The statistic explanation of the US crime data

YAN, Ningyu YAN, Bokai LAI, Yanming

April 9, 2023

Abstract

In this report, a statistical analysis of US crime rate is conducted. The goal is to find the main contributions to the average crime rate using the crime data. The original dataset includes 85 American cities' crime information from 1969 to 1992 with many missing values and useless data. Firstly, the dataset is simplified using preprocessing techniques. Then, principal component analysis (PCA), Isometric Mapping (Isomap) and uniform manifold approximation and projection (UMAP) methods are used to reduce the number of variables from 21 to 5 and 11. Visualization results show that the 85 cities are well separated into several groups by the three methods. Finally, the linear regression model is applied to find the relationship between 'principle components' and the average crime rate. In short, we find that the US crime rate is affected by many factors, such as the population structure, police number, mayoral term, society welfare, etc.

1 Introduction

The crime data in the United States has been recorded since its founding, which depends on various factors. The statistical analysis of these crime data may reveal significant information, such as the principal contributions to the crime rate and the prediction of it, and this may facilitate the government to manage the whole country and give us a safer society. In this paper, we will explore the dataset of the US crime data from 1969 to 1992 including 85 cities. This dataset contains various information, such as the population composition, election year and the number of police officers. Then we will solve 3 problems:

- The preprocessing of the data set.
- Conduct three methods, principal component analysis (PCA), Isometric Mapping (Isomap) and uniform manifold approximation and projection (UMAP), to the dataset (feature selection).
- Based on selected features, using the linear regression model to predict the crime rate.

2 Data Preprocessing

The original dataset 'crime2.xlsx' demonstrates the criminal records of 85 U.S. cities, which has a matrix size of 2361×41. Each row represents the names of cities, and each column represents detailed information of the corresponding city. There are many blanks in the data matrix. Some of the information (the total number of police officers, different types of crime, the election year, etc.) is valuable, but the other may be (web,jid, etc.) useless to the crime rate. So we need data preprocessing to remove unhelpful information and standardize the form of the input matrix for programming.

Firstly, we remove the data of different cities' names and years, if we consider these two factors, many dummy variables will be introduced, which will make a negative impact on PCA. Then the total police number and crime rate are calculated by the summation of the column 'sworn' and 'civil', and the ratio of total crime number and total population. In this report, our main goal is to analyze the average crime rate caused by generalized factors instead the detailed classifications of crimes,

so the columns of different kinds of crimes will be removed. Notice that this dataset contains some inconsistent data and hence we also eliminate them, such as the columns of 'termlim', 'jid', 'mayor', 'data_wa', 'data_my' and 'web'. Furthermore, we delete the samples with missing data. Finally, the data normalization is conducted.

After data preprocessing, we get a data matrix with the size of 1078×21. The rows represent the different samples, the columns represent the different factors relating to the average crime rate.

3 Data Reduction and Visualization

It is difficult for us to handle high-dimensional datasets, so data reduction methods can be used to transform complex datasets into a simplified form and conduct the feature selection. This procedure will also eliminate the collinearity of samples, which will benefit the linear regression (otherwise, the $\left(X^{\mathrm{T}}X\right)$ will not be invertible, and it will be inconvenient to estimate model parameters). PCA is a classical linear method for data reduction, which assumes the data lies in a low-dimensional linear subspace. However, it often misses the non-linear structure and sometimes may get inaccurate results. Thus, we can generalize the linear model to non-linear ones, i.e., Isomap and UMAP, to analyze high-dimensional datasets.

3.1 Methods

We give brief introductions of PCA, Isomap and UMAP.

- **Principal Component Analysis (PCA)**
  PCA is defined by a transformation which transforms the data to a new coordinate system such that the variance in descending order by some scalar projection of the data comes to lie on the coordinate in descending order. Thus the first few coordinates could store most information of the data.
- **Isometric Mapping (Isomap)**
  Isomap is a nonlinear dimensionality reduction method, and it is one extension of the classical MDS method by changing the Euclidean distance into the geodesic distance. The algorithm provides a simple method for estimating the intrinsic geometry of a data manifold based on a rough estimate of each data point's neighbors on the manifold. Isomap is highly efficient and generally applicable to a broad range of data sources and dimensionalities.
- **Uniform Manifold Approximation and Projection (UMAP)**
  UMAP is a novel manifold learning technique for dimension reduction, and it is also a nonlinear method. It is constructed from a theoretical framework based on Riemannian geometry and algebraic topology. The embedding is found by searching for a low-dimensional projection of the data that has the closest possible equivalent fuzzy topological structure.

3.2 Results

We first conduct PCA. The 21 variance ratios are shown in Table 1 and the information of the main principal components can be found in Figure 1 and 2. In Figure 2, it is shown that the first principal component has the most load on "a10_14", "a5_9" and "a15_19", which are the population percentage within age range 10-14, 5-9 and 15-19, respectively. Hence we can interpret the first component as an index reflecting the population information. The fourth principal component has the most load on "term2", "a20_24", "term4" and "term3", referring to the population percentage within age range 20-24 and indicator variables for cities with two-year, three years and four years mayoral terms. Hence the fourth component can be viewed as an index relating to the mayoral term.

In Figure 3-8, we use a color bar to indicate the crime rate. The lighter the color, the higher the crime rate.

In Figure 3, we find that San Francisco, Austin, Sanantonio and El Paso are isolated in the periphery while other cities gather in the middle. In Figure 4, we find that the cities isolated in the periphery are San Francisco, Austin, Sanantonio, El Paso and New York.

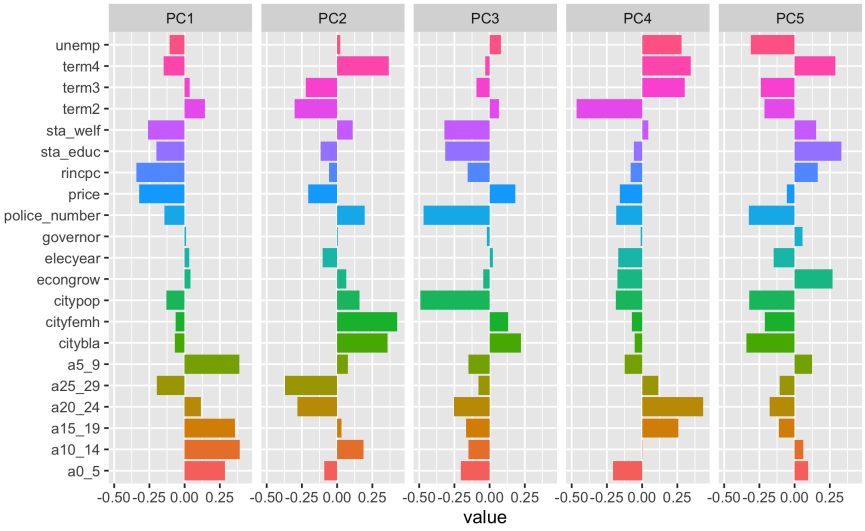

Figure 1: The first five principal components

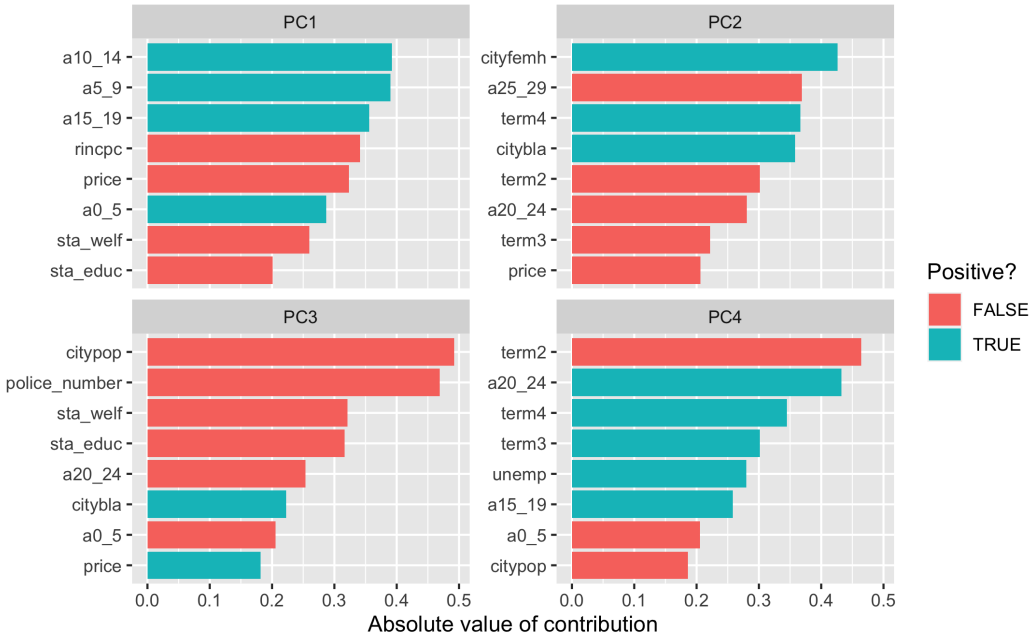

Figure 2: Indexes with large contributions to the first four principal components

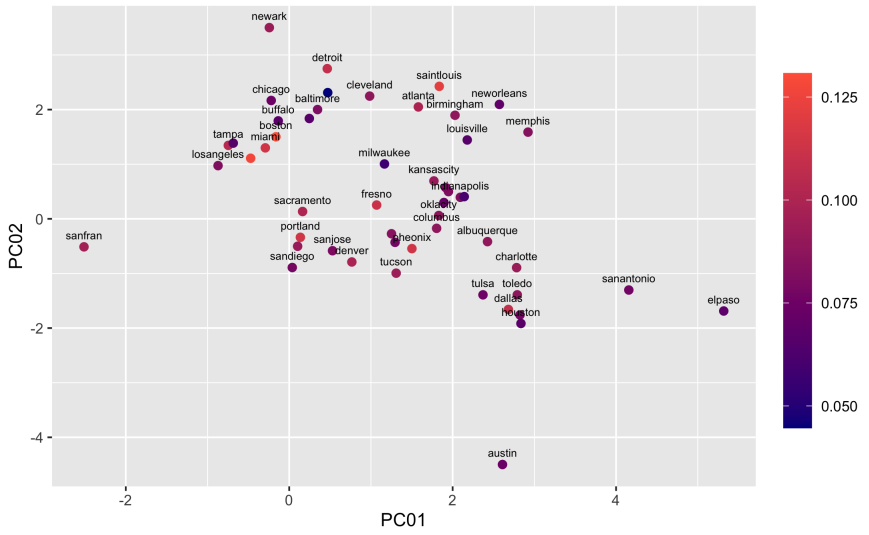

Figure 3: The centered and scaled crime data of 85 cities in 1975 projected onto the first principal component and the second principal component(instances of some cities are removed due to the incomplete data)

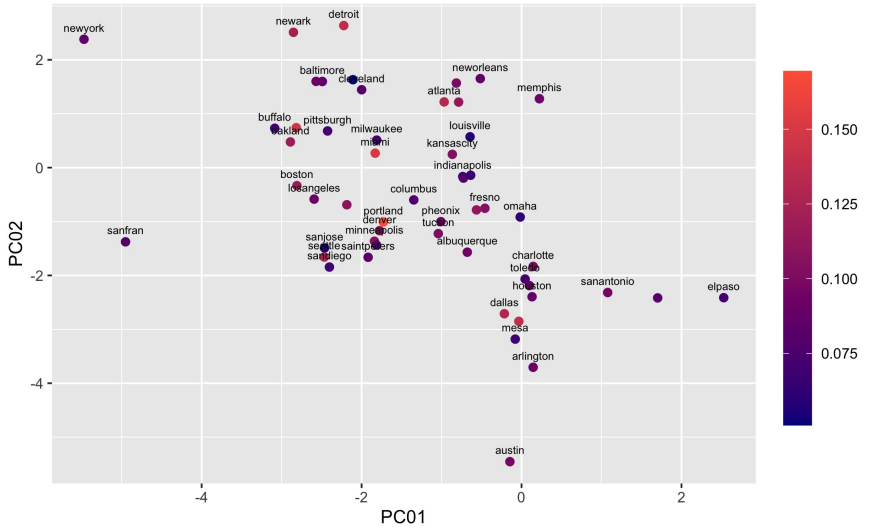

Figure 4: The centered and scaled crime data of 85 cities in 1985 projected onto the first principal component and the second principal component(instances of some cities are removed due to the incomplete data)

Table 1: Variance ratios

number	variance ratio	number	variance ratio
1	0.132386	12	0.043186
2	0.103446	13	0.034303
3	0.085942	14	0.028599
4	0.080506	15	0.024378
5	0.073758	16	0.016027
6	0.066845	17	0.013963
7	0.064977	18	0.009195
8	0.060950	19	0.006385
9	0.052591	20	0.006098
10	0.049650	21	6.412483e-17
11	0.046814		

In Figure s-1 and Figure s-2, we find that the coordinates of the instances under the first and the second principal components decrease as time goes on. The crime rate increases as time goes on. We conjecture that this is due to demoralization and economic depression.

We then conduct Isomap and UMAP. Figure 5, Figure s-3 and Figure s-4 show all the instances locations under the first and the second components of PCA, Isomap and UMAP, respectively. We can see that Isomap and UMAP perform better than PCA in data reduction, which is consistent with the theory. Especially, UMAP separates the 85 cities into 5 groups remarkably.

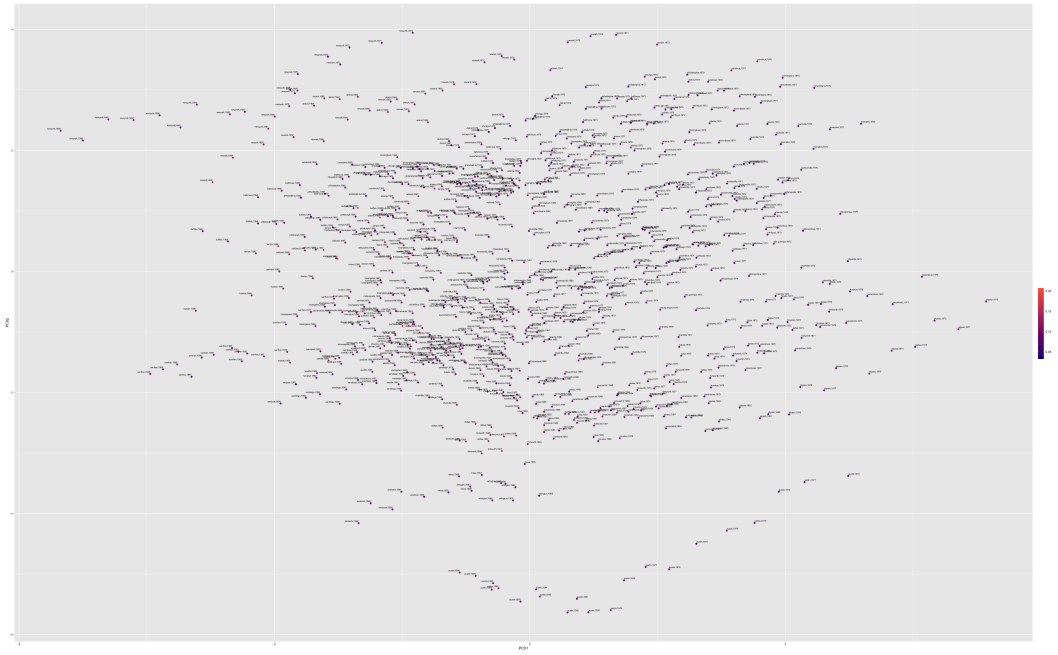

Figure 5: The centered and scaled crime data of 85 cities from 1969-1992 projected onto the first component and the second component of PCA(instances of some cities and years are removed due to the incomplete data)

The scatter diagrams in different planes of the three methods are plotted in Figure 6, 7 and 8. We find that Isomap and UMAP perform better than PCA once again.

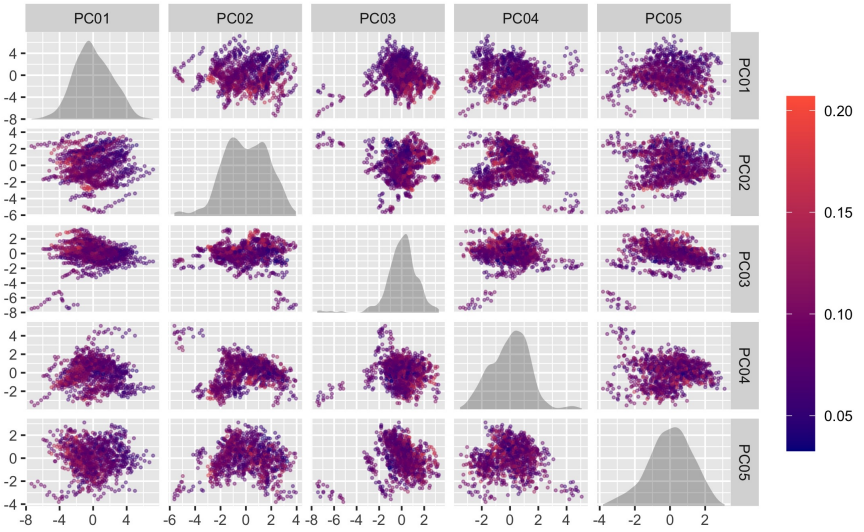

Figure 6: The scatter diagram of centered and scaled crime data of 85 cities from 1969-1992 in different planes of PCA(the diagonal represents the density function of the corresponding component)

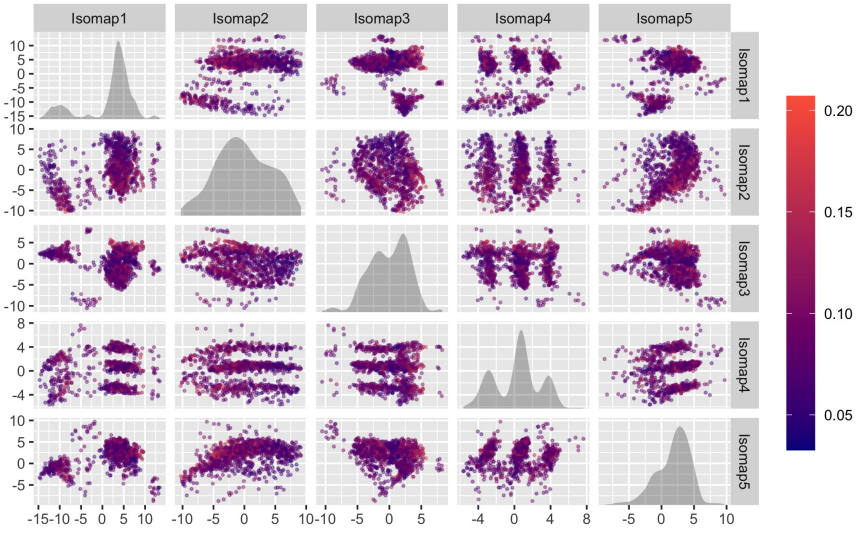

Figure 7: The scatter diagram of centered and scaled crime data of 85 cities from 1969-1992 in different planes of Isomap(the diagonal represents the density function of the corresponding component)

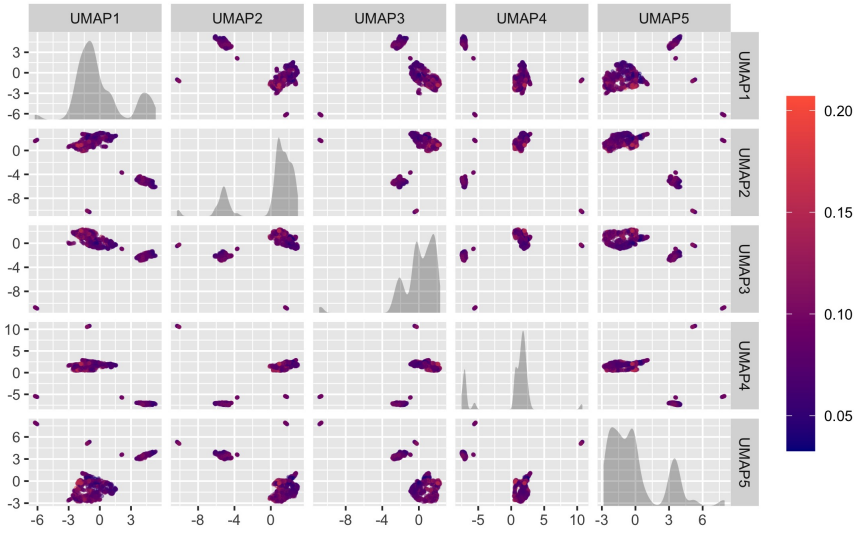

Figure 8: The scatter diagram of centered and scaled crime data of 85 cities from 1969-1992 in different planes of UMAP(the diagonal represents the density function of the corresponding component)

3.3 Discussion

Though nonlinear data reduction methods may give us more comprehensive results, there still exists some shortcomings.

As for Isomap, it may encounter "short-circuit errors" if the connectivity of each data point in the neighborhood graph k is too large with respect to the manifold structure or if the noise in the data moves the points slightly off the manifold. But if k is too small, the neighborhood graph may become too sparse to approximate geodesic paths accurately. So an appropriate k must be chosen carefully.

As for UMAP, Tara Chari[1] deemed that while the popular UMAP was intended to faithfully represent the local and global structure of high-dimensional data in two or three dimensions, there is evidence they fail in some situations. Because theorems of UMAP providing guarantees on the embeddings rely on numerous assumptions, which are unlikely to hold in practice and ignore the coupling of PCA to nonlinear methods[2].

4 Linear Regression Analysis

In this section, we will use the linear regression model to predict the average crime rate. Since we have eliminated the colinearity in the dataset, there will be less obstacles to use linear regression ($X^{\mathrm{T}}X$ now is invertible). There are four steps in general to conduct a linear regression model:

- Estimate model parameters using Least squares estimation. $\hat{\beta} = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}Y$, where $\hat{\beta}$ is the estimation of model parameters, X and Y represent samples and response (crime rate) respectively.
- Determine the distributions of parameter Estimator. $E\hat{\beta} = \beta, Cov(\hat{\beta}, \hat{\beta}) = \sigma^2\left(X^{\mathrm{T}}X\right)^{-1}$
- Evaluate performance statistics. For example, we can use the standard error of the regression model: $S_{yx} = \sqrt{\frac{SSE}{n-m-1}}$, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Or we can use R^2 goodness of fit: $S_{yx} = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, $SSR = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y}_i)^2$ $SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$.
- Test for Statistical Significance of Factors. We can construct a statistic: $T_j = \frac{\hat{\beta}_j}{\sqrt{C_{jj}Q/(n-m-1)}} \sim t(n-m-1)$ to conduct hypothesis test.

The results of linear regression with three data reduction methods are shown in the following tables.

Table 2: Linear regression with 5 PCA components

variable	PC1	PC2	PC3	PC4	PC5
coefficient	-0.0046899	-0.0015284	0.0045568	-0.0007056	-0.0015472

Table 3: Linear regression with 11 PCA components

variable	PC1	PC2	PC3	PC4	PC5	PC6
coefficient	-0.0046899	-0.0015284	0.0045568	-0.0007056	-0.0015472	0.0008104
variable	PC7	PC8	PC9	PC10	PC11	
coefficient	0.0005352	0.0049271	-0.0029860	0.0008374	-0.0004754	

Table 4: Linear regression with 5 UMAP components

variable	UMAP1	UMAP2	UMAP3	UMAP4	UMAP5
coefficient	0.0076540	-0.0021127	-0.0061266	-0.0008989	0.0027652

The result of linear regression without data reduction methods are shown in Table 5. The coefficient of "term4" is "NA" due to the collinearity of the high-dimensional data. We can see that linear regression equipped with a data reduction method doesn't suffer from such a problem.

Table 5: Linear regression without data reduction

variable	coefficient	variable	coefficient
elecyear	2.717e-04	a10_14	5.941e-01
governor	-4.597e-04	a15_19	-1.117e+00
rincpc	3.038e-03	a20_24	2.993e-01
econgrow	-1.006e-01	a25_29	3.684e-01
unemp	1.627e-01	citybla	2.959e-04
citypop	-4.943e-09	cityfemh	5.630e-04
term2	2.368e-03	sta_educ	3.925e-05
term3	-9.758e-03	sta_welf	-7.773e-05
term4	NA	price	1.492e-03
a0_5	6.069e-02	police_number	3.758e-07
a5_9	-5.057e-01		

5 Conclusion

In this project, we apply several data reduction methods, i.e., PCA, Isomap, UMAP, to handle the US crime data sets and successfully transform the high-dimensional problem into a lower-dimensional one. Specifically, we use the three methods to reduce the number of variables from 21 to 5 and 11. From the visualization results, we find that the 85 cities are well separated under the new coordinates with respect to the three methods. We also find that Isomap and UMAP perform better than PCA. We then conduct a linear regression analysis to the low-dimensional data and find the independence of the crime rate in 85 US cities and the corresponding components of the three methods. In short, the US crime rate is affected by many factors, such as the population structure, police number, mayoral term, society welfare, etc.

Contributions of Group Members

LAI Yanming and YAN Bokai studied the dataset and conducted the data preprocessing. YAN Bokai wrote the codes and performed the numerical simulations. YAN Ningyu studied the background of data reduction methods. LAI Yanming and YAN Ningyu wrote the report.

Supplementary Materials

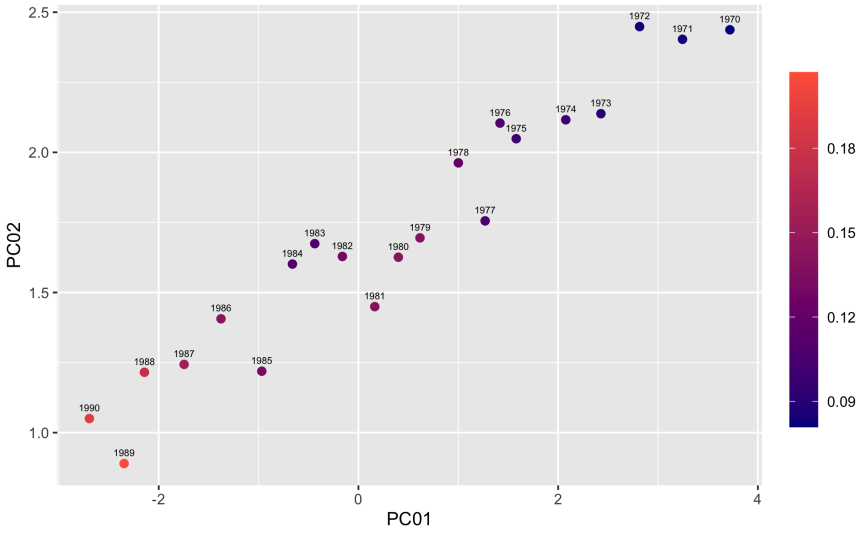

Figure s-1: The centered and scaled crime data of Atlanta from 1969-1992 projected onto the first principal component and the second principal component(instances of some years are removed due to the incomplete data)

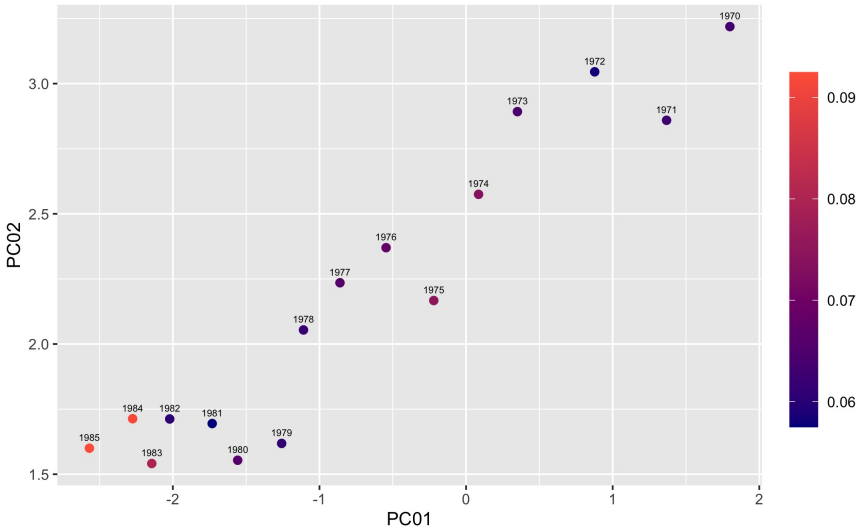

Figure s-2: The centered and scaled crime data of Chicago from 1969-1992 projected onto the first principal component and the second principal component(instances of some years are removed due to the incomplete data)

Figure s-3: The centered and scaled crime data of 85 cities from 1969-1992 projected onto the first component and the second component of Isomap(instances of some cities and years are removed due to the incomplete data)

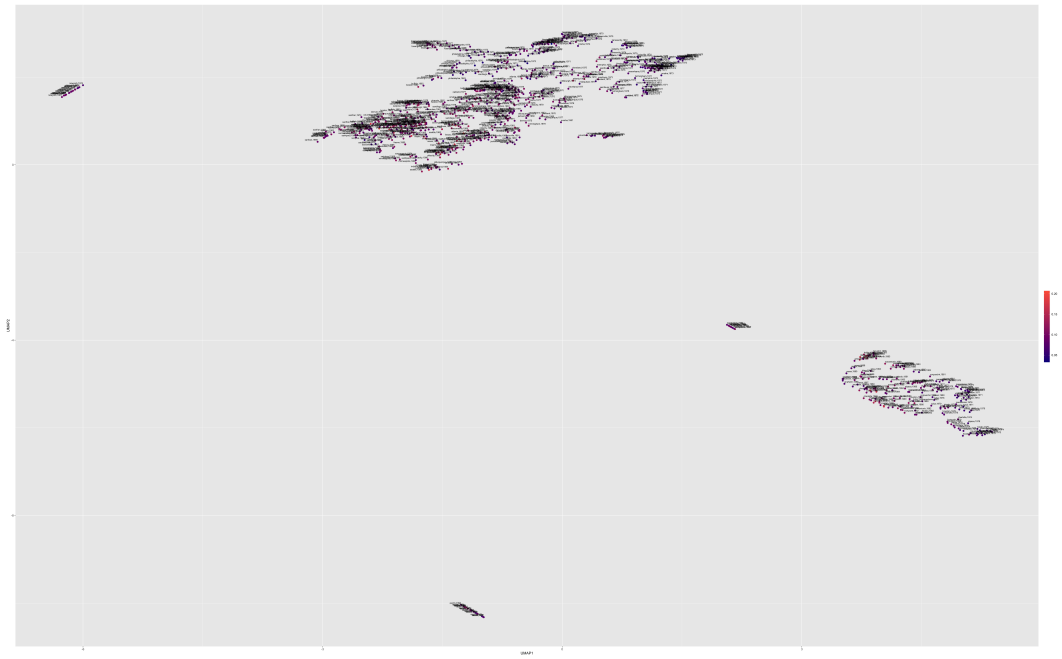

Figure s-4: The centered and scaled crime data of 85 cities from 1969-1992 projected onto the first component and the second component of UMAP(instances of some cities and years are removed due to the incomplete data)

References

[1] Tara Chari, Joeyta Banerjee, and Lior Pachter. The specious art of single-cell genomics. *BioRxiv*, pages 2021–08, 2021.

[2] George C Linderman and Stefan Steinerberger. Clustering with t-sne, provably. *SIAM Journal on Mathematics of Data Science*, 1(2):313–332, 2019.