Creating a cluster column and assigning corresponding cluster value to each row.**

Crime	District	Anti-social behaviour	Criminal damage and arson	Drugs	Possession of weapons	Robbery	Violence and sexual offences	Cluster
3	Arun	3403.0	1112.0	255.0	99.0	56.0	2651.0	1
74	Coventry	5989.0	2980.0	439.0	163.0	565.0	5553.0	2
230	Rochford	1319.0	479.0	63.0	25.0	21.0	970.0	1
276	Stevenage	2482.0	909.0	272.0	248.0	48.0	2006.0	1
57	Cheltenham	5104.0	981.0	171.0	37.0	68.0	1655.0	1

5. Results

Cluster	Anti-social behaviour	Criminal damage and arson	Drugs	Possession of weapons	Robbery	Violence and sexual offences
1	2729.3	906.6	193.7	51.0	47.8	1994.4
2	8277.9	2582.0	715.8	158.0	425.1	6398.2
3	21549.0	7628.0	1716.2	492.8	1310.5	17055.8

We can observe the following from the clusters generated:

- Cluster 1: Districts with low counts of crime for all crime types.
- Cluster 2: Districts with moderate counts of crime for all crime types.
- Cluster 3: Districts with very high counts of crime for all crime types.

Cluster	Attribute of Cluster	Count in Dataframe
1	Districts with low counts of crime for all crime types.	261
2	Districts with moderate counts of crime for all crime types.	81
3	Districts with moderate counts of crime for all crime types.	6

Sample result for cluster 3:

Crime	District	Cluster
18	Birmingham	3
28	Bradford	3
159	Leeds	3
165	Liverpool	3
170	Manchester	3
247	Sheffield	3

6. **Map Visualization**

    a. Technique: We acquired a chape file for the UK region at district level. We then assigned each district a cluster using the previously made dataframe. Then we assigned a color to each district based on the cluster value and then generated the maps below.