Erasmus University Rotterdam

Tatev Karen Aslanyan

Multivariate Statistics Case Study

Ranking U.S. cities based on a single linear combination of rating variables

---

1 Introduction

City ratings have been the interest of many studies for a long time as they represent the welfare of specific geographic areas and thus affect the attention of capital, urban growth and development. According to Rogerson (1999), by the introduction of "Places Rated Almanac" (Boyer and Savageau, 1985), city ratings started a new era with statistical rankings based on quality of life factors. Ever since, there have been many analyses using this data set. Becker et al. (1987) provides alternative ranking methods combined with population effects. Landis and Sawicki (1988) states that while "Places Rated Almanac" is useful for potential migrants, it's of little importance to local planning policy.
In our analysis, we try to find the rankings of the cities in United States based on a single combination of 9 rating variables using multivariate techniques: principal components analysis and factor analysis. Moreover, we will also use canonical correlation analysis to get more insight of this data and investigate the correlation between two sets of rating variables (if existing). We aim to find the linear combination of rating variables that would maximally explain the variation of the data and rank the U.S. cities according to this new rating criterion. The report is structured as follows. Section 2 briefly describe the data and data transformation. Section 3 explains the three methods mentioned above. Section 4 presents the analysis and the results. Finally, section 5 concludes our findings.

2 Data

The data"Places Rated Almanac" (Boyer and Savageau, 1985) contains 9 variables constructed for 329 metropolitan areas of the U.S. [1]. These composite variables are Climate, Housing Cost, Health care, Crime, Transportation, Education, Arts, Recreation and Personal Economy. Except for Housing Cost and Crime, higher scores of these variables represent better conditions with regard to the rating of that city. By Figure 1, the kernel densities for the distributions of nine variables have one peak and are mostly asymmetric. Since we are applying FA technique, we have to make sure that the assumptions of it are satisfied, that means that the data has to be symmetric. According to Mosteller and Tukey (1977), power transformations should be used to symmetrize these distributions. Therefore, we take log transformations for all variables except for Climate and Education because they are already symmetric. Even though PCA and CCA don't require this transformation but for comparison and consistency purposes we use the same transformed data for all approaches. Figure 2 shows that the transformed variables have symmetric distributions.[2]

[1] These areas are denoted by the U.S. Bureau of the Census.
[2] In the remaining parts of this analysis we will call the log-transformed variables by their original names.

3 Methods

3.1 Principal Component Analysis

We use Principal Component Analysis (PCA) to investigate which data fields in the city ranking determinants are responsible for the most variation in the cities ratings. With PCA, we aim to summarize our data and reduce the dimension of our problem while loosing as little information as possible. We first construct new independent variables which will be linear combinations of the original 9 rating variables (matrix X with covariance matrix Σ). We then calculate the eigenvalues (λ_1,λ_2,...,λ_p) and corresponding eigen vectors (e_1,e_2,$\cdots$,e_p) of Σ such that $\lambda_1 \geq \cdots \geq \lambda_p$. Then, the i^{th} principal component is:

$$Y_i = e_i^T X = e_{i1}x_1 + e_{i2}x_2 + \cdots e_{ip}x_p$$

Since "Places Rated Almanac" data consists of variables which are measured in different units and PCA is not scale-invariant, the scaling can effect the results. We will therefore use the correlation matrix instead of covariance matrix to avoid this indeterminacy.

3.2 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) aims to explore and measure association between two groups of variables. CCA might help us to identify how the ratings of cities based on first set of rating variables is correlated with ratings of the cities based on second set of rating variables. We put the first set of rating variables which in our opinion are closely related in X and we put the second sets of variables in the matrix Y. We define $\eta = a^T X$ and $\phi = b^T Y$ where a and b are vectors. With CCA we aim to find the these linear combinations such that the correlation between them (Cor(η,ϕ)) is maximized. So, instead of investigating the structure within rating variables group like PCA, CCA will explore the relation between two sets of rating variables.

3.3 Factor Analysis

Factor Analysis (FA) is another approach for reducing the dimension. With Exploratory FA we aim to describe the relationships among the variables without knowing in advance the number of factors. The model is: $\mathbf{X} - \mu = \mathbf{AF} + \epsilon$. X [pxN] is the matrix of 9 rating variables, μ is [pxN] population mean matrix, A [pxk] is common factor loadings matrix, F [kxN] is matrix of common factors and ϵ [pxN] is matrix of specific factors. Therefore, in our case FA can be seen as series of 9 regressions where we predict X_i's using unobserved common factors f_1,$\cdots$,f_m, that is:

$$X_i = \mu_i + a_{i1}f_1 + \cdots + a_{i1}f_m + \epsilon_i$$

The number of estimated parameters of covariance matrix of X is p*(p+1)/2 = 9*10/2 = 45. In FA model the number of parameter is p*m+p = 9(m+1). If m = 4, both number

of parameters are equal to 45 which is not a dimension reduction but still satisfies the identification criterion and if m = 3, 36 < 45 which does leads to dimension reduction. Therefore, we observe that m should be smaller than equal 4 to prevent unidentified case since but smaller than equal 3 to lead to dimension reduction.

4 Analysis and Results

Table 1 presents correlation matrix from which we see that Arts and Health Care have the highest correlation. Apart from that, Variable Arts also has some correlation with Housing Cost, and Transportation. These relations are verified by Figure 3.

4.1 Principal Component Analysis results

Figure 4 presents the biplot unscaled data from which we could conclude that Climate and Education contain the most variation but that is not the case and this result is consistent with the data transformation of Section 2, in which we didn't log-transform these two variables. Figure 5 presents biplot with scaled data and we observe something totally different: Crime and Education now have the smallest margins and the remaining variables show substantial variations. Table 2 we observe that the first three principal components explain 63.10% of the total variation in data. Applying the "elbow rule" for Figure 6 we see that we can optimally retain 3 components. The variances of 9 principal component are (**3.34 1.22 1.12** 0.91 0.82 0.55 0.49 0.31 0.25). So, when we apply the Kaiser rule and exclude principal components whose eigenvalues are less than average we again come to the same conclusion to retain 3 principal components. However, for the third approach which requires to include components which all together explain 90% of total variation" we have to retain 7 components. Since, two of the 3 criteria suggest the use of 3 principal components we have chosen to retain 3 principal components.

From Table 3 and Figure 7 we observe that first principal component corresponds to the variables Arts, Health Care, Housing Cost, Transportation and Recreation with all positive loadings. So, this component can be seen as a measure of the quality of Arts, Health Care, Transportation, Recreation, and Housing Cost. Second component corresponds mainly to Crime(positive), Education(negative) and Economy(positive). Finally, the third component corresponds to Climate(negative) and Economy(positive).

From Table 4 we observe that variables Arts, Health Care, Transportation, Housing Cost and Recreation are strongly and positively correlated with first principal component. From that we can conclude that areas with high level of arts have better health care system and also better recreation facilities but this comes at the cost of housing. Second component is positively correlated with Economy and Crime, negatively correlated with

Education. These suggest that cities with high income households usually experience high level of crime but low level of the education. From third component, which is mainly a measure of climate and economy, we observe that areas with high level of economy have also extreme weather conditions. Table 5 which presents the cities' ranks based on PC1 scores from highest to lowest. So, the cities like San-Francisco, New-York, Los-Angeles and Boston which are top-ranked cities have high level of arts and also have high level of health care, recreation and transportation but also high housing costs.

4.2 Canonical Correlation Analysis Results

From the correlation matrix, we see that Arts and Health Care are strongly correlated. Therefore, we assume that variables Arts and Health Care are in group X and the remaining variables are in group Y. We find two canonical correlations with corresponding values of 0.730 and 0.368 which suggests that there is a substantial correlation between these two groups X and Y. Since, the first canonical correlation is much larger than the second one, we can argue that the first one is important while second is not.

Table 7 presents the a_{ij} and b_{ij} coefficients corresponding to two canonical correlations. As argued above, we will look mainly at the first column of the table corresponding to the first canonical correlation. We observe that both Health Care and Arts are negatively correlated with Economy and positively correlated with everything else.

We can interpret η mainly as index of Arts and Health Care and ϕ as index of living conditions (housing cost, education, transportation, economy, ect.). Figure 8 shows clear correlation between X and Y scores for the first canonical correlation which is not the case in Figure 9 for the second factor, where we observe some correlation but not significant. This verifies earlier conclusion that the first canonical correlation is important but second one is not.

4.3 Factor Analysis Results

Table 8 and 9 present the FA results without and with varimax rotation, respectively. The latter has been done for better interpretation of the results by making some common factor loadings 0 and some approximately 1. First factor is mostly correlated with Arts and Health care but also with Education. So, it is mainly a measure of these three ratings and it distinguishes cities with high level of Health Care, Arts and Education. Second factor is highly correlated with Housing Cost, so it is mainly a measure of housing costs and distinguishes cities with very high housing costs. The third factors is highly correlated with Arts and Recreation. Thus, it is mainly a measure of these two variables.So, it distinguishes cities with high level of arts and recreation facilities.

One of the biggest advantages of FA over PCA is goodness of fit test which we can apply

to test the number of common factors what we want to retain in FA. As mentioned in Section 3.2, for identification purposes m need to be smaller or equal to 4. However, for all m's satisfying this criteria chi square statistics are very large (218.95, 134.44, 78.39 and 32.83) for m = 1,..,4 respectively. This means that we have to reject the null hypothesis in all cases [3] and we conclude that the FA model doesn't fit this data well and the assumed underlying factor model is wrong.

5 Conclusion

From Section 4.2, we observe that FA can't be adequately applied, this might be due to non-linearity of our data[4]. CCA results of Section 4.3 show that Arts and Health Care are negatively correlated with Economy and positively correlated with Climate, Housing Cost, Transportation, Education and Recreation. From PCA results in Section 4.1 we argue that first principal component explains the largest part of total variation (36.64%) and it represents a single linear combination of 9 original rating variables.

$$PC_1 = 0.469log(X_{Arts}) + 0.407log(X_{HlthCare}) + 0.387log(X_{HouseCost}) + 0.370log(X_{Transp}) + 0.351log(X_{Recreat}) + 0.277X_{Educ} + 0.253log(X_{Crime}) + 0.195X_{Climate} + 0.160log(X_{Econ})$$ [5]

From this we conclude that the most weighted rating criterion corresponds to Arts with museums, public radio stations, entertainment and knowledge environments. Arts is followed by Health Care, Housing Cost, Transportation, Recreation and positively related to them. When we compare the Figure 8 of CCA and Table 5 of PCA we observe some similarities. We see that top-ranked cities based on PC1 in the table are all in the right top corner of the graph of CCA. Summarizing, the top ranked cities like San-Francisco, New-York, Los-Angeles (Table 5) have high level of arts but also high level of health care, transportation and recreation facilities. However, these cities have also high level of housing cost while for the bottom ranked cities like Texarkana, Dothgan and Gadsden hold the exact opposite. In these cities the quantity and quality of museums, theaters and other entertainment and knowledge facilities is very low. Moreover, the level of health care, transportation, and recreation facilities is also very low. Finally, we expect that in these cities the housing cost will be low.

[3] Likelihood Ratio Test: $H_0 : \hat{\Sigma} = \hat{A}\hat{A}^T + \hat{\Phi}$

[4] One of the assumptions of FA is then violated.

[5] This is the result using PCA with scaling so the inputs here are all scaled variables

Appendix

Appendix 1: Graphs and Tables

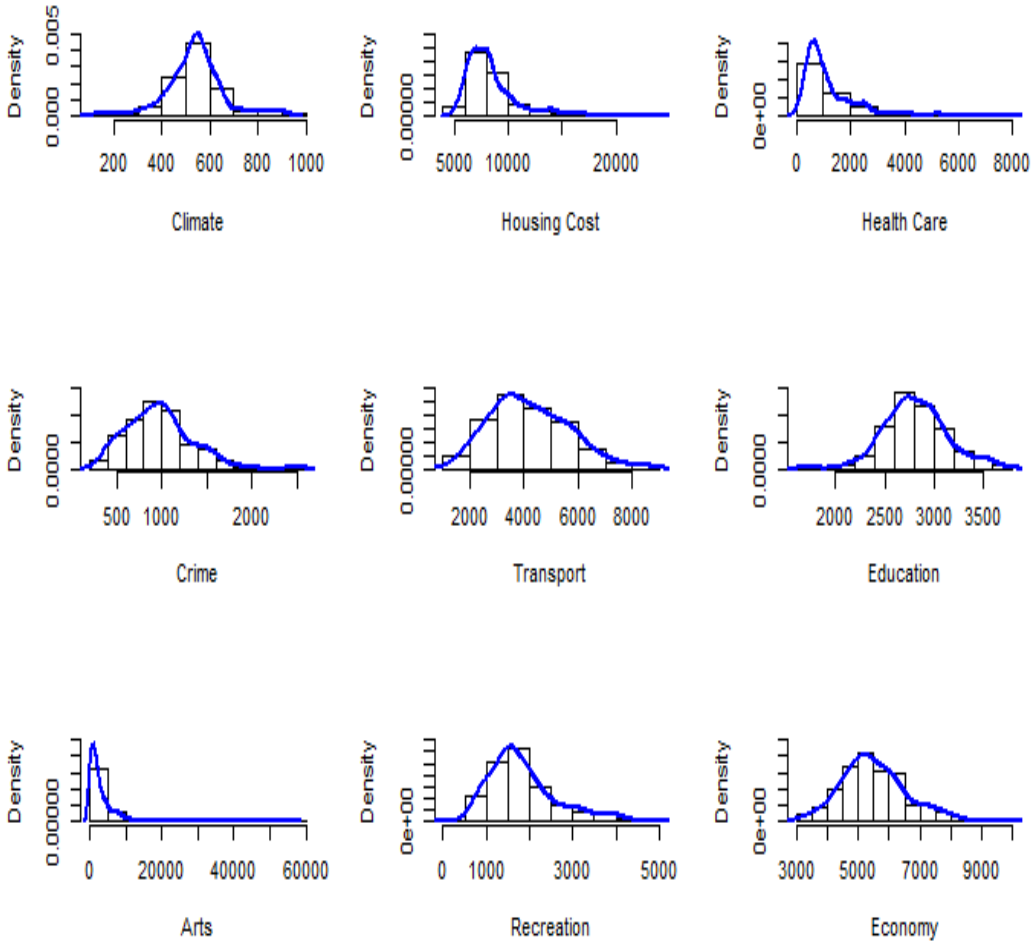

Figure 1: The Kernel densities for the distributions of the nine variables

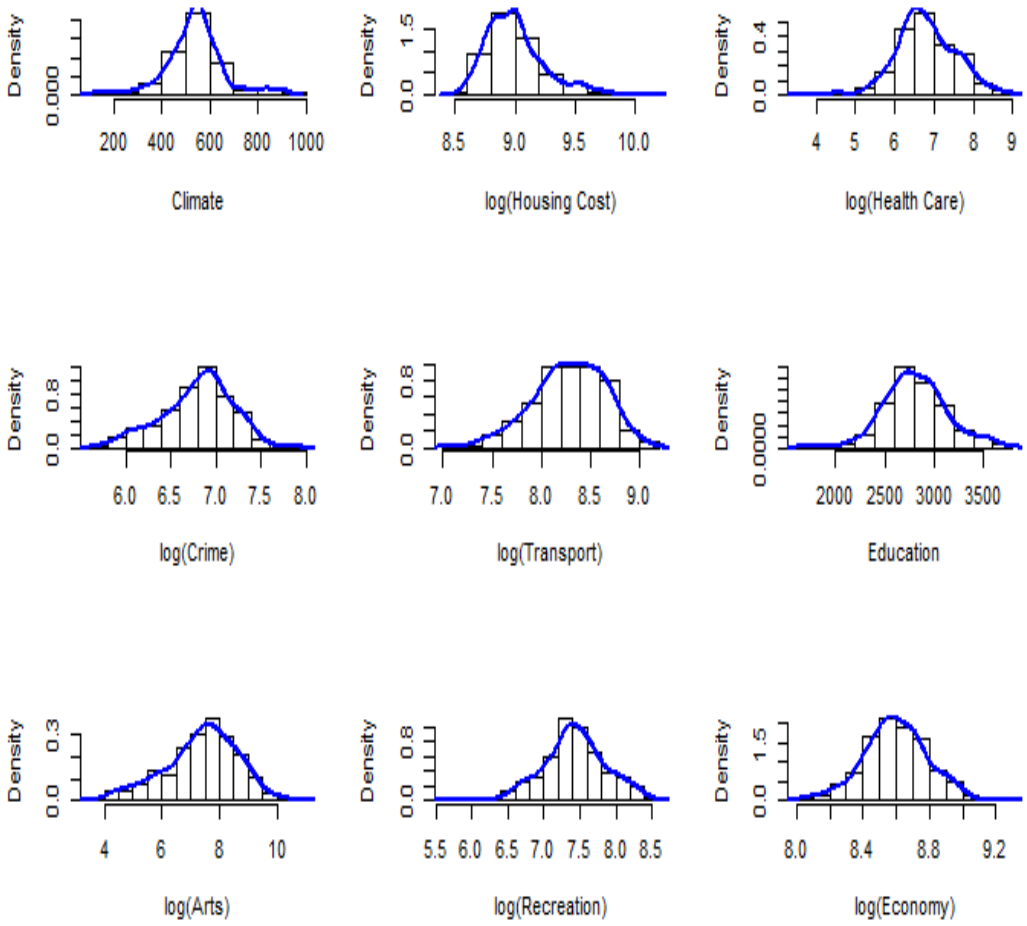

Figure 2: The Kernel densities of the distributions of the log-transformed variables

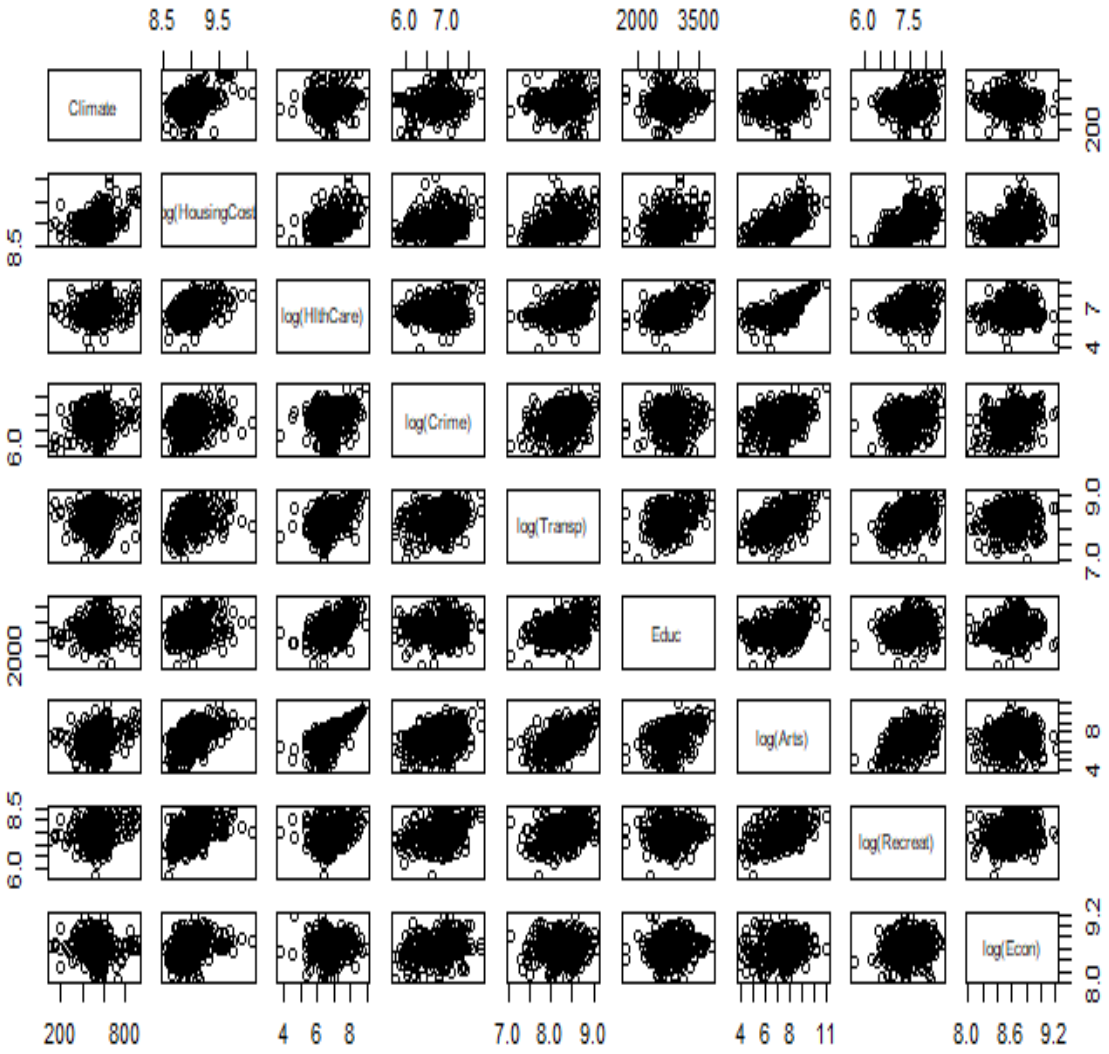

Figure 3: Scatterplot matrix displays showing scatterplots of all pairwise combination of the transformed data

	Climate	Housing Cost	Health Care	Crime	Transport	Education	Arts	Recreation	Economy
Climate	1	0.364	0.181	0.202	0.065	0.065	0.218	0.176	-0.092
Housing Cost	0.364	1	0.432	0.139	0.318	0.217	**0.509**	0.461	0.297
Health Care	0.181	0.432	1	0.184	0.419	0.480	**0.678**	0.254	0.054
Crime	0.202	0.139	0.184	1	0.274	0.056	0.347	0.292	0.276
Transport	0.065	0.318	0.419	0.274	1	0.320	**0.548**	0.391	0.063
Education	0.065	0.217	0.480	0.056	0.320	1	0.348	0.100	0.134
Arts	0.218	**0.509**	**0.678**	0.347	**0.548**	0.348	1	0.497	0.139
Recreation	0.176	0.461	0.254	0.292	0.391	0.100	0.497	1	0.176
Economy	-0.092	0.297	0.054	0.276	0.063	0.134	0.135	0.176	1

Table 1: The correlation matrix of log transformed data

Components	Eigenvalue	Proportion	Cumulative
1	3.345	0.372	0.372
2	1.215	0.135	0.507
3	1.119	0.124	**0.631**
4	0.907	0.101	0.732
5	0.817	0.091	0.823
6	0.549	0.061	0.884
7	0.487	0.054	0.938
8	0.309	0.034	0.972
9	0.250	0.028	1

Table 2: The results of the principal components analysis

Variable	PC1	PC2	PC3
Climate	0.194	0.114	**-0.737**
Housing Cost	**0.387**	0.151	-0.180
Health Care	**0.407**	-0.377	0.029
Crime	0.253	**0.469**	0.062
Transport	**0.370**	-0.147	0.134
Education	0.277	**-0.467**	0.302
Arts	**0.469**	-0.104	-0.006
Recreation	**0.351**	0.300	-0.092
Economy	0.160	**0.511**	**0.549**

Table 3: The variable loadings of the first three principal components

Variable	PC1	PC2	PC3
Climate	0.356	0.126	**-0.780**
Housing Cost	**0.708**	0.166	-0.191
Health Care	**0.743**	-0.416	0.031
Crime	0.463	**0.517**	0.066
Transport	**0.677**	-0.162	0.142
Education	0.507	**-0.515**	0.319
Arts	**0.859**	-0.115	-0.007
Recreation	**0.642**	0.330	-0.097
Economy	0.292	**0.564**	**0.581**

Table 4: The correlation between the first three components and the nine variables

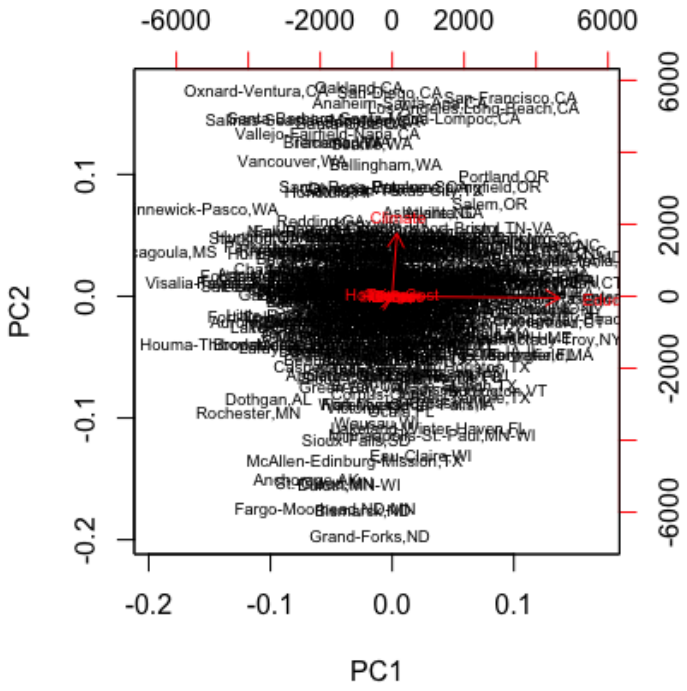

Figure 4: Biplot of PCA without scaling

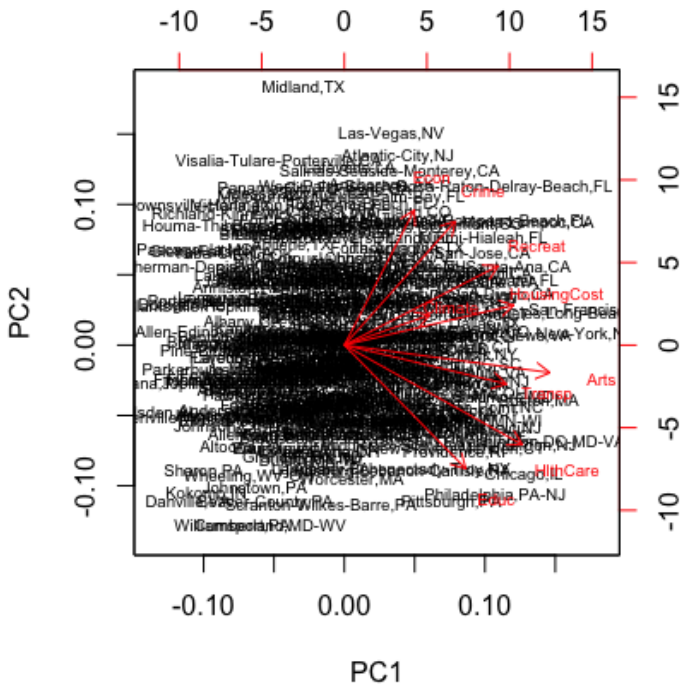

Figure 5: Biplot of PCA with scaling

Rank	City	PC1 score
1	San-Francisco,CA	5.924
2	New-York,NY	5.693
3	Los-Angeles,Long-Beach,CA	4.992
4	Boston,MA	4.593
5	Washington,DC-MD-VA	4.581
6	Chicago,IL	4.238
7	Baltimore,MD	3.770
8	San-Diego,CA	3.711
9	Seattle,WA	3.607
10	Philadelphia,PA-NJ	3.555
11	Newark,NJ	3.423
12	Denver,CO	3.408
13	Oakland,CA	3.396
14	Dallas,TX	3.310
15	Miami-Hialeah,FL	3.279
...	...	...
...	...	...
...	...	...
...	...	...
325	Steubenville-Weirton,OH-WV	-3.543
326	Danville,VA	-3.690
327	Texarkana,TX-Texarkana,AR	-3.702
328	Dothgan,AL	-3.848
329	Gadsden,AL	-4.531

Table 5: Rankings of cities with PC1 scores

Rank	City	PC2 score
1	Midland,TX	3.657
2	Las-Vegas,NV	3.013
3	Atlantic-City,NJ	2.681
4	Visalia-Tulare-Porterville,CA	2.624
5	Lafayette,LA	2.525
6	Salinas-Seaside-Monterey,CA	2.476
7	Anchorage,AK	2.276
8	West-Palm-Beach-Boca-Raton-Delray-Beach,FL	2.236
9	Panama-City,FL	2.223
10	Fresno,CA	2.205
11	Odessa,TX	2.169
12	Stockton,CA	2.154
13	Melbourne-Titusville-Palm-Bay,FL	2.107
14	Daytona-Beach,FL	1.999
15	Victoria,TX	1.989
...	...	...
...	...	...
...	...	...
...	...	...
325	Beaver-County,PA	-2.223
326	Pittsburgh,PA	-2.229
327	Scranton-Wilkes-Barre,PA	-2.277
328	Cumberland,MD-WV	-2.559
329	Williamsport,PA	-2.568

Table 6: Rankings of cities with PC2 scores

Group	Variable	CC1	CC2
X	Health Care	-0.387	1.803
	Arts	-0.524	-0.886
Y	Climate	-3.129	0.000
	Housing Cost	-1.784	1.313
	Crime	-6.159	-0.416
	Transport	-1.004	-0.459
	Education	-1.082	0.002
	Recreation	-4.829	-1.484
	Economy	6.925	-1.238

Table 7: Coefficients of Canonical Correlation Analysis

Variable	Factor 1	Factor 2	Factor 3
Climate	0.365		
Housing Cost	**0.997**		
Health Care	0.447	**0.742**	-0.260
Crime	0.147	0.282	0.397
Transport	0.329	0.475	0.253
Education	0.226	0.443	-0.183
Arts	**0.524**	**0.671**	0.214
Recreation	0.467	0.223	0.459
Economy	0.297		0.159

Table 8: The results of Factor Analysis without rotation

Variable	Factor 1	Factor 2	Factor 3
Climate	0.123	0.337	
Housing Cost	0.286	**0.949**	0.109
Health Care	**0.870**	0.172	0.176
Crime			0.497
Transport	0.367	0.168	0.485
Education	**0.520**		
Arts	**0.606**	0.288	**0.566**
Recreation	0.102	0.390	**0.563**
Economy		0.302	0.153

Table 9: The results of Factor Analysis with varimax rotation

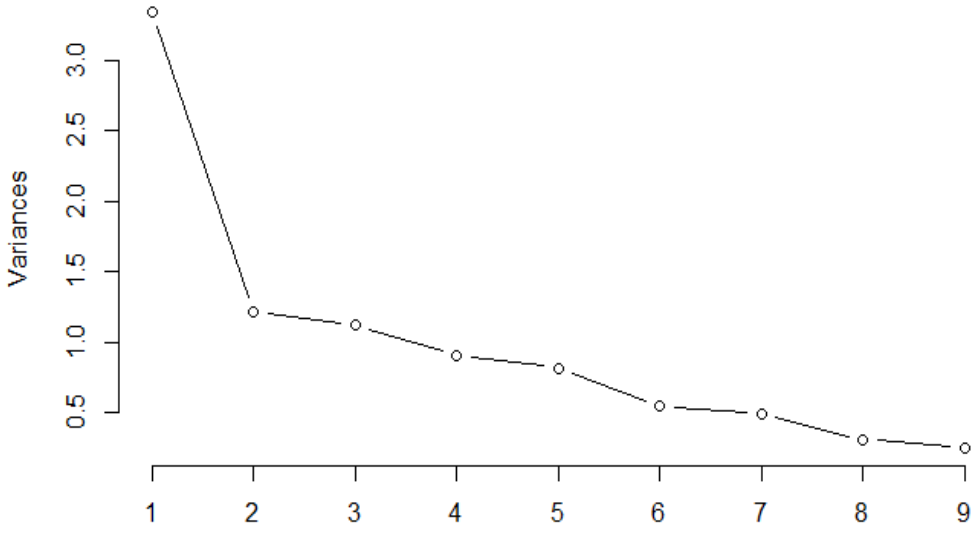

Figure 6: Scree plot: The variances explained by 9 PC's

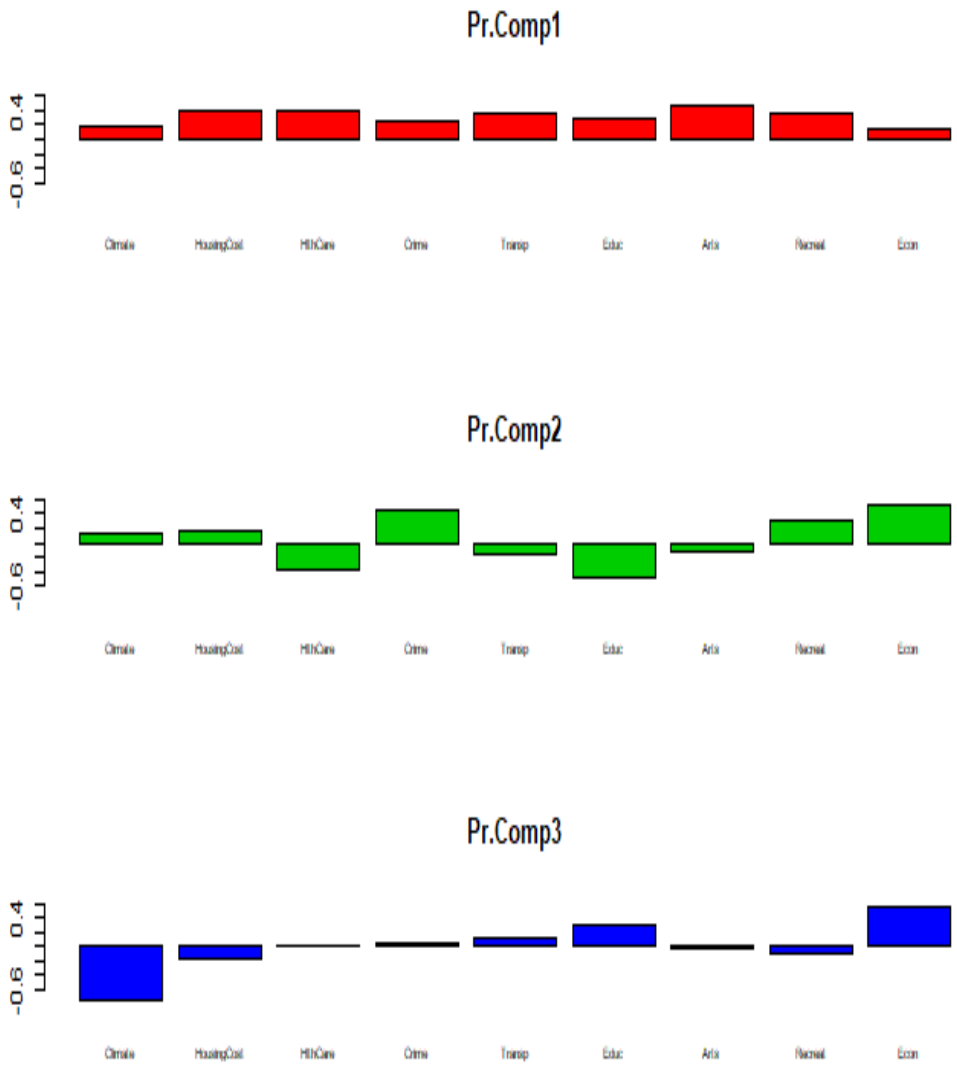

Figure 7: Barplots of 3 Principal Components

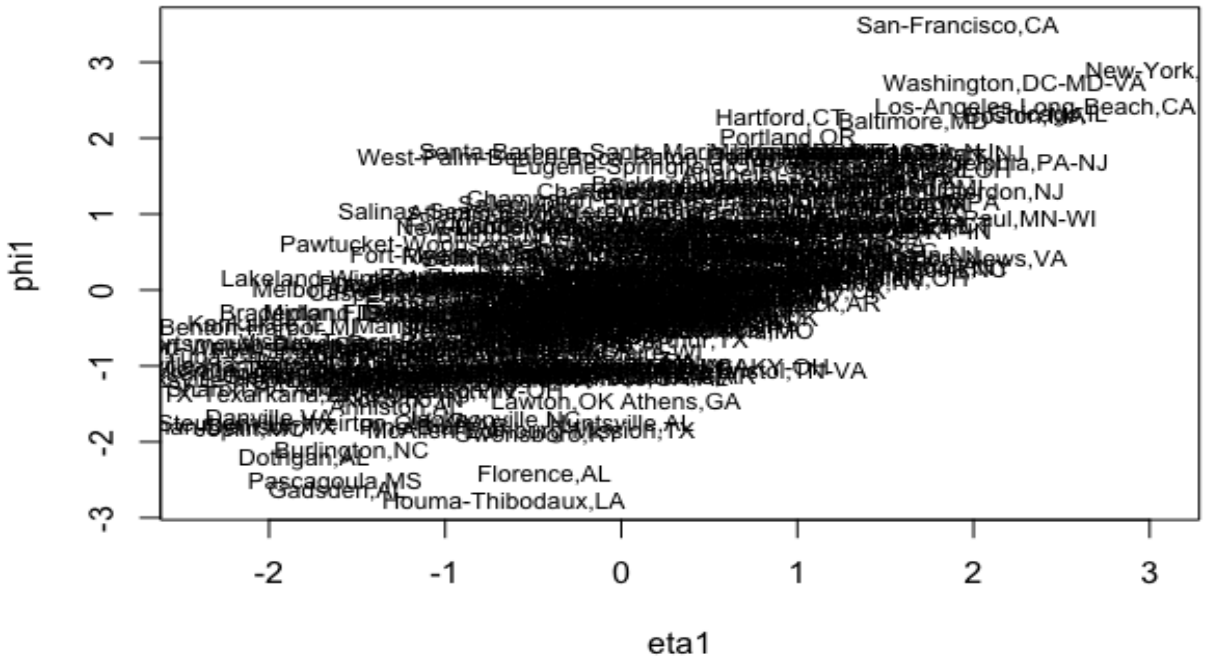

Figure 8: Canonical Correlation results 1

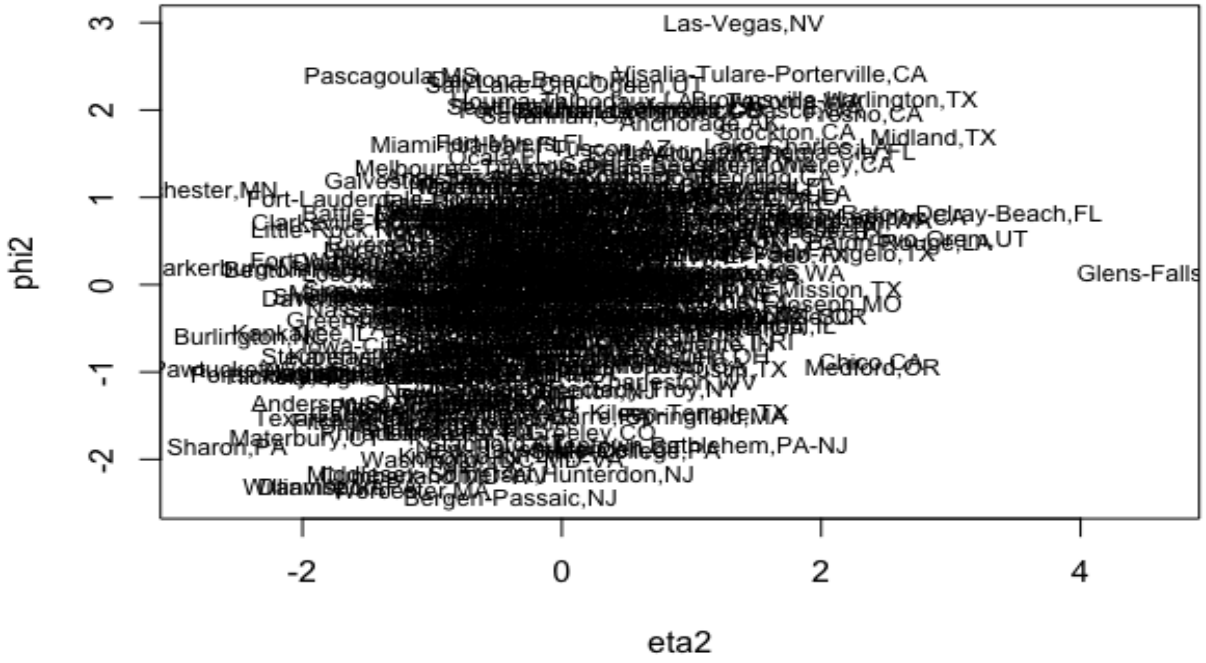

Figure 9: Canonical Correlation results 2

References

[1] Becker, R. A., Denby, L., McGill, R., and Wilks, A. R. (1987). Analysis of Data from the Places Rated Almanac. *The American Statistician, 41(3)*, 169–186. doi:10.2307/2685098

[2] Boyer, R. and Savageau D. (1981). *Places Rated Almanac: Your Guide to Finding the Best Places to live in America*. Chicago: Rand McNally.

[3] Johnson, R. A. and Wichern, D. W.(2007). *Applied Multivariate Statistical Analysis* (6th ed.), 430–480, New Jersey: Pearson Education

[4] Johnson, R. A. and Wichern, D. W.(2007). *Applied Multivariate Statistical Analysis* (6th ed.), 481–538, New Jersey: Pearson Prentice Hall.

[5] Landis, J. D. and Sawicki, D. S. (1988). A planner's guide to the Places Rated Almanac. *Journal of the American Planning Association, 54(3)*, 336–346.

[6] Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression* , Reading, MA: Addison-Wesley.

[7] Pierce, R. M. (1984), Rating America's Metropolitan Areas, *American Demographics, 7(7)*, 20–25.

[8] Rogerson, R. J. (1999). Quality of Life and City Competitiveness. *Urban Studies, 36(5–6)*, 969–985.