Optimizing Public Policy Planning using Data Analysis: A Socio-Economic Consulting Project for Argentina

Section 0: Introduction

Argentina is divided into 23 distinct provinces with many differences between them. We analysed provincial data for Argentina using ten global socio-economic indicators for each province. Our main contribution involved identifying provinces with similar socio-economic patterns across Argentina. Our findings provide public policy planners with a birds eye view of the socio-economic development in Argentina, which might help guide the optimization of delivery and planning of possible public policies. Additionally, appropriate changes to our methodology could make our work useful for analyzing development levels and optimizing the planning of public policy, globally and regionally, for other countries.

Section 1: Exploratory Data Analysis

The main inferences from the summary statistics (table 1) computed were the spread of Interquartile range values compared to the mean of the various indicator variables. This initially could suggest that provinces have varying levels of development; looking that the minimum/maximum values, particularly poor/good in some areas of socio-economic development. We also note that we have no data for the province `Tierra del Fueg`.

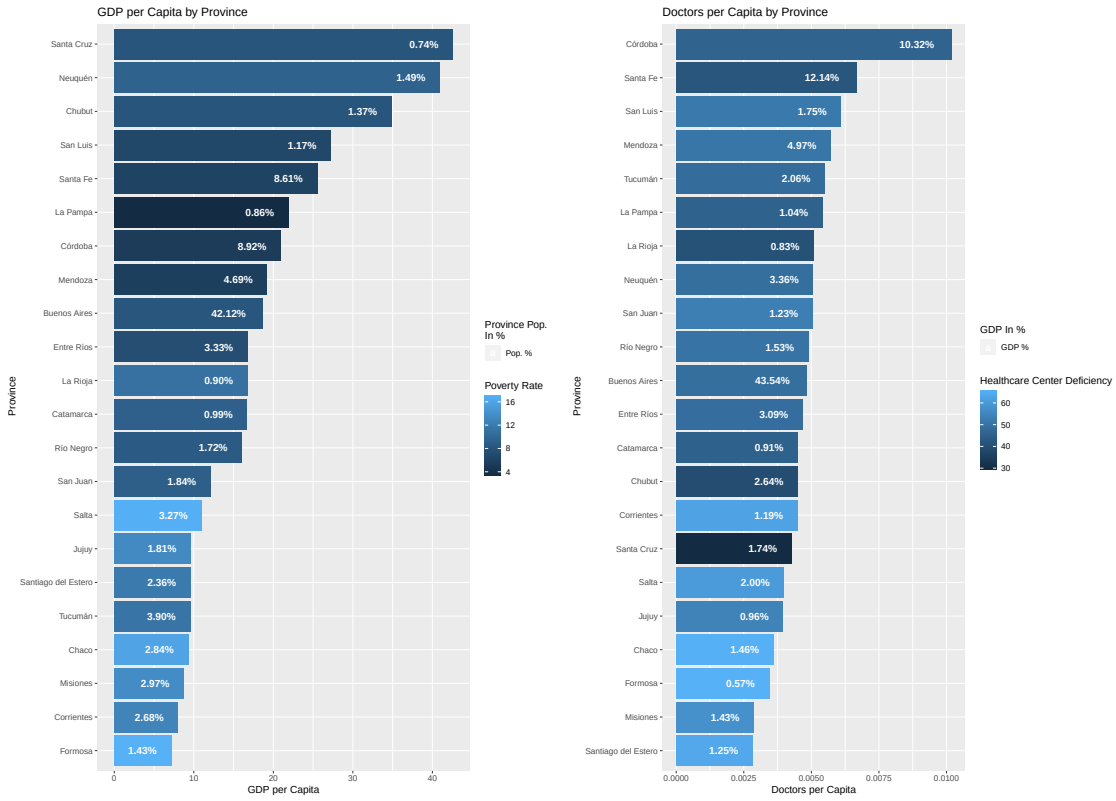

Figure 1: Left: How does GDP per capita change for each province in Argentina, and how it relates to the poverty rate and population in percent. Right: How does 'Doctors per capita' change for each province in Argentina, and how it relates to 'Healthcare Center Deficiency' and GDP in percent (proportion of the total GDP for the 22 provinces in Argentina, no data for Tierra del Fueg).

In the left plot of figure 1, we observe that the three provinces with the highest GPD per capita are `Santa Cruz`, `Neuquein` and `Chubut`; whereas, the three provinces with the lowest GPD per capita are `Formosa`, `Corrientes` and `Misiones`. In fact, the eight provinces with the lowest GPD per capita reside in the north of Argentina (figure 7), and all have a higher than average poverty rate. We observe that there is a negative correlation between poverty rate and GPD per capita. One also observes that the province `Buenos Aires` holds 42% and 44% Argentina's total

population and GDP respectively over the 22 provinces. Therefore we hypothesize that `Buenos Aries` is a true outlier. Despite `Buenos Aries` having an approximately average GDP per capita, it has an above average poverty rate, and due to population size, a huge number of people living in poverty. This preliminary evidence suggests that `Buenos Aires` might have a very diverse population in terms of socio-economic prosperity.

The right plot of Figure 1 highlights the provinces with the highest and lowest doctors per capita; namely, `Córdoba`, `Santa Fe` and `San Luis`, and `Formosa`, `Misiones` and `Santiago del Estero` respectively. Similarly, the six provinces with the lowest doctors per capita reside in the north of Argentina, and all have higher than average healthcare center deficiency (figure 8). Interestingly, the province with the highest GPD per capita, `Santa Cruz`, has a low doctors per capita and low healthcare center deficiency. One explanation may be due to `Santa Cruz` having a low poverty rate and population i.e. health standards may be higher so there is less demand for doctors compared to other provinces in Argentina. Interestingly, `Córdoba` and `Santa Fe` also have a high population and GPD, and a low poverty rate.

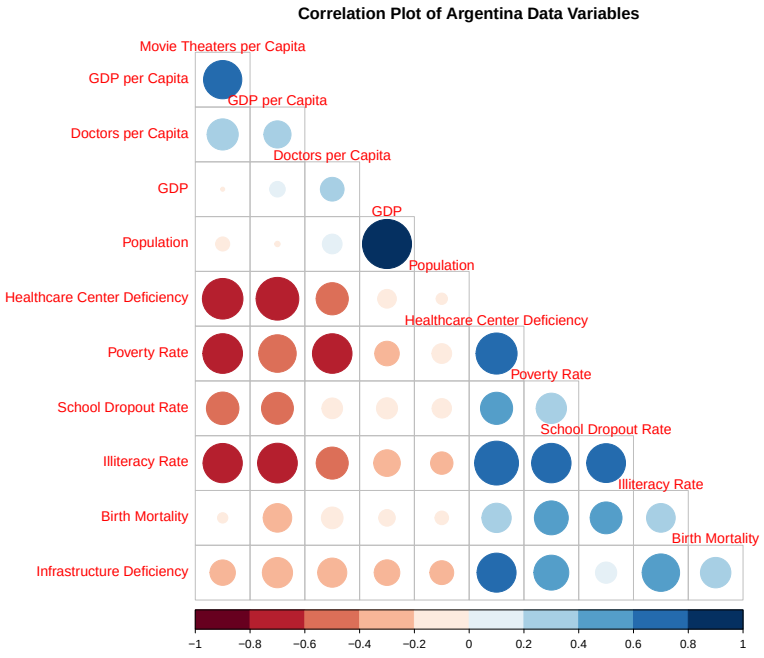

Figure 2: Correlation matrix of Argentina indicator variables. Blue and Red indicate positive and negative correlation respectively

As expected, from figure 2, we observe that socio-economic indicators were interrelated and correlated with each other. Figure 2 provides the relationships that one expects too see, for example,

- `Healthcare Centre Deficiency`, `Poverty Rate` and `Illiteracy Rate` are all significantly negatively correlated with `Movie Theatres per Capita`, `GPD per Capita` and `Doctors per Capita`.
- `Iliteracy Rate` is significantly positively correlated with `Healthcare Centre Deficiency`, `Poverty Rate` and `School Dropout Rate`.
- `Population` is strongly correlated with `GPD`. `Birth Mortality` is positively correlated to `School Dropout Rate`

From the exploratory data analysis conducted in section 1, we provide preliminary evidence that the provinces that will need highest priority are the northern provinces of Argentina (figures 1, 7, 8). More specifically, we suspect the group of high priority provinces includes `Corrientes`, `Chaco`, `Formosa`, `Jujuy`, `Misiones`, `Salta` and `Santiago del Estero` (figures 1, 7, 8).

Section 2: PCA and Clustering Analysis

Overview of Analysis: In figure 2 we observed that socio-economic indicators are highly correlated, which is an essential pre-requisite for PCA. In this section we use PCA and clustering to analyse the underlying patterns in the data, and categorise each province in terms of province development level. It is important to note that some of the indicators in the Argentina data set where not adjusted for the number of people living in each province. However, allowing the inclusion of these indicators allowed us to determine a reduced data set containing only four uncorrelated principle components (figures 9, 16). To avoid any bias with different numbers of residents across provinces, scaling was computed before applying PCA. In section 2 we hypothesized that `Buenos Aires` was a true outlier; `kmeans` and `hierarchical` clustering methods confirmed this (figures 10, 11, 12). Therefore, we decided to exclude `Buenos Aires` from further analysis with the aim to find a better cluster solution. Two options were investigated: one involved removing `Bueonos Aires` from both clustering algorithms (figures 13, 14, 15), the other involved removing `Bueonos Aires` before computing the principle components and both clustering algorithms (figures 18, 19, 4). The latter was chosen. One reason for this was that `Bueonos Aires` represents 42% and 43% of Argentina's population and GPD respectively (figure 1). Therefore, it is particularly important to treat this province carefully, so it was decided to analyse the development level of `Bueonos Aires` based on the results from section 1. In all scenarios considered, four principle components were enough to explain at least 80% of the variance (figure 16). This method of choosing the number of principle components was preferred due to the number of provinces being larger than the number of indicator variables.

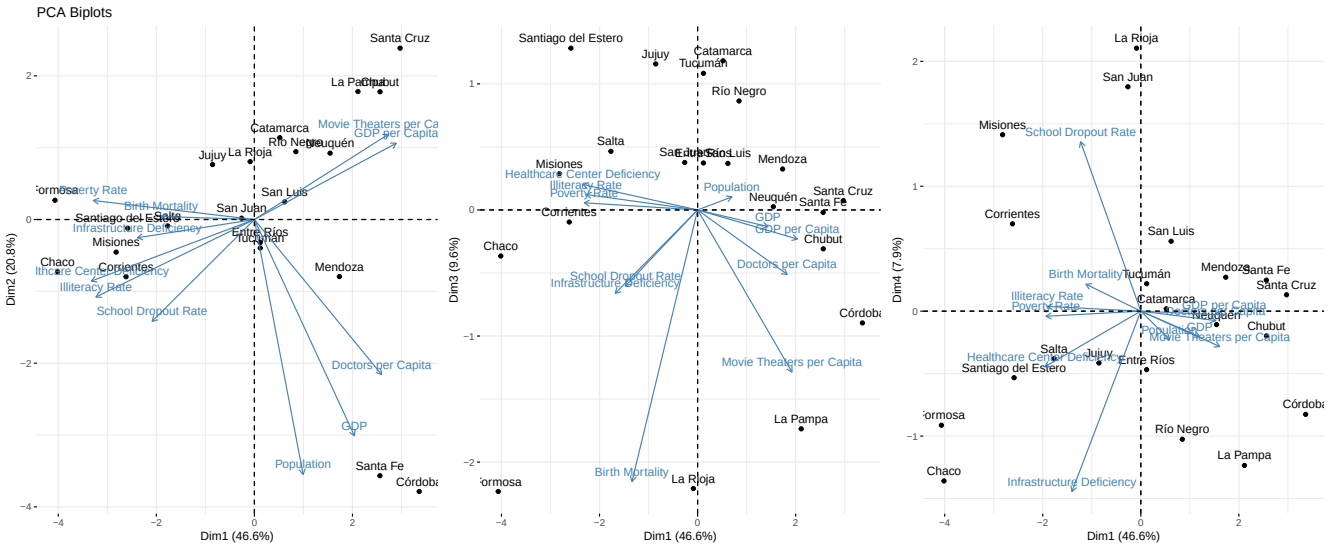

Figure 3: A selection of Bi-plots to visualize the patterns, not evident in the raw Argentina data, uncovered by PCA. Note: `Buenos Aires` excluded from PCA

Principle Component Pattens: The first principle component explains 45.3% of the variance and correlates the most with `Healthcare Center Deficiency` (-0.387), `Poverty Rate` (-0.383) and `Illiteracy Rate` (-0.377) (figures 3, 17 and table 2). In fact, we observe that the first principle component is strongly negatively correlated with the above three variables and `Infrastructure Deficiency`, `School Dropout Rate`, `Birth Mortality`; and, moderately positively correlated with `GPD per Captia` and `Movie Theaters per Capita`. The second principle component explains 20.3% of the variance, correlates the most with `GDP` (-0.530) and `Population` (-0.619), and `Doctors per Capita` (-0.376) (figures 3, 17 and table 2), and is strongly negatively correlated to `GPD`, `Population` and `Doctors per Capita`. We also observe that the main contributions to the third and fourth principle components are from `Birth Mortality` and `Movie Theatres per Capita`, and, `Infrastructure Deficiency` and `School Dropout Rate` respectively (figures 3, 17 and table 2).

The sample Bi-plots in figure 3 therefore aid visualization and identification of clusters of provinces with similar development levels. The first principle component seems to represent a 'social vulnerability score', with provinces such as `Chaco` and `Misiones` being significantly more socially vulnerable than provinces like `Santa Cruz` and `La Pampa`. We may interpret provinces similar to `San Juan` and `Santa Fe` being socially vulnerable for some characteristics. For example, `Santa Fe` has contrasting characteristics i.e. large `School Dropout Rate` and `Doctors`

`Per Capita`. Therefore, the first principle component indicates three different clusters: developed (`Santa Cruz`,`La Pampa`, etc.), emerging (`San Juan`, `San Luis`, etc.), and developing (`Misiones`,`Chaco` etc.) respectively. The second principle component is the driving factor between the segregation of `Santa Fe` and `Córdoba` from provinces such as `San Luis`, suggesting a fourth cluster. The conclusions drawn are also supported by analysis conducted in section 1 (figure 1)

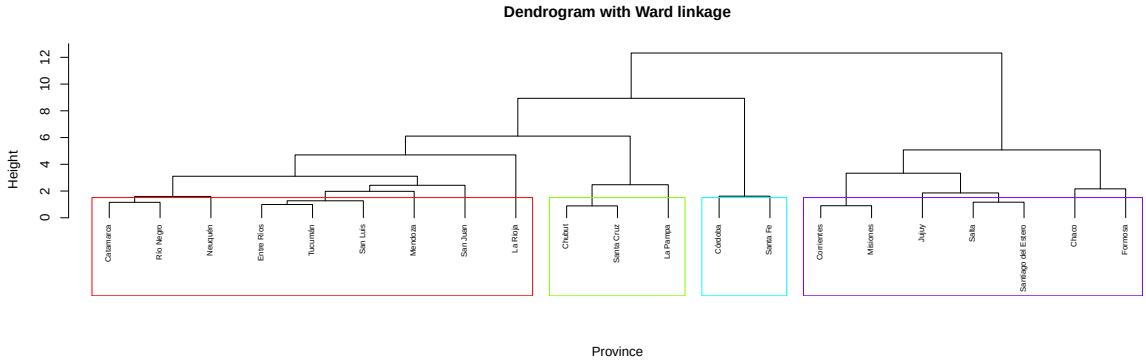

Figure 4: Hierarchical clustering using ward linkage cutting at four clusters. Note: Ward linkage is based on the multidimensional variance like PCA so is certainly theoretically the best choice for clustering with our compressed Argentina data.

One reason ward linkage was used is because it had a higher agglomerative coefficient (0.868) than complete (0.820), average (0.735) and single (0.520) linkage. We choose to cut the dendrogram into four clusters because this division yields distinct clusters that align well with the analysis completed in section 1 and offer meaningful insights. Looking at the branch heights, the purple clustering is very dissimilar from the other three clusters (figure 4). And Indeed, notice that if it was decided to cut at three clusters, merging red and green (figure 4) produces a new heterogeneous cluster. Despite the increment of internal cohesion (total WCV) of the clusters starting to be negligible after approximately 3 clusters (figure 18), a sharp increase in the gap statistic was observed at 4 clusters (figure 18). Therefore, we also consolidate these clusters by computing the `kmeans` algorithm and observing that only the provinces `Neuquén` and `Jujuy` change clusters (figure 19).

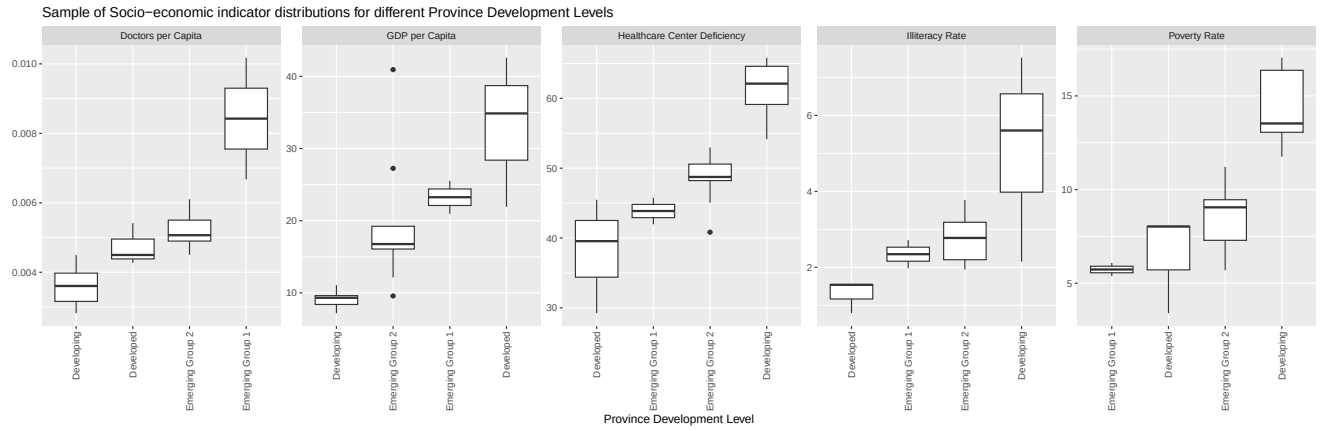

Figure 5: Sample of Socio-economic indicator distributions for different Province Development Levels

Development Groupings: With the exclusion of `Buenos Aires`, the remaining regions were clustered into four separate development levels: Developed, Emerging Group 1, Emerging group 2, and Developing (figure 6). It was decided for `Buenos Aires` to be classified as an emerging province. This was because excluding population and GPD, it holds sufficiently similar attributes to emerging group 2 (figure 1). From the dendrogram branch heights (figure 4), it is clear that there is a significant difference between the attributes of the developing provinces, and the emerging and developed provinces. This difference is much more significant than differences obtained between emerging and developed. Figure 5 illustrates that developed provinces have sufficient levels of most of the desired

attributes, and developing countries have none (figures 6, 20, 21). Whereas, emerging provinces hold a mixture of desirable and undesirable attributes, such as emerging group 1 having a high `Doctors per Capita`, but a high `illiteracy rate`. We note the significance of these findings in terms of population effected. Developed provinces hold 2.97% of the population compared to developing 17.4% and emerging 79.63%. By far the majority of people live in provinces with either very poor or mixed demographics in terms of socio-economic prosperity.

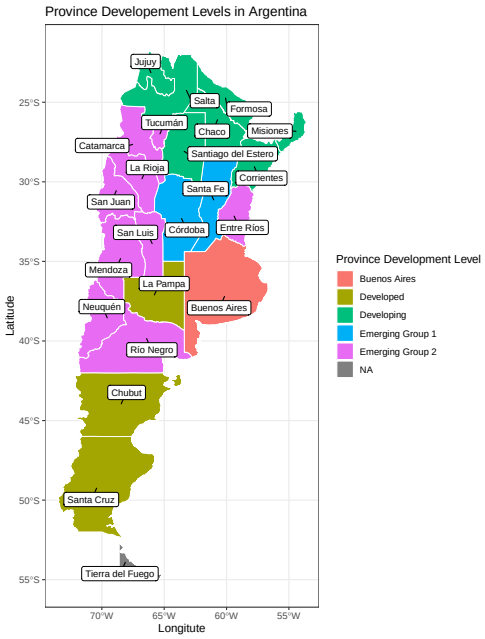

Figure 6: Province Development Levels across Argentina

Section 3: Conclusion

All developing (high priority) provinces are in the north of Argentina, with developed provinces in the south of Argentina. Developing provinces share characteristics such as high `Illiteracy Rate`, `Poverty Rate`, `Infrastructure Deficiency`, `Birth Mortality`, `Healthcare Center Deficiency`, and `School Dropout Rate`, and low `GDP per Capita`, and `Movie Theaters per Capita`. In contrast, developed provinces display the opposite characteristics: high `GDP per Capita` and `Movie Theaters per Capita`, and low `Illiteracy Rate`, `Poverty Rate`, `Infrastructure Deficiency`, `Birth Mortality` and `School Dropout Rate`. Emerging provinces have varying levels of the socio-economic indicators above, and are located in the middle of Argentina. Our study had some limitations. First, we did not consider the proportion of possible undocumented events. For example hypothetically, the socio-economic influence COVID-19 had on indicator variables for different provinces is not represented in the methodology. Furthermore, we cannot completely exclude the effects that `Tierra del Fueg` might have on the province development levels or our analysis. Therefore, further research and data should be completed and collected to refine our interpretations. Further analyses should include additional indicators that have the potential to influence the province development groups allocated, such as life expectancy, population density, crime rates etc. A natural extension to this consulting project would be to investigate spatial dependence with the aim of identify global and regional patterns in the data (global and local morans I tests). Our findings provide evidence for spatial dependencies for different socio-economic indicator variables across Argentina, particularly the northern provinces (figures 7, 8). Supervised learning techniques could be applied to assess previous impacts of implemented public policies. Hence, aid decision making for new public policies to help improve the life of Argentinians, particularly those living in developing provinces in the north. In terms of possible public policies, the report suggested direct investment into high priority provinces in education systems (e.g. figures 1,4,6,17,20). This intern should reduce the school dropout rate and illiteracy rate, fostering socio-economic development and reducing poverty. The report also recognized that major funding is needed in high priority provinces in healthcare centers (e.g. figures 1,4,6,17,20). This would not only attract more doctors, but also reduce the poverty rate.

Generative AI Statement

No generative AI was used in the creation of this report.

Appendix

Tables

Table 1: Initial Summary Statistics for Argentinian Province Data

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
GDP	22	30557028	61830995	3807057	8041587	19994520	292689868
Illiteracy Rate	22	3.2	1.9	0.79	2	3.7	7.5
Poverty Rate	22	9.9	3.8	3.4	7.5	13	17
Infrastructure Deficiency	22	13	7.2	3.8	7.6	16	31
School Dropout Rate	22	1.7	1.2	0.2	0.81	2.5	3.9
Healthcare Center Deficiency	22	51	9.2	29	46	57	66
Birth Mortality	22	5	3.5	0.8	3	5.9	16
Population	22	1686352	3219828	273964	514372	1230606	15625084
Movie Theaters per Capita	22	0.0000071	0.0000044	0.0000018	0.0000041	0.0000093	0.000019
Doctors per Capita	22	0.0049	0.0015	0.0028	0.0041	0.0053	0.01
GDP per Capita	22	18	10	7.2	9.6	22	43
Population Percentage	22	4.5	8.7	0.74	1.4	3.3	42

Table 1: Summary statistics

	PC1	PC2	PC3	PC4
GDP	0.2376507	-0.5249755	-0.0482861	-0.0978973
Illiteracy Rate	-0.3766381	-0.1880398	0.0446523	0.0181129
Poverty Rate	-0.3826203	0.0461370	0.0206992	-0.0196182
Infrastructure Deficiency	-0.2780186	-0.0442505	-0.2417211	-0.6920106
School Dropout Rate	-0.2429421	-0.2471443	-0.2196408	0.6500739
Healthcare Center Deficiency	-0.3874737	-0.1494041	0.0737959	-0.2137060
Birth Mortality	-0.2215270	0.0074084	-0.7868454	0.1035215
Population	0.1158030	-0.6187765	0.0378156	-0.1090935
Movie Theaters per Capita	0.3185384	0.2069644	-0.4697101	-0.1357136
Doctors per Capita	0.3021658	-0.3761657	-0.1866225	-0.0361054
GDP per Capita	0.3369884	0.1851460	-0.0843681	-0.0147779

Table 2: Raw contributions for principle components computed excluding `Buenos Aires`

Figures

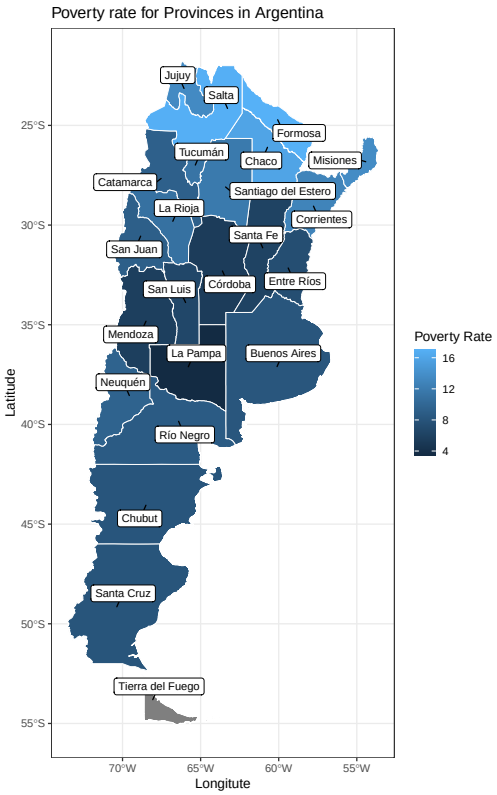

Figure 7: Poverty rate for Provinces in Argentina

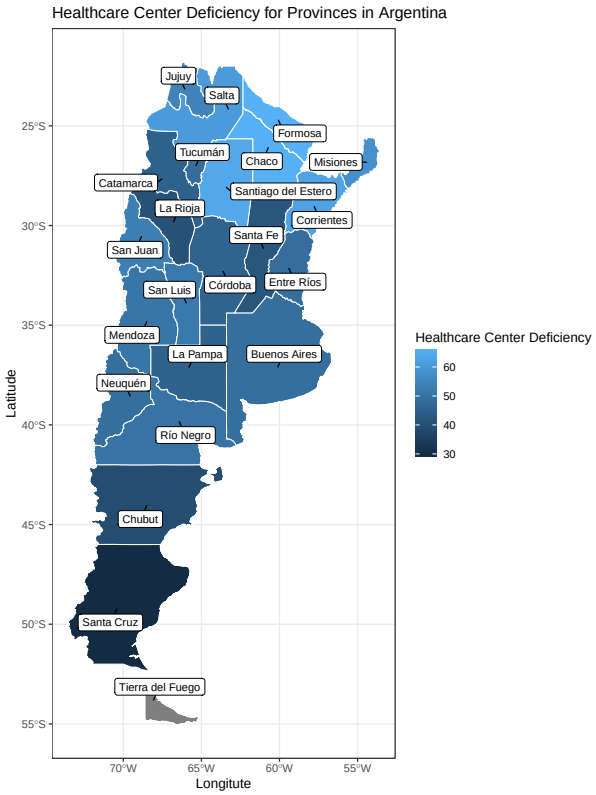

Figure 8: Healthcare Center Deficiency for Provinces in Argentina

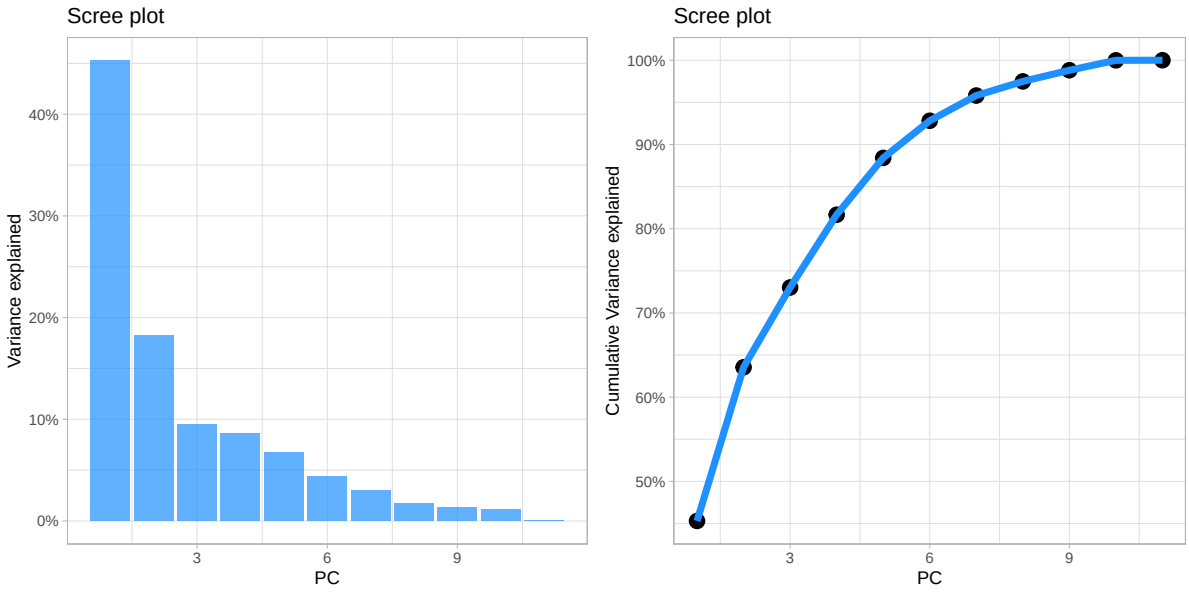

Figure 9: Note 'Buenos Aires' included in PCA for principle components (only 10 displayed). Left plot shows a scree plot for total variance explained by each principle component. Right plot shows the cumulative variance explained at each principle component.

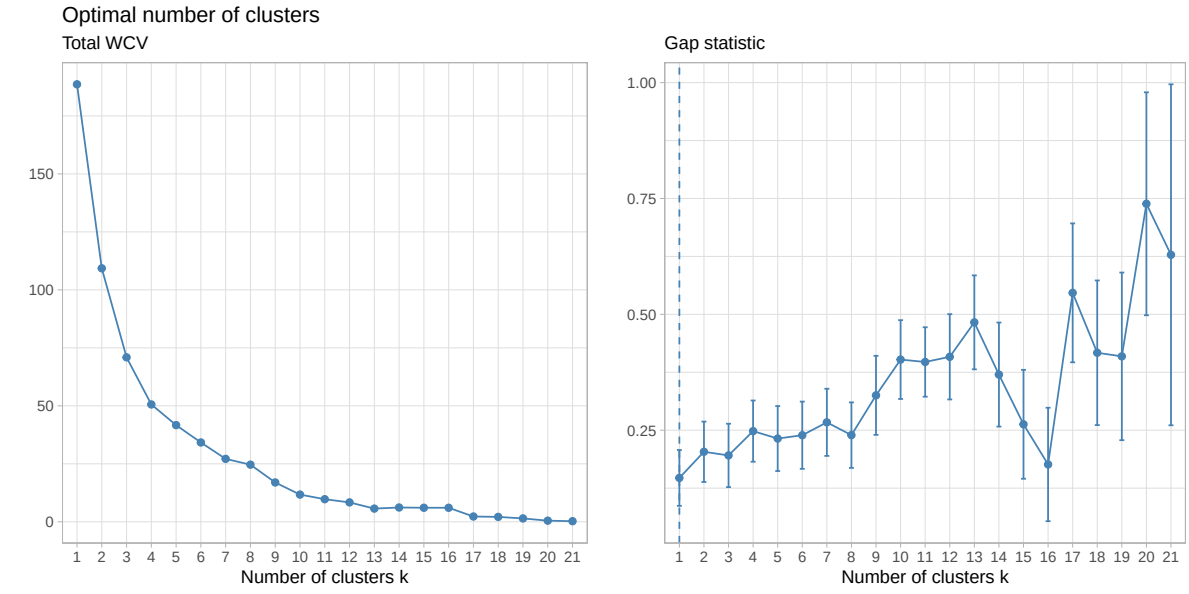

Figure 10: Note 'Buenos Aires' included in PCA. Left plot shows the increment of internal cohesion (total WCV) with respect to the number of clusters. Right plot shows the value of the gap-statistic with respect to the number of clusters

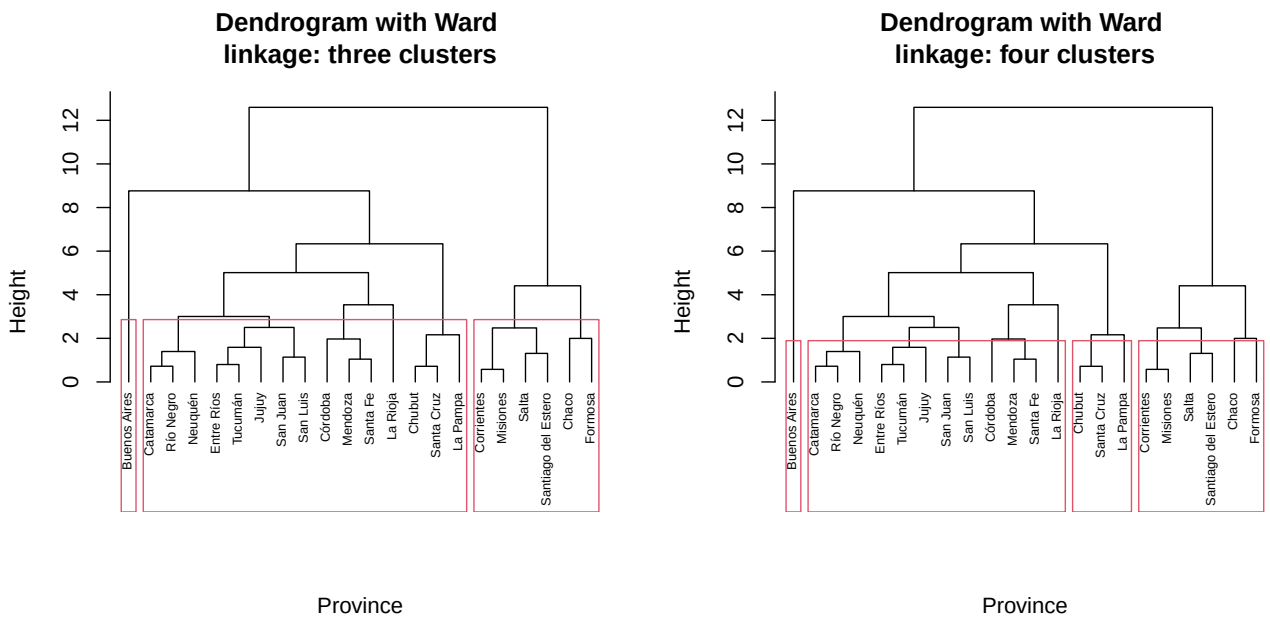

Figure 11: Note `Buenos Aires` included in PCA and clustering, and ward linkage agglomerative coefficient closest to one. Ward linkage dendrograms with cutting at three and four clusters. `Buenos Aries` remains alone.

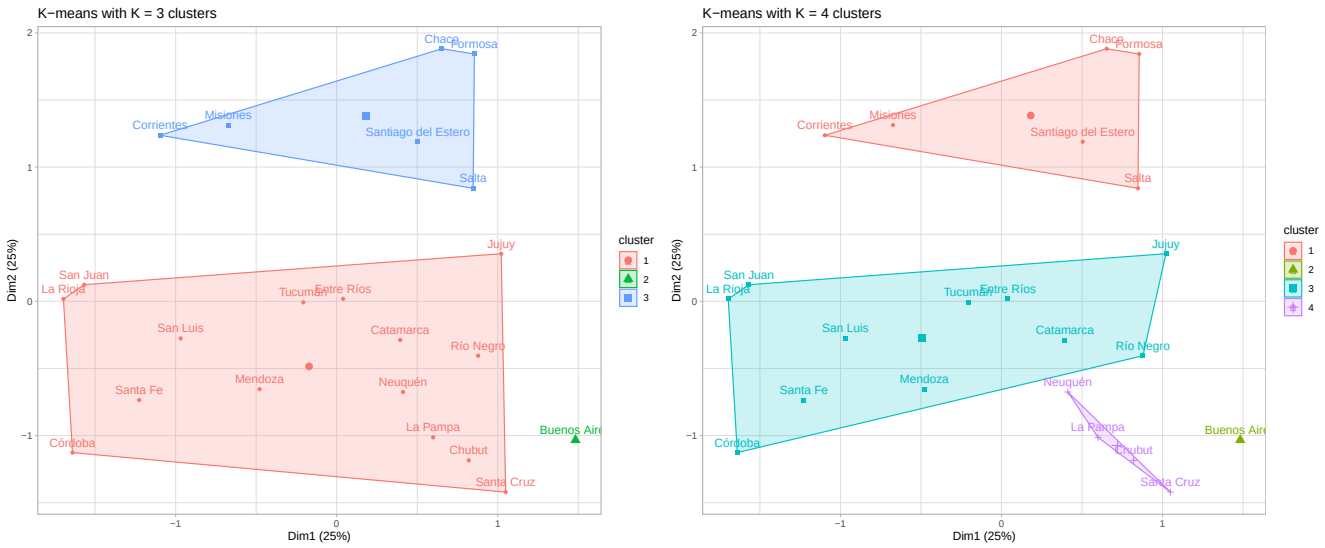

Figure 12: Note `Buenos Aires` included in PCA and clustering. K-means clustering for 3 and 4 pre-defined clusters. `Buenos Aries` remains alone.

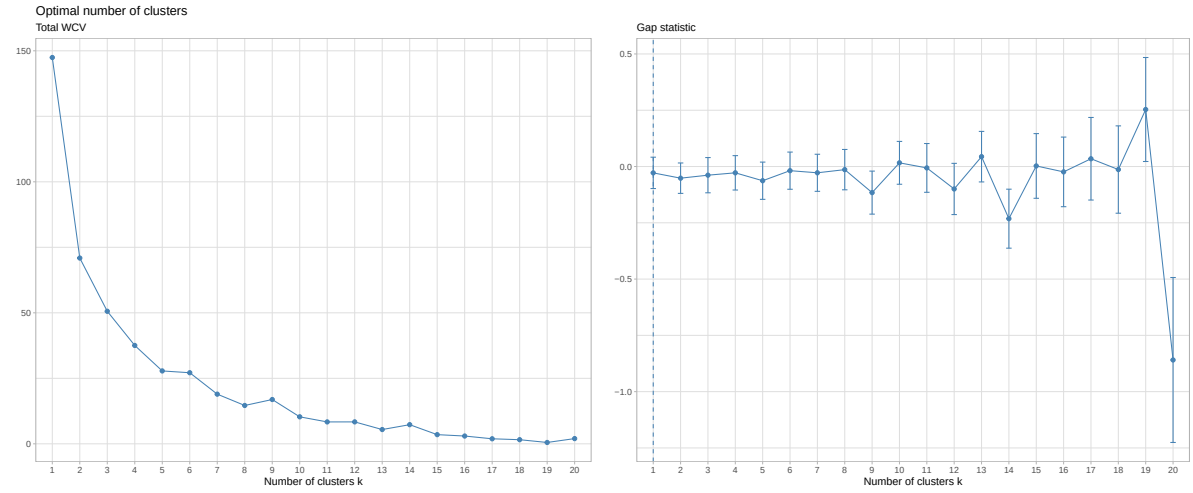

Figure 13: Note, 'Buenos Aires' in PCA but not in Clustering algorithms. Left plot shows the increment of internal cohesion (total WCV) with respect to the number of clusters. Right plot shows the value of the gap-statistic with respect to the number of clusters

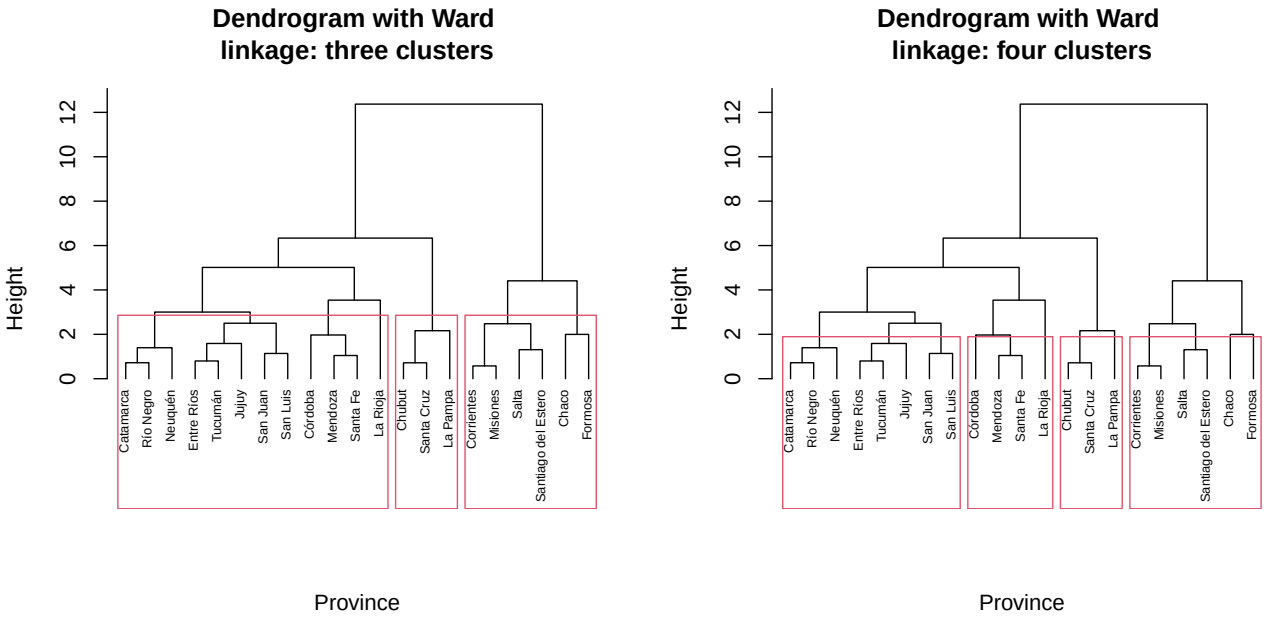

Figure 14: Note `Buenos Aires` included in PCA, not included in clustering, and ward linkage agglomerative coefficient closest to one. Ward linkage dendrograms with cutting at three and four clusters.

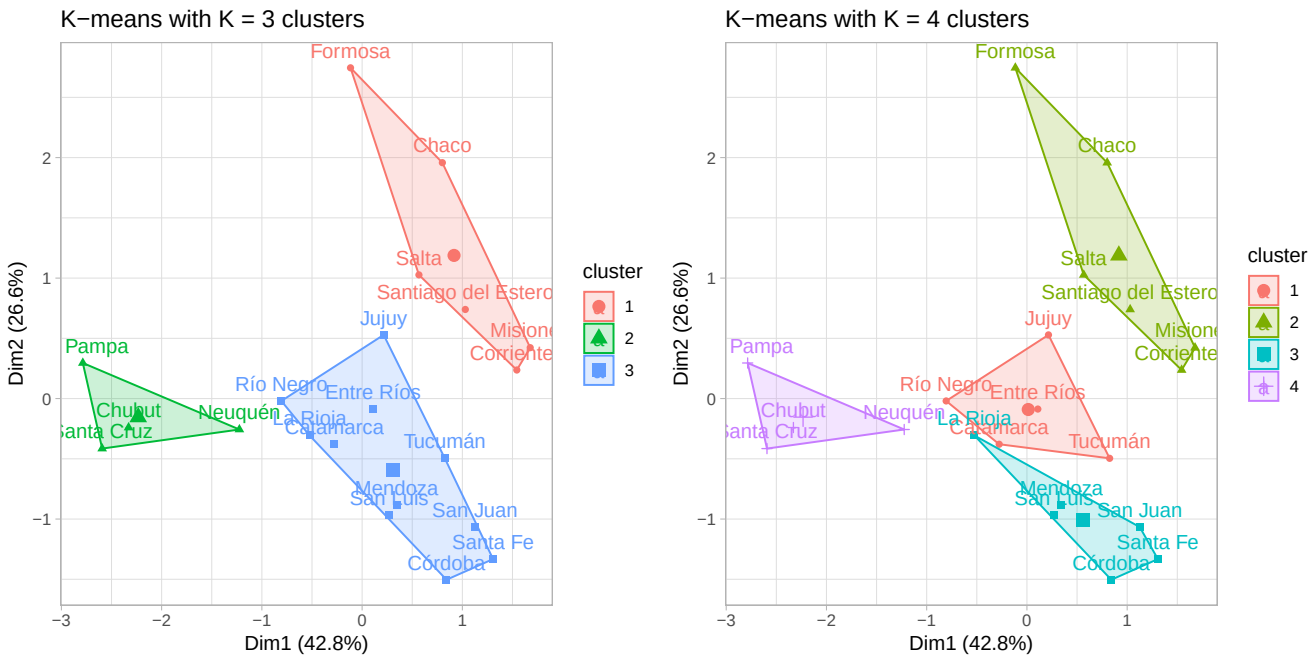

Figure 15: Note `Buenos Aires` included in PCA but not in clustering. K-means clustering for 3 and 4 pre-defined clusters.

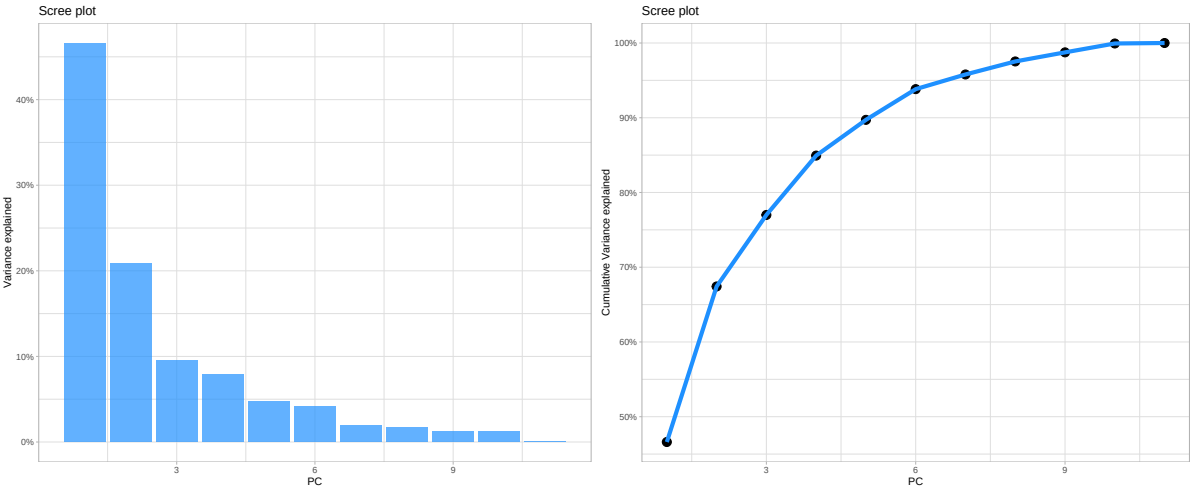

Figure 16: Note 'Buenos Aires' not in PCA for principle components (only 10 displayed). Left plot shows a scree plot for total variance explained by each principle component. Right plot shows the cumulative variance explained at each principle component.

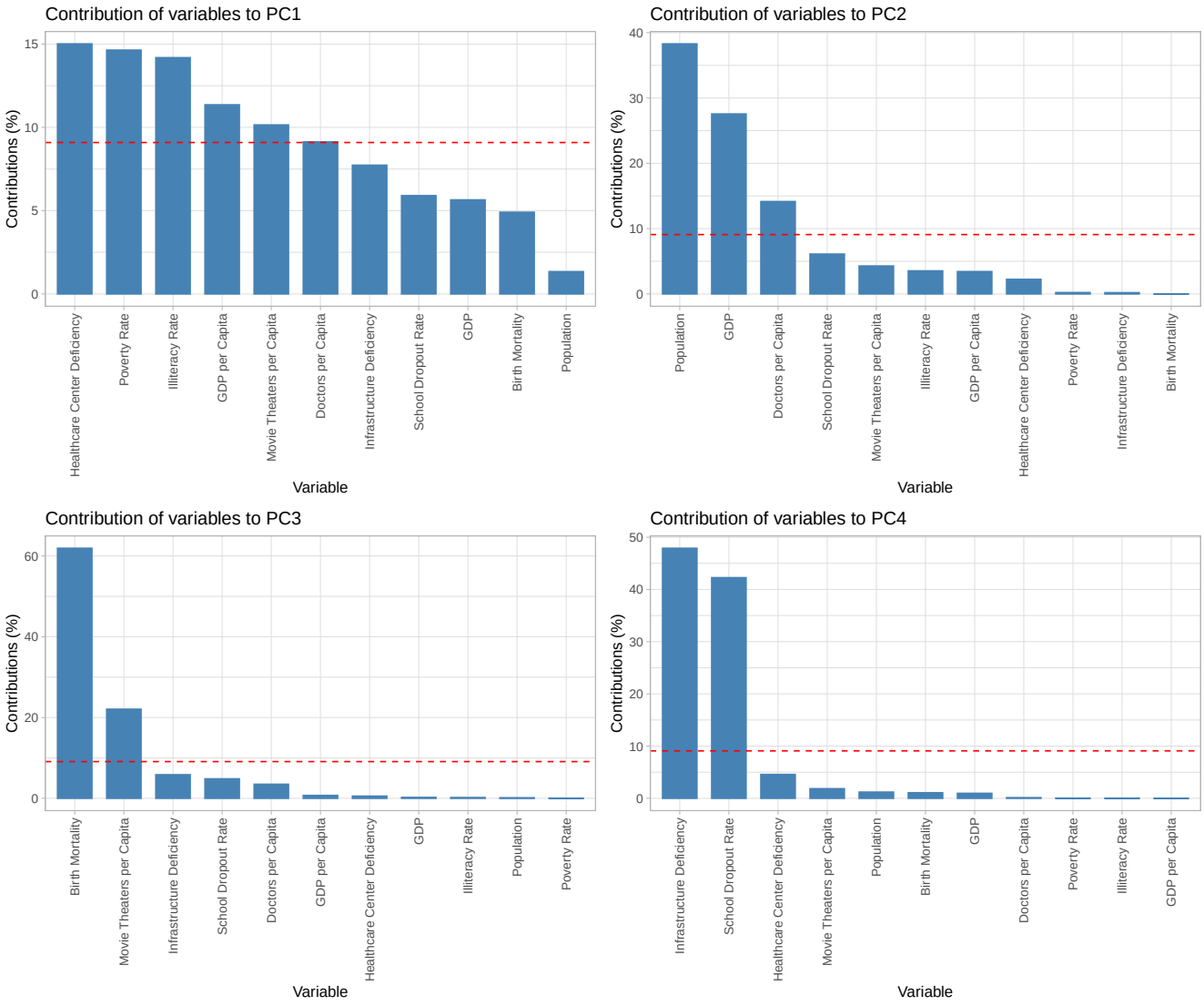

Figure 17: Note `Buenos Aires` not in PCA. Contribution of socio-economic indicator variables to principle components 1-4

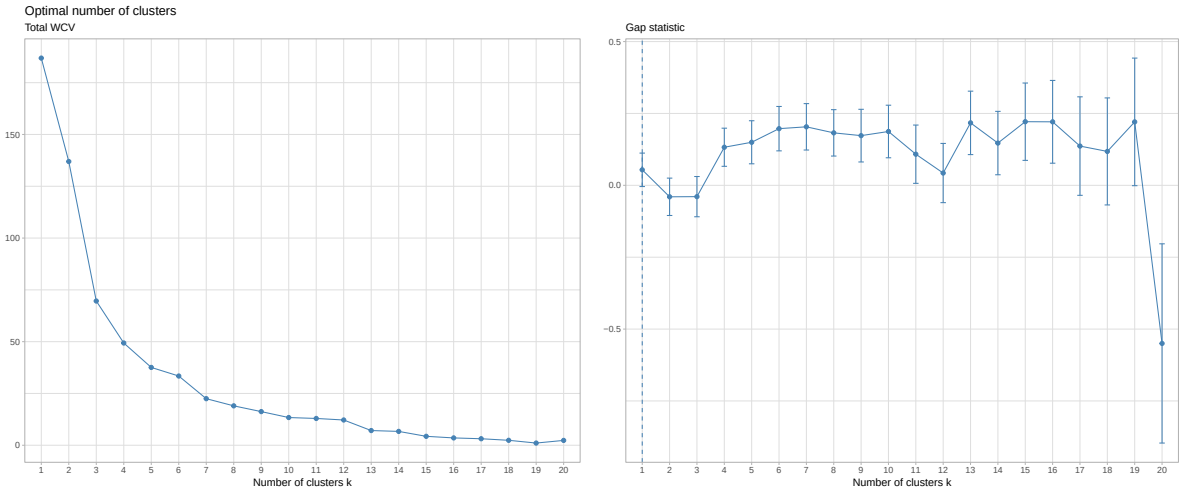

Figure 18: Note, 'Buenos Aires' not PCA. Left plot shows the increment of internal cohesion (total WCV) with respect to the number of clusters. Right plot shows the value of the gap-statistic with respect to the number of clusters

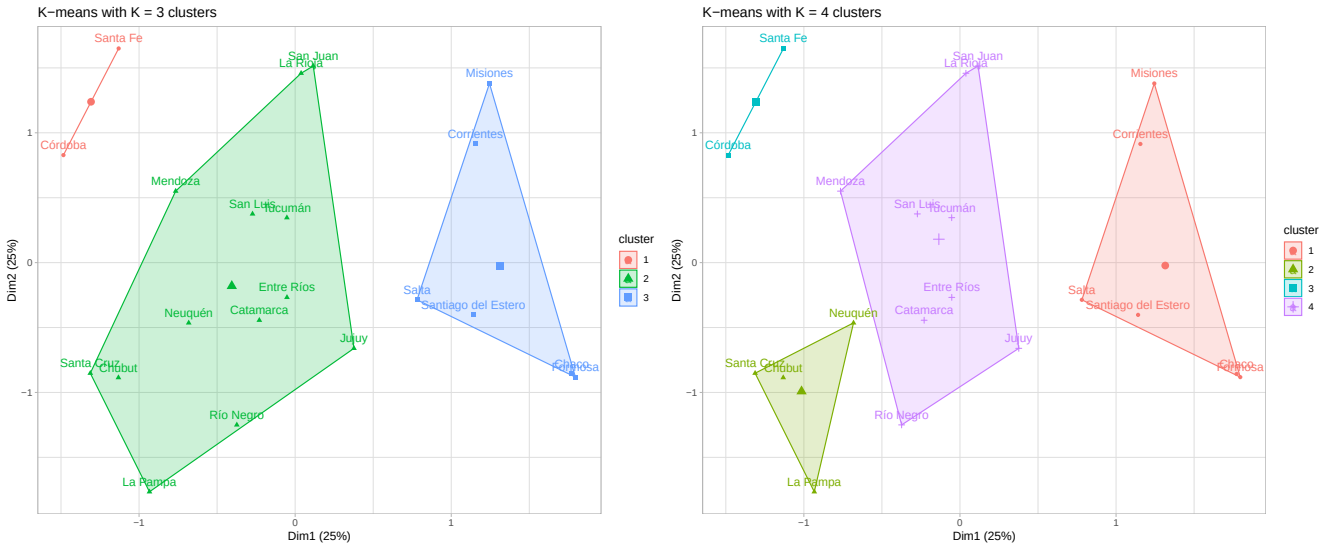

Figure 19: Note `Buenos Aires` not in PCA. K-means clustering for 3 and 4 pre-defined clusters.

Figure 20: A selection of Bi-plots to visualize the patterns, not evident in the raw Argentina data, uncovered by PCA for each province development level.

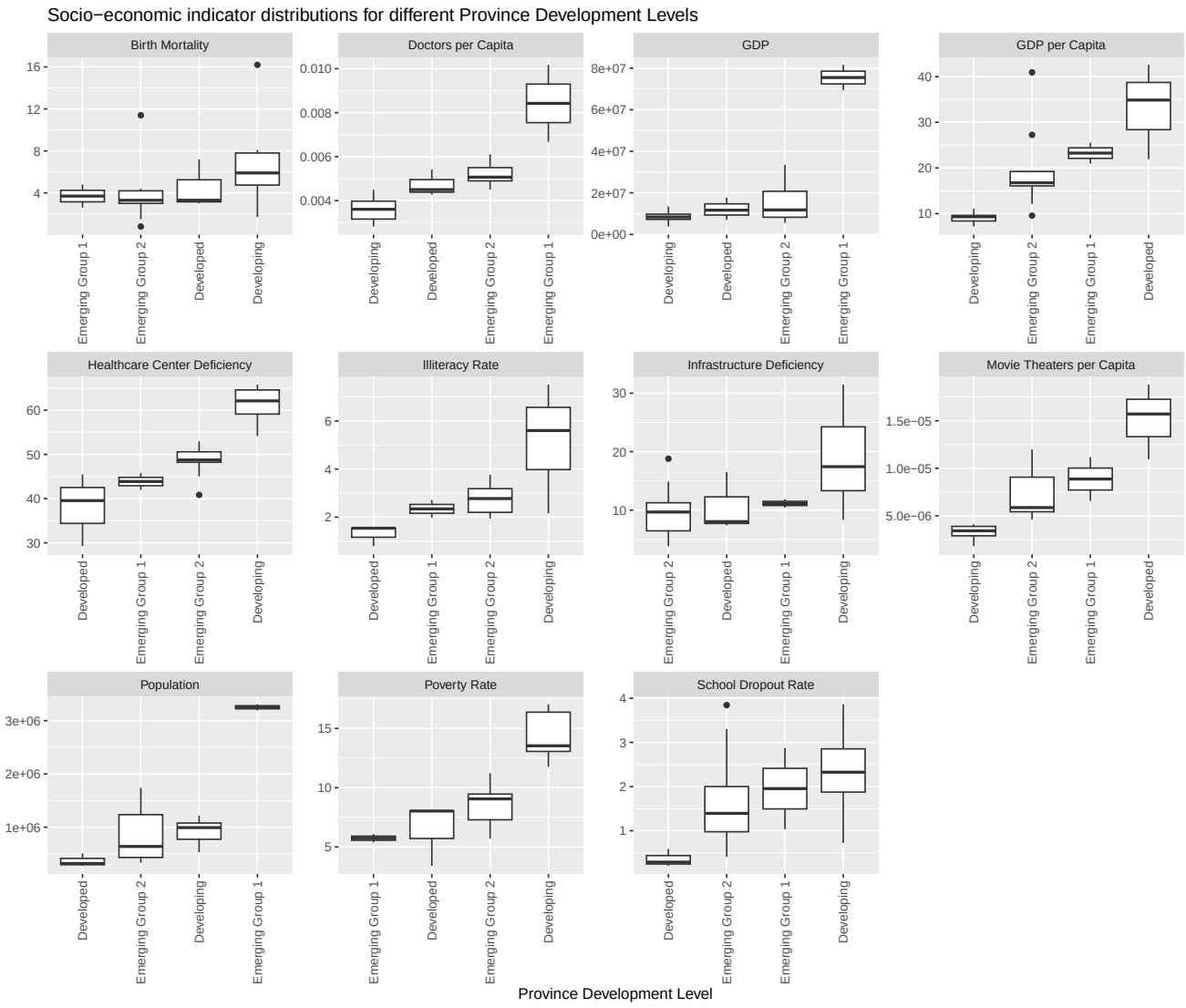

Figure 21: Complementory box plots for different indicator variables for each province development group