PCA on Finnish Municipalities

Dimensionality reduction with PCA on Finnish municipality demographics

Niko Miller

10.04.2022

Contents

1 Introduction

1.1 Data and Research Questions

Tilastokeskus provides data on 32 demographic variables of Finnish municipalities. The data can be freely obtained from Tilastokeskus' website in csv. format [Statistics Finland, 2022].

The variables provide information on e.g., the municipalities' population and its growth, age structure, economic sector structure, and the split between different housing types.

The purpose of this study is to reduce dimensionality in the data and assess whether differences between the municipalities can be explained by considerably less than 32 dimensions - namely by first few principal components (PCs). This study's contribution can be seen as exploratory data analysis where the municipalities are clustered in a lower dimension. The results can be applied in further work in e.g., more advanced clustering, prediction or economic decision making.

Clustering municipalities with demographic factors can be useful in a business context, e.g., marketing. Knowing how municipalities differ in demographics can help a company to improve its targeting and e.g., launch different marketing campaigns or product lines in different municipalities.

2 Univariate Analysis

2.1 Variable Descriptions

Table 1. *Variable Descriptions*

Original	**Modified**	**Explanation**
Taajama-aste, %	Prop.Urban.Areas	NA
Väkiluku	Pop	NA
Väkiluvun muutos edellisestä vuodesta, %	Pop.Change	NA
Alle 15-vuotiaiden osuus väestöstä, %	Prop.Below15	NA
15-64 -vuotiaiden osuus väestöstä, %	Prop.15to64	NA
Yli 64-vuotiaiden osuus väestöstä, %	Prop.Over64	NA
Ruotsinkielisten osuus väestöstä, %	Prop.Swedish	NA
Ulkomaan kansalaisten osuus väestöstä, %	Prop.Foreign	NA
Syntyneiden enemmyys, henkilöä	Pop.Growth	NA
Kuntien välinen muuttovoitto/-tappio, henkilöä	Migr.Gain	NA
Perheiden lukumäärä	Families	NA
Asuntokuntien lukumäärä	Households	NA
Rivi- ja pientaloissa asuvien asuntokuntien osuus, %	Prop.Households.rowSmall	NA
Vuokra-asunnoissa asuvien asuntokuntien osuus, %	Prop.Households.Rent	NA
Vähintään toisen asteen tutkinnon suorittaneiden osuus 15 vuotta täyttäneistä, %	Prop.Educ.Degree2	NA
Korkea-asteen tutkinnon suorittaneiden osuus 15 vuotta täyttäneistä, %	Prop.Educ.High	NA
Alueella asuvan työllisen työvoiman määrä	Employed	NA
Työllisyysaste, %	Empl.Rate	NA
Asuinkunnassaan työssäkäyvien osuus, %	Prop.Households.Empl	NA
Työttömien osuus työvoimasta, %	Prop.Unempl	NA
Eläkeläisten osuus väestöstä, %	Prop.Pension	NA
Taloudellinen huoltosuhde	Support.Ratio	NA
Alueella olevien työpaikkojen lukumäärä	Jobs	NA
Alkutuotannon työpaikkojen osuus, %	Prop.Primary.Sector	NA
Jalostuksen työpaikkojen osuus, %	Prop.Secondary.Sector	NA
Palvelujen työpaikkojen osuus, %	Prop.Services.Sector	NA
Työpaikkaomavaraisuus	Jobs.Self.Suff	NA
Vuosikate, euroa/asukas	Margin.Citizen	NA
Lainakanta, euroa/asukas	Loan.Citizen	NA
Konsernin lainakanta, euroa/asukas	Concern.Loan.Citizen	NA
Opetus- ja kulttuuritoiminta yhteensä, nettokäyttökustannukset, euroa/asukas	Educ.Cult.Citizen	NA
Sosiaali- ja terveystoiminta yhteensä, nettokäyttökustannukset, euroa/asukas	Soc.Health.Citizen	NA

2.2 Descriptive Statistics

Table 2. *Descriptive Statistics*

Var	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Std.
Prop.Urban.Areas	0.0	47.2	60.8	61.7	77.3	100.0	22.2
Pop	105.0	2673.8	6134.0	18039.0	15145.2	658457.0	49728.7
Pop.Change	-4.8	-1.2	-0.6	-0.5	0.2	4.0	1.2
Prop.Below15	4.0	12.1	14.4	14.8	16.9	30.8	3.9
Prop.15to64	46.5	52.3	55.5	55.6	58.4	68.1	4.3
Prop.Over64	10.8	24.8	29.3	29.6	35.1	44.6	7.0
Prop.Swedish	0.0	0.1	0.3	10.5	0.9	92.4	25.8
Prop.Foreign	0.2	1.4	2.2	3.2	3.6	25.7	3.3
Pop.Growth	-648.0	-66.0	-30.0	-28.8	-7.2	1384.0	151.0
Migr.Gain	-1044.0	-48.8	-14.0	-1.4	10.0	2094.0	199.1
Families	26.0	728.2	1646.0	4777.1	4219.2	161282.0	12430.1
Households	59.0	1311.8	2810.5	8974.2	7159.2	344898.0	25608.4
Prop.Households.rowSmall	13.1	74.7	88.5	82.0	94.1	98.8	16.4
Prop.Households.Rent	7.1	17.5	20.8	21.8	24.2	49.9	7.1
Prop.Educ.Degree2	57.3	66.8	69.4	69.7	72.5	82.1	4.5
Prop.Educ.High	12.6	19.8	23.1	24.3	27.6	59.3	6.5
Employed	42.0	965.2	2343.0	7436.5	5851.8	301908.0	22136.3
Empl.Rate	58.3	67.5	71.0	70.8	74.5	82.4	5.0
Prop.Households.Empl	20.1	41.8	59.6	57.9	73.2	91.7	18.3
Prop.Unempl	3.7	10.1	12.2	12.4	14.4	21.6	3.3
Prop.Pension	12.8	27.3	32.7	32.9	38.6	49.4	7.8
Support.Ratio	101.4	139.9	161.5	164.6	188.1	241.4	30.6
Jobs	18.0	847.2	2002.0	7722.1	5237.5	413677.0	28037.6
Prop.Primary.Sector	0.1	4.0	9.7	10.6	16.1	36.5	7.7
Prop.Secondary.Sector	2.1	17.7	23.2	24.1	29.5	67.0	10.0
Prop.Services.Sector	23.6	56.6	62.9	62.9	69.4	93.1	10.3
Jobs.Self.Suff	34.5	71.6	87.2	85.8	100.1	170.1	19.8
Margin.Citizen	-583.3	501.0	660.0	674.9	822.9	4897.1	378.2
Loan.Citizen	0.0	2035.9	3273.4	3331.0	4275.7	10897.8	1812.9
Concern.Loan.Citizen	0.0	3545.7	5374.8	5565.2	7156.7	18154.8	2936.7
Educ.Cult.Citizen	366.3	1790.4	1969.3	2012.4	2231.9	3088.5	340.6
Soc.Health.Citizen	1221.9	3437.8	3999.8	4046.8	4645.3	6778.7	891.7

2.3 Distribution Plots

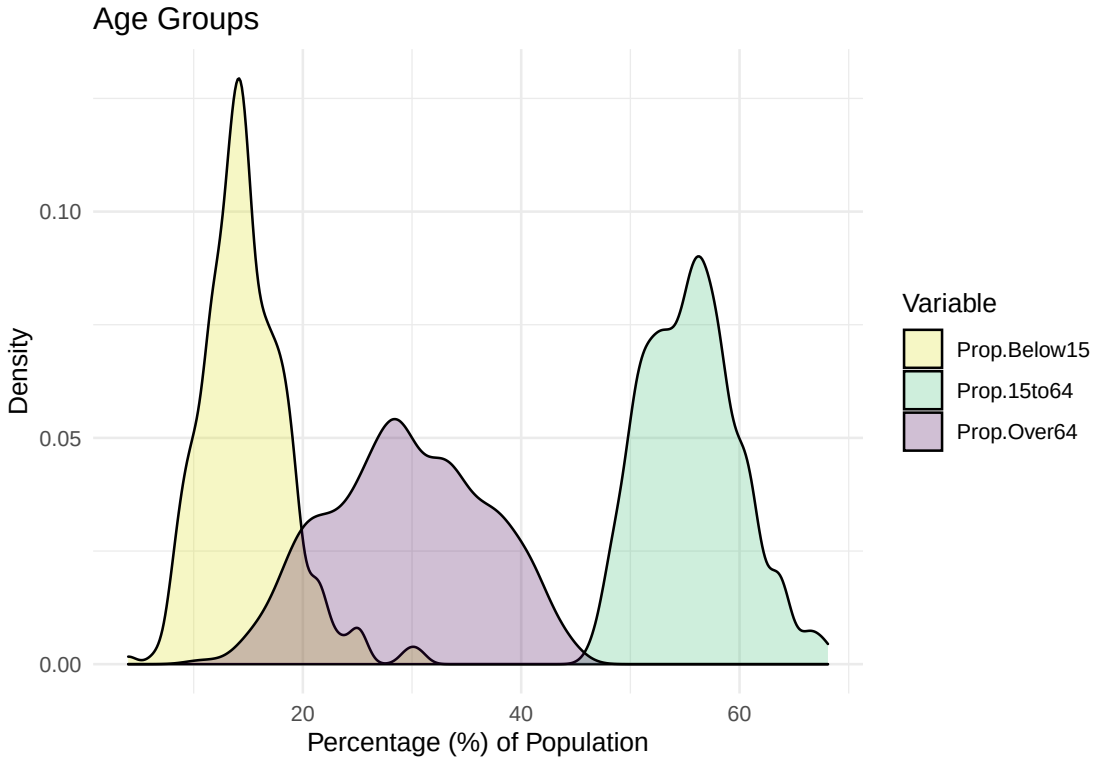

Figure 1. *Probability density estimates for age groups*

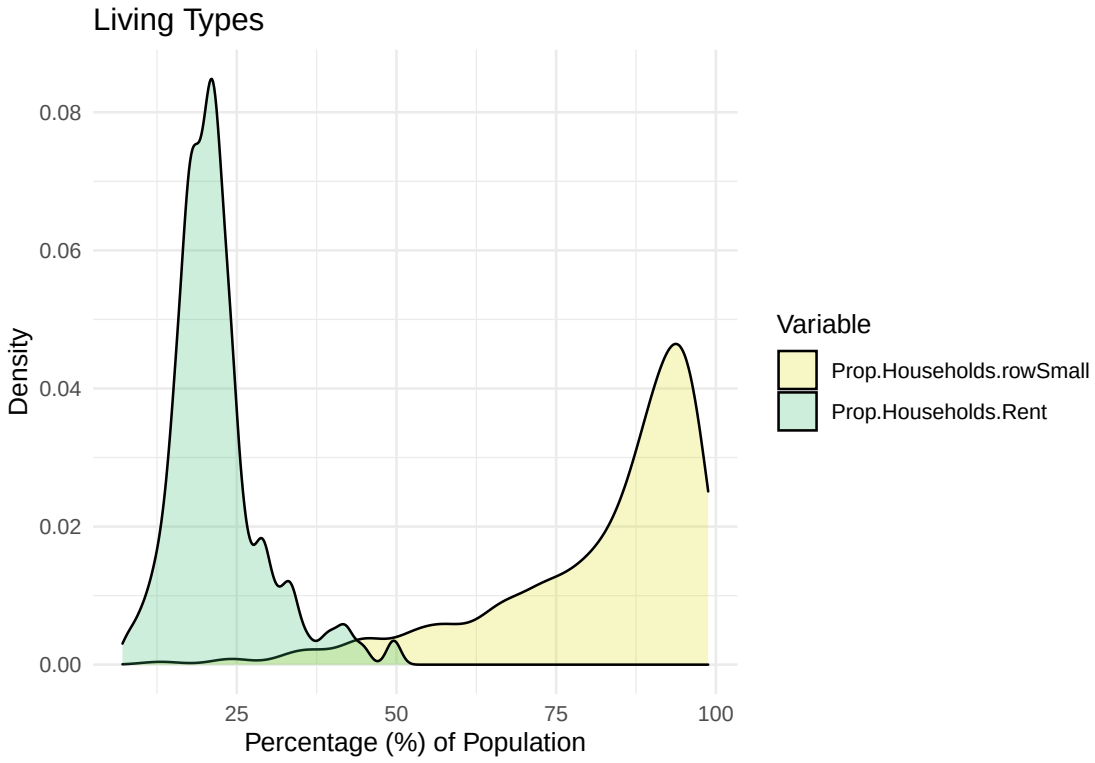

Figure 2. *Probability density estimates for living types*

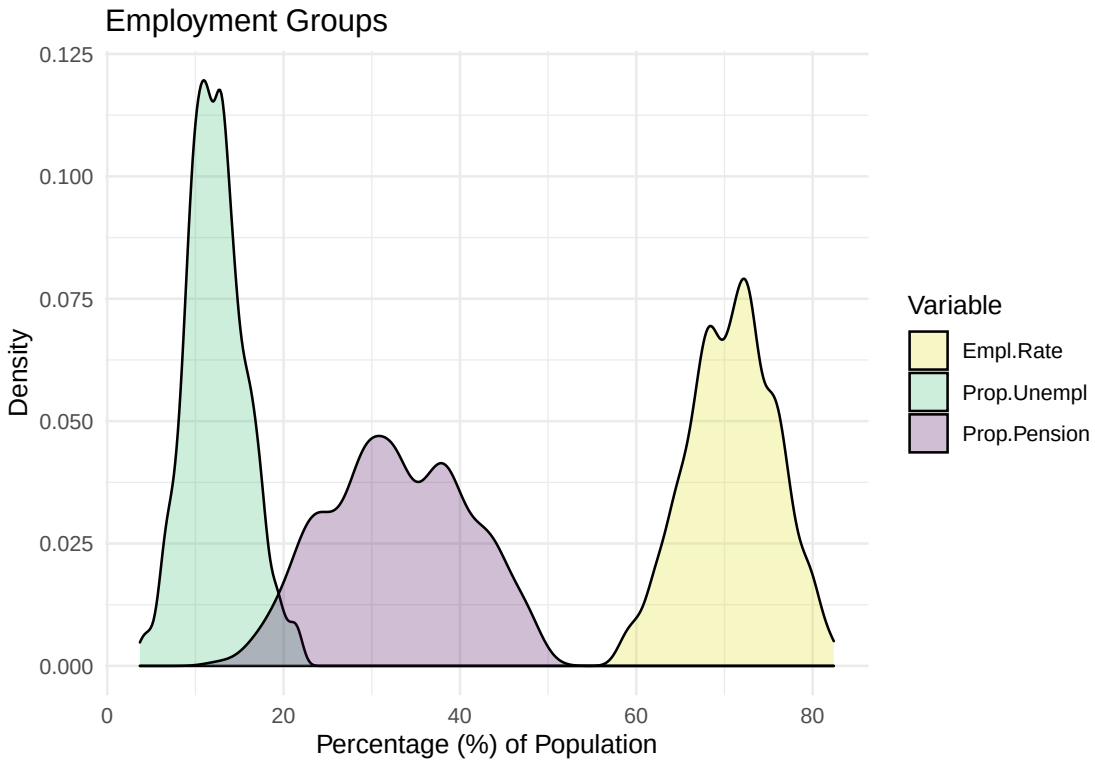

Figure 3. *Probability density estimates for employment groups*

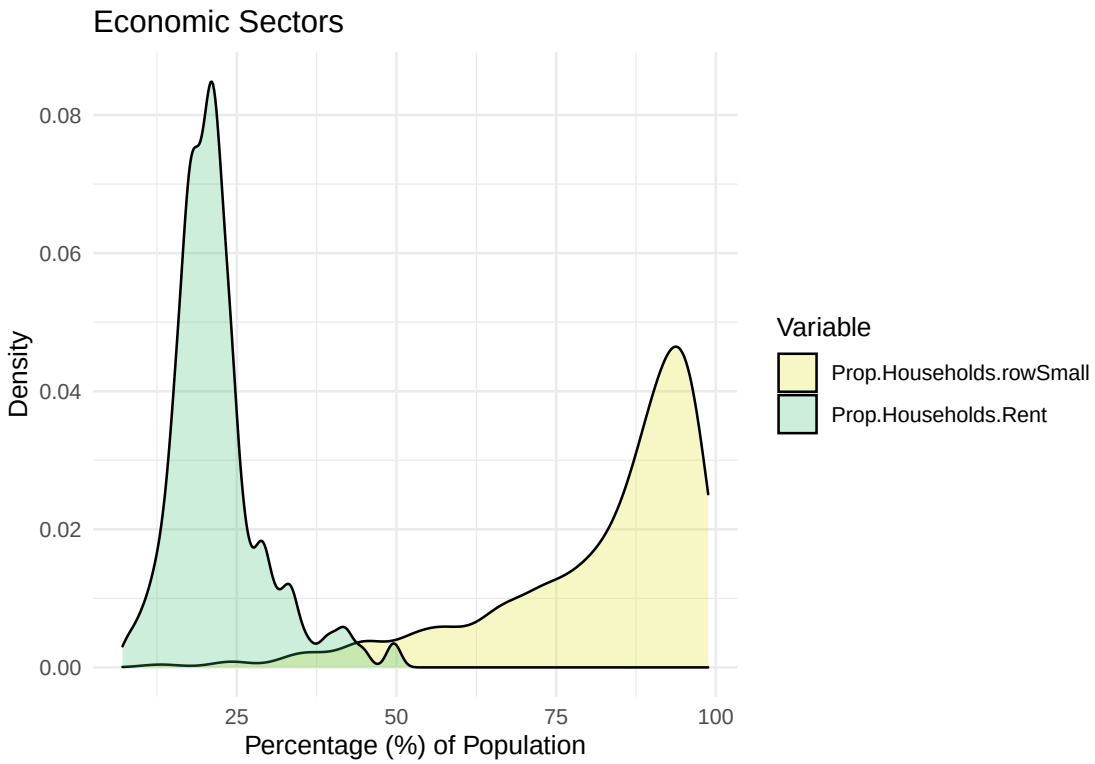

Figure 4. *Probability density estimates for economic sectors*

3 Bivariate Analysis

3.1 Linear Dependencies

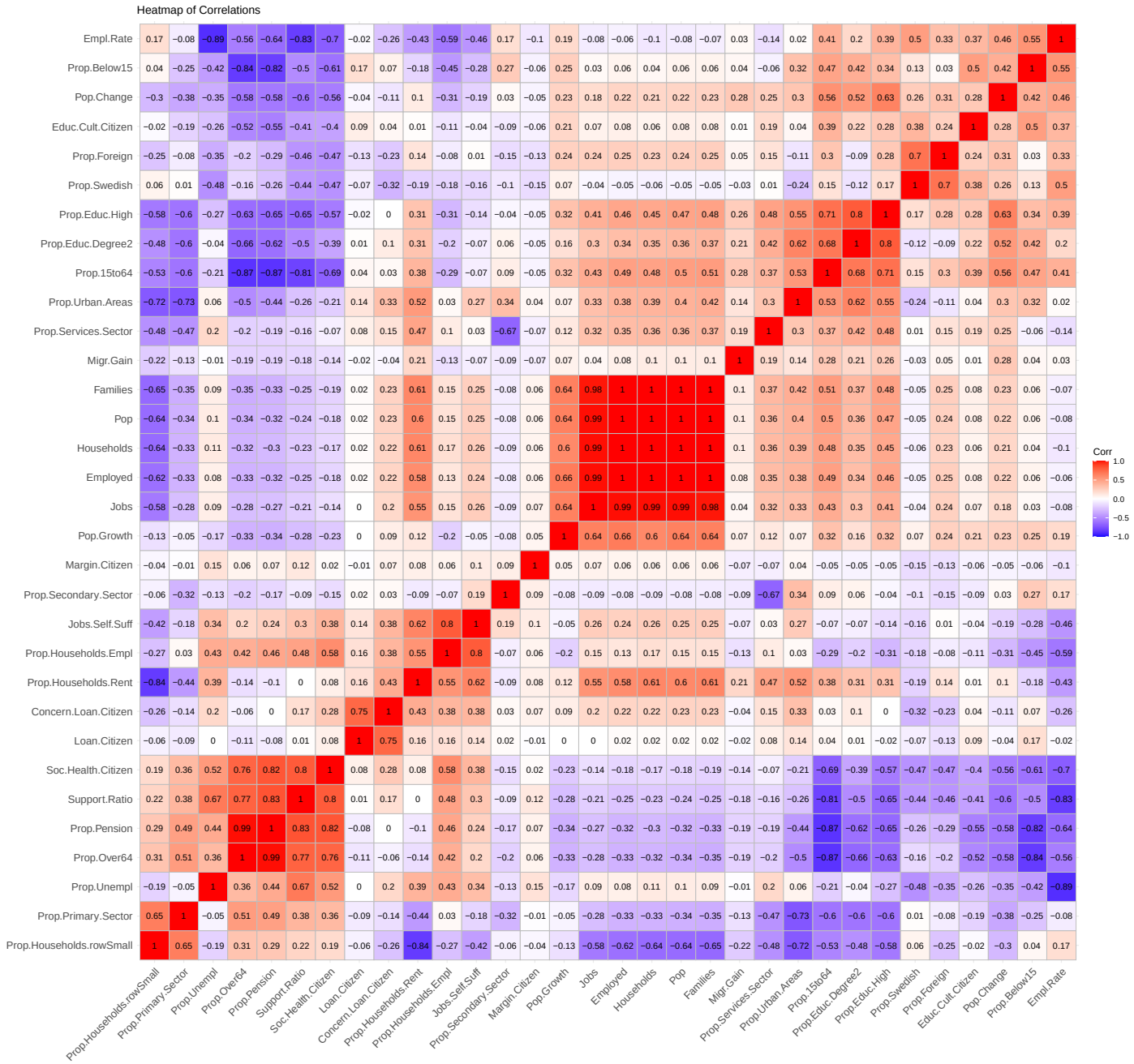

Figure 5. *Heatmap of Pearson'n correlation between all variables*

4 Principal Component Analysis (PCA)

4.1 Scree Plot

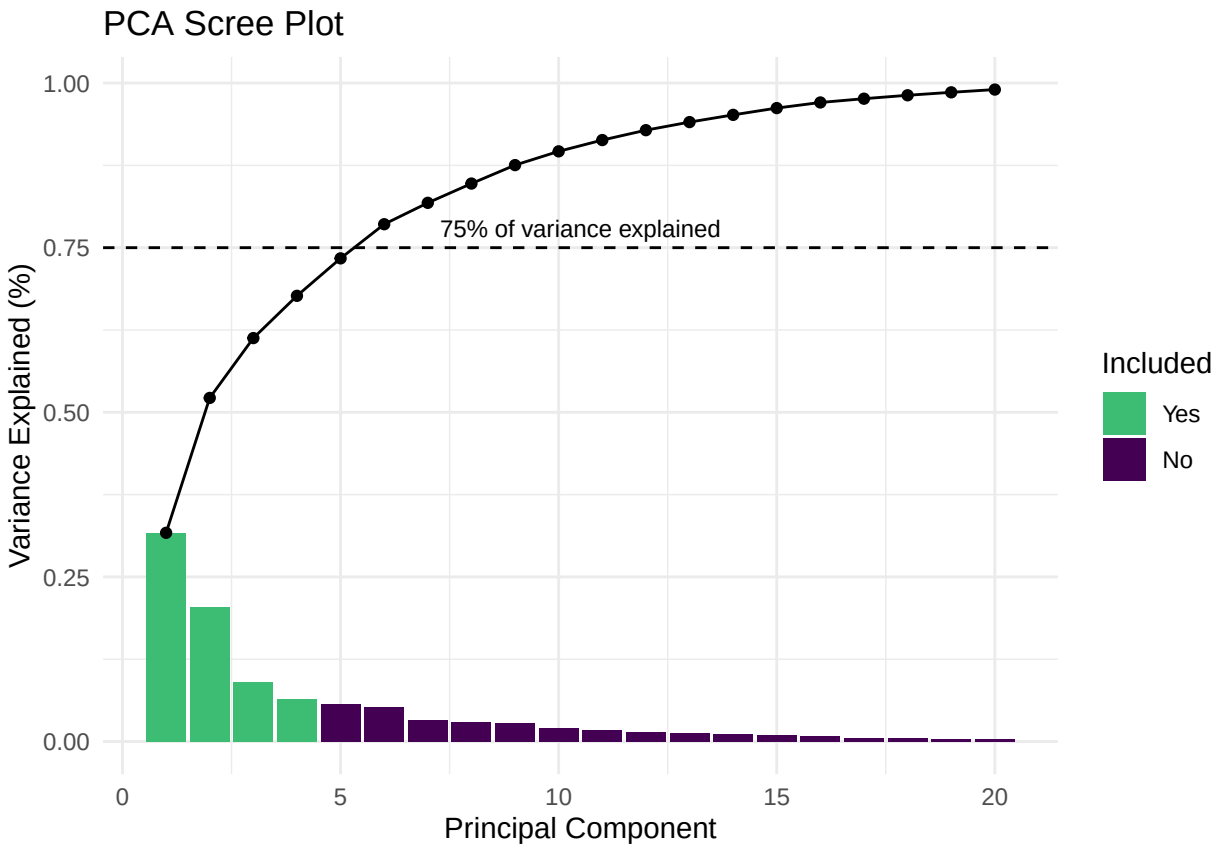

Figure 6. *Variance explained by each Princial Component (PC) up to the 20th PC. This type of a plot is also known as a Scree Plot*

4.2 Principal Components 1 and 2

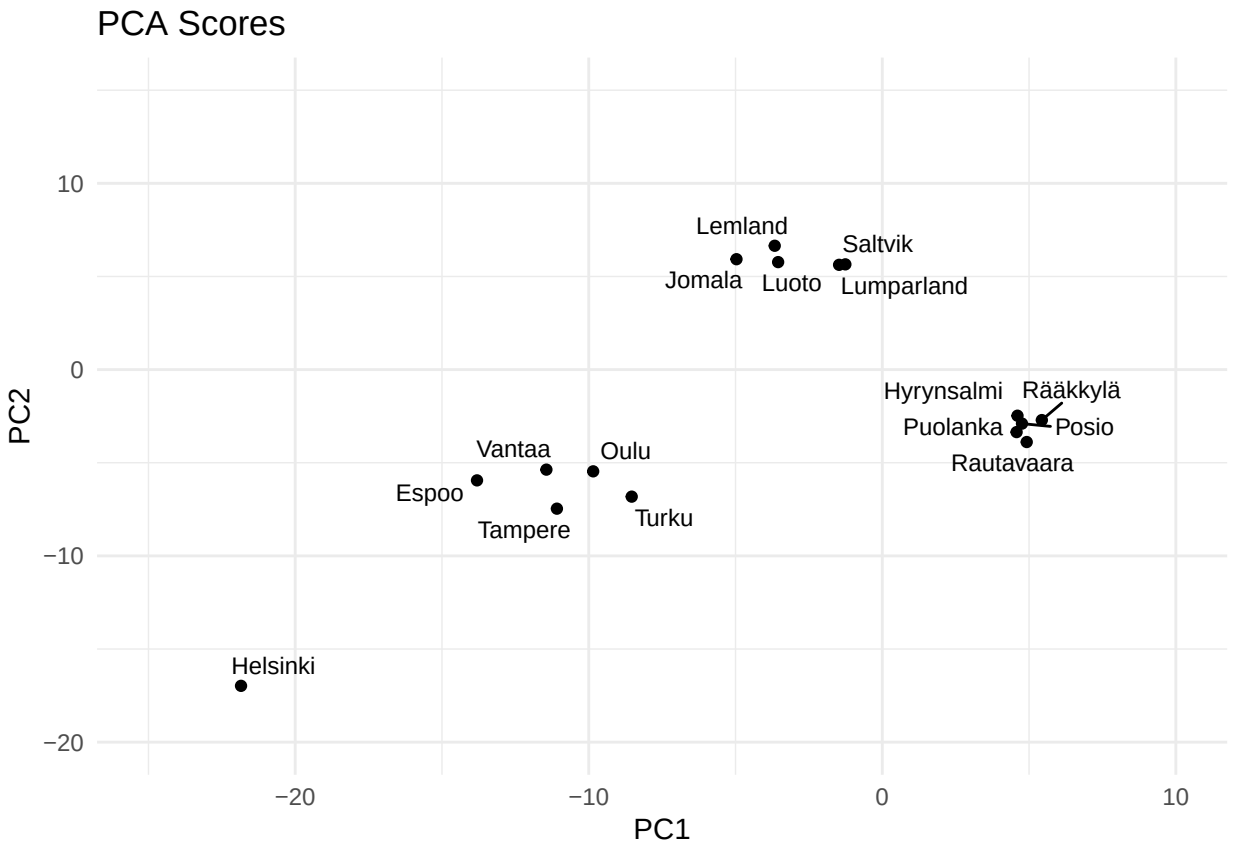

Figure 7. *Score plot for all municipalities that have a top 5 absolute loading in any of PC1 and PC2*

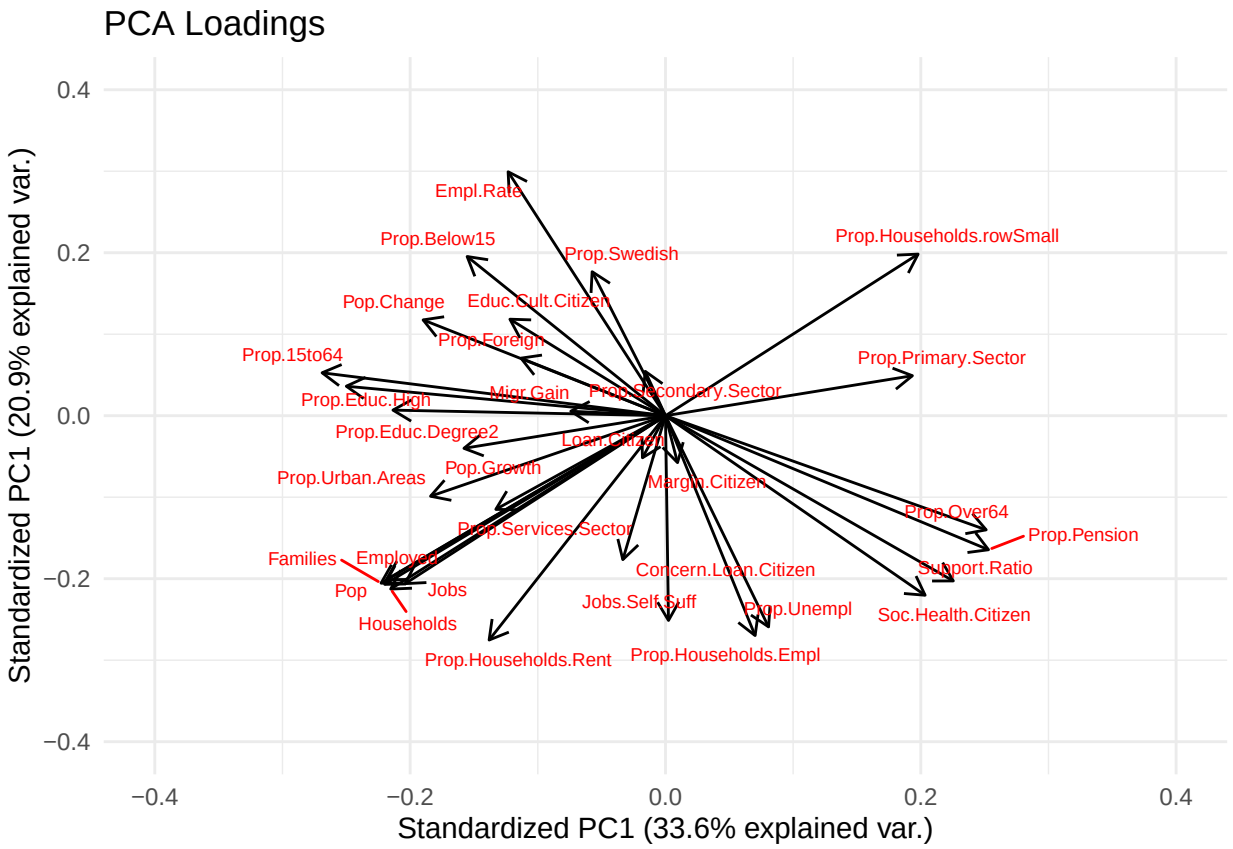

Figure 8. *Loading directions for PC1 and PC2*

4.3 Principal Components 3 and 4

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

```
## Warning: Removed 4 rows containing missing values (geom_text_repel).
```

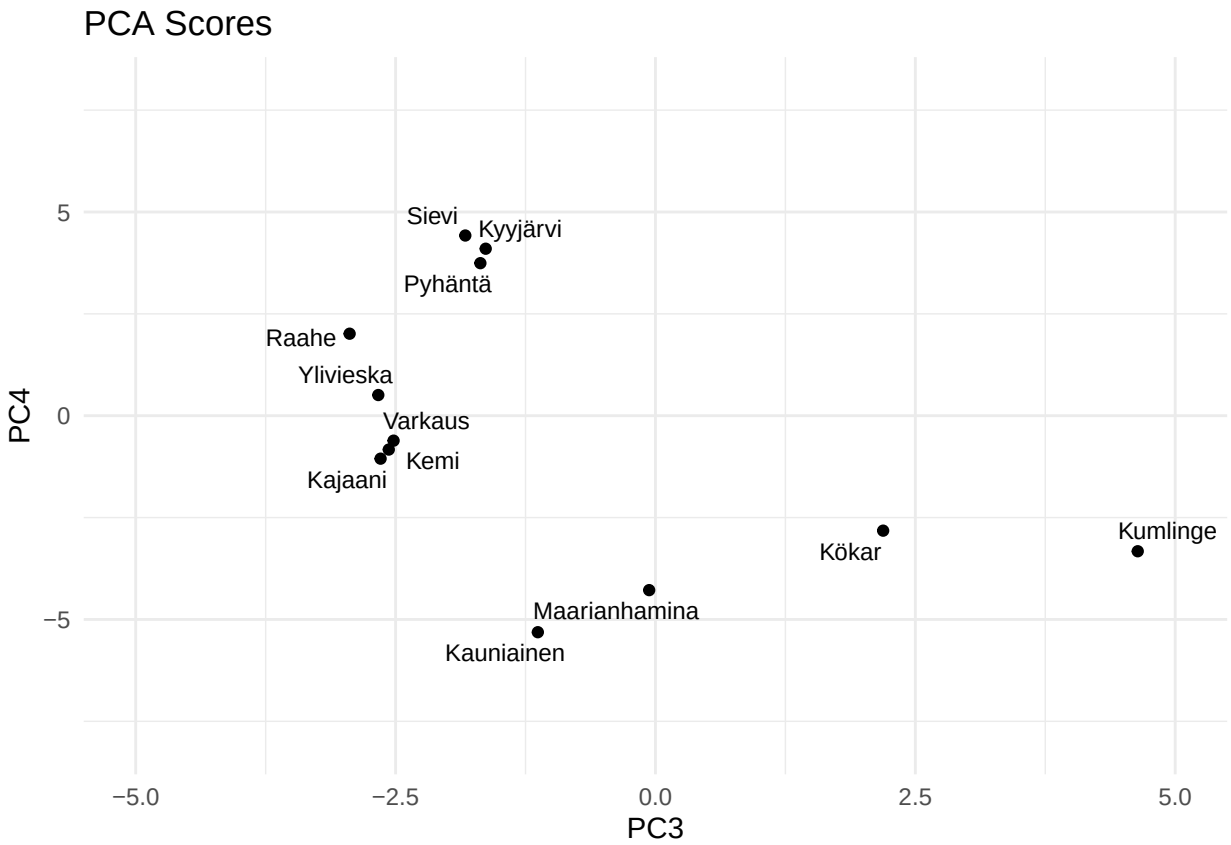

Figure 9. *Score plot for all municipalities that have a top 5 absolute loading in any of PC3 and PC4*

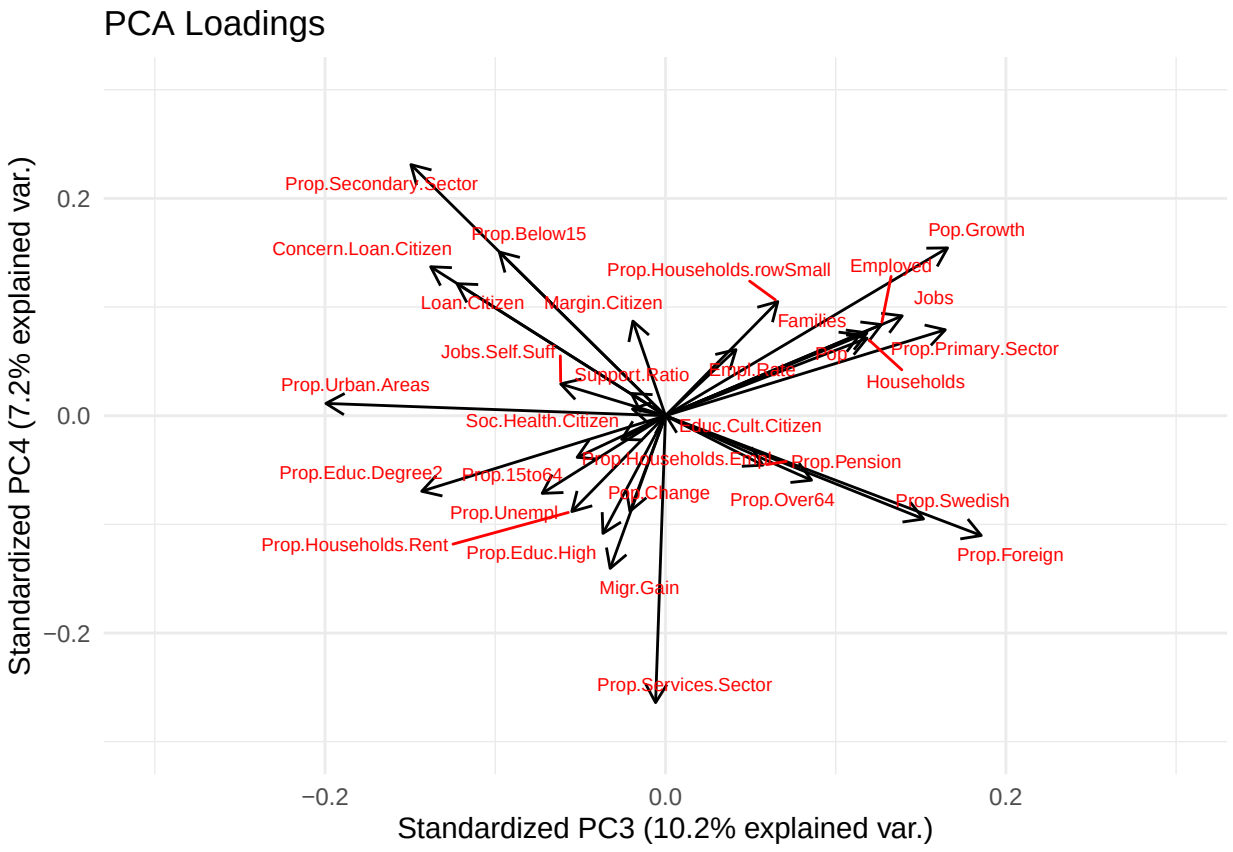

Figure 10. *Loading directions for PC3 and PC4*

5 Discussion and Conclusions

- PC1 and PC2 do not explain a high percentage of the variance (above 80%). Hence, PCA might not be so well suited
- Using population, employed persons and some other absolute measures might have alienated Helsinki from the other cities and biased the analysis a bit. Perhaps leaving out Helsinki would have been a wise choice
- Highly correlated variables could have been reduced to only one, i.e., remove unnecessary variables
- Something else?

References

Statistics Finland. Kuntien avainluvut muuttujina alue 2021, tiedot ja vuosi, 2022. URL https://pxnet2.stat.fi/PXWeb/pxweb/fi/Kuntien_avainluvut/Kuntien_avainluvut___2021/kuntien_avainluvut_2021_viimeisin.px/table/tableViewLayout1/. [Data accessed 2022-04-20 16:20:07].