Analysis of USArrest Dataset

Ayca Begum Tascioglu

February 2023

Data

The dataset used in this reports consists 50 rows, as states of USA, and corresponding observations for 4 variables, Murder, Assault, UrbanPop and Rape. The dataset contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973, also with the percent of population living in urban areas.

General Information about Data

Index: 50 entries, Alabama to Wyoming

Memory Usage: 2.0+ KB

Data Columns (total 4 columns)

Column	Non-Null Count	Data Type
Murder	50	float64
Assault	50	int64
UrbanPop	50	int64
Rape	50	float64

Table 1 : General Information about Data

	Murder	Assault	UrbanPop	Rape
Count	50	50	50	50
Mean	7.788	170.76	65.54	21.232
Standard Deviation	4.35551	83.337661	14.474763	9.366385
Variance	18.970465	6945.165714	209.518776	87.729159
Minimum	0.8	45	32	7.3
25%	4.075	109	54.5	15.075
50%	7.25	159	66	20
75%	11.25	249	77.75	26.175
Maximum	17.4	337	91	46

Table 2: Description of Data columns

As we can see the data in different columns have extremely different means and variances; if we do not perform scaling operation, most of the principal components will be driven by the variable with higher mean and variance.

	Murder	Assault	UrbanPop	Rape
Count	50	50	50	50
Mean	-7.105427e-17	1.387779e-16	-4.396483e-16	8.593126e-16
Standard Deviation	1.010153e+00	1.010153e+00	1.010153e+00	1.010153e+00
Variance	1.020408	1.020408	1.020408	1.020408
Minimum	-1.620693e+00	-1.524362e+00	-2.340661e+00	-1.502548e+00
25%	-8.611383e-01	-7.486054e-01	-7.704502e-01	-6.640245e-01
50%	-1.247758e-01	-1.425453e-01	3.210209e-02	-1.220847e-01
75%	8.029251e-01	9.483628e-01	8.521012e-01	5.330962e-01
Maximum	2.229265e+00	2.015028e+00	1.776781e+00	2.671197e+00

Table 3: Description of Data columns after Scaling

As shown in the table, now our data is more scaled in terms of mean and variance therefore we can observe each variable's effect on Principal Component Analysis fairly. Scaling can be performed by $\frac{df - df.mean}{df.std()}$ formula.

Pearson Correlation Coefficient is the measure of linear relationship between two sets of data. In this project, it is used to investigate correlations between columns.

Pearson Correlation Coefficient: $r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$

Where x_i is the values of x-variable in a sample, and $\bar{x}$ is the mean of the x variable, and y_i is the values of y-variable in a sample, and $\bar{y}$ is the mean of the y variable.

	Murder	Assault	UrbanPop	Rape
Murder	1	0.801873	0.069573	0.563579
Assault	0.801873	1	0.258872	0.665241
UrbanPop	0.069573	0.258872	1	0.411341
Rape	0.563579	0.665241	0.411341	1

Table 4: Pearson Correlation

By looking the correlation graph, we can conclude that three crimes, murder, assault and rape are correlated with each other, wheras urban population has no meaningful correlation with these crimes.

Dimensionality Reduction Techniques

Dimensionality Reduction Techniques are useful for transforming data from high dimensional space to low dimensional space; in that case we can still have meaningful features with reduced memory space and calculation in a time efficient way. When dimensionality increases, the space needed to perform an algorithm also increases proportionally which leads to the curse of dimensionality.

Dimensionality Reduction Techniques is an advantageous way to avoid the curse of dimensionality. Feature selection methods like filters, wrappers, embedded methods and feature extraction methods like Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Linear Discriminant Analysis (LDA) are the examples of Dimensionality Reduction Techniques. In this report, USArrest dataset is analyzed with SVD and PCA.

Singular Value Decomposition

By using SVD we can decompose a matrix into 3 matrices $U,\ S,\ and\ V$. If we have a $n\ \times\ p$ matrix A, we will end up with a complex unitary $n\ \times\ n$ matrix U, a rectangular diagonal $n\ \times\ p$ matrix S, and a complex unitary $p\ \times\ p$ matrix V.

SVD can be formulated as $A\ =\ USV^T$

$u,\ s,\ v\ =\ np.linalg.svd(scale(df),\ full\_matrices\ =\ True)$

$Variance\ =\ np.round(s\ **\ 2/np.sum(s\ **\ 2),\ decimals\ =\ 3)$

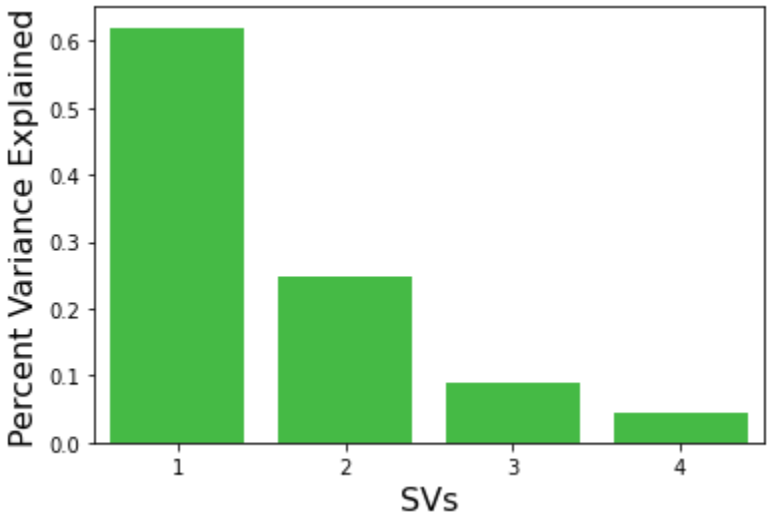

Figure 1: Percentage of variation in each column explained by each SVD

Principal Component Analysis [1],[2]

Principal Component Analysis is a linear dimensionality reduction method to avoid the curse of dimensionality while keeping meaningful features. In our project, we will try to inspect explained variance with principal components in order to understand correlations.

$def\ PCA(df)$:

$features\ =\ pd.DataFrame(scale(df),\ index\ =\ df.index,\ columns\ =\ df.columns)$

$covariance\ matrix\ =\ features.cov()$

$evalue,\ evect\ =\ np.linalg.eig(features.cov())$

$return\ evalue,\ evect,\ pd.DataFrame(evect.T.dot(features.T))$

	Murder	Assault	UrbanPop	Rape
Murder	1.020408	0.818238	0.070992	0.575080
Assault	0.818238	1.020408	0.264155	0.678818
UrbanPop	0.070992	0.264155	1.020408	0.429736
Rape	0.575080	0.678818	0.419736	2.020408

Table 5: Correlation Matrix

	PC1	PC2	PC3	PC4
Alabama	0.985566	1.133392	0.156267	-0.444269
Alaska	1.950138	1.073213	-0.438583	2.040003
Arizona	1.763164	-0.745957	-0.834653	0.054781
Arkansas	-0.141420	1.119797	-0.182811	0.114574
California	2.523980	-1.542934	-0.341996	0.598557
Colorado	1.514563	-0.987555	0.001465	1.095007
Connecticut	-1.358647	-1.088928	-0.118469	-0.643258
Delaware	0.047709	-0.325359	-0.881978	-0.718633

Florida	3.013042	0.039229	-0.096285	-0.576829
Georgia	1.639283	1.278942	1.076797	-0.342460
Hawaii	-0.912657	-1.570460	0.902807	0.050782
Idaho	-1.639800	0.210973	-0.499104	0.259801
Illinois	1.378911	-0.681841	-0.122021	-0.677496
Indiana	-0.505461	-0.151563	0.424666	0.228055
Iowa	-2.253646	-0.104054	0.017556	0.164564
Kansas	-0.796881	-0.270165	0.206496	0.025553
Kentucky	-0.750859	0.958440	0.670557	-0.028369
Louisiana	1.564818	0.871055	0.454728	-0.783480
Maine	-2.396829	0.376392	-0.330460	-0.065682
Maryland	1.763369	0.427655	-0.559070	-0.157250
Massachusetts	-0.486166	-1.474496	-0.179599	-0.609497
Michigan	2.108441	-0.155397	0.102372	0.384869
Minnesota	-1.692682	-0.632261	0.067317	0.153070
Mississippi	0.996494	2.393796	0.215508	-0.740808
Missouri	0.696787	-0.263355	0.225824	0.377444
Montana	-1.185452	0.536874	0.123742	0.246889
Nebraska	-1.265637	-0.193954	0.015893	0.175574
Nevada	2.874395	-0.775600	0.314515	1.163380
New Hampshire	-2.383915	-0.018082	-0.033137	0.036855
New Jersey	0.181566	-1.449506	0.243383	-0.764454
New Mexico	1.980024	0.142849	-0.339534	0.183692
New York	1.682577	-0.823184	-0.013484	-0.643075

North Carolina	1.123379	2.228003	-0.954382	-0.863572
North Dakota	-2.992226	0.599119	-0.253987	0.301277
Ohio	-0.225965	-0.742238	0.473916	-0.031139
Oklahoma	-0.311783	-0.287854	0.010332	-0.015310
Oregon	0.059122	-0.541411	-0.237781	0.939833
Pennsylvania	-0.888416	-0.571100	0.359061	-0.400629
Rhode Island	-0.863772	-1.491978	-0.613569	-1.369946
South Carolina	1.320724	1.933405	-0.131467	-0.300538
South Dakota	-1.987775	0.823343	-0.109572	0.389293
Tennessee	0.999742	0.860251	0.652864	0.188083
Texas	1.355138	-0.412481	0.643195	-0.492069
Utah	-0.550565	-1.471505	-0.082314	0.293728
Vermont	-2.801412	1.402288	-0.144890	0.841263
Virginia	-0.096335	0.199735	0.211371	0.011713
Washington	-0.216903	-0.970124	-0.220848	0.624871
West Virginia	-2.108585	1.424847	0.131909	0.104775
Wisconsin	-2.079714	-0.611269	0.184104	-0.138865
Wyoming	-0.629427	0.321013	-0.166652	-0.240659

Table 6: Principal Components for each state

$$Eigen\ values\ = array([2.53085875,\ 1.00996444,\ 0.17696948,\ 0.36383998])$$

	v1	v2	v3	v4
Murder	0.535899	0.418181	0.649228	-0.341233
Assault	0.583184	0.187986	-0.743407	-0.268148
UrbanPop	0.278191	-0.872806	0.133878	-0.378016
Rape	0.543432	-0.167319	0.089024	0.817778

Table 7: Eigen Vectors

	PC1	PC2	PC3	PC4
Explained Variance Ratio	0.62	0.24	0.089	0.043

Table 8: Explained Variance Ratio for each Principal Component

Since we scaled the data, we ended up with a covariance matrix, similar to correlation matrix. As we can conclude from the table, PC1 explains 62% of data, PC2 explains 24% of data, PC3 explains 8% of data and PC4 explains 4% of data.

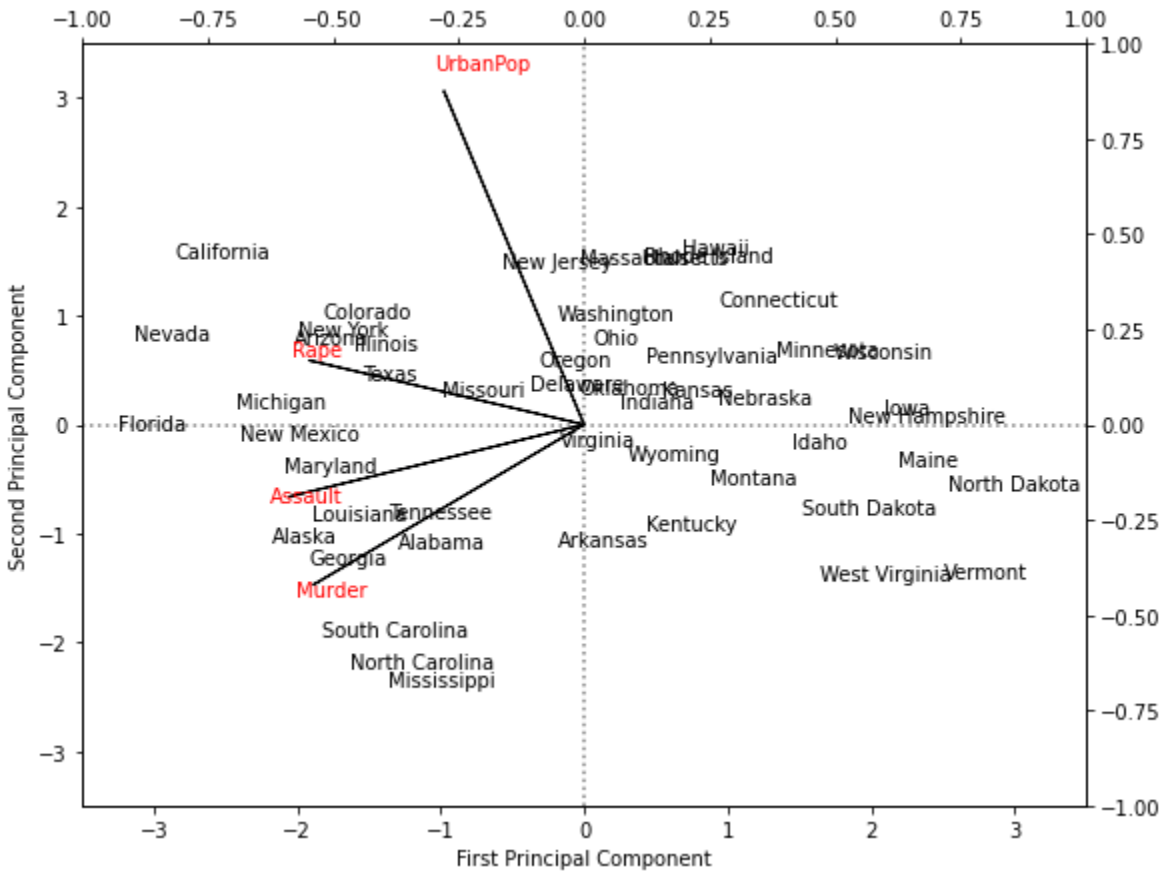

Figure 2 : Biplot of first two Principal Components(PC1 and PC2)

PC1 loading vector gives similar weights on Assault, Murder and Rape and less weight on urban population. By looking at the biplot, we can conclude that PC1 is a measure of overall rates of the serious crimes, Rape, Assault and Murder; whereas PC2 is the level of the urbanization of the state. Crime variables are related to each other, and they are located closer to each other. We can conclude that states with larger positive results on PC1 have higher crime rates whereas states with larger positive results on PC2 have higher urbanization rate.

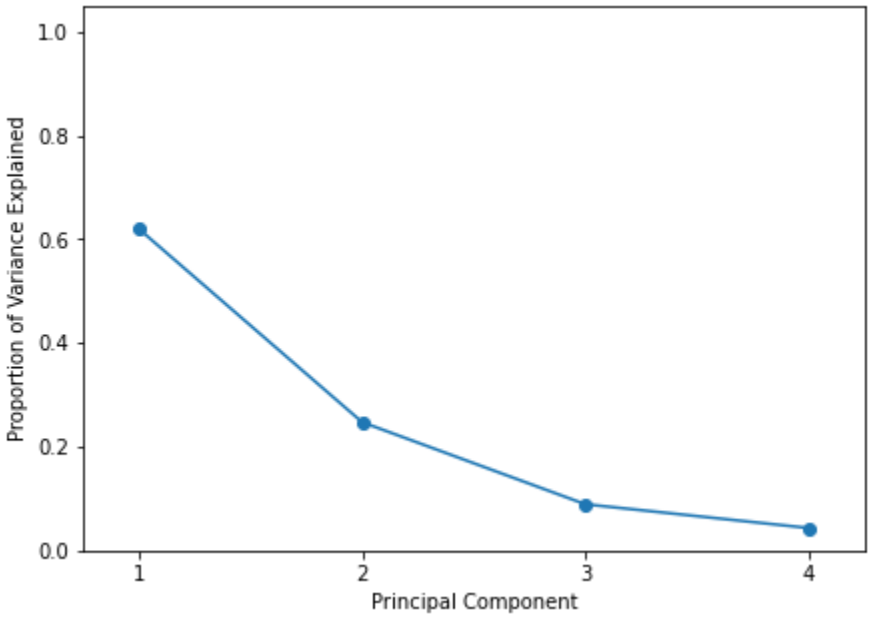

Figure 3: Proportion of Variance and Principal Components (PCA1, PCA2, PCA3, PCA4)

By looking at the elbow in the scree plot, and conditioning on 80% threshold, we can select first two Principal Components where summation of PC1 (62%), and PC2 (24%) is equal to 86%; which means they can explain 86% of variance in the dataset.

References

[1] “SDS 293 - machine learning.” [Online]. Available: https://www.science.smith.edu/~jcrouser/SDS293/. [Accessed: 22-Feb-2023].

[2] H. Goswami, “Understanding USARRESTS data using PCA.” [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/377338_75ed92a8463d482a80045abcae0e395d.html. [Accessed: 22-Feb-2023].