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 describes the variables used in this study.

Table 1. Description of all used variables. Variable is the name of the variable that is used in this report. Original Variable is the original variable given in Tilastokeskus' dataset, Year is the timestamp of the data point, and Unit is the unit in which the variable is measured in.

Variable	Original Variable	Year	Unit
Deg.Urbanisation	Degree of urbanisation	2020	%
Popul	Population	2021	Number
Popul.Growth	Population change from the previous year	2021	%
Prop.Under15	Share of persons aged under 15 of the population	2021	%
Prop.15to64	Share of persons aged 15 to 64 of the population	2021	%
Prop.Over64	Share of persons aged over 64 of the population	2021	%
Prop.Swedish	Share of Swedish-speakers of the population	2021	%
Prop.Foreign	Share of foreign citizens of the population	2021	%
Excess.Births	Excess of births	2020	Persons
Migr.Gain	Intermunicipial migration gain/loss	2020	Persons
Families	Number of families	2020	Number
Households	Number of household-dwelling units	2020	Number
Prop.Households.Terr.Det	Share of household-dwelling units living in terraced houses and detached houses	2020	%
Prop.Households.Rental	Share of household-dwelling units living in rental dwellings	2020	%
Prop.Educ.Degree2	Share of persons aged 15 or over with at least upper secondary qualifications	2020	%
Prop.Educ.Degree3	Share of persons aged 15 or over with tertiary level qualifications	2020	%
Labour.Force	Employed labour force resident in the area	2020	Number
Empl.Rate	Employment rate	2020	%
Prop.Empl.Muni	Share of persons working in their municipality of residence	2019	%
Prop.Unempl	Proportion of unemployed among the labour force	2020	%
Prop.Pensioners	Proportion of pensioners of the population	2020	%
Depend.Ratio	Economic dependency ratio	2020	Ratio
Jobs.Muni	Number of workplaces in the area	2019	Number
Prop.Primary.Sector	Share of workplaces in primary production	2019	%
Prop.Secondary.Sector	Share of workplaces in secondary production	2019	%
Prop.Services.Sector	Share of workplaces in services	2019	%
Jobs.Self.Suff	Workplace self-sufficiency	2019	Ratio
Contr.Margin	Annual contribution margin	2020	EUR per capita
Loan.Stock	Loan stock	2020	EUR per capita
Group.Loan.Stock	Group loan stock	2020	EUR per capita
Educ.Cult.Activity	Educational and cultural activities	2020	EUR per capita
Soc.Health.Activity	Social and health care activities	2020	EUR per capita

2.2 Descriptive Statistics

Table 2 shows descriptive statistics for each variable.

Table 2. Descriptive Statistics of the variables. Variable is the name of the variable, Min is the minimum, 1st Qu. is the 1st quartile, Median is the median, Mean is the arithmetic mean, 3rd Qu. is the 3rd quartile, Max is the maximum and Std. is the sample standard deviation.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Std.
Deg.Urbanisation	0.0	47.2	60.8	61.7	77.3	100.0	22.2
Popul	105.0	2673.8	6134.0	18039.0	15145.2	658457.0	49728.7
Popul.Growth	-4.8	-1.2	-0.6	-0.5	0.2	4.0	1.2
Prop.Under15	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
Excess.Births	-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.Terr.Det	13.1	74.7	88.5	82.0	94.1	98.8	16.4
Prop.Households.Rental	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.Degree3	12.6	19.8	23.1	24.3	27.6	59.3	6.5
Labour.Force	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.Empl.Muni	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.Pensioners	12.8	27.3	32.7	32.9	38.6	49.4	7.8
Depend.Ratio	101.4	139.9	161.5	164.6	188.1	241.4	30.6
Jobs.Muni	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
Contr.Margin	-583.3	501.0	660.0	674.9	822.9	4897.1	378.2
Loan.Stock	0.0	2035.9	3273.4	3331.0	4275.7	10897.8	1812.9
Group.Loan.Stock	0.0	3545.7	5374.8	5565.2	7156.7	18154.8	2936.7
Educ.Cult.Activity	366.3	1790.4	1969.3	2012.4	2231.9	3088.5	340.6
Soc.Health.Activity	1221.9	3437.8	3999.8	4046.8	4645.3	6778.7	891.7

2.3 Distribution Plots

To analyze how variables are distributed, we use Kernel density estimation to estimate the probability densities for several variable groups.

Figure 1 shows the probability density estimates for age groups. Age groups are shares of population aged below 15, 15 to 64, and over 64. We can see that . . .

Figure 2 shows the probability density estimates for housing types. Housing types are shares of population living in either terraced or detached houses and in rental apartments. The Figure revels that . . .

Figure 3 shows the probability density estimates for employment measures. Employment measures are the employment rate, the proportion of population that are unemployed, and the proportion of population that are pensioners. The distributions show that . . .

Figure 4 shows the probability density estimates for the weights of economic sectors in the municipalities. Sectors are the primary-, secondary-, and services sectors. We can interpret that . . .

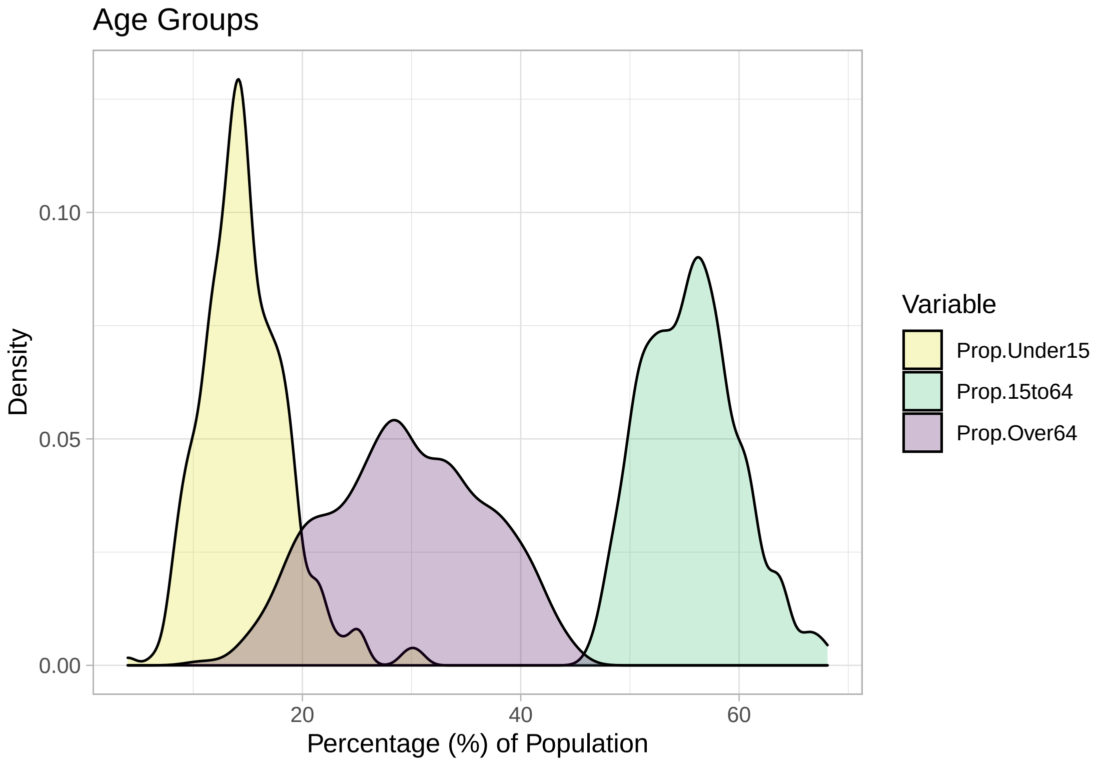

Figure 1. Probability density estimates for age groups

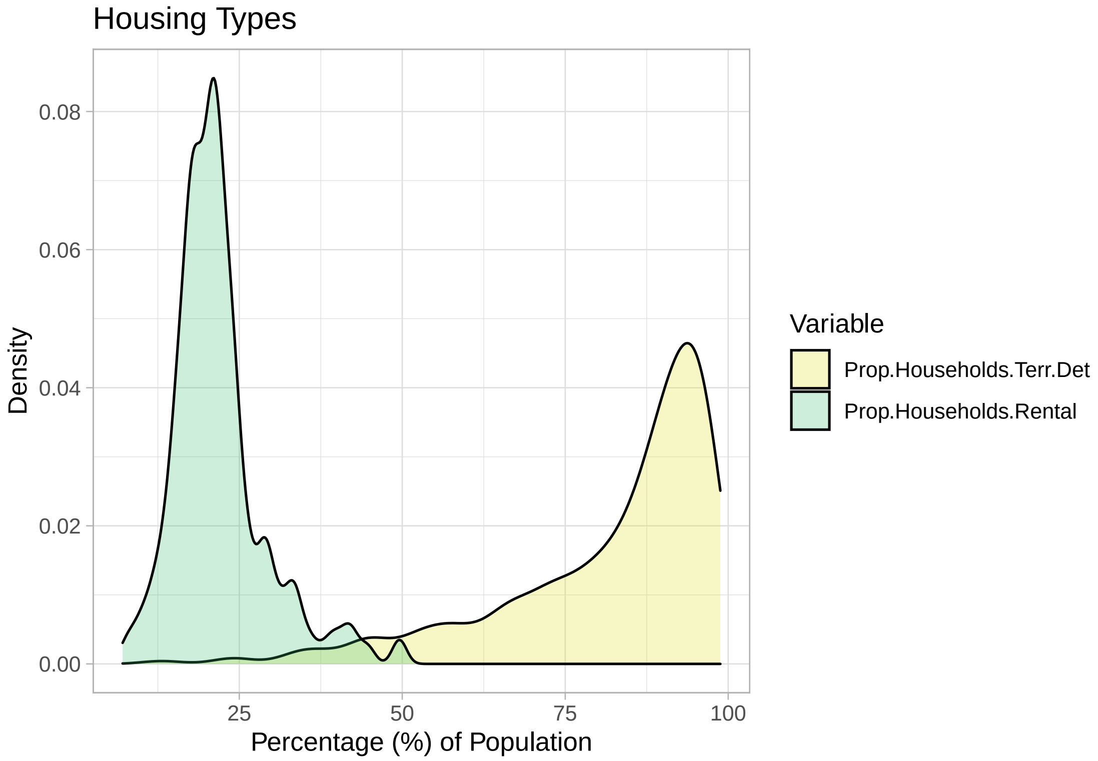

Figure 2. Probability density estimates for housing types

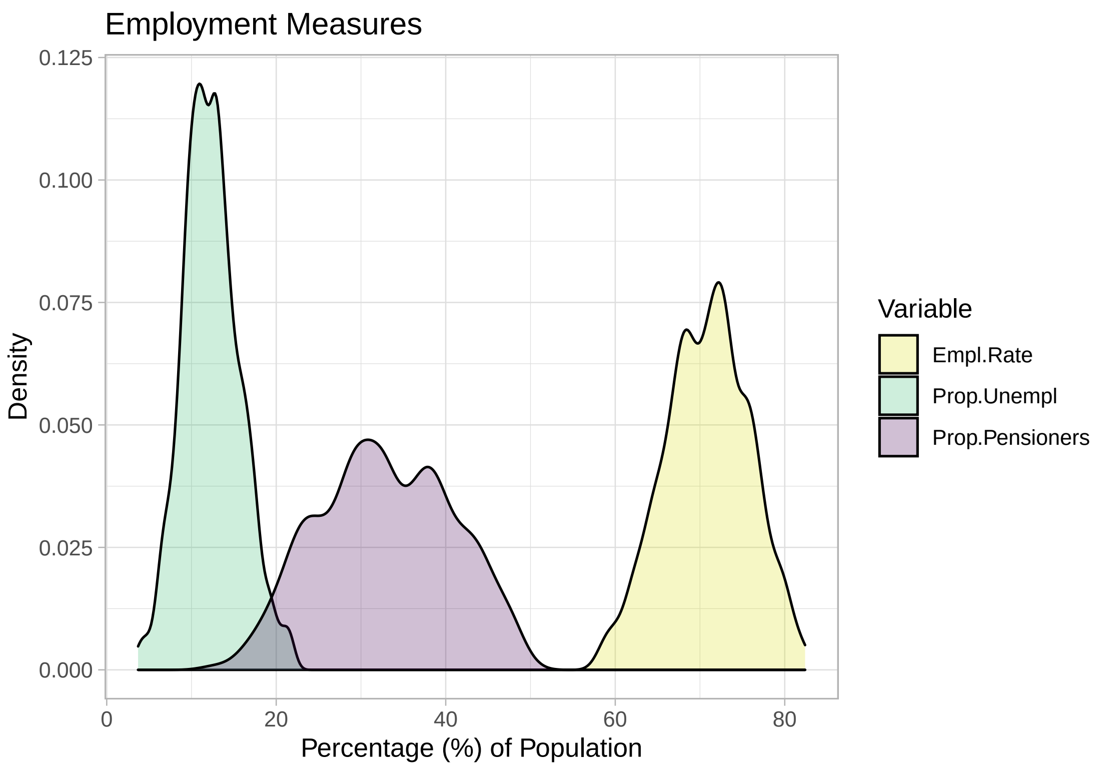

Figure 3. Probability density estimates for employment related measures

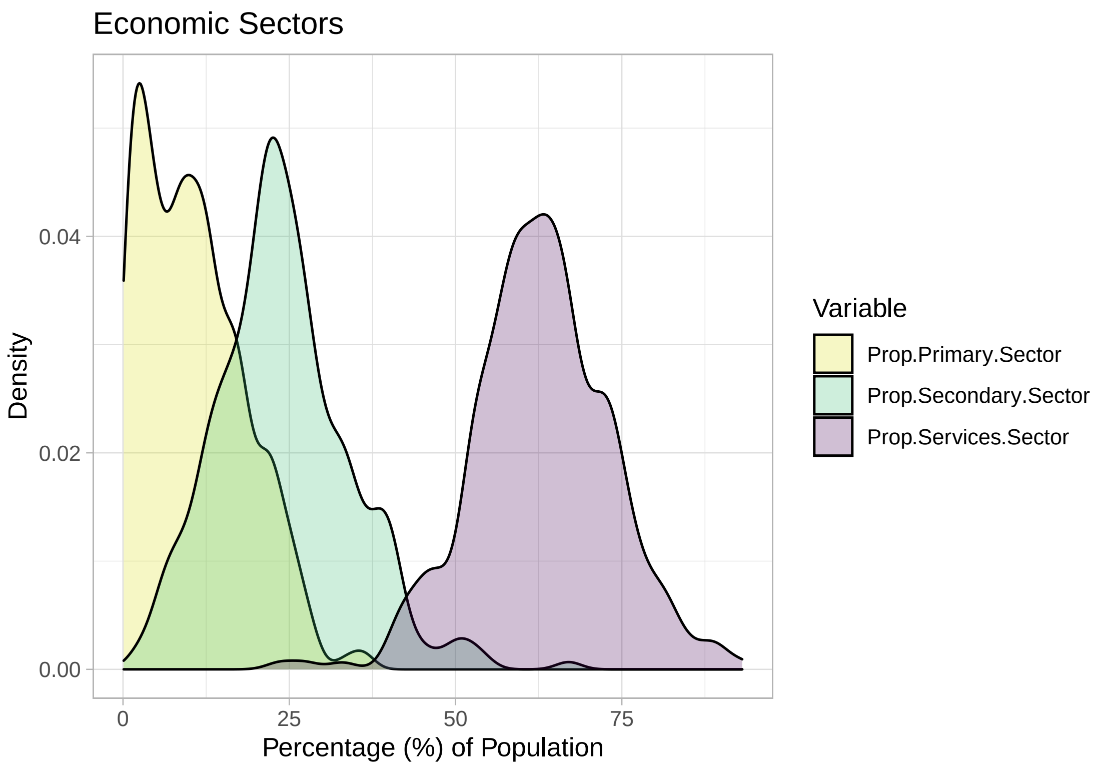

Figure 4. Probability density estimates for economic sectors

3 Bivariate Analysis

3.1 Linear Dependencies

To analyze bivariate dependencies in the data, we assess linear relationships between the variables with Pearson's correlation analysis. Figure 5 shows a heatmap of correlations between all variables.

The heatmap shows that . . .

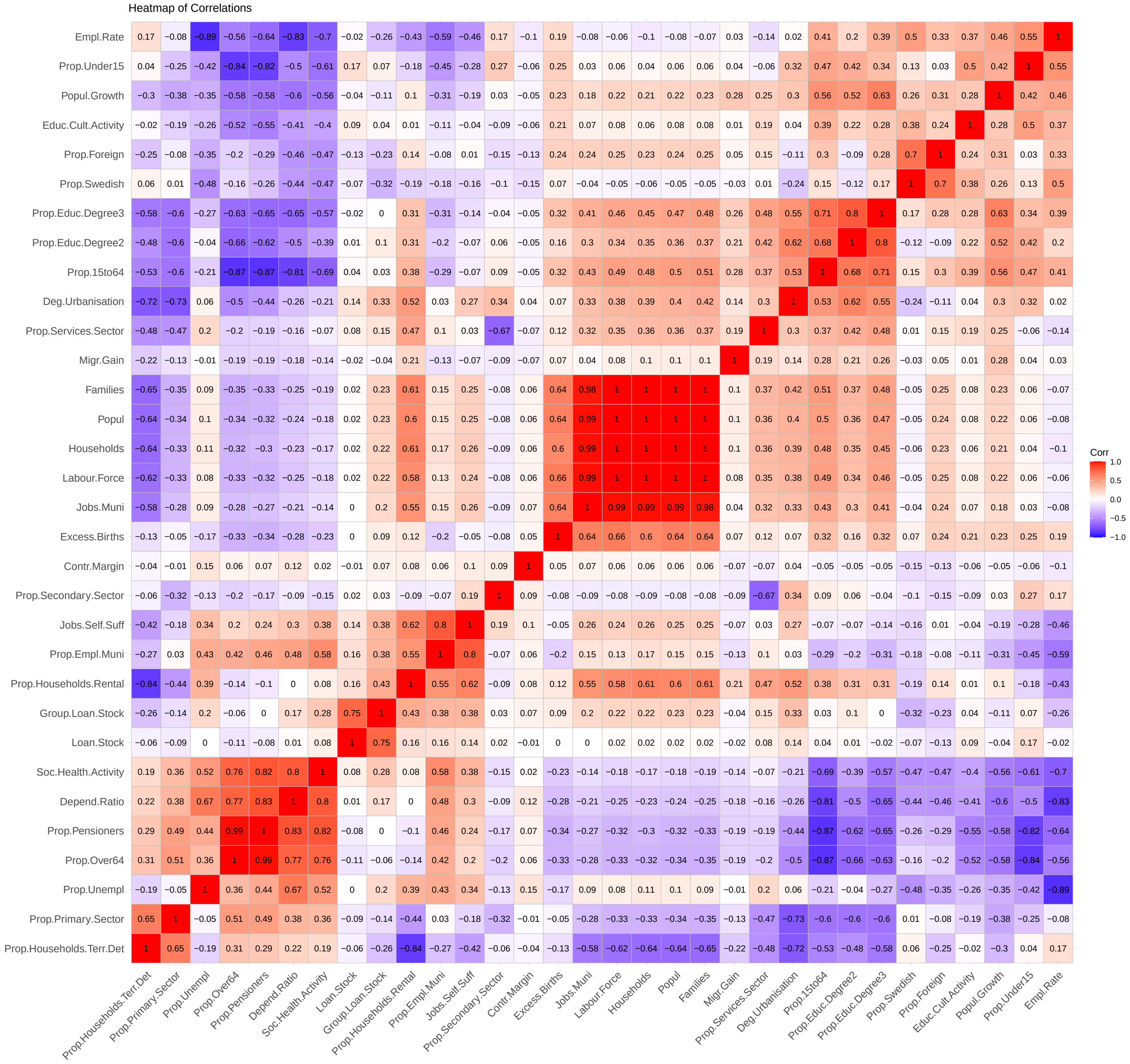

Figure 5. Heatmap of Pearson'n correlation between all variables

4 Principal Component Analysis (PCA)

To assess how much proportion of variance can be explained with the first few PCs, we create a scree plot. Figure 6 shows the resulting plot. We can see that we can explain close to 55% of all variation using only the first two PCs. Furthermore, close to 70% of the variation can be explained using the first four PCs.

4.1 Scree Plot

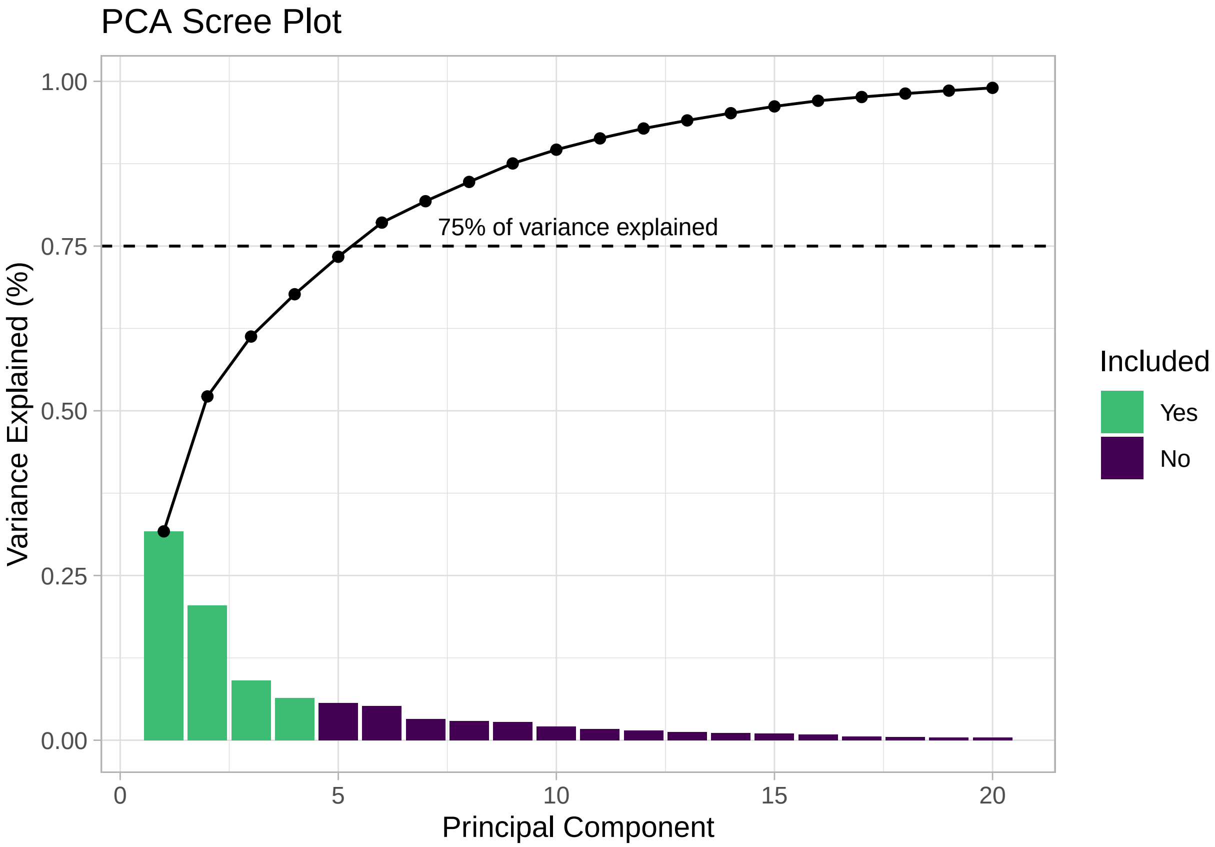

Figure 6. Variance explained by each Princial Component (PC) up to the 20th PC. This type of a plot is also known as a Scree Plot

4.2 Principal Components 1 and 2

Figure 7 shows the scores for PCs 1 and 2.

Figure 8 shows the loading directions for PCs 1 and 2.

Figure 9 shows the scores for PCs 3 and 4.

Figure 10 shows the loading directions for PCs 3 and 4.

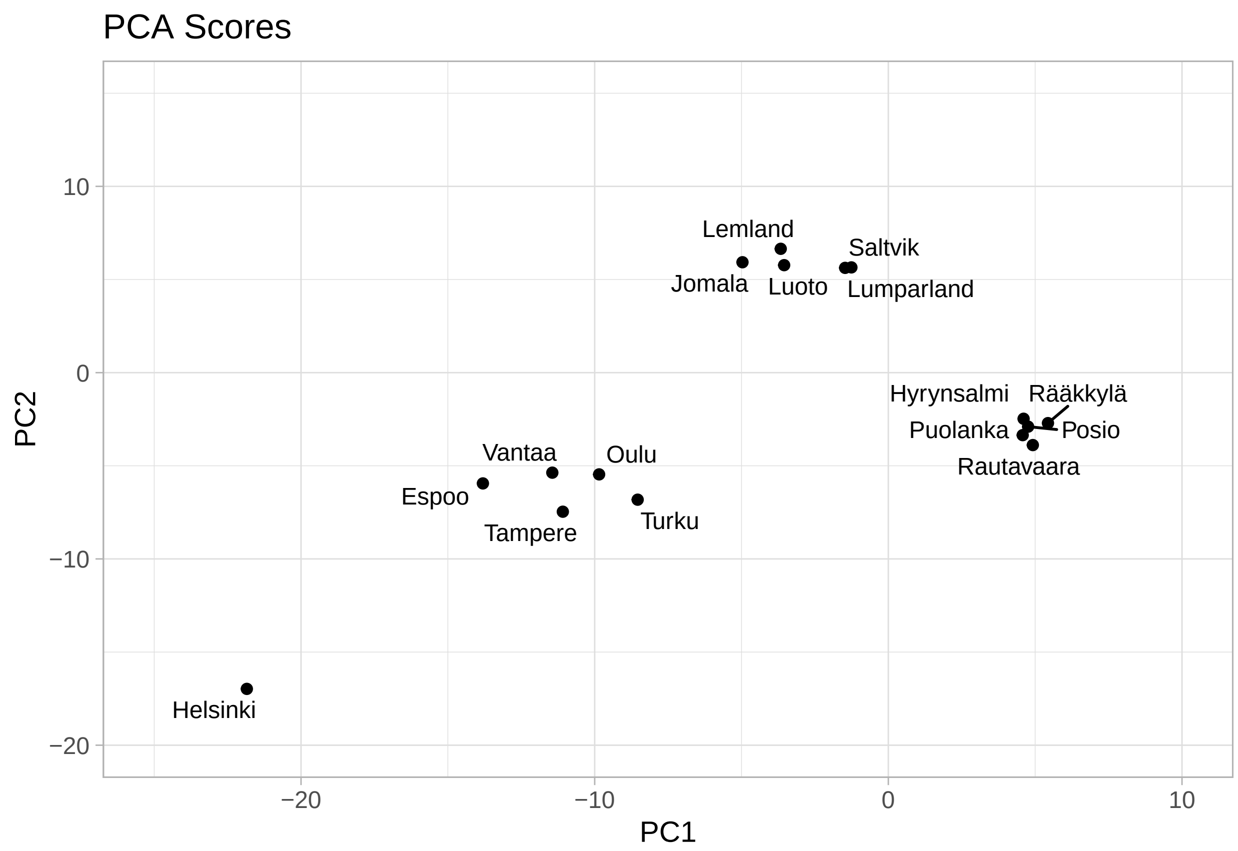

Figure 7. Score plot for all municipalities that have a top 5 absolute loading in any of PC1 and PC2

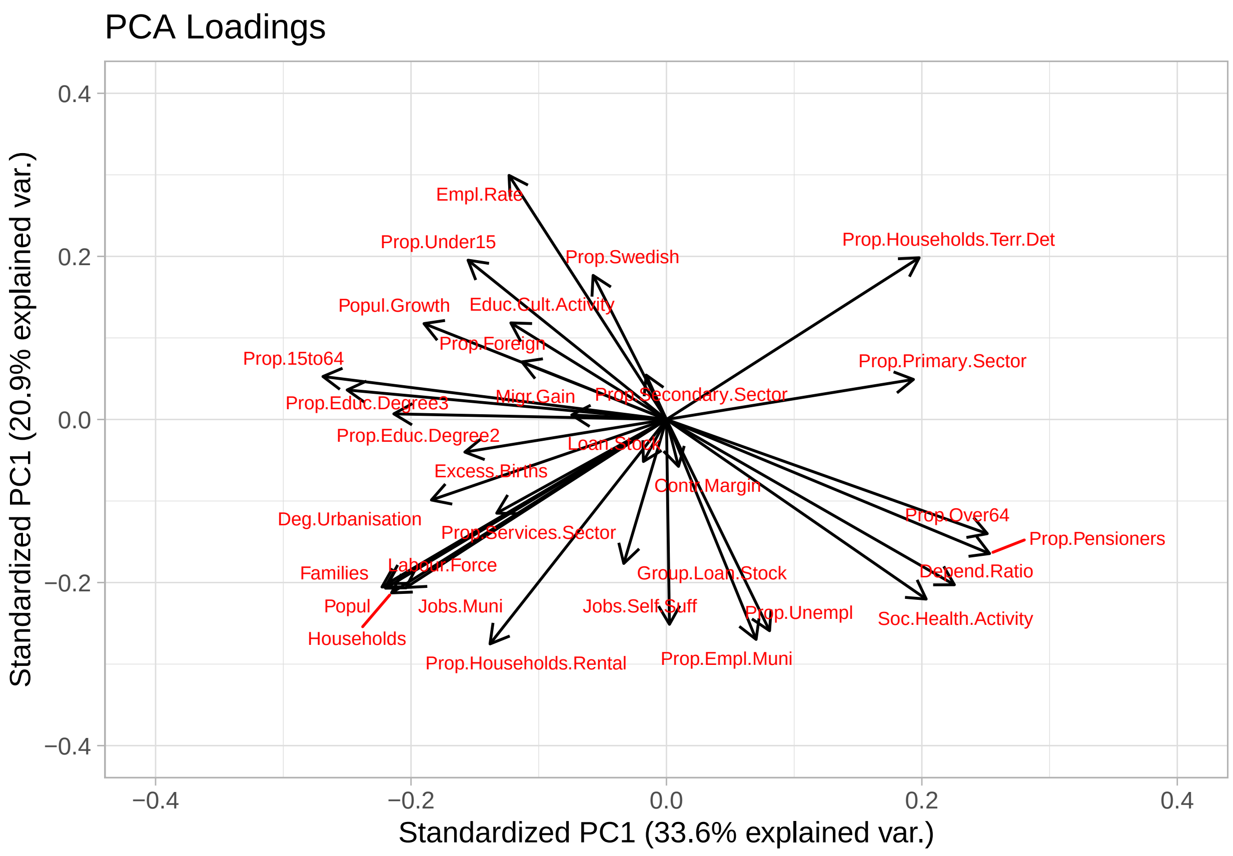

Figure 8. Loading directions for PC1 and PC2

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).
```

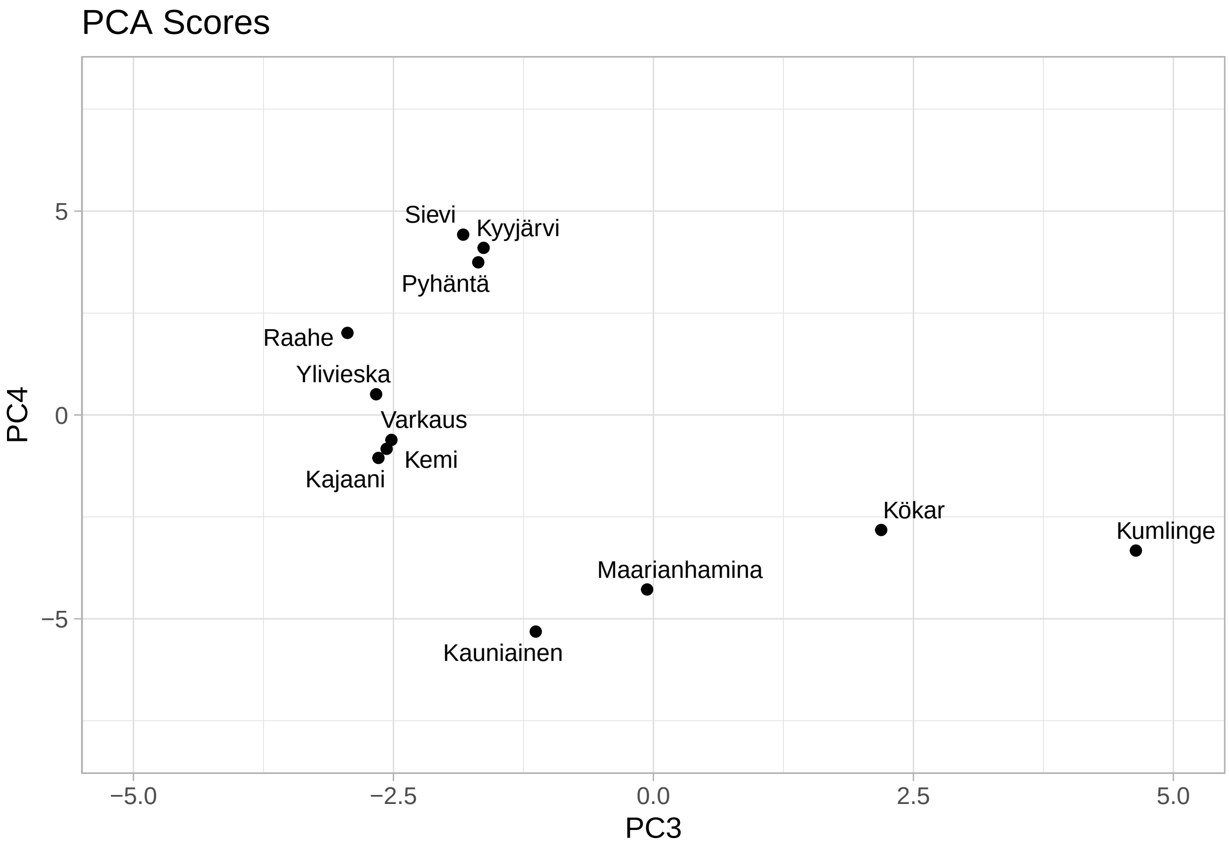

Figure 9. Score plot for all municipalities that have a top 5 absolute loading in any of PC3 and PC4

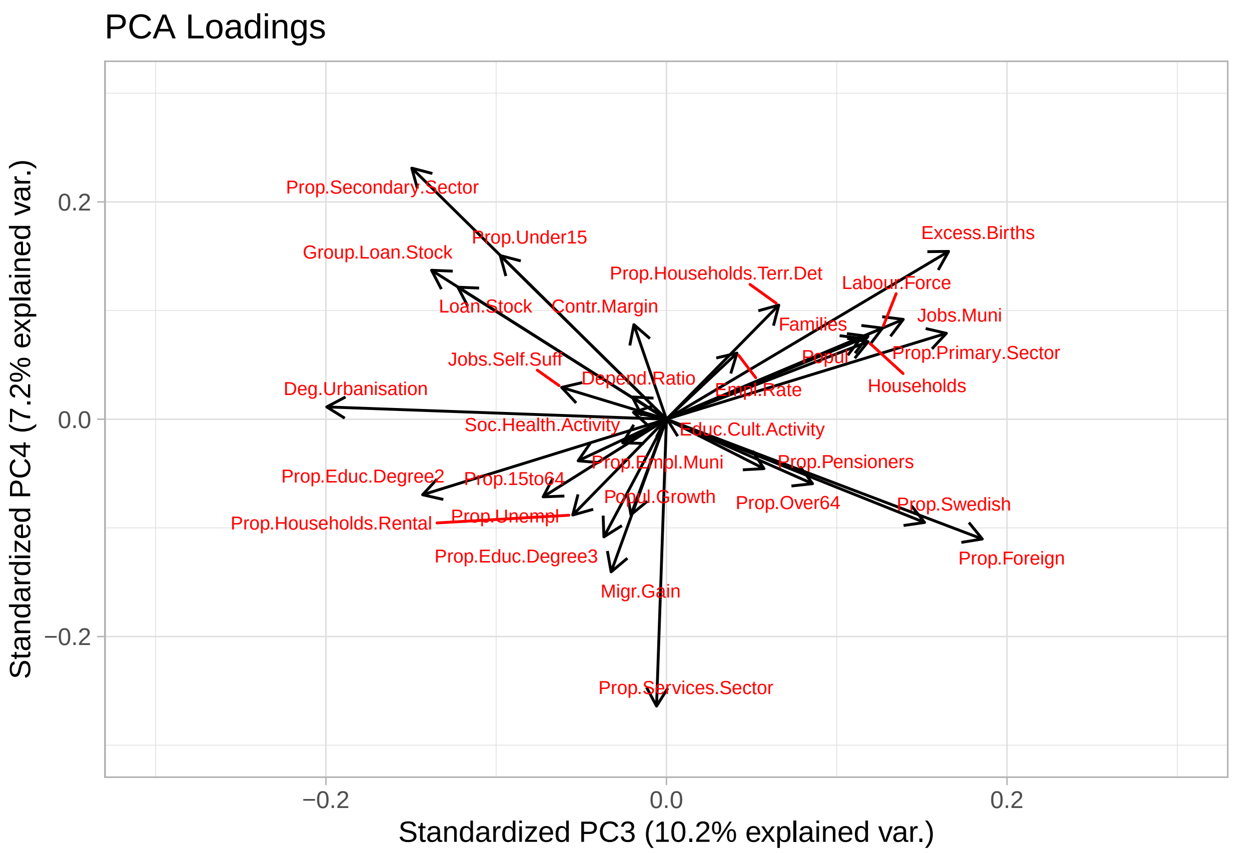

Figure 10. Loading directions for PC3 and PC4

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].