MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

Mini Project Report of

Introduction to Data Analytics (CSE 2126)

Crime and Demography

SUBMITTED
BY

Aaryan Takayuki Panigrahi – 220962366
Tanay Srivastava - 220962432

Department of Computer Science and Engineering
Manipal Institute of Technology, Manipal.
Oct 2023

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Manipal
00/00/2023

CERTIFICATE

This is to certify that the project titled **Crime and Demography** is a record of the Bonafide work done by **Aaryan Takayuki Panigrahi – 220962366 and Tanay Srivastava - 220962432** submitted in partial fulfilment of the requirements of **Introduction to Data Analytics (CSE 2126)** course of Manipal Institute of Technology, Manipal, Karnataka, (A Constituent Institute of Manipal Academy of Higher Education), during the academic year 2023-2024.

Name and Signature of Examiner:

Dr. Roopalakshmi R,
AssociateProfessor,
CSE Dept.

TABLE OF CONTENTS

INTRODUCTION

Crime is a nuisance to society. We may try to solve crime by brute-force. Recruiting more police officers to respond to crime and arrest criminals. While this is a great approach that has brought order to society for decades, it has its own limitations.

In the recent history of modern societies, with developments in computer science and data analysis, people have tried to solve crime by trying to understand it better through the statistical lenses of data analysis. CompStat one is a good example of this.

CompStat, short for "computer statistics" or "comparative statistics," is a management and accountability system used by police departments to track and analyze crime data. It was first developed and implemented by the New York City Police Department (NYPD) in the 1990s.

Our project aims to leverage the transformative impact of data analysis in the realm of law enforcement. By exploring a crime dataset enriched with demographic and housing fields, we strive to uncover meaningful patterns, correlations, and actionable insights. Through robust data analysis techniques, we aspire to provide law enforcement agencies with the tools to make informed decisions, contributing to the enhancement of public safety. This project underscores the potential of data analysis as a powerful tool for crime reduction and community well-being, building upon the principles that have proven effective in initiatives like CompStat.

It was found that the crime rate went down drastically because now the department could take informed decisions. Appropriate measures were taken depending on the factors like locality, traffic, etc., which led to the city being safer than it ever was.

Inspired by the above case study, we decided to go for a similar crime dataset which had demographic as well as housing fields.

PROBLEM STATEMENT & OBJECTIVES

Problem – Identifying the patterns of crimes in relation to different demographic attributes for a given population and region. We also aim to predict the category of crimes and their causation based on certain parameters.

Objective:

1. Identifying Patterns

We aim to analyze the data and visualize it using various tools available through various python libraries. We can also analyze the correlation between different attributes to gain insights, which can be used for further analysis with clustering models, etc.

2. Predicting Crimes based on related attributes:

We aim to perform multivariate analysis(regression) so that we can so how the crime rates vary with the parameters concerned. We take the correlated parameters together and try to predict what type of crime could be committed given a set of demographic circumstances.

3. Analyzing the causation of crimes:

We also aim to form clusters of diverse types of crimes with closely related attributes which could help in understanding the cause of certain crimes and what measures could be taken to prevent such crimes.

METHODOLOGY

1. Data Preparation:

a. The dataset which is originally a csv file is imported as a pandas **dataframe**. We can also get an overview of the dataset using the *.describe()* function for a pandas dataframe.

b. Missing Values:
   i. First, we **identified** the attributes with many missing values. These attributes having more than 1000 missing values are **dropped** from the dataset to minimize errors in our analysis.
   ii. However, there are certain attributes like murders/burglaries committed which cannot be dropped from the dataset. The missing values in such attributes are replaced by the **mean** values.

c. Data Normalization: We have also normalized the data so that machine learning models can be trained more efficiently, as algorithms like gradient descent and k-means clustering converge faster, with scaled data.

2. Data Transformations:

a. Grouping attributes: We also decided to group similar attributes together so that we can work on further analysis fluently.
b. Correlations: pandas provide an inbuilt function to work out the correlation matrix for all the attributes in dataframe. The correlation between crime and demographic and housing attributes helps us identify patterns and relations between different attributes.

3. Data Visualization:

a. We are using several types of plotting techniques to get deeper insights into our data before proceeding for Data Modelling.
b. These include heatmaps to visualize the correlation matrix, bar plots to compare the relation between different groups of attributes, etc.
c. Python libraries such as matplotlib, and seaborn are very useful for this.
d. We also visualized the clusters formed by k-means clustering using PCA (Principal Component Analysis) as implemented by sk-learn.

4. Data Modelling:

a. Multivariate Linear Regression:

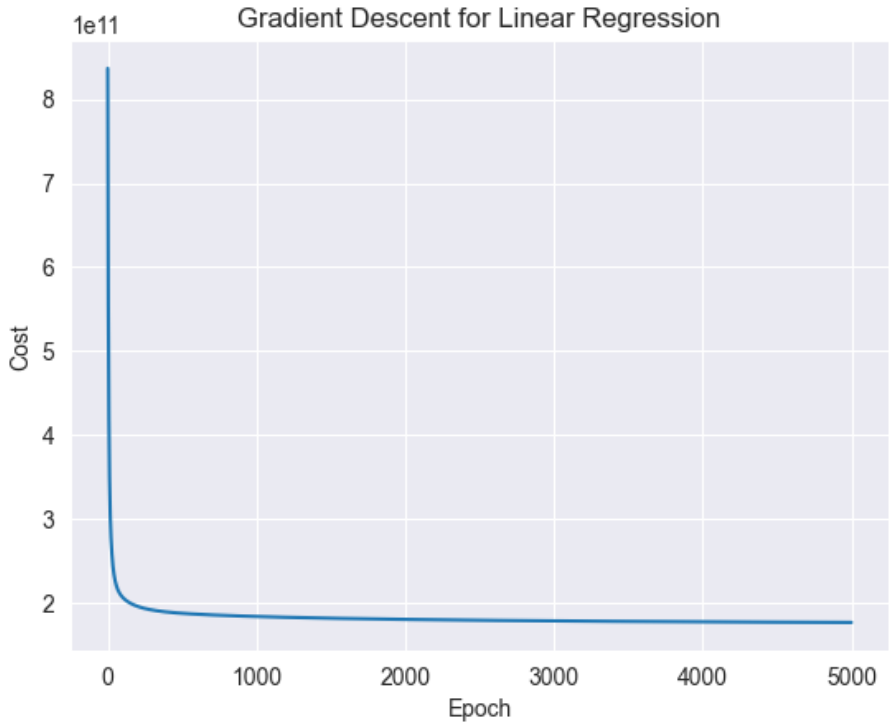
A linear function can be fit to the data using linear regression. Such a function predicts an attribute (Y), in this case the rate of violent crimes, given a multidimensional input vector of other attributes.

$$Y = W.(X.T) + B + e$$

where:

W – Weights
B – Bias
e – Residual Error
X – Input Matrix
Y – Prediction (Rete of Crime)

The weights (W) and bias (B) are determined through the process of training the model, typically using an algorithm like Gradient Descent, which tries to minimize the difference between the predicted values and the actual values in the training data.



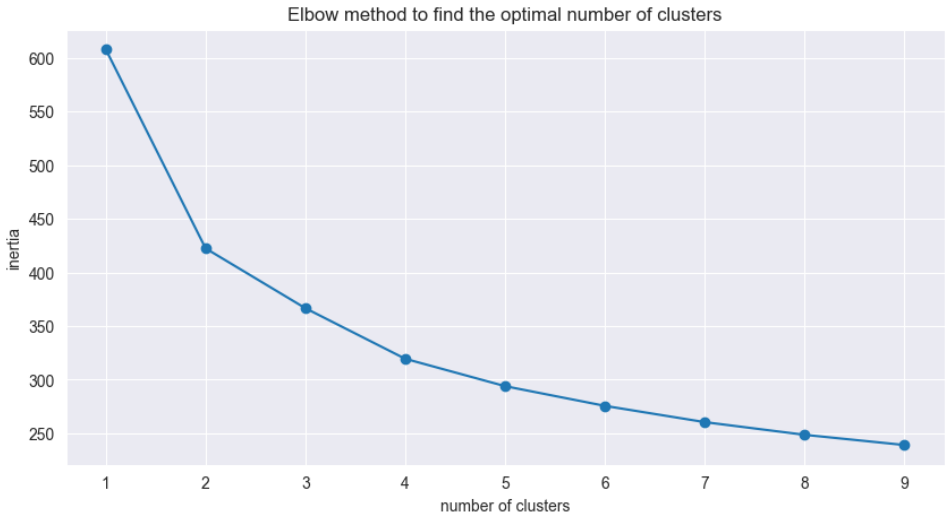
The above graph shows how the residual error (cost) decreases and converges as the model trains for multiple epochs.

After the model has been trained and validated, we have an optimal (W), and (B) that accurately represent the relation between the various attributes in X and the output parameter Y (crime rate).
This relation can be analyzed by looking at the weights corresponding to each attribute, as in W. The **higher the weight** of a particular attribute, the **stronger its influence** on the crime rate (Y). This allows us to identify the factors that are most responsible for the crime rates, so that we can take measures to improve them.

b. K-Means Clustering:

This is an unsupervised learning algorithm used to find patterns in the data, and group similar datapoints together.
We need to know the number of clusters we want to form, for the K-Means algorithm. As we want our clusters to be as distinct as possible, i.e. have as little overlap as possible, we can use the Elbow Method to find the optimal number of clusters.

A suitable value of K (number of clusters) is chosen using the Elbow Method and the value of K for which the change in variance is maximum is taken for further clustering of attributes.

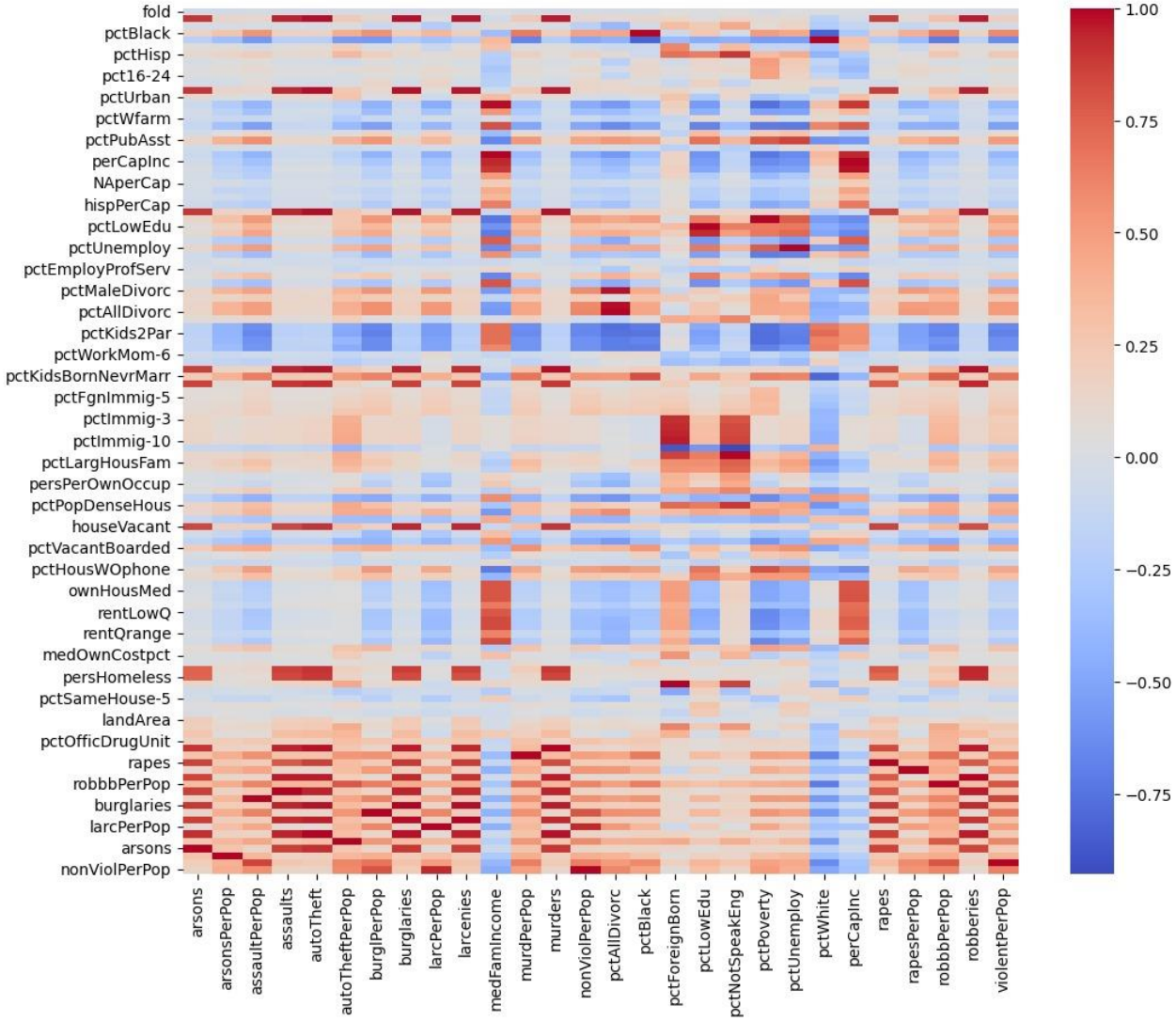


The above graph plots the number of clusters against the sum of squared distances of each point from its cluster centroid. The elbow point is the point where the sum of squared distances stops decreasing rapidly.

The results of all these models are discussed in the following section –

RESULTS & SNAPSHOTS

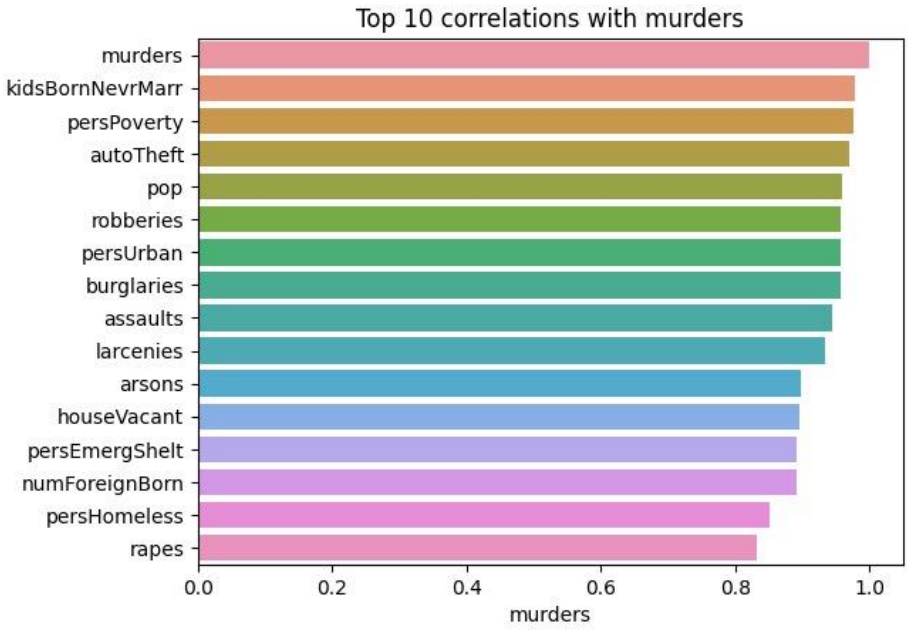
Correlations Matrix Heat-map



The above heatmap shows a pictorial representation of the correlation matrix of relevant attributes

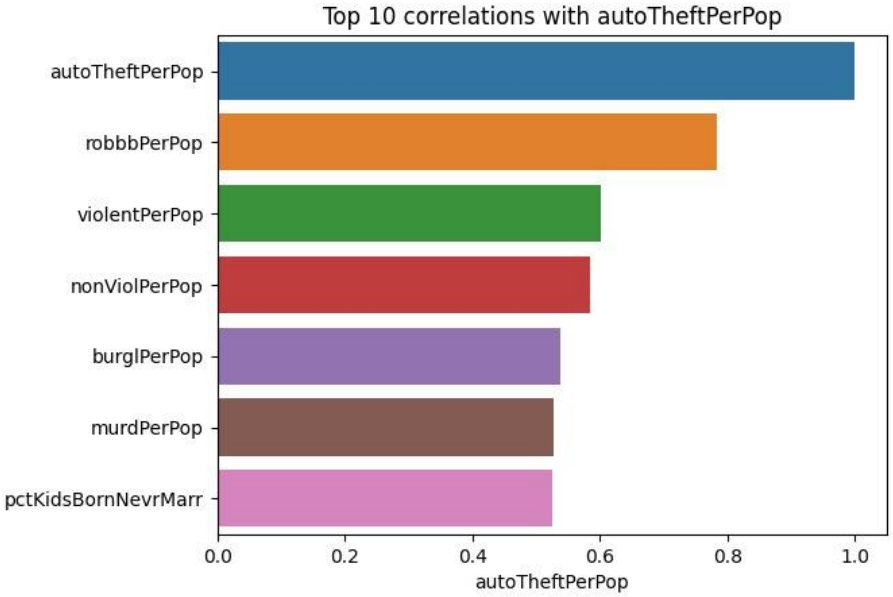
The above heatmap can be studied to get a general idea of which attributes are strongly correlated with each other.

But as there are too many variables, the correlations are not very easy to spot.
Hence, we have further identified a few attributes of interest from the above, and analyzed their correlations with the following bar plots -



This Graph shows the top 10 attributes that have a strong correlation with murders

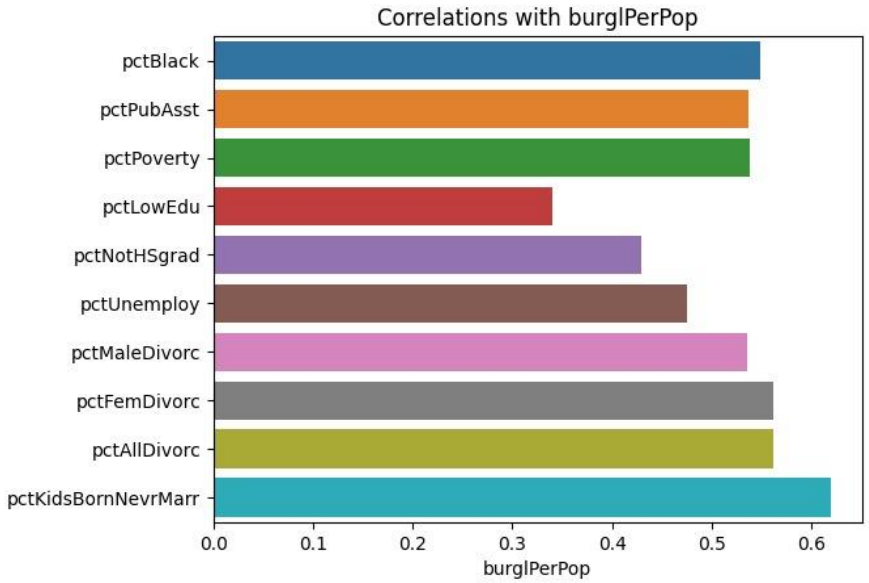
- We can see that **persPoverty** and **persHomeless** are in the above list of top 10. This suggests that poverty could be an important root cause of murders being committed in the first place.
- We can also see what other crimes are likely to be committed along with murders. These include **robberies**, **assaults**, etc.



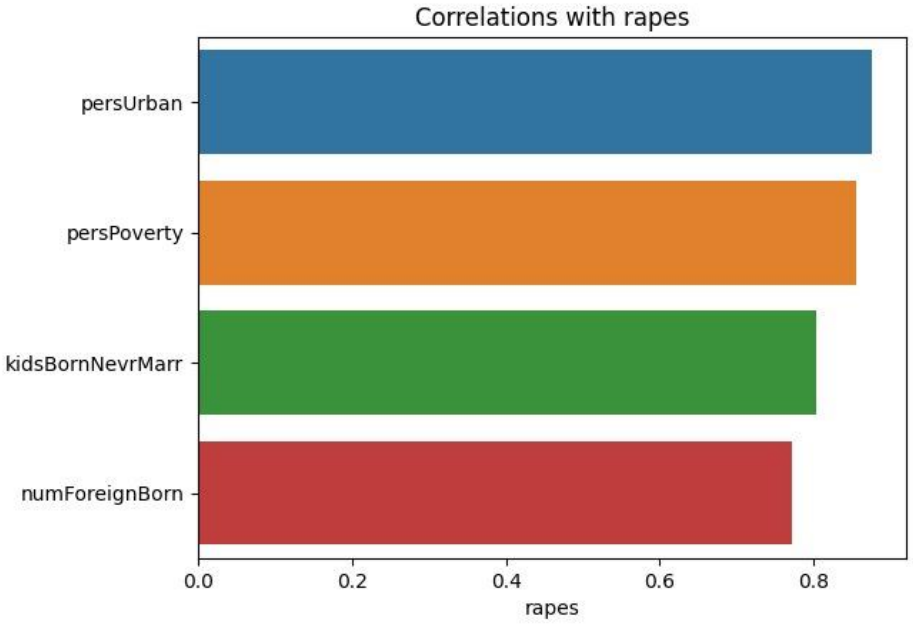
This Graph shows the top 10 attributes that have a strong correlation with murders

- It can be useful to analyze the patterns between different types of crimes with each other.
- Especially crimes like **autoTheft.** Because these crimes are often committed in conjunction with other crimes, such as **robberies**, and **murders.**

Our main aim is to analyze the correlation between demographics and crimes.
Below are the plots of the correlations of certain crimes exclusively with demographic attributes (we are not plotting correlations between different crimes here)



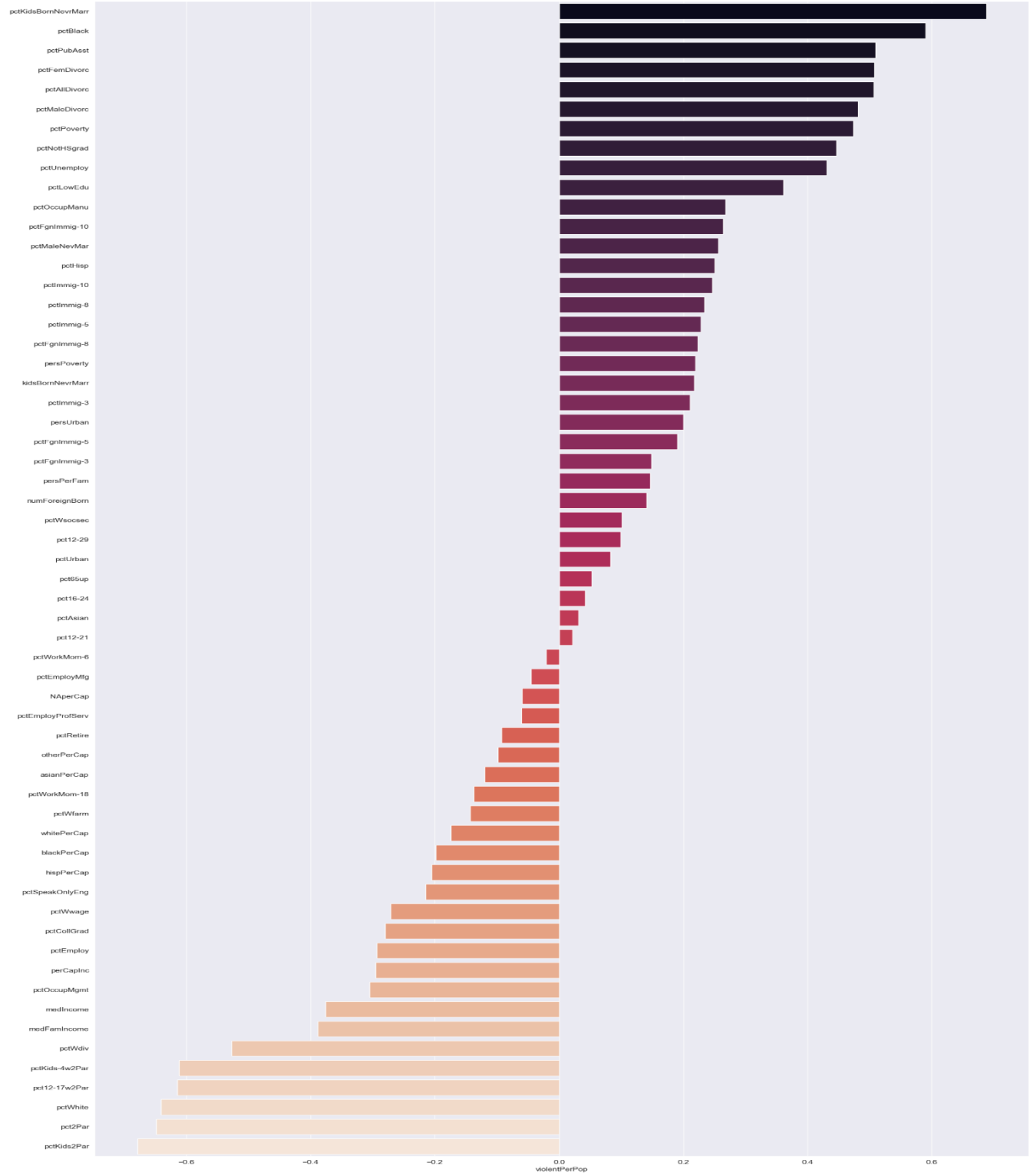
This Graph shows the top 10 attributes that have a strong correlation with murders



This Graph shows the top 10 attributes that have a strong correlation with murders

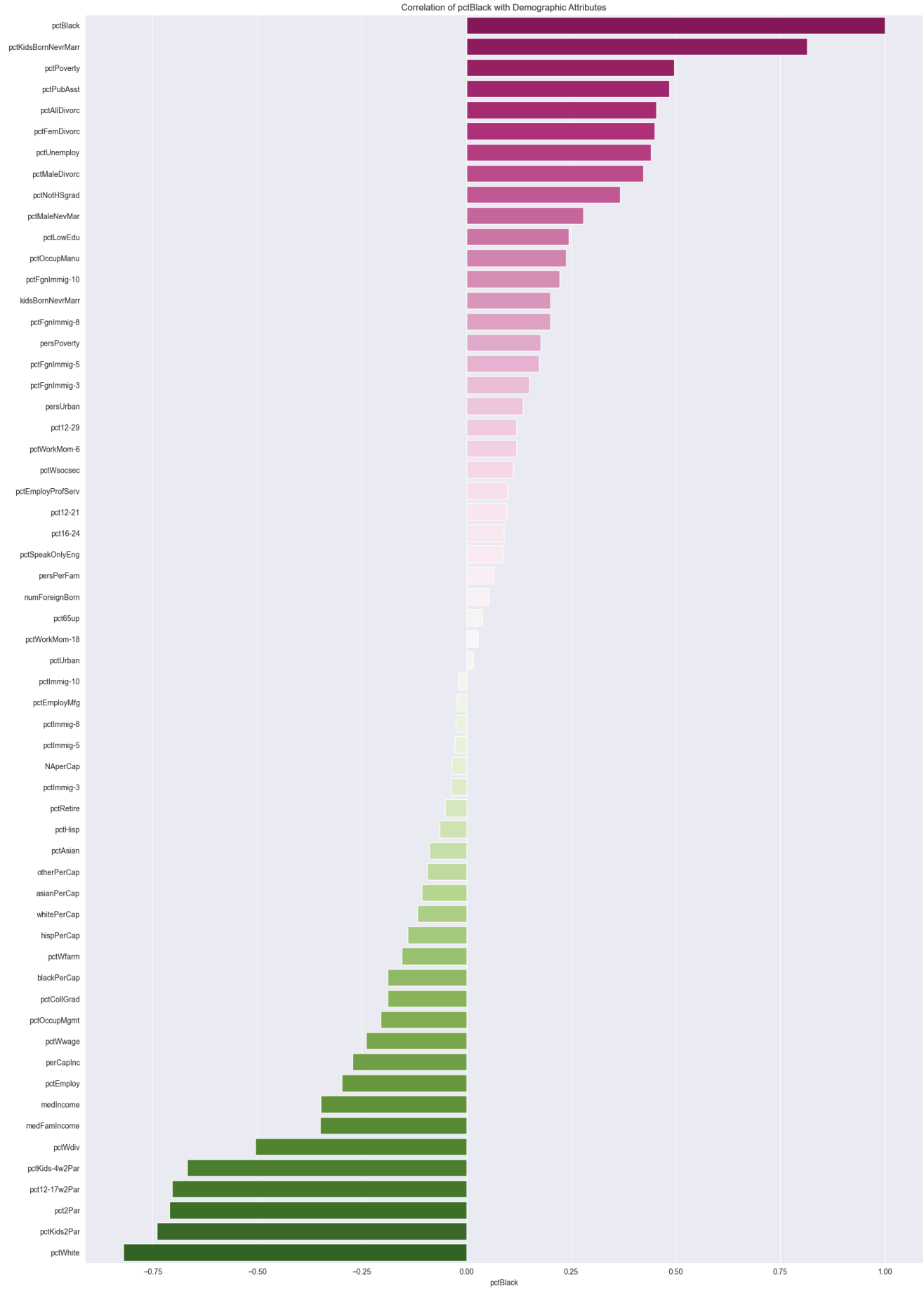
The above graphs plot only those correlations which are above a certain threshold (0.3). More such graphs can be plotted using the function defined in our Jupiter notebook.

Now let's see the correlations of violentPerPop with other attributes -



Upon first glance, it may seem as though there exists a strong correlation between race and crime.

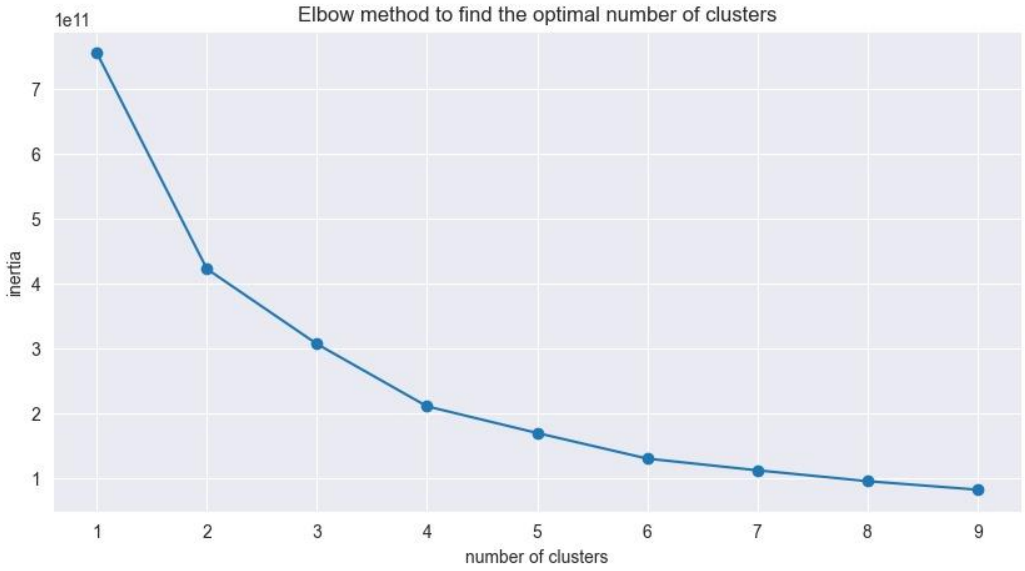
However, this is a univariate analysis, and we must look deeper and consider other factors such as poverty, education, etc.
In the following page, we plot the correlations between demographics and race to check for the same -



Correlation of **pctBlack** with Demographic Attributes

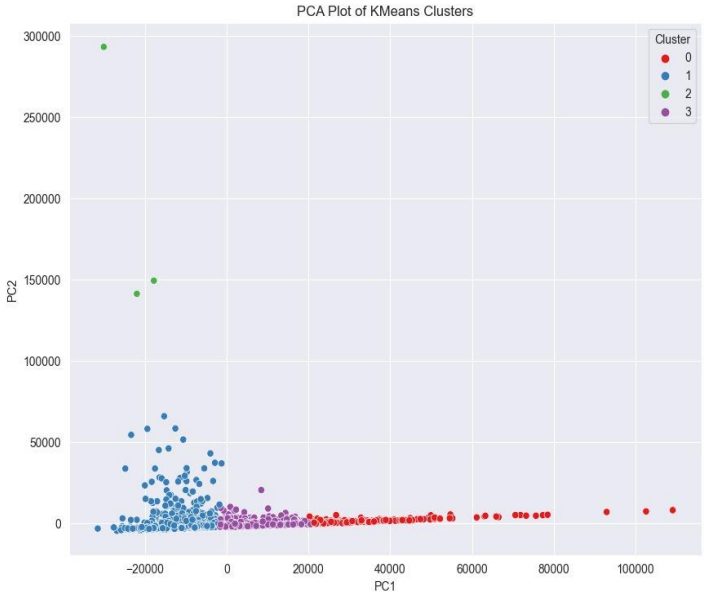
From the above, we can see that pctBlack is strongly correlated with some of the other attributes that were strongly correlated with Violent Crimes
These include poverty, divorce, unemployment, and illiteracy rates.
So while a certain race may be correlated with higer crime rates... It is important to consider the other factors as well.
This shows the shallow nature of a univariate analysis, and the need for a multivariate analysis.

K-Means Clustering Results -



	pctBlack	pctWhite	pctLowEdu	pctPoverty	perCapInc	pctUnemploy	medFamIncome	pctNotSpeakEng	pctForeignBorn	pctAllDivorc	...	burglaries	burglPerPop
0	2.647104	91.195837	3.098462	2.892896	29235.909502	3.263982	70255.660633	1.485339	10.337647	7.344615	...	178.846154	574.034299
1	13.688550	78.815420	12.473749	17.375564	11775.888889	7.605954	29713.463953	2.901968	6.371510	12.232841	...	1019.288507	1345.648890
2	4.772256	89.636720	6.046276	5.613613	17437.508015	4.525413	46295.545006	1.902676	7.851825	9.694883	...	307.955086	704.176665
3	27.256667	50.160000	15.923333	19.923333	15122.666667	9.546667	33143.666667	11.570000	27.883333	12.340000	...	65036.333333	1488.466667

We decided the optimal value of K using this elbow curve. K value with maximum variance is chosen and K-Means is implemented accordingly.



- From our analysis of the main attributes most responsible for crime are – Income, Education, and Unemployment.
- But certain communal differences may also appear to be correlated with crime rates.
- **A univariate analysis is incomplete – we do multivariate

Multivariate Linear Regression - Results

```
[121]
...  pctUrban              : -27.025687760143022
     pctRetire             : -13.17711395252773
     pctAsian              : -3.089615424415033
     pct16-24              : -2.6686955768450584
     pctWsocsec            : -2.598302453089338
     pctWhite              : -1.5137968618230615
     pctWwage              : -0.16464756254439789
     persUrban             : -0.04956104435589775
     pct12-29              : -0.001146005677895522
     pctPubAsst            : 0.00012694714536479727
     pct65up               : 0.000255325036414768
     pctHisp               : 0.002129104162784676
     pct12-21              : 0.04621579510745174
     pctBlack              : 3.10271387688076
     pctWfarm              : 23.46125297982643
     medIncome             : 26.543652599503396
     pctWdiv               : 47.26736767159472
```

- The above are the weights of the relevant attributes from our Multivariate Linear regression.
- We can see that medIncome has a big direct correlation with crime. This means that the economic state of society influences crime.
- We can also see that Education, Unemployment, and povery affect crime rates

```
Final ERROR & ACCURACY Analysis of the model ~

RMSE (Root Mean Square) =        13.930398324927298

NMRMS (Normalised RMSE) =        0.0036430199654085916

R2 (Squared Correlation) =       0.5872997887645499

Accuracy of model =              59.93 %
```

Our accuracy seems to be a bit low. We tried to verify this using sklearn's Linear Regression model, but that was also around the same range
This is probably because the data is not linearly separable.
We can try to improve the accuracy by using a non-linear model, such as a Decision Tree Regressor, or a Random Forest Regressor in the future.

But this is still valid for the purpose of analysing which attributes have a higher influence, as shown above.

Conclusion

- Demographics play a significant role in influencing crime patterns: Our data analysis project revealed a clear and compelling connection between demographic factors and crime rates.

- Data analysis techniques allow us to (do the below basically)

- Crime patterns can be identified, and crime rates can be reduced by addressing the root causes of various crimes in various locations.

Limitations and Future Scope

- Missing Data – For some attributes such as the ones related to the police force almost 60% of the fields were missing values. For this reason, these attributes had to be dropped and could not be used for analysis.
- Linear Models are insufficient to produce a model with enough predictive power. Hence, we only saw 50-60% of squared correlation (r2 accuracy) across KNNs and Linear regression with and without prebuilt libraries.
- We can try to improve the accuracy by using a non-linear model, such as a Decision Tree Regressor, or a Random Forest Regressor in the future.
- Models that allow for more complexity such as neural networks can be used for future improvements and further analysis.

REFERENCES

1. sklearn.neighbors.KNeighborsClassifier — scikit-learn 1.3.2 documentation
2. sklearn.linear_model.LinearRegression — scikit-learn 1.3.2 documentation
3. sklearn.cluster.KMeans — scikit-learn 1.3.2 documentation
4. Jupyter Notebook
5. UCI Machine Learning Repository