Project Report

on

Crime Behaviour Analysis and Segmentation

Ishwar Patel, 11707807, KM031, Roll no 41

1. Collection of Database

I took the dataset from github.com. The dataset consists of all the districts in India and the number of each type of crime committed. This dataset is of the year 2014. There are total of 91 columns or features. The URL for the dataset is:

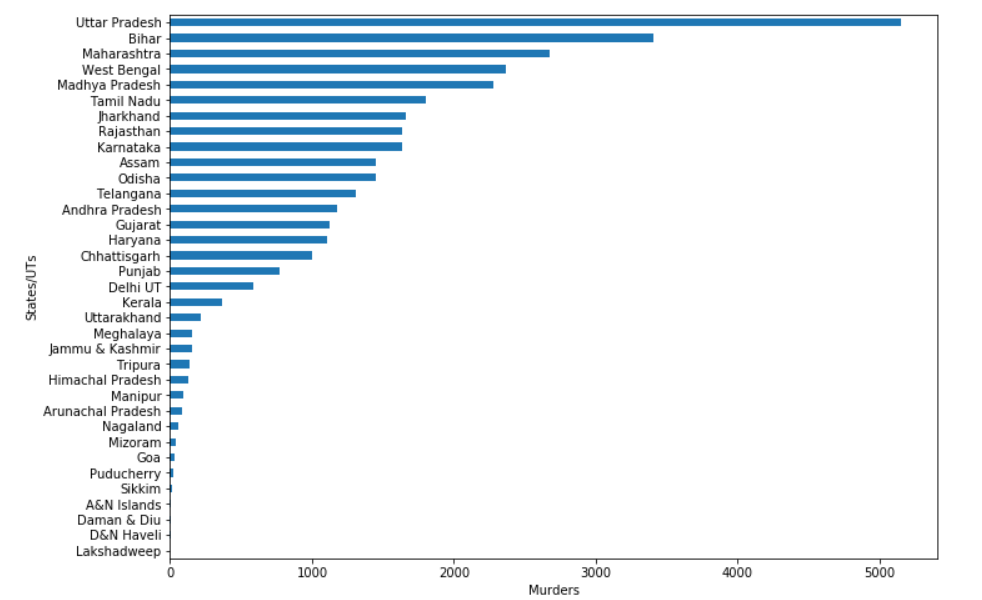
<https://github.com/15hu/Criminal-Behaviour-Analysis-and-Segmentation/blob/master/crime_data.csv>

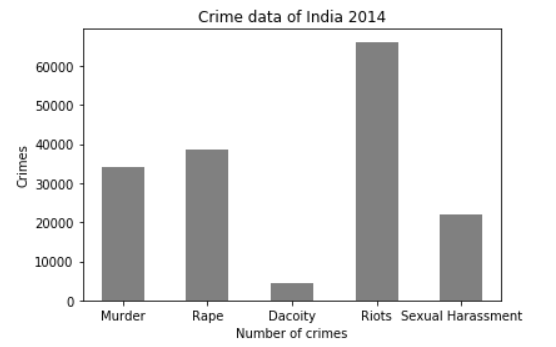
1. Implementation Code with Comment Lines

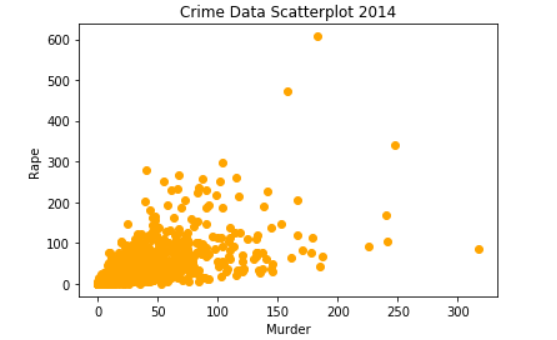
|  |
| --- |
|  |
|  | import pandas as pd # used to deal with data in csv file |
|  | from sklearn.cluster import KMeans #used for segmentation |
|  | import matplotlib.pyplot as plt # used to visualise the data |
|  | %matplotlib inline |
|  |  |
|  | #reading the dataset crime\_data.csv using pandas module |
|  | crime\_df=pd.read\_csv('crime\_data.csv') |
|  | print(crime\_df.head()) #print the top 5 rows of the dataset |
|  | print(crime\_df.columns) #print all the columns in the dataset |
|  |  |
|  | #Selecting some of the important features |
|  | crime\_data=crime\_df[['States/UTs','District','Year','Murder','Rape','Dacoity','Riots','Forgery','Acid attack','Sexual Harassment','Incidence of Rash Driving','HumanTrafficking','Total Cognizable IPC crimes']] |
|  | print(crime\_data.head()) |
|  |  |
|  | #Renaming one of the columns |
|  | crime\_data=crime\_data.rename(columns={'Total Cognizable IPC crimes':'Total IPC Crimes'}) |
|  | print(crime\_data.head()) |
|  |  |
|  | #Visualing the data using barplot, scatterplot and histogram |
|  | #Murder committed in each States/UTs |
|  | x = crime\_data.groupby('States/UTs')['Murder'].sum().sort\_values() |
|  | plt.figure(figsize=(11,8)) |
|  | plt.xlabel("Murders") |
|  | plt.ylabel("States/UTs") |
|  | print(x.plot(kind='barh')) |
|  |  |
|  | #Total IPC Crimes in each State |
|  | y = crime\_data.groupby('States/UTs')['Total IPC Crimes'].sum().sort\_values() |
|  | plt.figure(figsize=(11,8)) |
|  | plt.xlabel("Murders") |
|  | plt.ylabel("States/UTs") |
|  | print(x.plot(kind='barh',color='y')) |
|  |  |
|  | #Total number of each type of Crime |
|  | crimes=['Murder','Rape','Dacoity','Riots','Sexual Harassment'] |
|  | number=[crime\_data[crime].sum() for crime in crimes] |
|  | plt.title("Crime data of India 2014") |
|  | plt.xlabel("Number of crimes") |
|  | plt.ylabel("Crimes") |
|  | plt.bar(crimes,number,0.5,color='grey') |
|  | print(plt.show()) |
|  |  |
|  | #Scatterplot |
|  | fig, ax = plt.subplots() |
|  | ax.scatter(crime\_data["Murder"], crime\_data["Rape"],c='orange') |
|  | plt.xlabel("Murder") |
|  | plt.ylabel("Rape") |
|  | plt.title("Crime Data Scatterplot 2014") |
|  | print(plt.show()) |
|  |  |
|  | #Histogram |
|  | fig, ax = plt.subplots() |
|  | ax.hist(crime\_data['Murder'], range=(0,200), align='mid', histtype='stepfilled',color='r') |
|  | print(plt.show()) |
|  |  |
|  | #K-Means Clustering |
|  | X=crime\_data.values[:,3:] |
|  | print(X[:10]) |
|  |  |
|  | clusterNum=3 |
|  | k\_means=KMeans(init = "k-means++", n\_clusters = clusterNum, n\_init = 12) |
|  | k\_means.fit(X) |
|  | labels = k\_means.labels\_ |
|  | print(labels) |
|  |  |
|  | crime\_data["Clus\_km"] = labels |
|  | print(crime\_data.head(5)) |
|  | print(crime\_data.groupby('Clus\_km').mean()) |
|  |  |
|  | #Visualising the clusters |
|  | #Distribution of districts based on Murder and Total IPC Crimes |
|  | plt.scatter(X[:, 0], X[:, 9],c=labels.astype(np.float), alpha=0.5) |
|  | plt.xlabel('Murder', fontsize=18) |
|  | plt.ylabel('Total IPC Crimes', fontsize=16) |
|  | print(plt.show()) |
|  |  |
|  | #Distribution of districts based on Rape and Total IPC Crimes |
|  | plt.scatter(X[:, 1], X[:, 9],c=labels.astype(np.float), alpha=0.5) |
|  | plt.xlabel('Rape', fontsize=18) |
|  | plt.ylabel('Total IPC Crimes', fontsize=16) |
|  | print(plt.show()) |
|  |  |
|  | #Distribution of districts based on Sexual Harassment and Total IPC Crimes |
|  | plt.scatter(X[:, 6], X[:, 9],c=labels.astype(np.float), alpha=0.5) |
|  | plt.xlabel('Sexual Harassment', fontsize=18) |
|  | plt.ylabel('Total IPC Crimes', fontsize=16) |
|  | print(plt.show()) |

1. Data and Result Visualisation

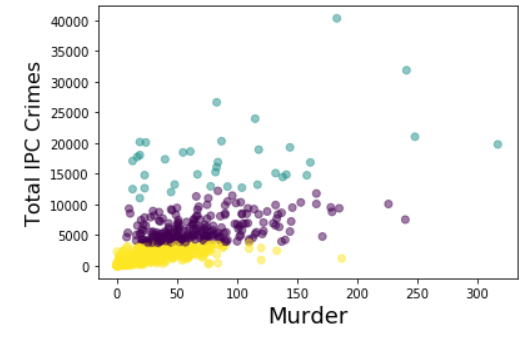
Data Visualisation:

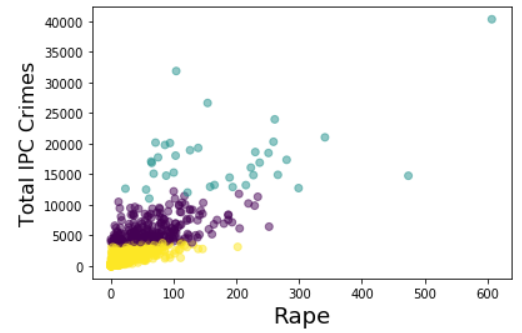


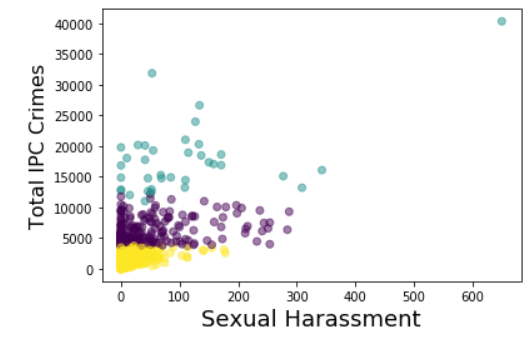




Data After Segmentation:







1. Github Repository

The link of my project repository is:

<https://github.com/Lovely-Professional-University-CSE/INT247-Machine-Learning-Project-KEM031-rollno_41>