Data Analysis Project

Study of the influence of earnings, education and police work on crime rates in Portuguese municipalities

Victor Malheiro

Introduction

With this project, the main goal is to study the impact of several variables such as earnings by gender, education, police work, unemployment and purchasing power on crime for all the Portuguese municipalities.

To work on the univariate and bivariate analysis we used the SPSS software that allows us to analyze the table with descriptive statistics of all variables, box plots and histograms for univariate analysis, correlation and covariance tables and scatter plots for the bivariate analysis.

For the multivariate analysis was used the SPAD software to do the principal components analysis and the cluster analysis.

To build this dataset, the information was gathered from the Pordata website, with all the variables that will be used for the analysis.

The dataset has three variables for earnings, AMEE – Average monthly earnings of employees, AMEE_M – Average monthly earnings of male employees and AMEE_F – Average monthly earnings of female employees . The variables were divided by gender because this will allow to have an analysis divided by gender to see if that has some impact on the crime levels of each municipality.

The variable AARU – Average annual rate of unemployed registered at the employment center, was to understand if crime is related to the unemployment level of each municipality.

In order to study the impact of the gender on crime, the dataset has the following two variables, AARRP_M – Average annual rate of men in the resident population and ARRP_F – Average annual rate of women in the resident population.

The dataset also has three variables related with crime and judicial system, ACR – Annual crime rate registered by the police per 1000 inhabitants, ER – Effectiveness rate (processes resolved / processes awaiting decision) and RR – Resolution rate (solved processes / new processes). The goal is to understand the relation between more efficient judicial systems in the different municipalities, the criminal rate and the other variables.

There are two variables for the education levels, P_15_LE – Rate of resident population over 15 years old with lower education level (lower than higher education) and P_15_HE – Rate of resident population over 15 years old with higher education level. With these variables the objective is to understand the relation between the level of education and the criminal rate.

Finally, the dataset has the variable PCPP – Per capita purchasing power (where the national average purchasing power value is equal to 100). This variable is useful to understand how the purchasing power relates to crime and the other variables that the dataset has that can have influence on crime.

Univariate Analysis

The first step in the univariate analysis was to produce a table with the descriptive statistics for all variables where several indicators could be analyzed for each variable to understand how to better classify the municipalities.

Descriptive Statistics Table

		AMEE	AMEE_M	AMEE_F	AARU	AARRP_M	AARRP_W	ACR	ER	RR	P_15_LE	P_15_HE	PCPP
N	Válido	308	308	308	278	308	308	308	287	296	308	308	308
	Omisso	0	0	0	30	0	0	0	21	12	0	0	0
Média		925,5168	1014,0185	845,1162	4,0692	47,6067	52,3933	28,0429	71,3094	100,6314	89,9588	10,0412	80,1058
Erro de média padrão		8,89432	13,36430	5,80607	,08918	,06998	,06998	,53709	1,05634	1,46452	,25038	,25038	1,04589
Mediana		892,5863	966,8000	824,4500	3,7815	47,5614	52,4386	26,9223	68,8000	100,0000	91,1000	8,9000	76,8500
Moda		731,63[a]	831,60[a]	728,20[a]	1,86[a]	42,00[a]	43,29[a]	8,62[a]	50,00	100,00	90,00[a]	7,30[a]	61,90[a]
Erro Desvio		156,09460	234,54244	101,89603	1,48697	1,22808	1,22808	9,42584	17,89556	25,19651	4,39416	4,39416	18,35536
Variância		24365,523	55010,154	10382,802	2,211	1,508	1,508	88,846	320,251	634,864	19,309	19,309	336,919
Assimetria		3,229	4,639	2,271	1,194	1,490	-1,490	1,793	-,181	1,534	-1,892	1,892	2,382
Erro de assimetria padrão		,139	,139	,139	,146	,139	,139	,139	,144	,142	,139	,139	,139
Curtose		16,859	36,313	8,789	1,454	13,104	13,104	6,735	-,118	7,161	5,178	5,178	11,944
Erro de Curtose padrão		,277	,277	,277	,291	,277	,277	,277	,287	,282	,277	,277	,277
Intervalo		1403,53	2622,10	791,10	8,36	14,71	14,71	71,83	90,20	220,80	29,30	29,30	164,40
Mínimo		731,63	739,80	693,70	1,86	42,00	43,29	8,62	9,60	37,50	67,80	2,90	55,20
Máximo		2135,17	3361,90	1484,80	10,22	56,71	58,00	80,46	99,80	258,30	97,10	32,20	219,60
Soma		285059,17	312317,70	260295,80	1131,23	14662,86	16137,14	8637,21	20465,80	29786,90	27707,30	3092,70	24672,60
Percentis	25	829,7638	885,5500	783,2500	2,9075	47,0713	51,8048	21,7349	60,2000	88,4000	88,2000	7,2000	67,2250

50	892,5863	966,8000	824,4500	3,7815	47,5614	52,4386	26,9223	68,8000	100,0000	91,1000	8,9000	76,8500
75	976,0231	1075,8000	885,6500	4,8202	48,1952	52,9287	32,4613	86,7000	109,6500	92,8000	11,8000	89,0750

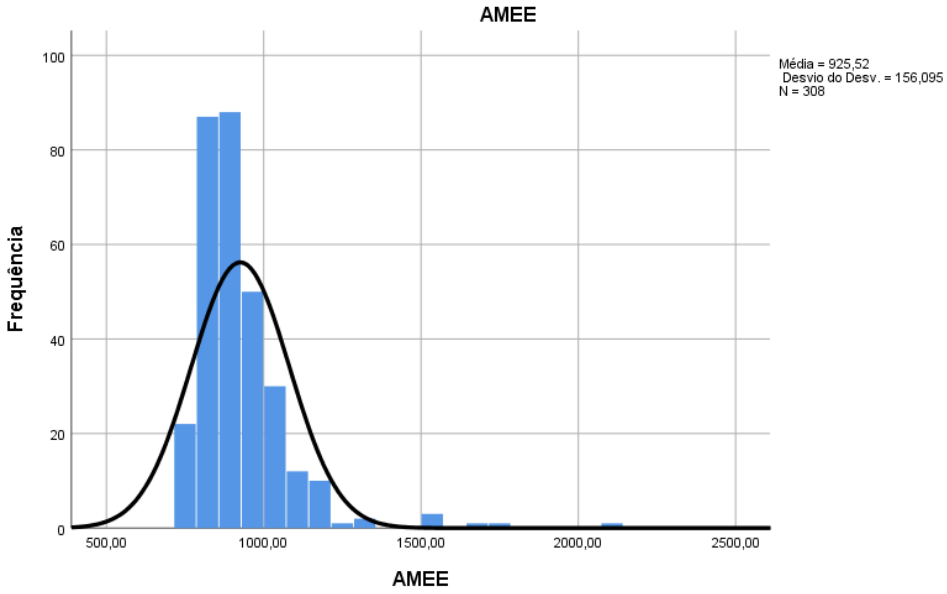
After having collected all the location, dispersion and shape indicators for all variables, the analysis was made. In addition to the indicators collected in the table of descriptive statistics, SPSS software was used to create the box plots and histograms for all the variables in order to identify outliers and analyze their dispersion.

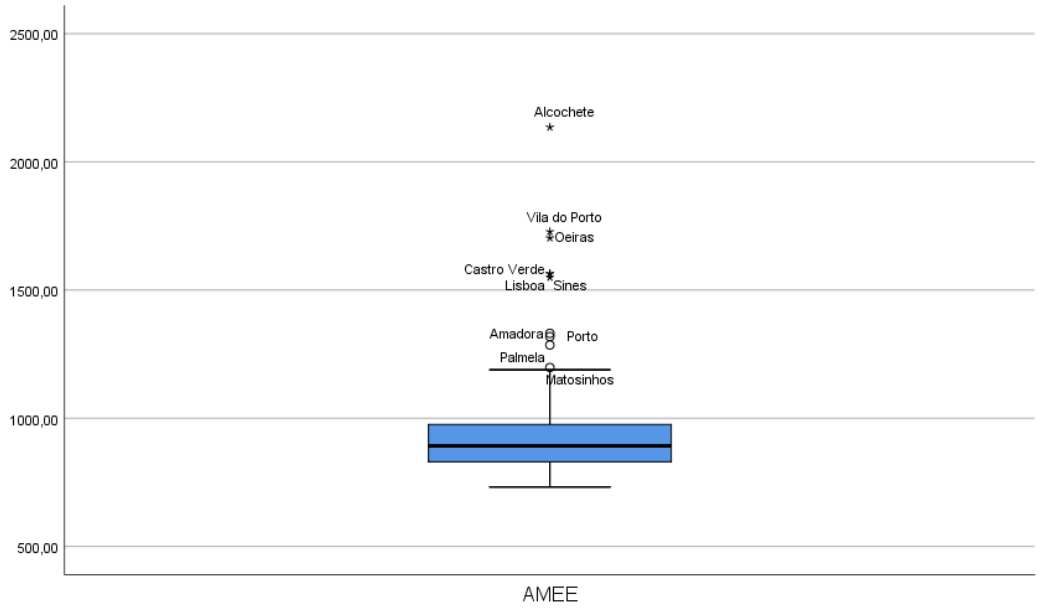
AMEE

Average monthly earnings of employees

The mean average monthly earning of an employee in Portugal is € 925,52. The maximum value for this variable is in Alcochete with a value of € 2135,17 and the minimum is € 731,63 in Celorico de Basto.

The distribution is asymmetric to the left with a value of 3,229 and the kurtosis is 16,859, that means the distribution is peaked. The box-plot indicates 6 upper severe outliers in Alcochete, Lisboa, Oeiras, Sines Castro Verder and Vila do Porto and 4 upper moderate outliers in Matosinhos, Porto, Amadora e Palmela.





AMEE_M and AMEE_F

Average monthly earnings of employees – Male

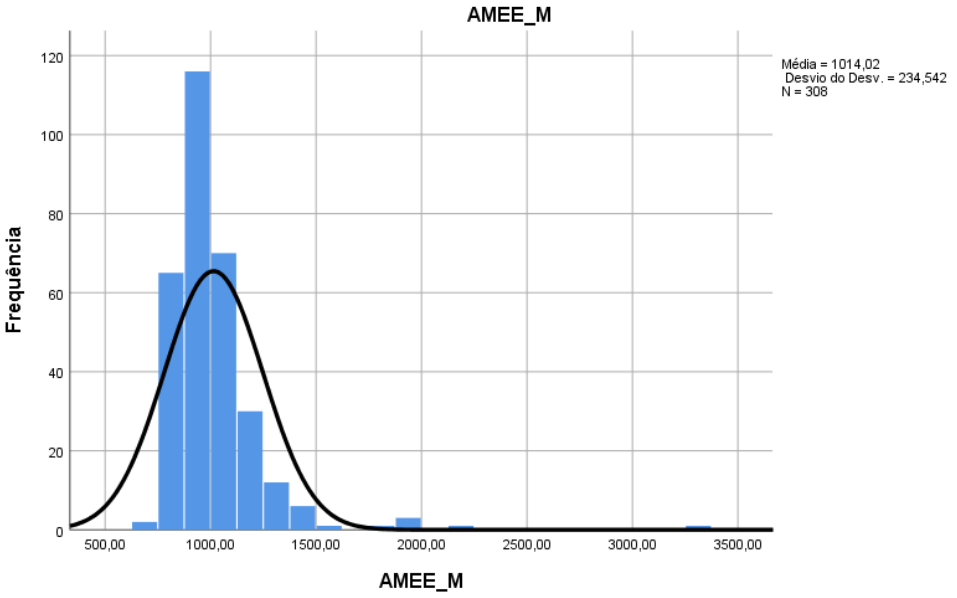
Average monthly earnings of employees – Female

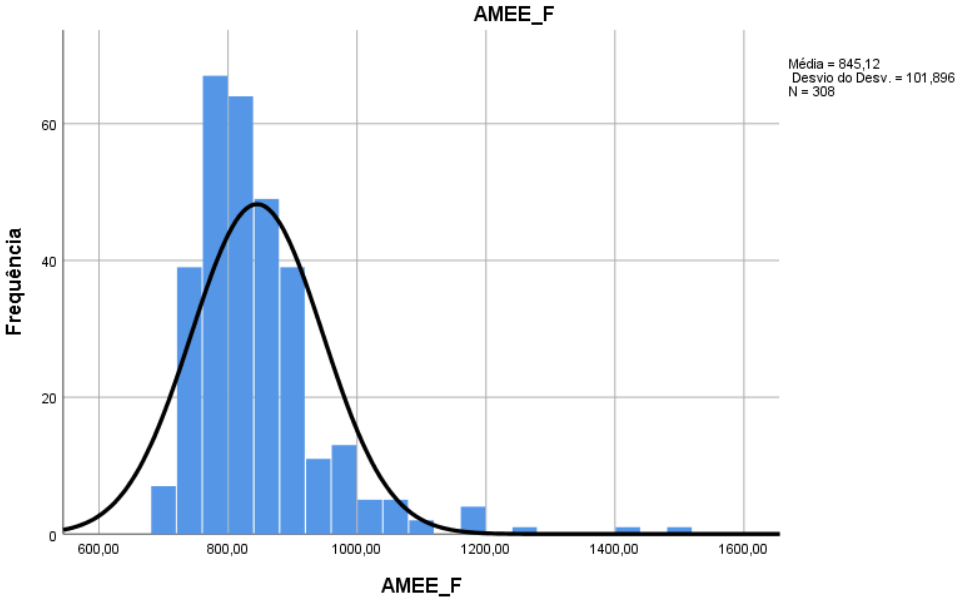
The maximum value for variable AMEE_M is 3361,9€ in Alcochete, the minimum is in Mondim de Basto with 739,80€ and the mean is 1014,02€. The distribution is asymmetric to the left with a value of 4,639 and the kurtosis is 36,313.

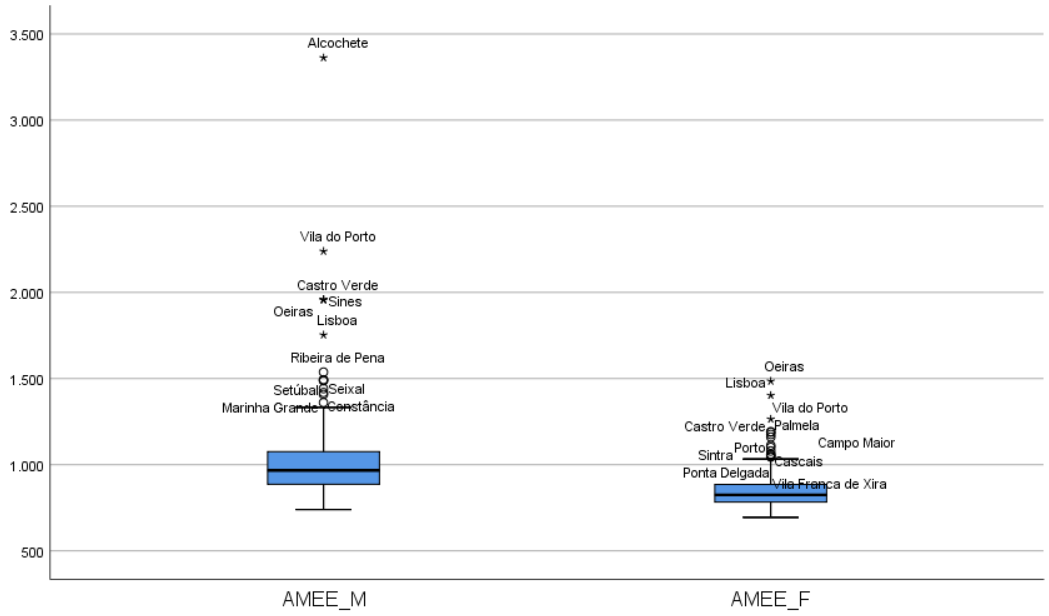
AMEE_F has a maximum of 1.484,8€ in Oeiras, the minimum is € 693,7 in Gavião and the mean is 845,12€. The kurtosis is 8,789 and the distribution is asymmetric to the left with a value of 2,271.

AMEE_M has 6 upper severe outliers in Alcochete, Vila do Porto, Sines, Oeiras, Castro Verde and Lisboa and 5 upper moderate outliers in Ribeira de Pena, Setúbal, Seixal, Marinha Grande and Constância.

AMEE_F has 3 upper sever outliers in Oeiras, Lisboa and Vila do Porto and 8 upper moderate outliers Palmela, Castro Verde, Campo Maior, Porto, Sintra, Cascais, Ponta Delgada and Vila Franca de Xira.


AMEE_M
Média = 1014,02
Desvio do Desv. = 234,542
N = 308
Frequência
120
100
80
60
40
20
0
500,00
1000,00
1500,00
2000,00
2500,00
3000,00
3500,00
AMEE_M


AMEE_F
Média = 845,12
Desvio do Desv. = 101,896
N = 308
Frequência
60
40
20
0
600,00
800,00
1000,00
1200,00
1400,00
1600,00
AMEE_F



AARU

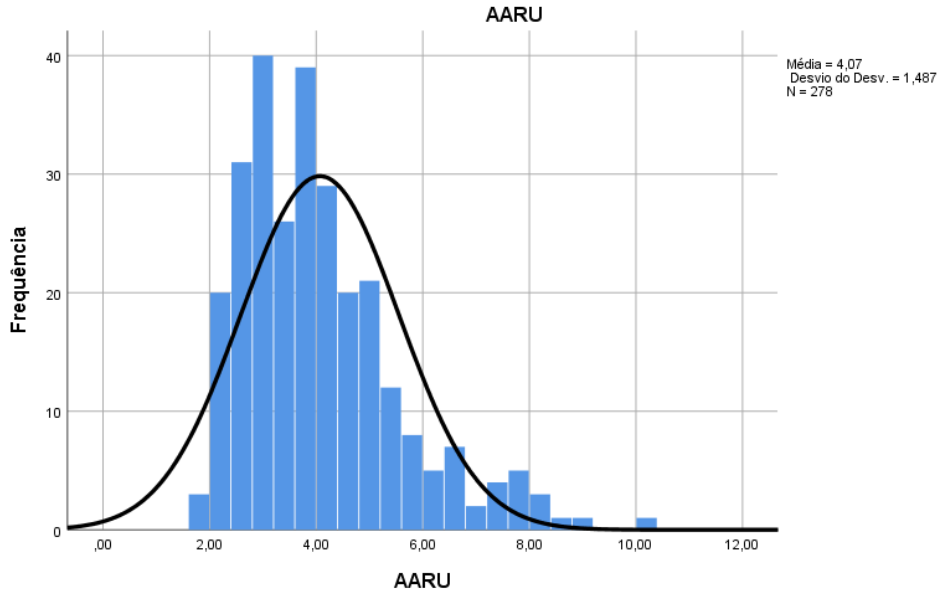
Average annual rate of unemployed registered at the employment center in the resident population

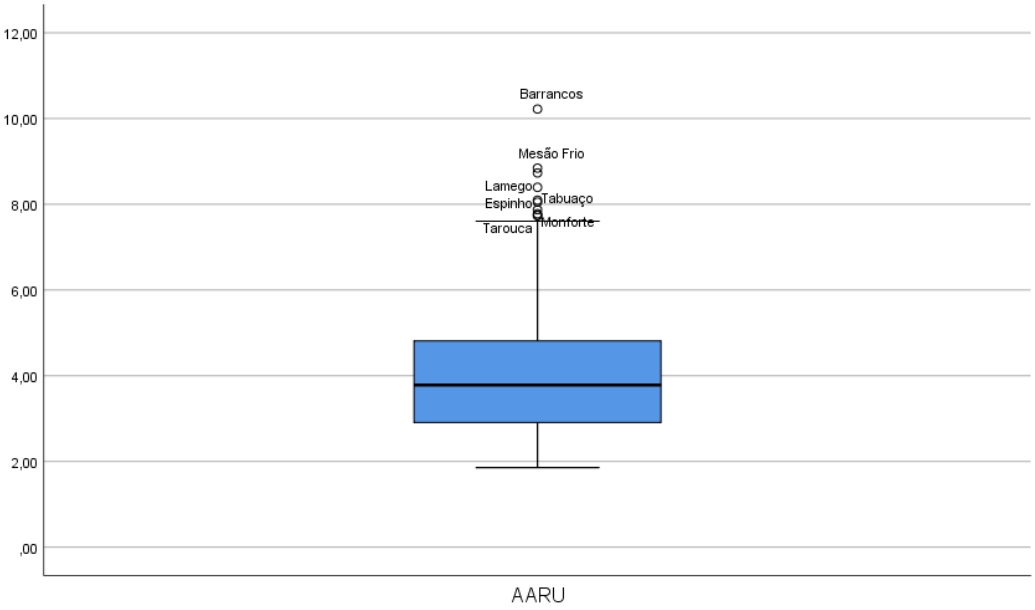
The mean average annual rate of unemployed resident population in Portugal is 4,07%. The maximum value for this variable is in Barrancos with a value of 10,22% and the minimum is 1,86% in Melgaço.

The distribution is asymmetric to the left with a value of 1,194 and the kurtosis is 1,454, with some of the data not concentrated.

The box-plot indicates seven upper moderate outliers in Barrancos, Mesão Frio, Lamego, Tabuaço, Espinho, Monforte and Tarouca.

This variable presents 30 missing values in Madeira and Açores municipalities.





AARRP_M and AARRP_W

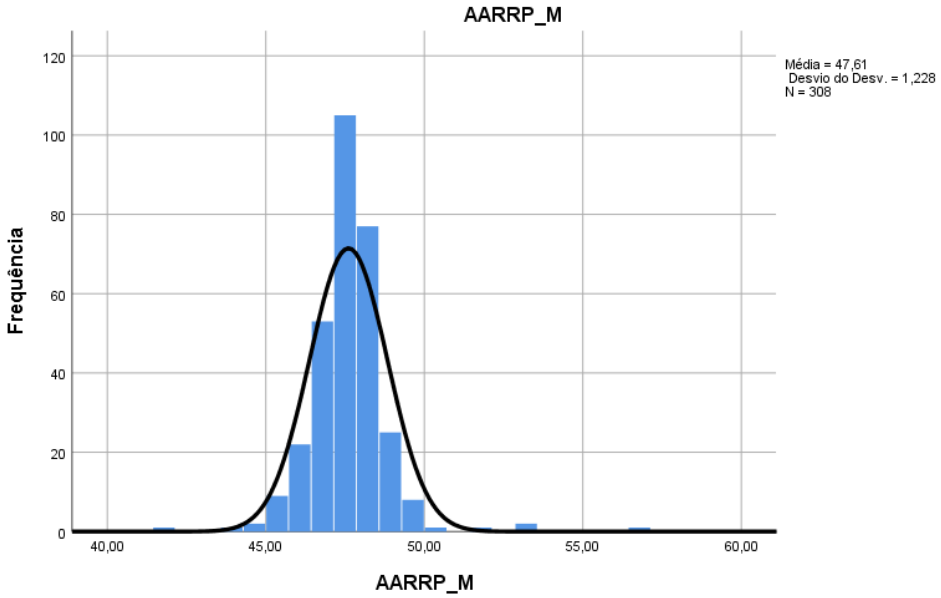
Average annual rate of men in the resident population
Average annual rate of women in the resident population

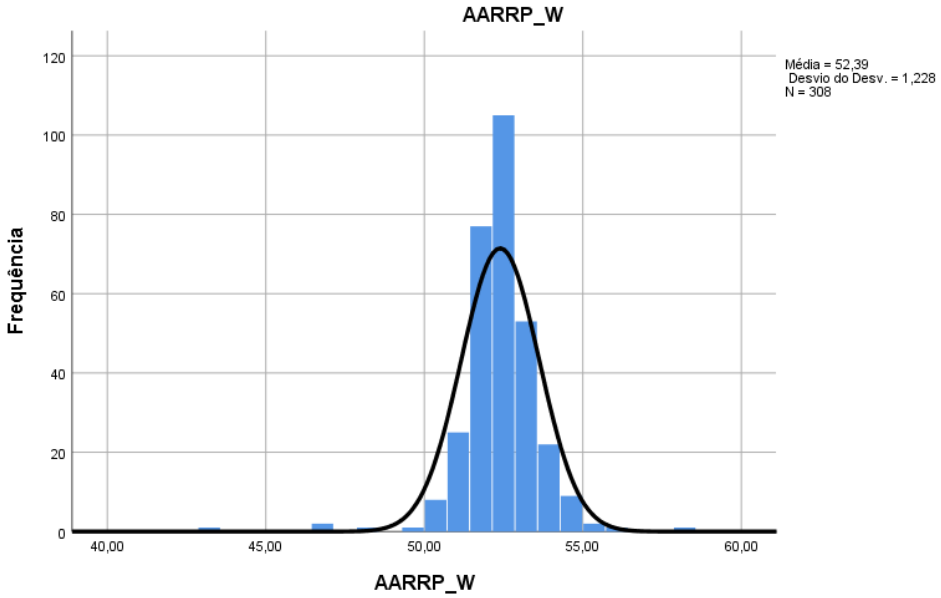
The maximum value for variable AARRP_M is 56,71% in Corvo, the minimum is in Porto Moniz with 42% and the mean is 47,61%. The distribution is asymmetric to the right with a value of 1,490 and the kurtosis is 13,104.

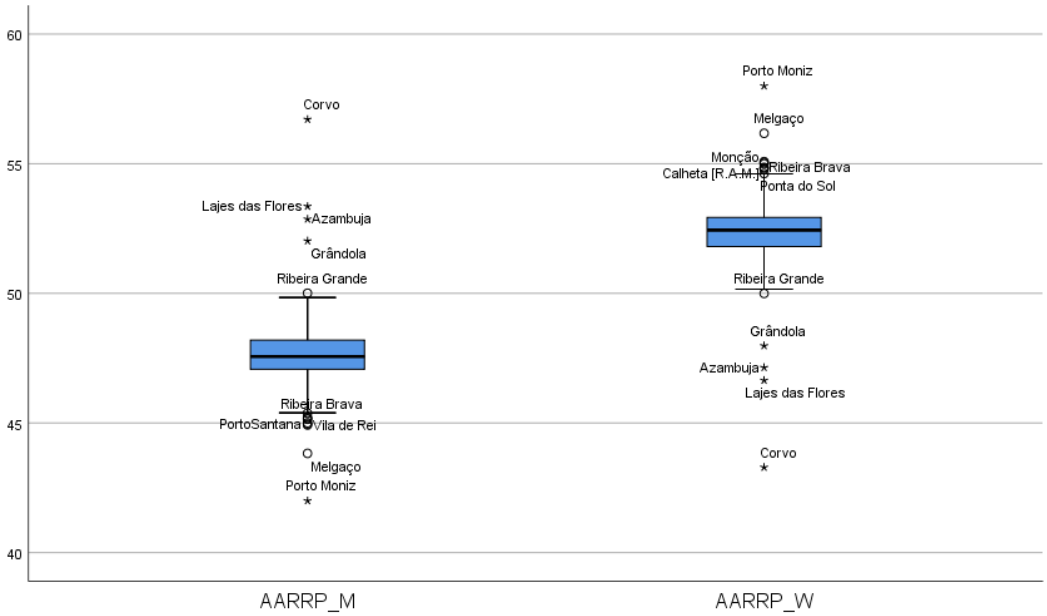
AARRP_W has a maximum in 58% in Porto Moniz, the minimum is 43,29% in Corvo and mean is 52,39%. The distribution is asymmetric to the right with a value of -1,490 and the kurtosis is 13,104.

AARRP_M has 4 upper severe outliers in Corvo, Lajes das Flores, Azambuja and Grândola and 1 upper moderate outlier in Ribeira Grande. This variable has also lower moderate outliers in Melgaço, Vila de Rei, Ribeira Brava and Porto and Santana and 1 lower severe outlier in Porto Moniz.

AARRP_W has upper moderate outliers in Melgaço, Monção, Ribeira Brava, Calheta, Ponta do Sol and has 1 upper severe outlier in Porto Moniz. This variable has also lower severe outliers in Corvo, Lajes das Flores, Azambuja and Grândula and 1 lower moderate outlier in Ribeira Grande.







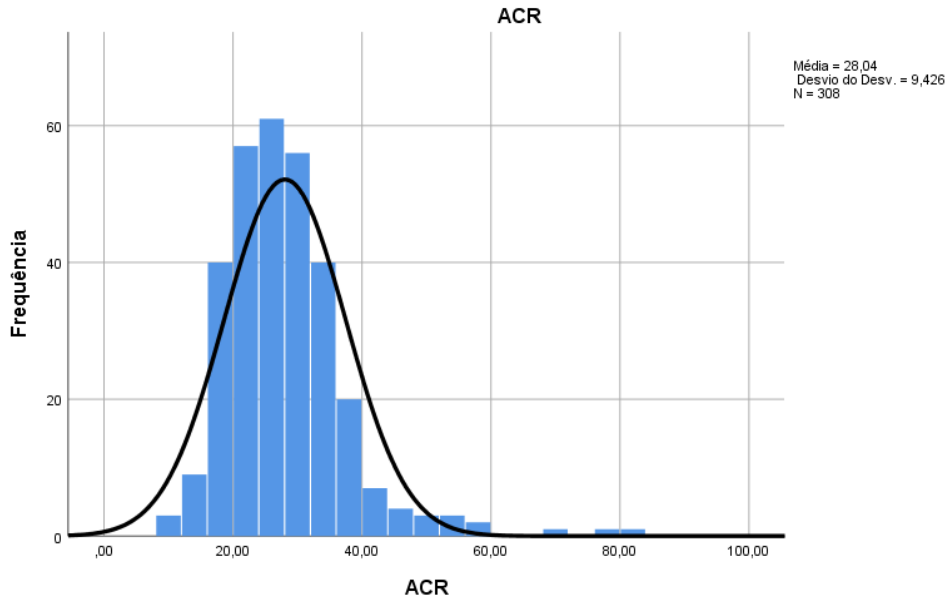
ACR

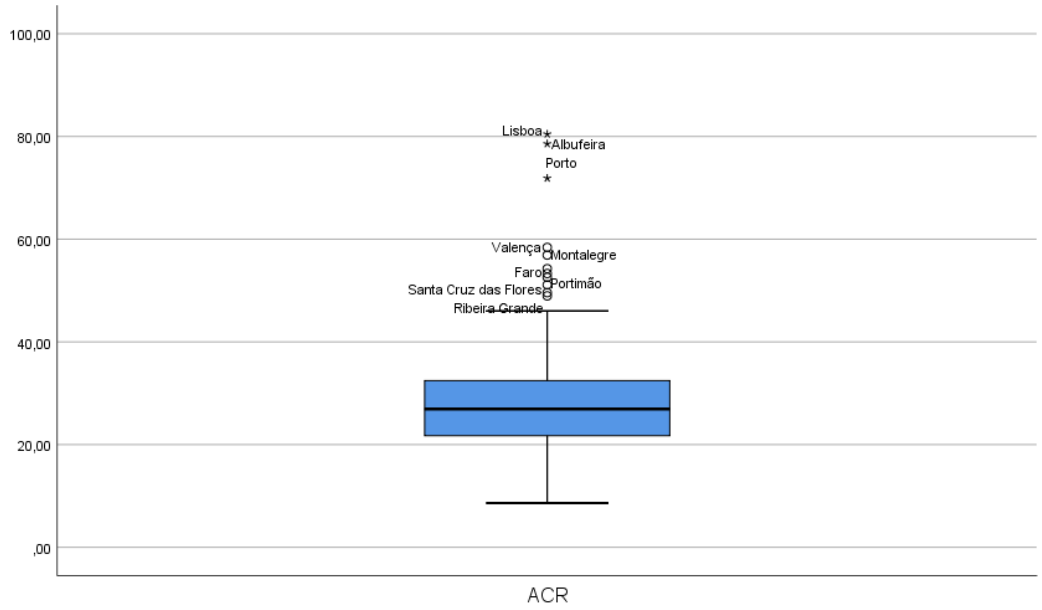
Annual crime rate registered by the police per 1000 inhabitants

This variable has a maximum of 80,46 in Albufeira, a minimum of 8,62 in Alandroal and the mean is 28,04.

The distribution is asymmetric with value 1,793, with a kurtosis value of 6,735.

The outliers are several, the extreme are Lisboa, Albufeira and Porto and the moderate ones are Valença, Montalegre, Faro, Portimão, Santa Cruz das Flores and Ribeira Grande.





RR and ER

Resolution rate (solved processes / new processes)

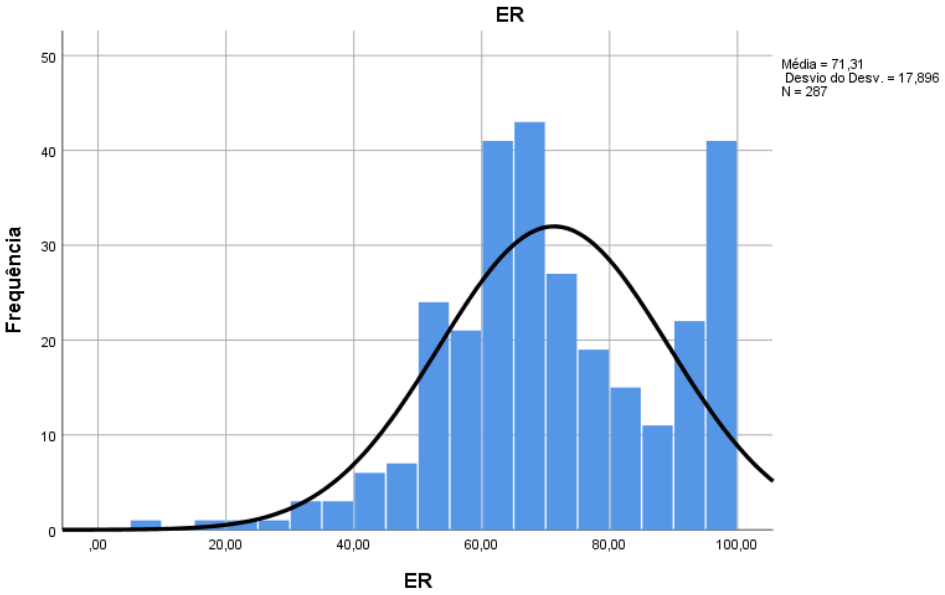
Effectiveness rate (processes resolved / processes waiting decision)

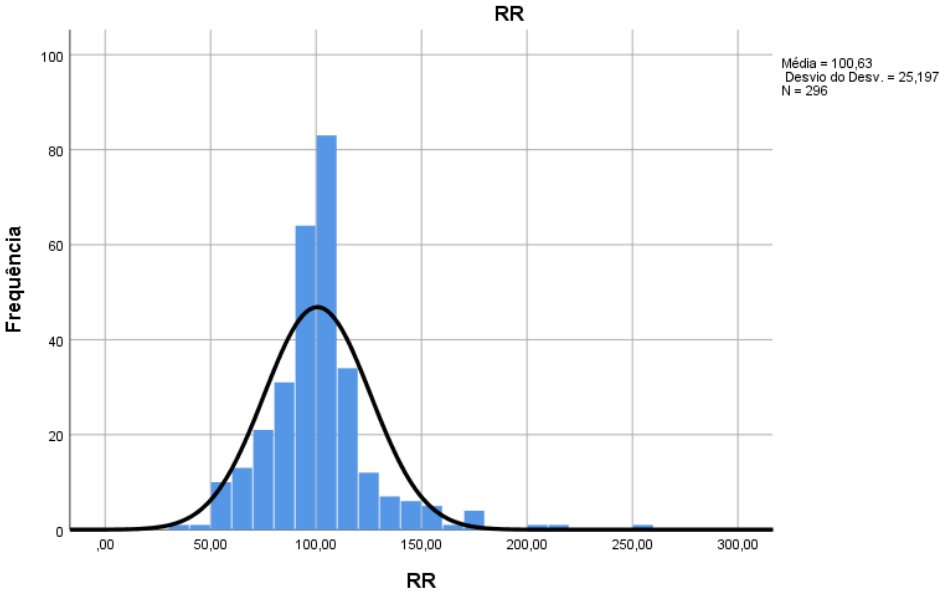
The maximum value for variable RR is 258,3 in Cadaval, the minimum is in Mortágua with 37,5, the mean is 100,63 and this variable has 12 missing values. The asymmetry is 1,534 the kurtosis is 7,161.

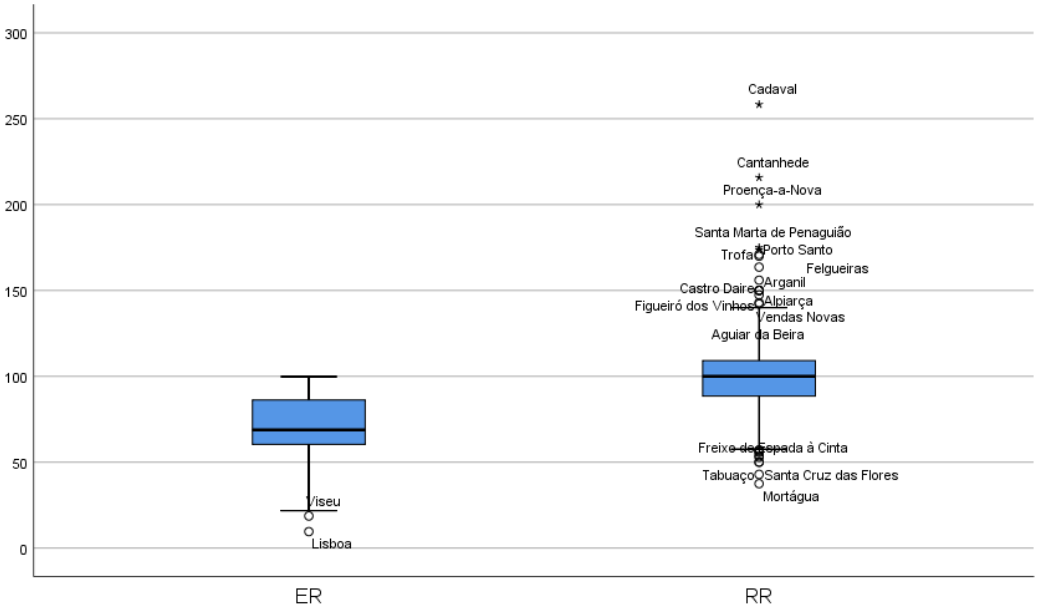
ER has a maximum in 99,8 in Felgueiras, Pinhel and Montemor-o-Novo, the minimum is 9,6 in Lisboa, the mean is 71,309 and has 21 missing values. The kurtosis is -0,118 and the symmetry is -0,181.

RR has several higher and lower outliers, the lower are all moderate in Mortágua, Tabuaço, Santa Cruz das Flores and Freixo de Espada à Cinta. The higher outliers are 13 and the 3 severe are Cadaval, Cantanhede and Proença-a-Nova, the moderate ones are Santa Marta de Penaguião, Porto Santo, Trofa, Felgueiras, Arganil, Castro Daire, Alpiarça, Figueiró dos Vinhos, Vendas Novas and Aguiar da Beira.

The outliers of ER are only two lower moderate outliers in Viseu and Lisboa.


ER
Média = 71,31
Desvio do Desv. = 17,896
N = 287
Frequência
ER


RR
Média = 100,63
Desvio do Desv. = 25,197
N = 296
Frequência
RR


Cadaval
Cantanhede
Proença-a-Nova
Santa Marta de Penaguião
Trofa
Porto Santo
Felgueiras
Arganil
Castro Daire
Alpiarça
Figueiró dos Vinhos
Vendas Novas
Aguiar da Beira
Freixo de Espada à Cinta
Tabuaço
Santa Cruz das Flores
Mortágua
Viseu
Lisboa
ER
RR

P_15_LE and P_15_HE

Rate of resident population over 15 years old with lower and higher education levels

With these two variables, we are trying to understand the impact of education on criminality and earnings.

For the first one the mean is 89,96% of the population that have lower education and for the high education level the mean is 10,04%.

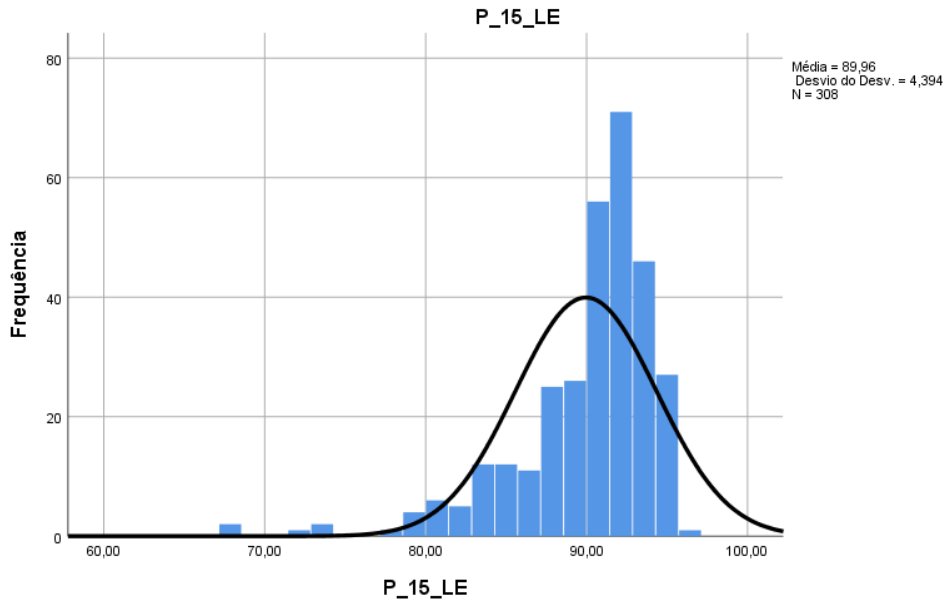
The municipality with the maximum value for P_15_LE is Pampilhosa da Serra with 97,1% and the minimum is in Lisboa with 67,8%.

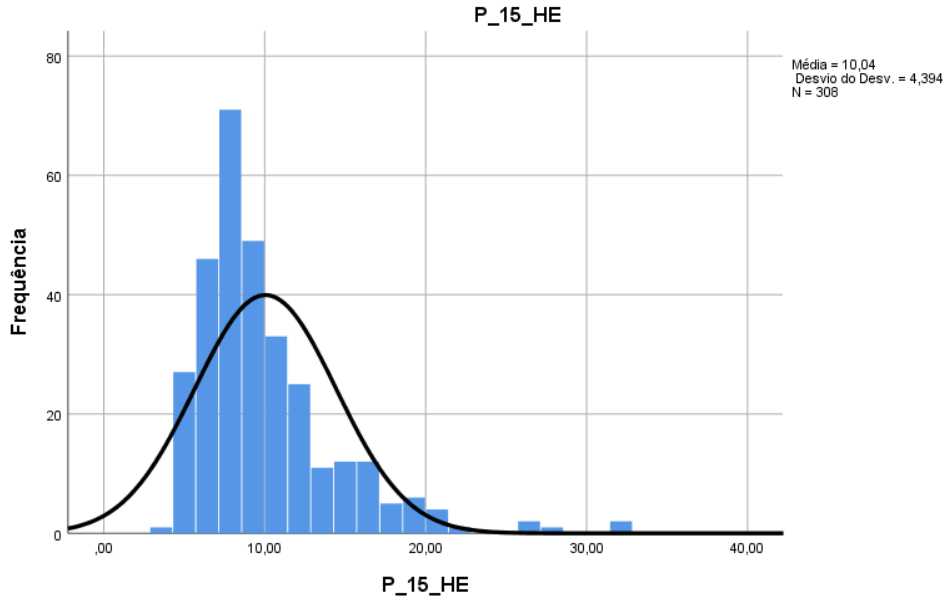
For P_15_HE the maximum is in Lisboa with 32,2% and the minimum is in Pampilhosa da Serra with 2,9%.

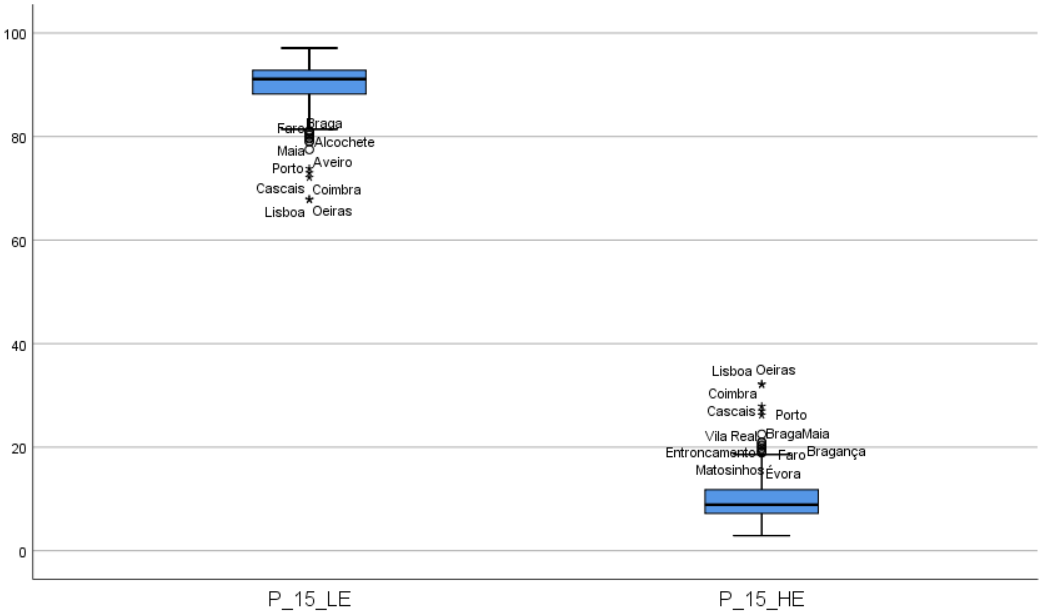
The asymmetry value for P_15_LE is -1,892 and P_15_HE has 1,892, so they are asymmetric in opposing sides. About the weight of the tails, the kurtosis values are the same for both at 5,178.

Regarding the outliers, P_15_LE has lower severe outliers in Lisboa, Coimbra, Cascais, Porto and Aveiro and the moderate outliers are Maia, Alcochete, Faro and Braga, we can see that they are in big urban areas.

P_15_HE has higher severe outliers in Lisboa, Oeiras, Coimbra, Cascais and Porto, the moderate areVila Real, Braga, Maia, Entroncamento, Faro, Bragança, Matosinhos and Évora.







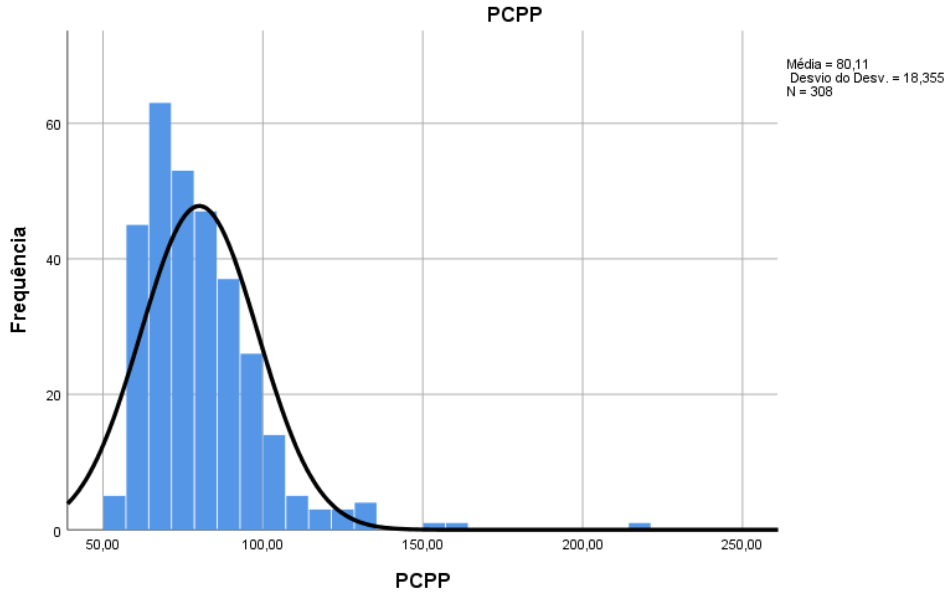
PCPP

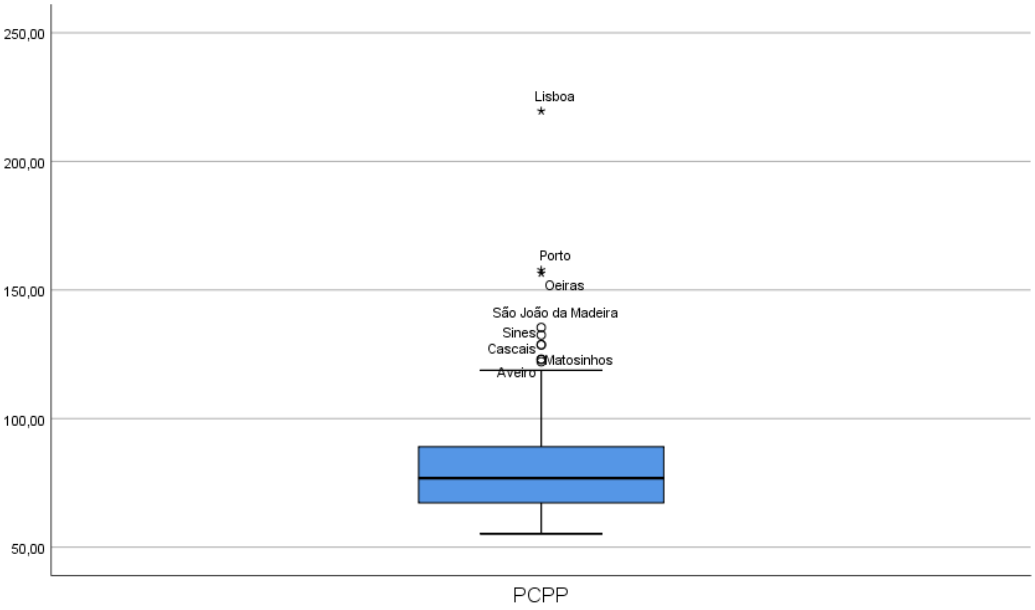
Per capita purchasing power (national average purchasing power equals 100)

The mean is 80,11, the maximum is in Lisboa with 219,6 and the minimum is 55,2 in Ponta do Sol in Azores.

The distribution is asymmetric with a value of 2,382 and the kurtosis is 11,944, with some of the data not concentrated.

The box plot indicates three severe outliers in Lisboa, Porto and Oeiras and the moderate outliers are São João da Madeira, Sines, Aveiro, Cascais and Matosinhos.





Bivariate Analysis

For the bivariate analysis, since all variables are numeric, the table has the correlation and covariance values between all the variables to check the associations between the variables.

Correlações

		AMEE	AMEE_M	AMEE_F	AARU	AARRP_M	AARRP_W	ACR	ER	RR	P_15_LE	P_15_HE	PCPP
AMEE	Correlação de Pearson	1	,975**	,882**	-,127*	,018	-,018	,279**	,029	-,053	-,640**	,640**	,713**
	Sig. (2 extremidades)		,000	,000	,034	,755	,755	,000	,622	,363	,000	,000	,000
	Soma dos quadrados e produtos cruzados	7480215,657	10958017,485	4305217,000	-8057,765	1052,210	-1052,210	126066,336	23671,005	-62265,886	-134760,455	134760,455	626828,163
	Covariância	24365,523	35693,868	14023,508	-29,089	3,427	-3,427	410,640	82,766	-211,071	-438,959	438,959	2041,786
	N	308	308	308	278	308	308	308	287	296	308	308	308
AMEE_M	Correlação de Pearson	,975**	1	,755**	-,133*	,002	-,002	,227**	,013	-,051	-,547**	,547**	,618**
	Sig. (2 extremidades)	,000		,000	,027	,973	,973	,000	,822	,384	,000	,000	,000
	Soma dos quadrados e produtos cruzados	10958017,485	16888117,245	5540340,837	-12675,951	168,494	-168,494	153810,665	16283,124	-89636,317	-173192,375	173192,375	816280,727
	Covariância	35693,868	55010,154	18046,713	-45,762	,549	-,549	501,012	56,934	-303,852	-564,145	564,145	2658,895
	N	308	308	308	278	308	308	308	287	296	308	308	308
AMEE_F	Correlação de Pearson	,882**	,755**	1	-,093	,011	-,011	,344**	,058	-,050	-,735**	,735**	,796**
	Sig. (2 extremidades)	,000	,000		,123	,848	,848	,000	,331	,392	,000	,000	,000
	Soma dos quadrados e produtos cruzados	4305217,000	5540340,837	3187520,179	-3860,998	422,422	-422,422	101556,856	30375,067	-38071,050	-101094,574	101094,574	456957,931
	Covariância	14023,508	18046,713	10382,802	-13,939	1,376	-1,376	330,804	106,207	-129,054	-329,298	329,298	1488,462
	N	308	308	308	278	308	308	308	287	296	308	308	308
AARU	Correlação de Pearson	-,127*	-,133*	-,093	1	,026	-,026	,002	-,038	-,034	,111	-,111	-,092
	Sig. (2 extremidades)	,034	,027	,123		,668	,668	,969	,545	,575	,066	,066	,127
	Soma dos quadrados e produtos cruzados	-8057,765	-12675,951	-3860,998	612,468	10,096	-10,096	9,117	-250,929	-336,455	205,603	-205,603	-705,895
	Covariância	-29,089	-45,762	-13,939	2,211	,036	-,036	,033	-,958	-1,255	,742	-,742	-2,548

	N	278	278	278	278	278	278	278	263	269	278	278	278
AARRP_M	Correlação de Pearson	,018	,002	,011	,026	1	-1,000**	-,011	-,051	-,005	,142*	-,142*	-,015
	Sig. (2 extremidades)	,755	,973	,848	,668		,000	,853	,386	,926	,012	,012	,791
	Soma dos quadrados e produtos cruzados	1052,210	168,494	422,422	10,096	463,010	-463,010	-37,676	-263,558	-42,617	235,841	-235,841	-105,059
	Covariância	3,427	,549	1,376	,036	1,508	-1,508	-,123	-,922	-,144	,768	-,768	-,342
	N	308	308	308	278	308	308	308	287	296	308	308	308
AARRP_W	Correlação de Pearson	-,018	-,002	-,011	-,026	-1,000**	1	,011	,051	,005	-,142*	,142*	,015
	Sig. (2 extremidades)	,755	,973	,848	,668	,000		,853	,386	,926	,012	,012	,791
	Soma dos quadrados e produtos cruzados	-1052,210	-168,494	-422,422	-10,096	-463,010	463,010	37,676	263,558	42,617	-235,841	235,841	105,059
	Covariância	-3,427	-,549	-1,376	-,036	-1,508	1,508	,123	,922	,144	-,768	,768	,342
	N	308	308	308	278	308	308	308	287	296	308	308	308
ACR	Correlação de Pearson	,279**	,227**	,344**	,002	-,011	,011	1	,064	-,033	-,328**	,328**	,462**
	Sig. (2 extremidades)	,000	,000	,000	,969	,853	,853		,283	,573	,000	,000	,000
	Soma dos quadrados e produtos cruzados	126066,336	153810,665	101556,856	9,117	-37,676	37,676	27275,862	3094,038	-2319,035	-4172,586	4172,586	24530,988
	Covariância	410,640	501,012	330,804	,033	-,123	,123	88,846	10,818	-7,861	-13,591	13,591	79,905
	N	308	308	308	278	308	308	308	287	296	308	308	308
ER	Correlação de Pearson	,029	,013	,058	-,038	-,051	,051	,064	1	,400**	-,142*	,142*	,107
	Sig. (2 extremidades)	,622	,822	,331	,545	,386	,386	,283		,000	,016	,016	,071
	Soma dos quadrados e produtos cruzados	23671,005	16283,124	30375,067	-250,929	-263,558	263,558	3094,038	91591,805	51426,767	-3235,099	3235,099	10212,547

	Covariância	82,766	56,934	106,207	-,958	-,922	,922	10,818	320,251	179,814	-11,312	11,312	35,708
	N	287	287	287	263	287	287	287	287	287	287	287	287
RR	Correlação de Pearson	-,053	-,051	-,050	-,034	-,005	,005	-,033	,400**	1	-,012	,012	-,005
	Sig. (2 extremidades)	,363	,384	,392	,575	,926	,926	,573	,000		,836	,836	,928
	Soma dos quadrados e produtos cruzados	-62265,886	-89636,317	-38071,050	-336,455	-42,617	42,617	-2319,035	51426,767	187284,918	-399,269	399,269	-727,651
	Covariância	-211,071	-303,852	-129,054	-1,255	-,144	,144	-7,861	179,814	634,864	-1,353	1,353	-2,467
	N	296	296	296	269	296	296	296	287	296	296	296	296
P_15_LE	Correlação de Pearson	-,640**	-,547**	-,735**	,111	,142*	-,142*	-,328**	-,142*	-,012	1	-1,000**	-,856**
	Sig. (2 extremidades)	,000	,000	,000	,066	,012	,012	,000	,016	,836		,000	,000
	Soma dos quadrados e produtos cruzados	-134760,455	-173192,375	-101094,574	205,603	235,841	-235,841	-4172,586	-3235,099	-399,269	5927,746	-5927,746	-21193,076
	Covariância	-438,959	-564,145	-329,298	,742	,768	-,768	-13,591	-11,312	-1,353	19,309	-19,309	-69,033
	N	308	308	308	278	308	308	308	287	296	308	308	308
P_15_HE	Correlação de Pearson	,640**	,547**	,735**	-,111	-,142*	,142*	,328**	,142*	,012	-1,000**	1	,856**
	Sig. (2 extremidades)	,000	,000	,000	,066	,012	,012	,000	,016	,836	,000		,000
	Soma dos quadrados e produtos cruzados	134760,455	173192,375	101094,574	-205,603	-235,841	235,841	4172,586	3235,099	399,269	-5927,746	5927,746	21193,076
	Covariância	438,959	564,145	329,298	-,742	-,768	,768	13,591	11,312	1,353	-19,309	19,309	69,033
	N	308	308	308	278	308	308	308	287	296	308	308	308
PCPP	Correlação de Pearson	,713**	,618**	,796**	-,092	-,015	,015	,462**	,107	-,005	-,856**	,856**	1
	Sig. (2 extremidades)	,000	,000	,000	,127	,791	,791	,000	,071	,928	,000	,000	
	Soma dos quadrados e produtos cruzados	626828,163	816280,727	456957,931	-705,895	-105,059	105,059	24530,988	10212,547	-727,651	-21193,076	21193,076	103434,209

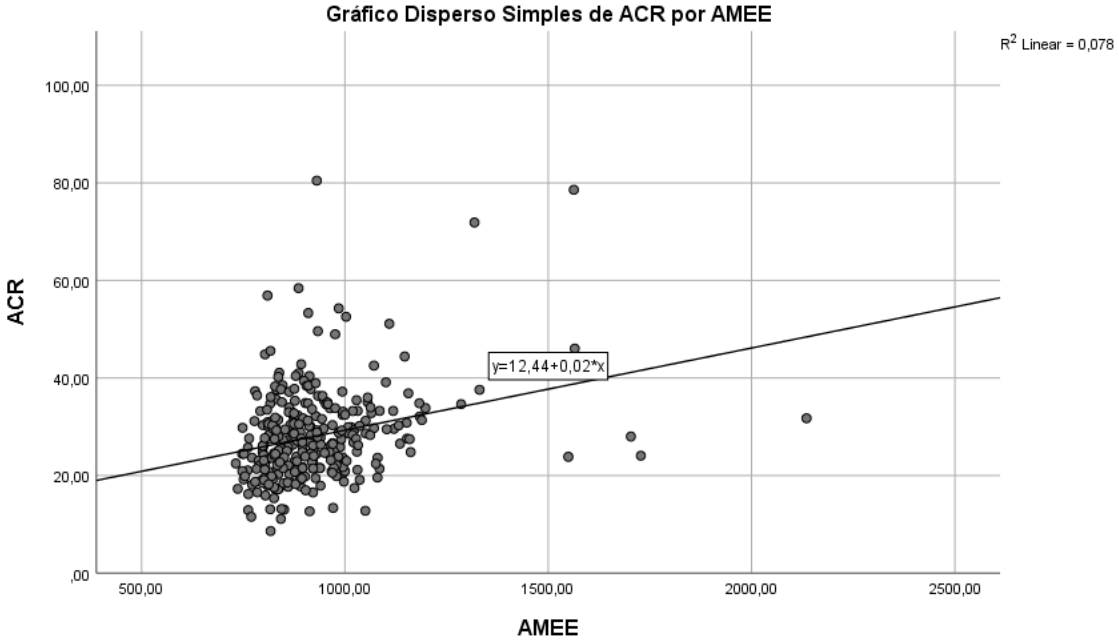
Covariância	2041,786	2658,895	1488,462	-2,548	-,342	,342	79,905	35,708	-2,467	-69,033	69,033	336,919
N	308	308	308	278	308	308	308	287	296	308	308	308

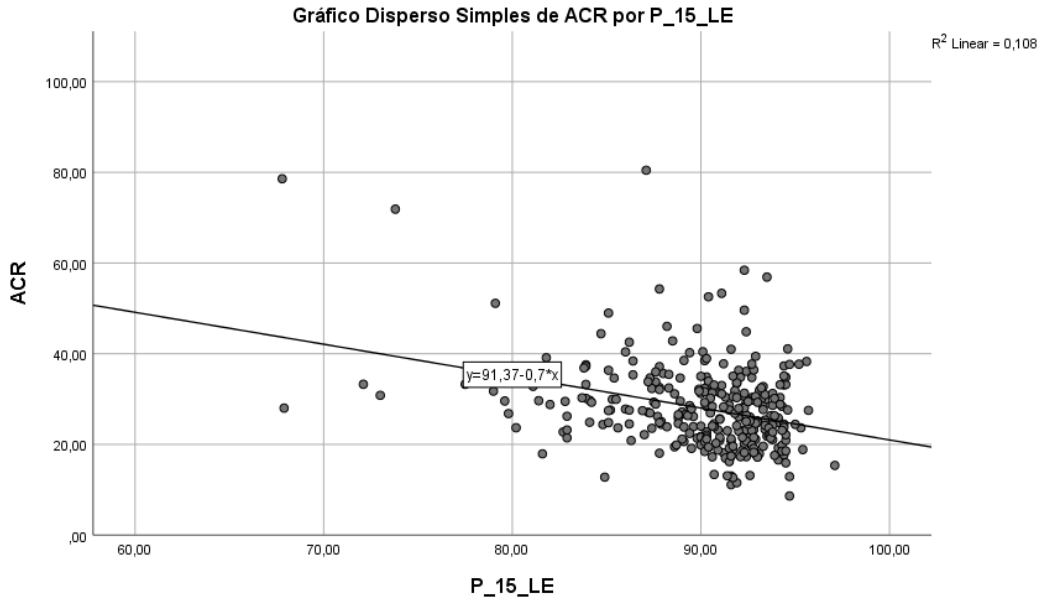
ACR has a positive but low correlation (0,279) with variable AMEE, which means that criminal rate tends to grow (even if it is slow) in municipalities with high average monthly earnings. This can be caused due to the increase in white collar crimes, which tend to happen more in municipalities with well-paid jobs, but it can also be related to the fact that in these municipalities there is more crimes of theft and assault. People with higher monthly earnings are more likely to file a complaint with the authorities and follow up with the cases in court, which can be quite expensive.

When analyzing the correlation between ACR and AMEE_M or AMEE_F we can see that there is a stronger correlation between ACR and AMEE_F than between ACR and AMEE_M. It is not clear why this is happening, but we can clearly conclude that municipalities where women have higher monthly earnings tend to have a higher crime rate.

The correlation between ACR and P_15_LE is negative (-0,199) while the correlation between ACR and with P_15_HE is positive (0,348). These figures lead us to conclude that the higher the level of education of the population of a municipality, the higher the crime rate, and this can happen because the higher levels of education are in the richest municipalities that are the ones with more crimes registered by the police

We can see that ACR has a high value for the covariance for the earnings variables, a positive covariance value for the variable representing the lower education and a negative value for the one representing the higher education level.





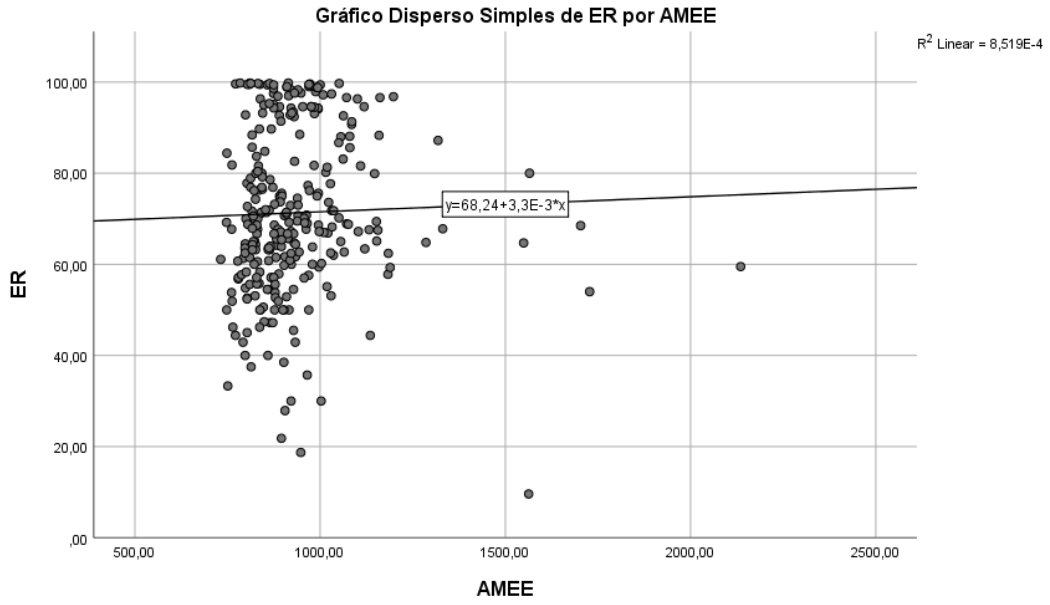
ER has a positive correlation with the earning variable (AMEE), but that correlation is very low.

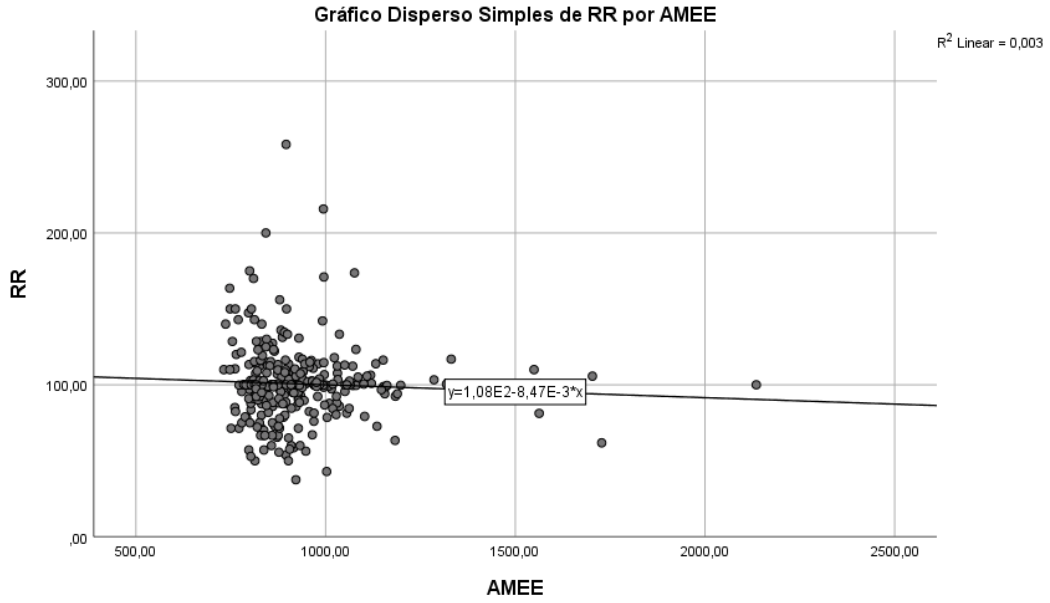
The correlation between RR and the earning variables is also very low but with a negative value so we cannot take any conclusions regarding these variables.

ER has a positive correlation with PCPP and that can come from the fact that the municipalities with higher PCPP have more resources to deal with the processes.

ER has a positive covariance with the earning variables and with P_15_HE and a negative value for P_15_LE. And RR has negative covariance values with the earnings and higher education and a positive value for the lower education variable.

This is probably because the RR rate depends on the new and solved processes and in the richest municipalities despite having more new processes they have better resources to deal with them. And because ER depends on the processes resolved and the ones waiting for decision, we can say that in the richest municipalities they have more complex processes that take more time to be resolved.





P_15_LE has a negative correlation with the three earnings variables. However, the higher negative correlation is with AMEE_F (-0,735). The municipalities with lower education levels tend to have employees with lower average monthly earnings, we also concluded that municipalities with lower education tend to have women with lower earnings than man.

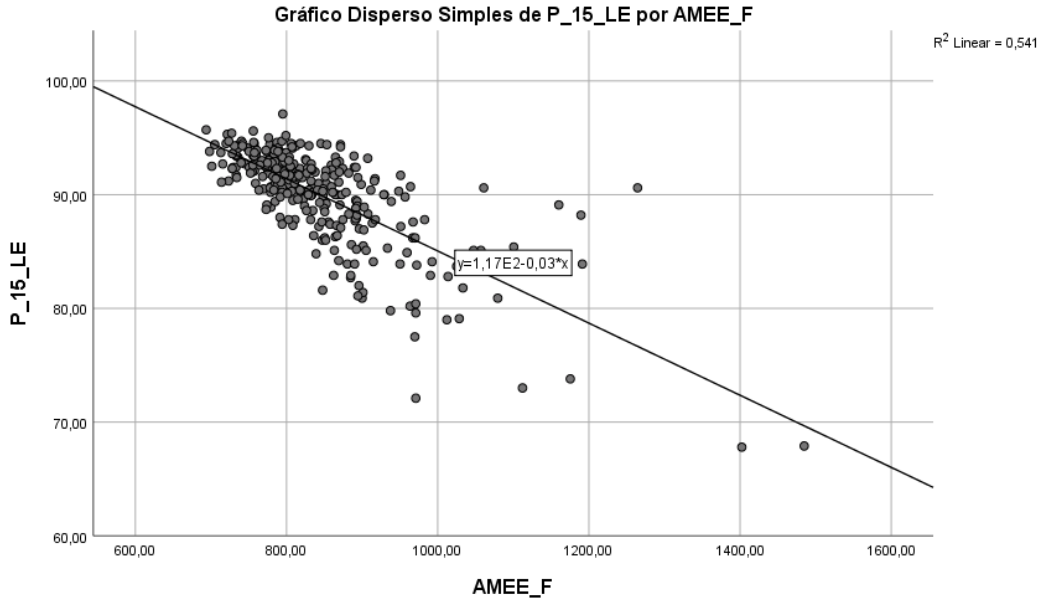
The correlation with ACR is negative maybe because residents of municipalities with higher values of lower education are from less urban municipalities that have less crime and have people with fewer resources to deal with the processes.

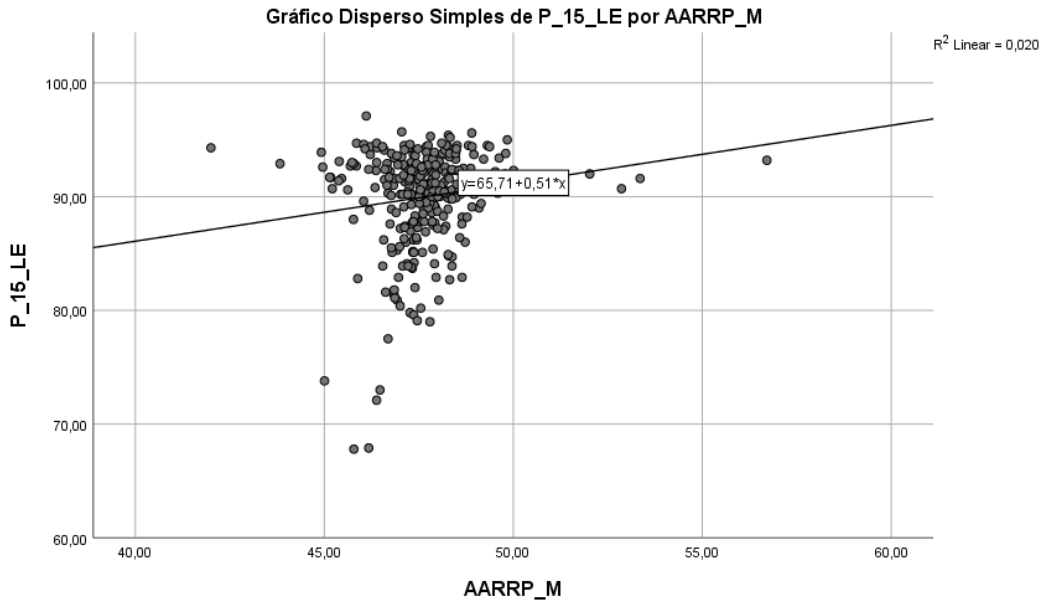
The variable that has the strongest correlation with P_15_LE is PCPP. This is a clear indicator that the lower the education levels of a municipality, the lower the per capita purchasing power.

When analyzing the correlation between P_15_LE and AARRP_M or AARRP_W we obtained symmetric correlation values, positive for AARRP_M (0,142) and negative for AARRP_W (-0,142). We can conclude that municipalities with more man than woman tend to have higher

levels of low education and the opposite can be concluded for municipalities with more resident woman.

The covariance between P_15_LE is strong and negative with the earning variables and ACR and ER, this might be due to the fact that the poorer municipalities are the ones with lower levels of education, less populated and with lower crimes registered by the police.

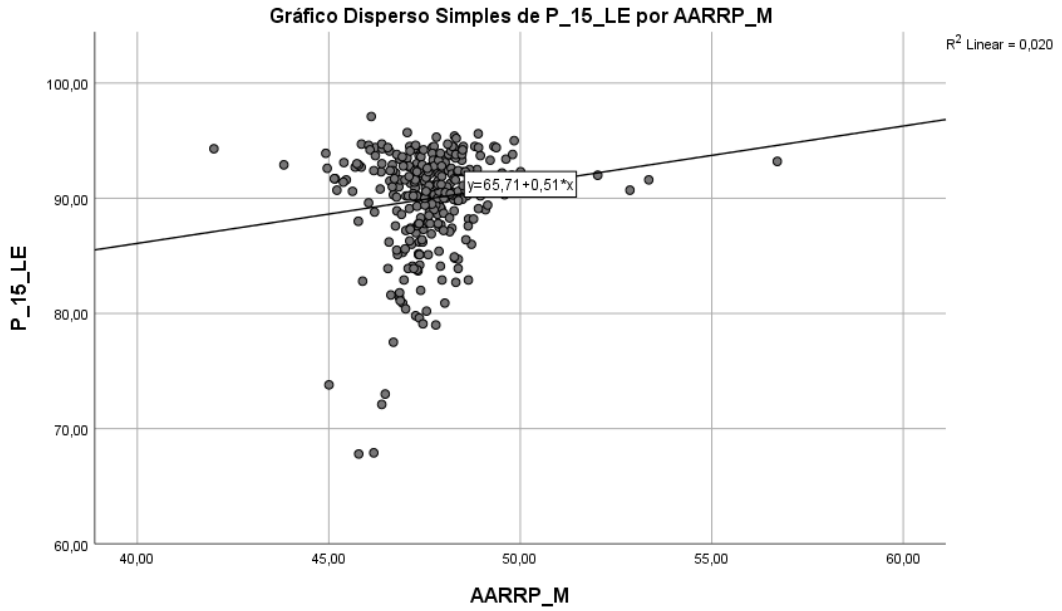




P_15_HE has positive correlations with the earnings variables. Like the variable P_15_LE, the variable P_15_HE has a strong correlation with the variable AMEE and with a higher value for females. Therefore, we can conclude through another variable that municipalities with higher levels of higher education tend to have women with higher incomes.

The correlation between P_15_HE and AARU is a negative correlation but with a low value, but we can conclude that the higher the levels of higher education the lower the unemployment levels.

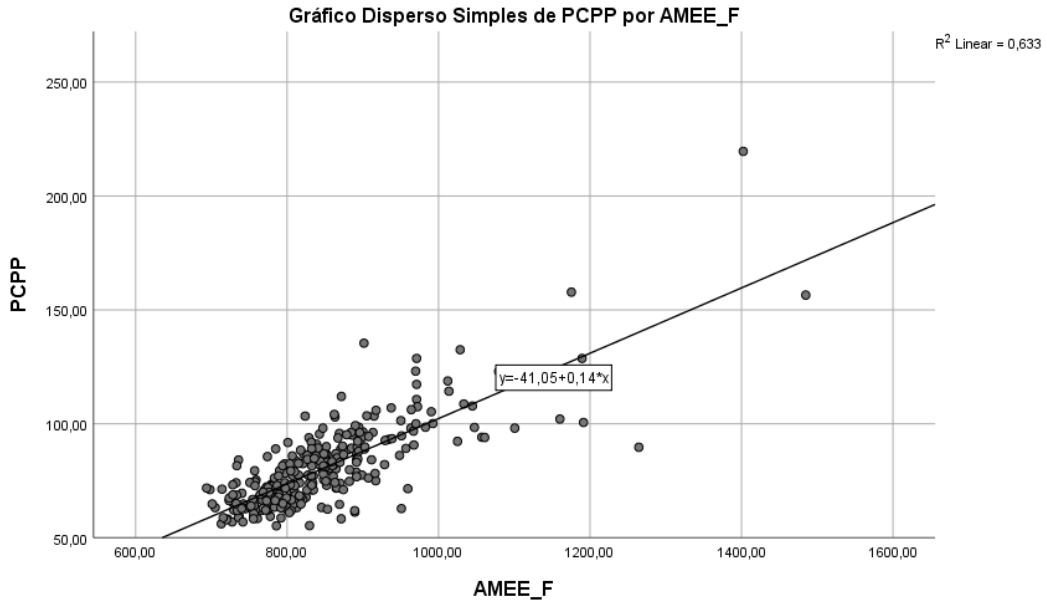
This variable has good correlations with the same variables that P_15_LE does but in contrary signs and this was expected by the contrary reasons explained before for the other variable.

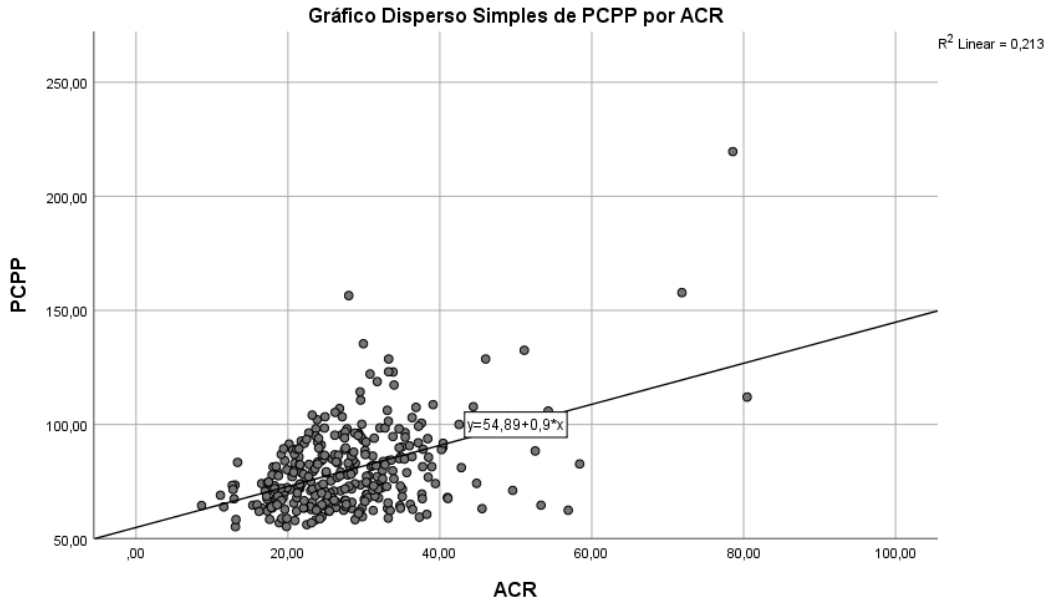


As expected, the PCPP variable has a strong positive correlation with the earnings variables but a higher value for the females variable. We can conclude that municipalities with higher rates of purchasing power also have a higher rate of women with higher monthly income when compared to men.

We can also verify that PCPP has a high negative correlation with P_15_LE and a high positive correlation with P_15_HE. This means that a higher rate of residents with higher education leads to an higher purchasing power.

PCPP has also an interesting positive correlation with ACR (0,462). This means that in municipalities with more purchasing power we also find an increase in the crime rate registered by the police.
The covariance is good with the earning variables, as expected, ACR and ER.





Multivariate Analysis

To start the multivariate analysis a principal component analysis was performed on the dataset and after that, we did a cluster analysis.

Like was seen on the univariate analysis, three variables have missing values, AARU has 30 missing values, ER and RR have 21 and 12 missing values, respectively.

Because the municipalities with missing values have at most missing values in two variables, we decided to keep them and compute a value to fill in, that is the mean value of the variable without the outliers identified in the univariate analysis.

Principal Component Analysis

Taking this dataset without missing values for the 308 municipalities and 12 variables, a normed PCA was performed on SPAD to analyze the multivariate structure of the data to obtain a smaller number of variables to describe the municipalities.

To have a homogeneous group of variables, regarding what was the objective on describing each municipality, it was decided to set the variable AMEE as supplementary because the interest is in describing the municipalities regarding the monthly earnings of employees by gender and for that the dataset has the variables AMEE_M and AMEE_F. The other 11 variables were set as active, like all the 308 municipalities that have the same weight.

Applying the PCA, we obtained the eigenvalues, the percentage of inertia and the cumulated percentage of inertia that are shown below.

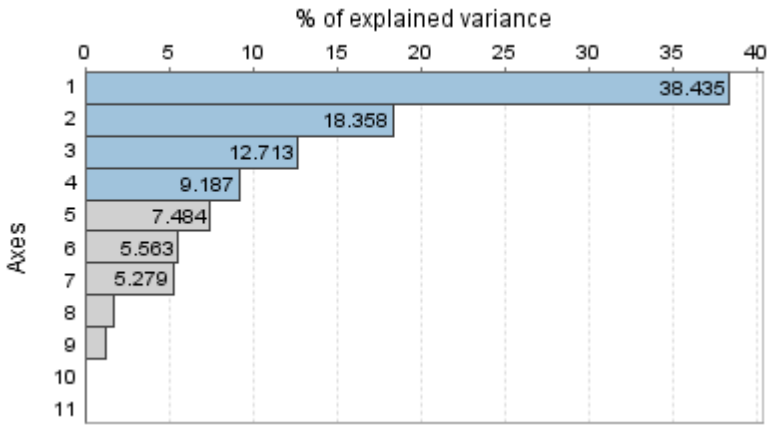
Axis	Variance of the axis (eigenvalue)	% of explained variance	Cumulated % of explained variance
1	4,228	38,4	38,4
2	2,019	18,4	56,8
3	1,398	12,7	69,5
4	1,011	9,2	78,7
5	0,823	7,5	86,2
6	0,612	5,6	91,7
7	0,581	5,3	97,0
8	0,192	1,7	98,8
9	0,136	1,2	100,0
10	0,000	0,0	100,0
11	0,000	0,0	100,0
Total	11,000	100,0	100,0

To decide how many principal components should be retained, these values were analyzed by different criterions.

By the Kaiser's criterion, the eigenvalues above one have the principal components that are more informative than the original variables and in this case, they are the first four.

Using the Pearson's criterion, can be seen on the table that the first five principal components explain more than 80% of the total inertia but the first four explain almost 80%.

And for the Cattel's criterion, can be seen on the graphical representation below, that the eigenvalues that have a larger difference between them are the first four.



With the analysis of these three criterions, it was decided to do the principal component analysis for the first four principal components.

To make the interpretation the individuals and variables relative and absolute contributions, the graphical representations of them in the different plans using the four axis chosen before, were analyzed.

For the variables, the data below was obtained and for the municipalities only has the values for the ones analyzed for each axis because the table with all of them was too big.

Active variables coordinates

Label of the variable	Axis 1	Axis 2	Axis 3	Axis 4
AMEE_M	-0,740	0,137	0,117	-0,144
AMEE_F	-0,883	0,148	0,078	-0,021
AARU	0,142	0,005	0,085	0,927
AARRP_M	0,126	0,986	-0,087	-0,002
AARRP_W	-0,126	-0,986	0,087	0,002
ACR	-0,472	0,069	0,022	0,346
ER	-0,141	-0,095	-0,814	0,076
RR	0,001	-0,073	-0,831	0,006
P_15_LE	0,932	0,052	0,034	-0,011
P_15_HE	-0,932	-0,052	-0,034	0,011
PCPP	-0,931	0,103	0,000	0,066

Contributions of the active variables to the axes (%)

Label of the variable	Axis 1	Axis 2	Axis 3	Axis 4
AMEE_M	13,0	0,9	1,0	2,0
AMEE_F	18,5	1,1	0,4	0,0
AARU	0,5	0,0	0,5	85,1
AARRP_M	0,4	48,1	0,5	0,0
AARRP_W	0,4	48,1	0,5	0,0
ACR	5,3	0,2	0,0	11,8
ER	0,5	0,4	47,4	0,6
RR	0,0	0,3	49,4	0,0
P_15_LE	20,6	0,1	0,1	0,0
P_15_HE	20,6	0,1	0,1	0,0
PCPP	20,5	0,5	0,0	0,4

Cosines of angles between main axes and initial axes (normed eigenvectors)

Label of the variable	Axis 1	Axis 2	Axis 3	Axis 4
AMEE_M	-0,360	0,096	0,099	-0,143
AMEE_F	-0,430	0,104	0,066	-0,021
AARU	0,069	0,003	0,071	0,922
AARRP_M	0,061	0,694	-0,074	-0,002
AARRP_W	-0,061	-0,694	0,074	0,002
ACR	-0,230	0,049	0,018	0,344
ER	-0,068	-0,067	-0,688	0,075

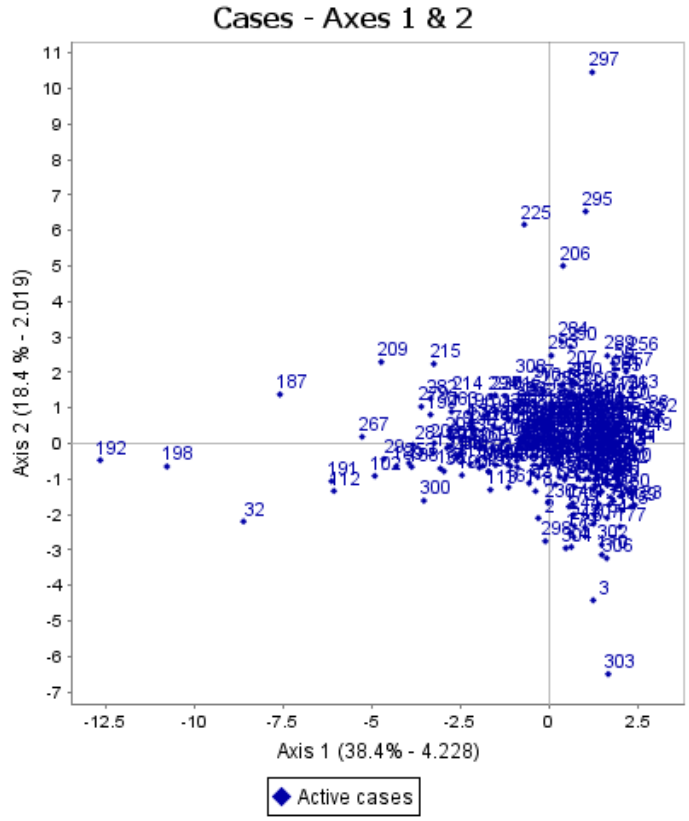
RR	0,001	-0,052	-0,703	0,006
P_15_LE	0,453	0,037	0,028	-0,011
P_15_HE	-0,453	-0,037	-0,028	0,011
PCPP	-0,453	0,072	0,000	0,066

For the first principal component, the municipalities that contribute more to the first axis were chosen.

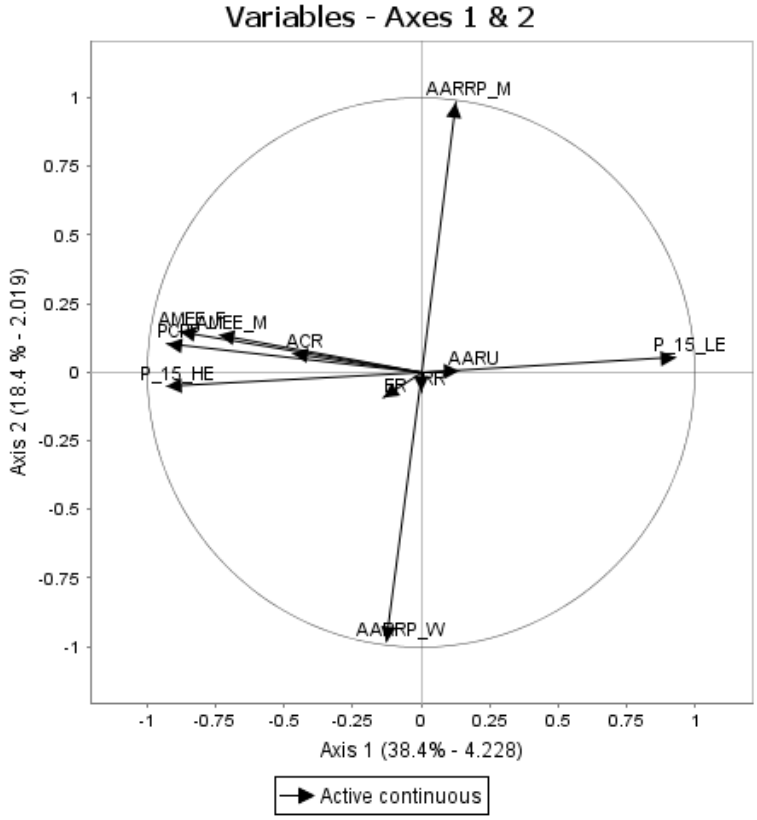
The municipalities with lower coordinates on this axis are Lisboa (192), Oeiras (198), Porto (32), Alcochete (187), Cascais (191) and Coimbra (112). On the positive side of the axis, the municipalities with higher coordinates are Cinfães (52), Baião (49) and Vinhais (86) but their coordinates are not very high when compared with the municipalities with coordinates on the negative side. All of these municipalities have a relative contribution higher than 0,5, so they are well represented on the axis, except Alcochete that has a relative contribution of 47,6% and a relative contribution on the plan of the first two axis of 49,2%.

The absolute contributions of all the municipalities are higher than the global mean absolute contribution that is 0,3, the value of all of them together is not close to 80% because this contribution is very disperse on all the municipalities.

Municipality	Relative contribution	Absolute contribution
Lisboa (192)	0,818	12,3
Oeiras (198)	0,915	8,9
Porto (32)	0,765	5,7
Alcochete (187)	0,476	4,4
Cascais (191)	0,839	2,9
Coimbra (112)	0,798	2,8
Cinfães (52)	0,507	0,7
Baião (49)	0,501	0,6
Vinhais (86)	0,788	0,6



To choose the variables more correlated with the first principal component the graphic below was analyzed and the absolute contributions of the variables. Can be seen below that the variables that are more correlated with the first principal component are P_15_LE that is positively correlated with a value of 0, 932 and P_15_HE with – 0,932, PCPP with – 0,931, AMEE_F with 0,883 and AMEE_M with – 0,740, that are negatively correlated.



The first principal component opposes the variables P_15_LE to variables P_15_HE, PCPP, AMEE_F and AMEE_M.

The first principal component opposes municipalities with a large amount of the population with lower education to municipalities with a large amount of the population with higher earnings and superior education.

The municipalities with higher coordinates on the second principal component on the positive side are Corvo (297), Lages das Flores (295), Azambuja (225) and Grândula (206). On the negative side are Porto Moniz (303), Melgaço (3), Santana (306) and Vila de Rei (170). All of them are well represented on the axis, analyzing the relative contributions below, and have an absolute contribution higher than the average absolute contribution of 0,3.

Municipality	Relative contribution	Absolute contribution
Corvo (297)	0,975	17,6
Lages das Flores (295)	0,926	6,9
Azambuja (225)	0,899	6,1
Grândula (206)	0,863	4,0
Porto Moniz (303)	0,878	6,8
Melgaço (3)	0,783	3,1
Santana (306)	0,681	1,7
Vula de Rei (170)	0,659	1,6

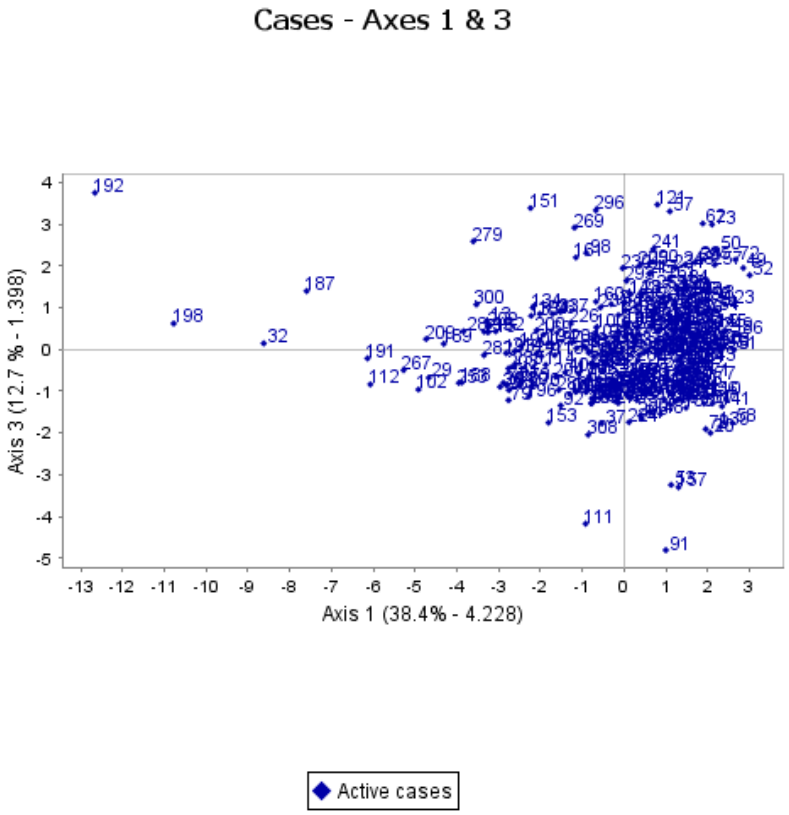
The variable more positively correlated with the second axis is AARRP_M, with a correlation value of 0,986, and negatively is AARRP_W with – 0,986 of correlation.

The second principal component opposes municipalities with high values for these two variables, so it opposes municipalities more populated by man to municipalities more populated by women.

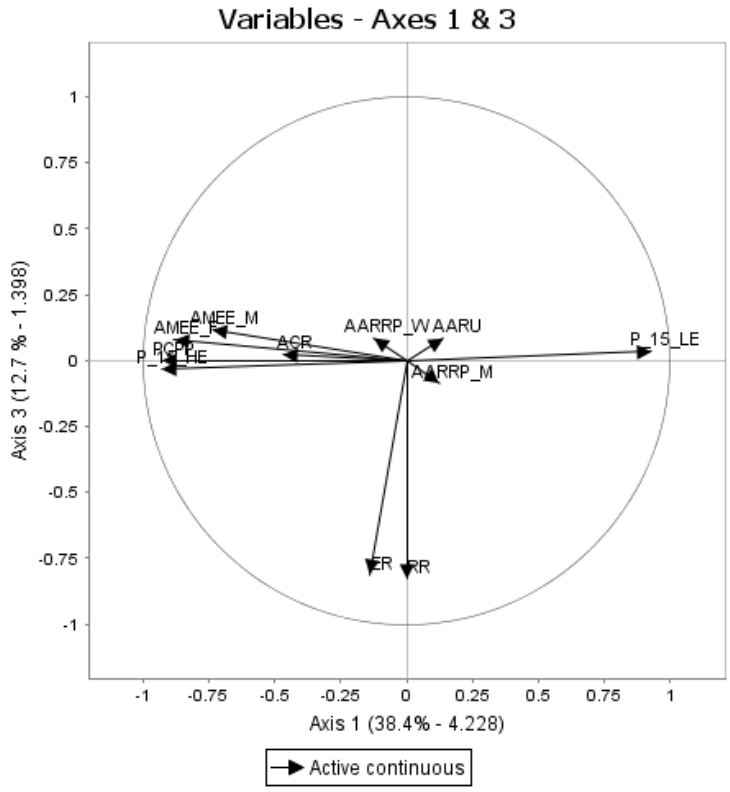
The municipalities with higher coordinates on the axis 3 are Lisboa (192), Viseu (151), Santa Cruz das Flores (296), Mortágua (121) and Penafiel (57). With lower coordinates are Cadaval (91), Cantanhede (111), Proença-a-Nova (157) and Felguerias (53).

As can be seen below, all the municipalities are well represented on the axis, except Lisboa that is well represented on the plan of axis 1 and 3 with a CTR=0,818+0,072=0,89, and Proença-a-Nova with a CTR of the same plan of 0,575.

Municipality	Relative contribution	Absolute contribution
Lisboa (192)	0,072	3,3
Viseu (151)	0,524	2,7
Santa Cruz das Flores (296)	0,574	2,6
Mortágua (121)	0,700	2,8
Penafiel (57)	0,771	2,5
Cadaval (91)	0,532	5,4
Cantanhede (111)	0,696	4,0
Proença-a-Nova (157)	0,497	2,5
Felguerias (53)	0,765	2,4



Regarding the correlation of the variables with the axis, can be seen that ER and RR are negatively correlated, with values of - 0,814 and – 0,831, and none variable have a good positive correlation with the axis.



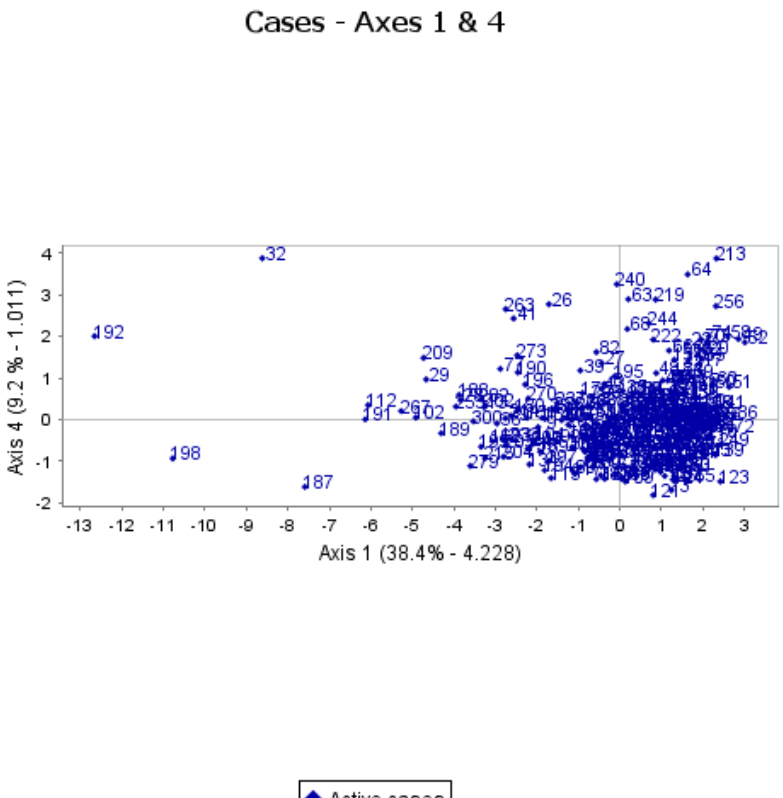
This axis opposes municipalities with high values on resolution rate (RR) and effectiveness rate (ER) to municipalities with low values on those two variables.

Finally, for the fourth axis has Porto (32), Barrancos (213), Mesão Frio (64) and Elvas (240) with the highest coordinates. With the lowest are Mortágua (121), Melgaço (3) and Alcochete (187).

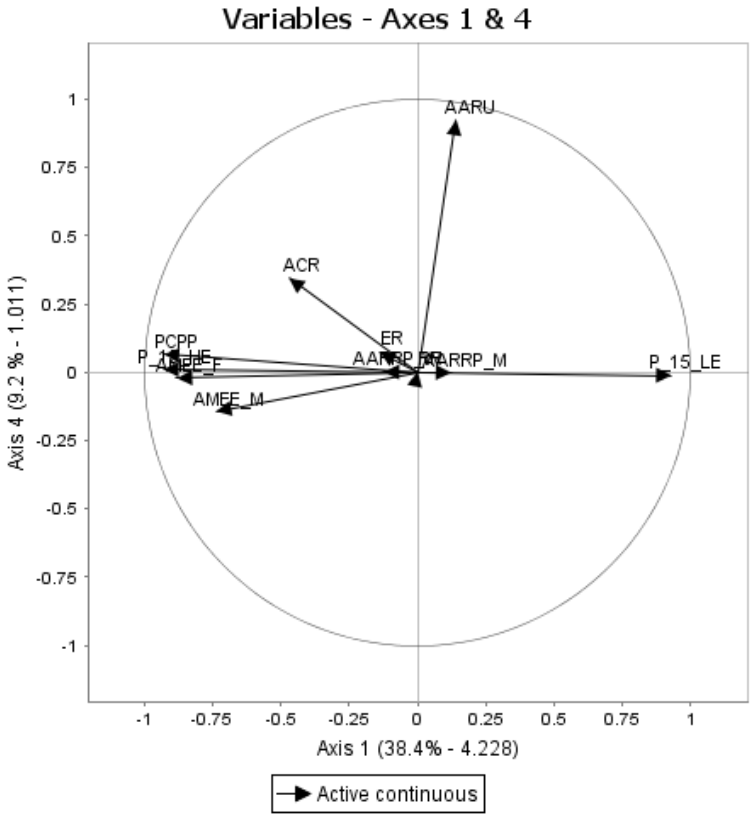
The municipalities of Barrancos, Mesão Frio and Elvas are well represented on the axis but the others are not because their CTR is lower than 0,5.

Porto on the plan of axis 1 and 4 has a CTR=0,155+0,765=0,92 and Alcochete has a CTR=0,476+0,022=0,498. For Mortágua and Melgaço their representation is still very week on the plan.

Municipality	Relative contribution	Absolute contribution
Porto (32)	0,155	0,9
Barrancos (213)	0,593	4,8
Mesão Frio (64)	0,530	3,9
Elvas (240)	0,905	1,1
Mortágua (121)	0,193	0,8
Melgaço (3)	0,116	4,8
Alcochete (187)	0,022	3,4



The variables best correlated with this principal component are unemployment rate (AARU), with 0,927, and the annual crime rate registered by the police (ACR) can be considered too with 0,346 of correlation.



This axis opposes the municipalities with high values of unemployment and annual crime rate registered by the police, to the municipalities with low values on these variables.

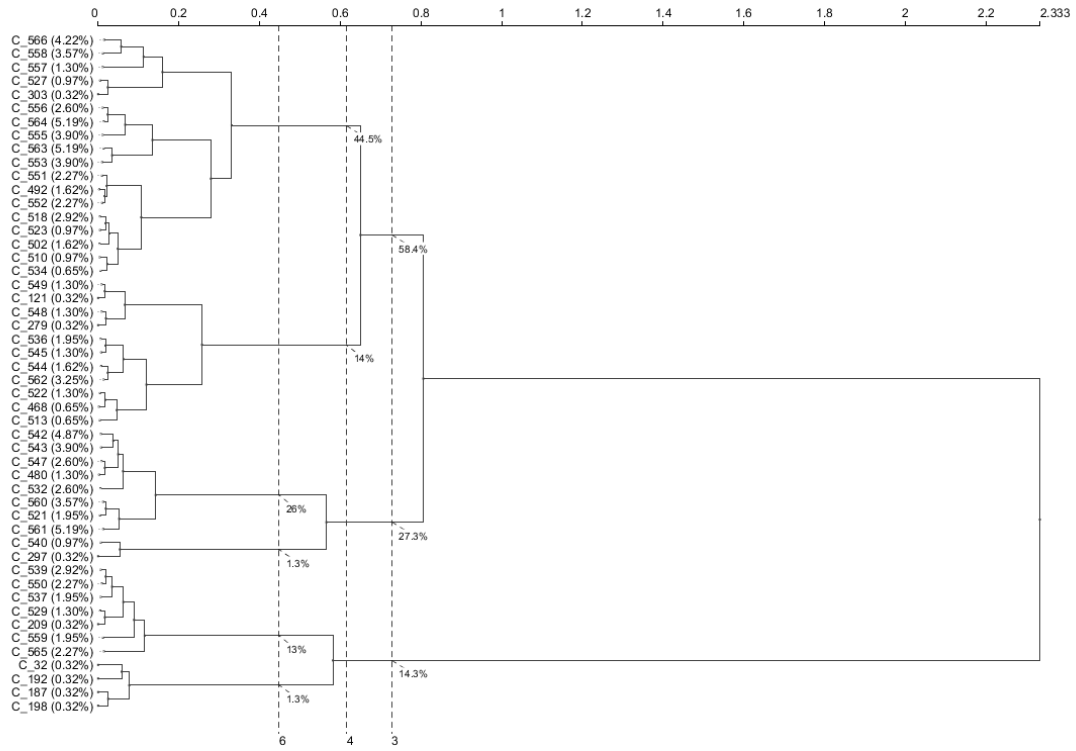
Cluster Analysis

To structure the municipalities in classes to see the elements that are similar with each other and to see the ones distinct from each other we performed a cluster analysis.

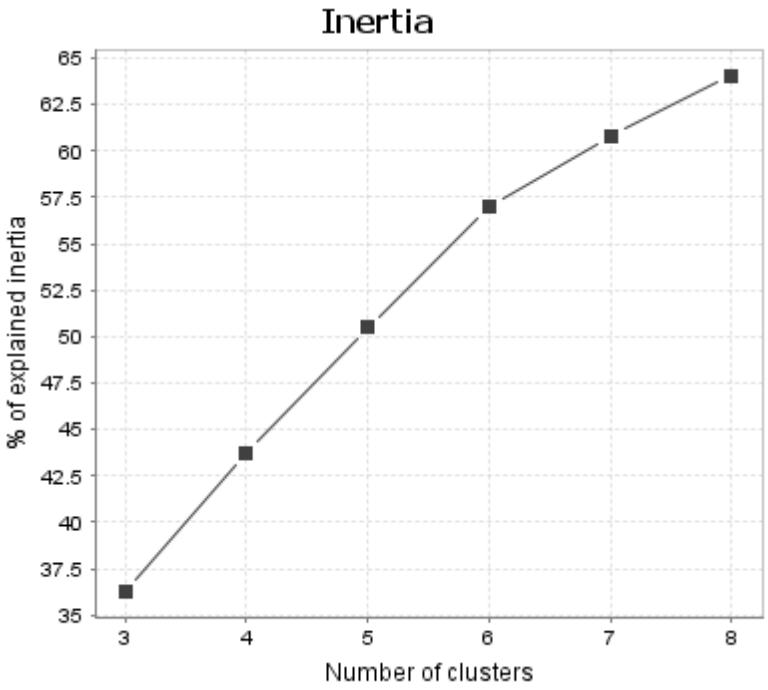
We performed a cluster analysis to the results of the principal component analysis on the numerical data that we have.

To do that we choose the hierarchical agglomerative clustering as the classification type, we used the four axis identified before with all the municipalities selected, the Ward's criterion as the aggregation indice and finding automatically the optimal number of clusters as the method.

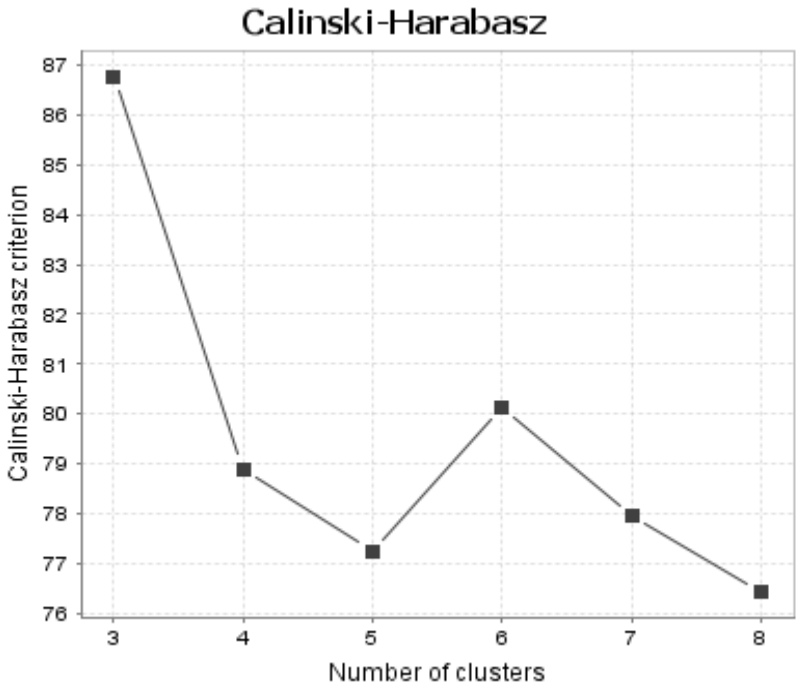
The result on SPAD is the dendrogram below for the maximum number of terminal elements of fifty. The best partitions selected are with three, four or six clusters.



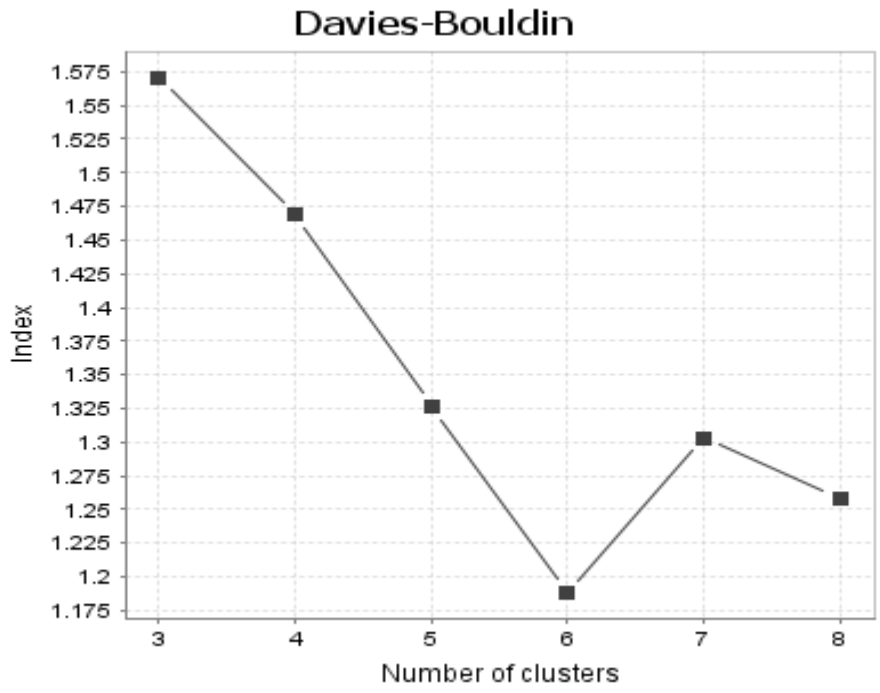
Next, is presented the graphical representations for the explained Inertia from three to eight cluster, chosen as the default options.



The Calinski-Harabasz index that has a maximum on three clusters and the second highest point is at six clusters.



Davies-Bouldin index graphical representation is below and with a minimum on six clusters.



With these results, can be said with confidence that the partition of the dendrogram on six clusters is the better way to analyze the results.

When SPAD tried to make a consolidation on the results, they indeed improved so below are the quality indicators for the three suggested partitions after the maximum number of consolidation iterations of ten has been reached.

As can be seen, the within cluster variance is the lowest one on six clusters and the between cluster variance is the highest one with six clusters as well.

Quality indicator (after consolidation)

Criteria	3 clusters	4 clusters	6 clusters
Within cluster variance	5,214	4,538	3,402
Between cluster variance	3,442	4,119	5,254
Between variance rate (η²)	39,763	47,580	60,695
Calinski-Harabasz (pseudo F) criterion	100,669	91,978	93,271
Davies-Bouldin's index	1,495	1,396	1,130

The table below is to make a comparison of the quality indicators before and after the consolidation and can be seen see that there was indeed an improvement of them.

Quality indicators

Name	Before consolidation	After consolidation
Within cluster variance	3,720	3,402
Between cluster variance	4,936	5,254
Between variance rate (η²)	57,025	60,695
Calinski-Harabasz (pseudo F) criterion	80,146	93,271
Davies-Bouldin's index	1,188	1,130

To have a better idea about the improvement of the information about the count and inertia on the six clusters before and after the consolidation the table below is presented.

Informations on the clusters

Cluster	Before consolidation			After consolidation		
	Count	Percentage	Inertia	Count	Percentage	Inertia
1	137	44,481	1,784	73	23,701	0,946
2	43	13,961	0,751	102	33,117	1,150
3	80	25,974	0,521	82	26,623	0,578
4	4	1,299	0,065	4	1,299	0,065
5	40	12,987	0,438	43	13,961	0,503
6	4	1,299	0,160	4	1,299	0,160
Overall	308	100,000	3,720	308	100,000	3,402

To complete the information about the clusters, below are the clusters centers on the four principal components after the consolidation and the distances between the centers.

Cluster center after consolidation

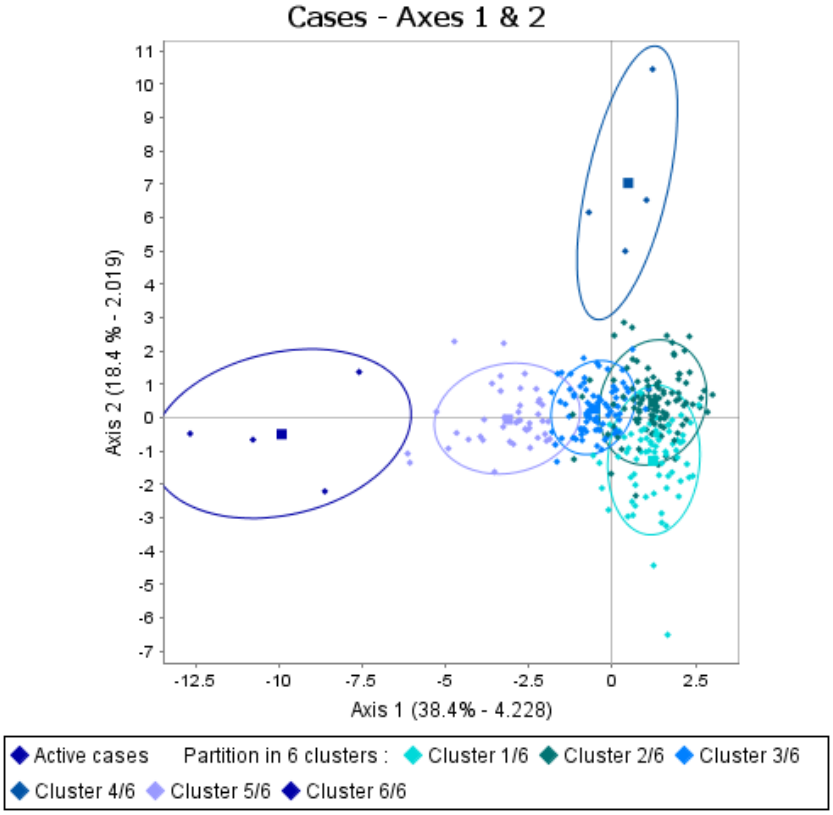
Variables	Center 1	Center 2	Center 3	Center 4	Center 5	Center 6
Axis_1	1,224	1,226	-0,518	0,480	-3,122	-9,911

Axis_2	-1,281	0,458	0,273	7,039	-0,041	-0,485
Axis_3	-0,572	0,903	-0,623	-1,060	-0,021	1,479
Axis_4	0,210	0,069	-0,397	-0,438	0,200	0,834

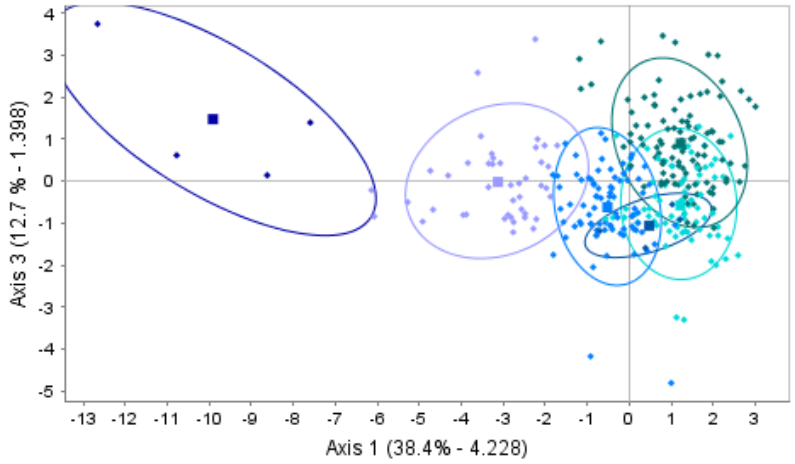
Distances between centers after consolidation

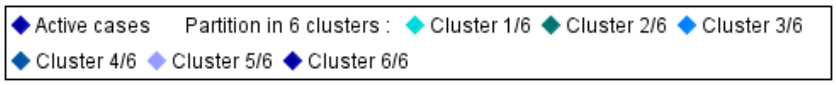
Centers	Center 1	Center 2	Center 3	Center 4	Center 5	Center 6
Center 1	0,000	2,285	2,412	8,393	4,553	11,368
Center 2	2,285	0,000	2,370	6,926	4,475	11,218
Center 3	2,412	2,370	0,000	6,853	2,757	9,734
Center 4	8,393	6,926	6,853	0,000	8,037	13,140
Center 5	4,553	4,475	2,757	8,037	0,000	6,996
Center 6	11,368	11,218	9,734	13,140	6,996	0,000

A graphical representation can be seen below for the clustering on the four axis, this number of cluster is the number that allowed us to have the better separation between the elements and to cover as much elements as possible.

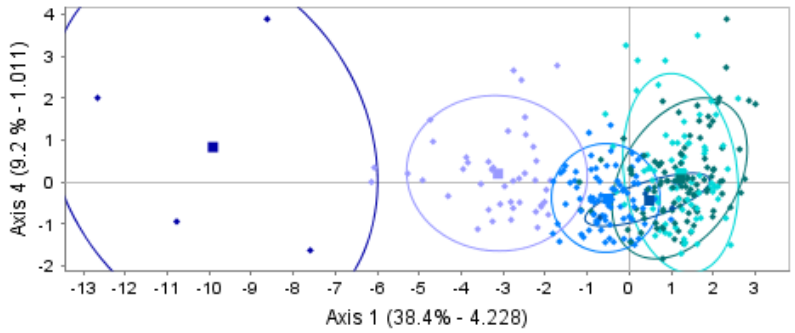


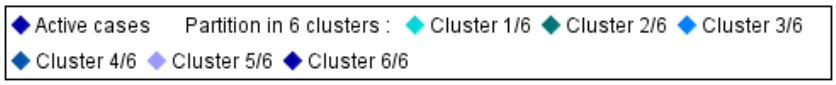
Cases - Axes 1 & 3





Cases - Axes 1 & 4





Finally, can be seen below that this solution in six clusters, after the consolidation, explains about 60% of the total inertia, while the solution in four clusters explained 47,6% and in three clusters explained only 39,8%.

Consolidation of the partition

Iterations	Total variance	Between cluster variance	Ratio
0	8,656	4,936	0,570
1	8,656	5,160	0,596
2	8,656	5,184	0,599
3	8,656	5,201	0,601
4	8,656	5,221	0,603
5	8,656	5,240	0,605
6	8,656	5,247	0,606
7	8,656	5,250	0,606
8	8,656	5,251	0,607
9	8,656	5,252	0,607
10	8,656	5,254	0,607

To have a direct characterization of the clusters in terms of the original variables, a characterization of the clusters of the typology was done in SPAD and that allowed to analyze the characteristics of the variables in the six clusters portioning.

For the cluster 1 with 73 municipalities, the variables with mean superior to the global mean are the ones for municipalities with more woman in the resident population (AARRP_W), with more people with lower education (P_15_LE), high resolution rate (RR), high effectiveness rate (ER) and the municipalities with more unemployment (AARU).

The ones with mean inferior to the global mean are the variables that represent the municipalities with lower monthly earnings for man (AMEE_M) and woman (AMEE_F), low high education (P_15_HE), low per capita purchasing power (PCPP) and lower proportion of male residents (AARRP_M).

Classe : Cluster 1/6 (Count : 73 - Percentage : 23.70%)

Characteristic variable	Mean in the partition	Global mean	Standard deviation (N) on the partition	Global standard deviation (N)	Count in the partition	Count	Test-Value	Probability
AARRP_W	53,360	52,393	1,026	1,226	73	308	7,691	0,000
P_15_LE	92,460	89,959	1,757	4,387	73	308	5,559	0,000
RR	113,460	100,509	23,610	24,666	73	308	5,119	0,000
ER	78,029	71,336	14,274	17,245	73	308	3,784	0,000
AARU	4,358	4,056	1,657	1,411	73	308	2,087	0,018
AMEE_M	892,368	1014,019	78,409	234,161	73	308	-5,065	0,000
P_15_HE	7,540	10,041	1,757	4,387	73	308	-5,559	0,000
PCPP	68,126	80,106	7,053	18,326	73	308	-6,374	0,000

AMEE_F	775,358	845,116	40,008	101,730	73	308	-6,686	0,000
AARRP_M	46,640	47,607	1,026	1,226	73	308	-7,691	0,000

It looks like this is a cluster of municipalities with lower earnings that are more represented by woman that are in a more difficult socio-economic situation but a in good situation when dealing with crime, probably because they are smaller and more rural municipalities with lower crime rates.

On the second cluster with 102 municipalities, the variables with a high mean comparing to the global mean are for the municipalities with high unemployment (AARU), more male residents (AARRP_M) and more people with lower education (P_15_LE).

The ones with lower mean are the municipalities that have lower earnings for man and woman (AMEE_M and AMEE_F), low high education (P_15_HE), low per capita purchasing power (PCPP), lower proportion of female residents (AARRP_W) and low values on the resolution rate (RR), effectiveness rate (ER) and crime rate registered by the police (ACR).

Classe : Cluster 2/6 (Count : 102 - Percentage : 33.12%)

Characteristic variable	Mean in the partition	Global mean	Standard deviation (N) on the partition	Global standard deviation (N)	Count in the partition	Count	Test-Value	Probability
P_15_LE	92,481	89,959	1,651	4,387	102	308	7,078	0,000
AARRP_M	48,012	47,607	0,830	1,226	102	308	4,069	0,000
AARU	4,351	4,056	1,473	1,411	102	308	2,580	0,005
ACR	25,950	28,043	7,582	9,411	102	308	-2,738	0,003
AARRP_W	51,988	52,393	0,830	1,226	102	308	-4,069	0,000
AMEE_M	928,562	1014,019	117,832	234,161	102	308	-4,492	0,000
AMEE_F	802,578	845,116	46,536	101,730	102	308	-5,147	0,000
PCPP	70,510	80,106	8,067	18,326	102	308	-6,446	0,000
P_15_HE	7,519	10,041	1,651	4,387	102	308	-7,078	0,000
RR	84,387	100,509	18,960	24,666	102	308	-8,045	0,000
ER	58,508	71,336	12,603	17,245	102	308	-9,157	0,000

Appears that this is a cluster for the poorer municipalities with more man, a low education level and higher unemployment but with low levels of criminality

On clusters 3 with 82 municipalities, the variables with higher mean are for the municipalities with higher resolution rate (RR) and higher effectiveness rate (ER), higher per capita purchasing

power (PCPP), more proportion of people with high education (P_15_HE), with more male residents than females (AARRP_M) but where the females have the highest earnings (AMEE_F).

On the other hand, the ones with the lowest mean are for municipalities with the lower proportion of female residents (AARRP_W), lower levels of low education (P_15_LE) and low unemployment (AARU).

Classe : Cluster 3/6 (Count : 82 - Percentage : 26.62%)

Characteristic variable	Mean in the partition	Global mean	Standard deviation (N) on the partition	Global standard deviation (N)	Count in the partition	Count	Test-Value	Probability
ER	78,577	71,336	14,748	17,245	82	308	4,424	0,000
RR	110,395	100,509	26,205	24,666	82	308	4,223	0,000
PCPP	85,198	80,106	7,303	18,326	82	308	2,928	0,002
P_15_HE	11,173	10,041	2,193	4,387	82	308	2,719	0,003
AARRP_M	47,864	47,607	0,597	1,226	82	308	2,208	0,014
AMEE_F	865,010	845,116	48,508	101,730	82	308	2,061	0,020
AARRP_W	52,136	52,393	0,597	1,226	82	308	-2,208	0,014
P_15_LE	88,827	89,959	2,193	4,387	82	308	-2,719	0,003
AARU	3,420	4,056	0,943	1,411	82	308	-4,750	0,000

This cluster seems to be a cluster for some of the richest municipalities with more resources to deal with crime, with high education level, low unemployment and where the females have the highest earnings.

For the fourth cluster, with only four municipalities, we could see that the variable with higher mean is for the municipalities with more male residents (AARRP_M) than female residents (AARRP_W) and that is the variable with the lower mean value.

Classe : Cluster 4/6 (Count : 4 - Percentage : 1.30%)

Characteristic variable	Mean in the partition	Global mean	Standard deviation (N) on the partition	Global standard deviation (N)	Count in the partition	Count	Test-Value	Probability
AARRP_M	53,737	47,607	1,781	1,226	4	308	10,033	0,000
AARRP_W	46,263	52,393	1,781	1,226	4	308	-10,033	0,000

This is a cluster for the municipalities with more man than woman.

The cluster five, with 43 municipalities, has the following variables with highest mean comparing to the global mean, high education (P_15_HE), per capita purchasing power (PCPP), highest earnings for man and woman (AMEE_F and AMEE_M), highest annual crime rate registered by the police (ACR) and highest effectiveness rate (ER). Contrasting with the variable for municipalities with more people with low education (P_15_LE) that has lower mean value than the global mean.

Classe : Cluster 5/6 (Count : 43 - Percentage : 13.96%)

Characteristic variable	Mean in the partition	Global mean	Standard deviation (N) on the partition	Global standard deviation (N)	Count in the partition	Count	Test-Value	Probability
P_15_HE	16,633	10,041	3,653	4,387	43	308	10,587	0,000
PCPP	105,242	80,106	11,853	18,326	43	308	9,665	0,000
AMEE_F	982,928	845,116	95,625	101,730	43	308	9,546	0,000
AMEE_M	1250,893	1014,019	255,401	234,161	43	308	7,128	0,000
ACR	33,938	28,043	10,342	9,411	43	308	4,414	0,000
ER	77,488	71,336	16,829	17,245	43	308	2,514	0,006
P_15_LE	83,367	89,959	3,653	4,387	43	308	-10,587	0,000

Appears that this cluster is for the richest municipalities that has more crimes but more resources to deal with them and higher education levels.

For the last cluster, with four municipalities, for the variables with high mean we have the ones with the highest earnings (AMEE_M and AMEE_F), more per capita purchasing power (PCPP), highest education level (P_15_HE), highest annual crime rate registered by the police (ACR) and highest proportion of female residents (AARRP_W).

Opposing the variables with lowest mean for the municipalities with lowest male proportion of residents (AARRP_M) and lower education levels (P_15_LE).

Classe : Cluster 6/6 (Count : 4 - Percentage : 1.30%)

Characteristic variable	Mean in the partition	Global mean	Standard deviation (N) on the partition	Global standard deviation (N)	Count in the partition	Count	Test-Value	Probability
AMEE_M	2141,500	1014,019	723,580	234,161	4	308	9,662	0,000
PCPP	163,175	80,106	36,147	18,326	4	308	9,096	0,000

AMEE_F	1268,650	845,116	186,478	101,730	4	308	8,354	0,000
P_15_HE	27,875	10,041	4,654	4,387	4	308	8,157	0,000
ACR	52,549	28,043	22,834	9,411	4	308	5,225	0,000
AARRP_W	53,812	52,393	1,018	1,226	4	308	2,321	0,010
AARRP_M	46,188	47,607	1,018	1,226	4	308	-2,321	0,010
P_15_LE	72,125	89,959	4,654	4,387	4	308	-8,157	0,000

This is a cluster for the richest municipalities with more woman and high criminality, but they doesn't stand when it comes to dealing with crime.

For the supplementary variable, average monthly earnings for employees (AMEE), it has a low mean on clusters one and two that are the clusters representing the poorer municipalities. However, it has high mean values comparing to the global mean for the clusters five and six because these represent the richest municipalities.

Doing the clustering by the means k-means classification type, the same six clusters were found, with similar number of municipalities per cluster and almost the same characteristic variables per cluster so it is not necessary to repeat all the information that we already talked about on the hierarchical clustering explained before.

Conclusion

After performing the univariate and bivariate analysis of this dataset, it was possible to reach the following conclusions:

Municipalities with a high per capita purchasing power tend to have employees with higher average monthly earnings, have a higher rate of effectiveness, which may indicate that a resident population with more purchasing power has more resources to deal with judicial costs and higher crime rates and those municipalities has more proportion of the female employees with higher average monthly earnings than male employees with higher average monthly.

Municipalities with high rates of lower education tend to have lower levels of criminal rates, this can be related to lower resources of the people to deal with crime and because they are from less urban municipalities that has less crime.

Unemployment rate and criminal rate are not directly related, with a correlation of only 0.002.

From the multivariate analysis can be concluded that there are differences between municipalities regarding earnings and education levels, as can be seen in clusters 6 and 5, and this is linked to the crime rates because these are the municipalities more urban with more inequalities among the residents, and that leads to have some proportion of the population that tend to have more delinquent behaviors, and because of that it would be interesting to have some social indicators that could be analyzed to see this differences better but that didn't came to mind when the dataset was built.

The poorer and less populated municipalities have less inequalities probably because in most of them the level of life is lower than in the more populated municipalities, and population level is a good indicator of development because people go to the more developed areas searching for jobs and bigger earnings, and because of that the crime rate is lower and when it exists people don't go the police as easy as in the bigger municipalities because the judicial system is expensive and in the less developed municipalities the means to deal with crime are less.