PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO

Programa de Estudos Pós-Graduados em Administração

Mestrado em Administração

PESQUISA SOCIO-ECONOMICA AO NIVEL MUNICIPAL NO BRASIL

focando principalmente indicadores relacionados a Habitação, Educação, Trabalho e muito particularmente SAÚDE

MÉTODOS QUANTITATIVOS DA PESQUISA EMPÍRICA

Professor Dr. Arnoldo Jose de Hoyos

Elaine Pinheiro Palmeira

1 – INTRODUÇÃO

O objetivo deste trabalho é efetuar diversas análises dos dados da Pesquisa Firjan/FGV sobre o Desenvolvimento dos Municípios nos períodos de 2000 e 2010. Iniciamos com o entendimento dos dados, incluindo a definição dos indivíduos e das variáveis, suas classificações em variáveis categóricas ou quantitativas, os significados e unidades de medida, além da apresentação da tabela de dados.

Em seguida, será analisada cada uma das variáveis separadamente quanto a sua forma de distribuição, os valores atípicos, medidas de centro e dispersão. Neste momento faremos uso de gráficos (*pie chart*, barras, histogramas, gráficos de ramos, box-plot, dot-plot e curvas de densidade) e de medidas numéricas (média, mediana, quartis, desvio-padrão, variância, intervalo de confiança e teste de normalidade de Anderson-Darling).

Na sequência, faremos comparações entre as diversas variáveis analíticas, utilizando técnicas como relações entre as variáveis, regressões múltiplas, comparações, amostragem dos dados, análise multivariada, análise de conglomerados, análise discriminante, regressão logística, análise de correspondência e arvores de classificação. O software estatístico utilizado é o **MINITAB 16**

Não será possível, a partir destes dados, efetuarmos a análise de tendência pois não existem séries temporais de dados, requisitos para esta técnica.

2 – OS DADOS

2.1 – OS INDIVÍDUOS

Os indivíduos deste trabalho são compostos pelas médias ponderadas dos indicadores das dimensões Habitação (H6), Renda (R1), Trabalho (T1_2), Saúde (S1_1) e Educação (E e E2_4), padronizados pela média do Brasil para os diferentes municípios. Ao todo são 5565 municípios considerados brasileiros, incluindo o Distrito Federal. Os dados analíticos foram extraídos do IBGE e possibilitam uma comparação entre os dados colhidos em 2000 com 2010. O foco da análise deste trabalho são os dados referentes à 2010.

O Brasil encontra-se política e geograficamente dividido em cinco regiões distintas, que possuem traços comuns referentes aos aspectos físicos, humanos, econômicos e culturais. Os limites de cada região - Norte, Nordeste, Sudeste, Sul e Centro-Oeste - coincidem sempre com as fronteiras dos Estados que as compõem.

2.2 – AS VARIÁVEIS

As variáveis desta pesquisa incluem os 3 principais índices sintéticos que são ISDM, IFDM e IFGF, que são médias ponderadas dos dados analíticos globais da pesquisa, e variáveis analíticas, referente à educação, saúde, renda, emprego e habitação.

Tabela 1. Comparativo entre as Variáveis ISDM e IFDM

O QUE O ISDM (FGV) MEDE	O QUE O IFDM (Firjan) MEDE
Educação: taxa de analfabetismo e taxa de crianças e jovens que frequentam a escola em cada etapa, desempenho na Prova Brasil (MEC)	**Educação:** taxa de matrícula infantil, abandono, distorção idade-série, desempenho no Ideb, taxa de docentes com ensino superior
Saúde e Segurança: taxa de mortalidade infantil, gravidez precoce e mortalidade por causas evitáveis; homicídios	**Saúde:** número de consultas pré-natal, óbitos por causa mal definidas e óbitos infantis evitáveis
Renda: presença de pobreza e extrema pobreza	**Emprego e renda:** geração, estoque e salários médios dos empregos formais
Trabalho: taxa de ocupação e formalização	
Habitação: coleta de lixo, energia elétrica, água canalizada, esgotamento sanitário, domicílio próprio	

Tabela 2. A definição das Variáveis

Variável	Significado	Tipo	Unidade de Medida
REGIÃO	Nome da Região do Brasil	Texto	Na
UF	Unidade da Federação	Texto	Na
MUNICÍPIO	Nome do Município	Texto	Na

ISDM	Índice Social de Desenvolvimento Municipal: Média ponderada dos indicadores das dimensões Habitação, Renda, Trabalho, Saúde e Segurança e Educação (H, R, T, S e E) padronizada pela média do Brasil.	Numérico	Escala convertida para intervalo entre 0 e 1.
EDUCAÇÃO	Média ponderada dos indicadores da dimensão Educação (E1_1, E1_2, E2_1, E2_2, E2_3, E2_4, E2_5, E2_6, E3_1, E3_2 e E3_3) padronizada pela média do Brasil.	Numérico	Escala convertida para intervalo entre 0 e 1.
EMPREGO E RENDA	Geração, estoque e salários médios dos empregos formais (IFDM).	Numérico	Escala convertida para intervalo entre 0 e 1.
LIQUIDEZ	Índice de liquidez dos municípios.	Numérico	Escala convertida para intervalo entre 0 e 1.
H6	Percentual de pessoas que vivem em domicílio que tem densidade de moradores por dormitório inferior a 2.	Numérico	Escala convertida para intervalo entre 0 e 1.
R1		Numérico	Escala convertida para intervalo entre 0 e 1.
T1_2	Taxa de formalização entre os empregados	Numérico	Escala convertida para intervalo entre 0 e 1.
S1_1	Taxa de sobrevivência infantil no primeiro ano de vida, representada pela diferença entre o número de nascidos vivos e o número de óbitos até um ano de idade.	Numérico	Escala convertida para intervalo entre 0 e 1.
E2_4	Percentual de crianças de 7 a 14 anos que estão na série correta segundo a idade	Numérico	Escala convertida para intervalo entre 0 e 1.

3. ANÁLISE DAS VARIÁVEIS

3.1 VARIÁVEIS CATEGÓRICAS

Para este tipo de variável, as pesquisas concentram-se nas análises de gráficos do tipo *pie chart* e barras.

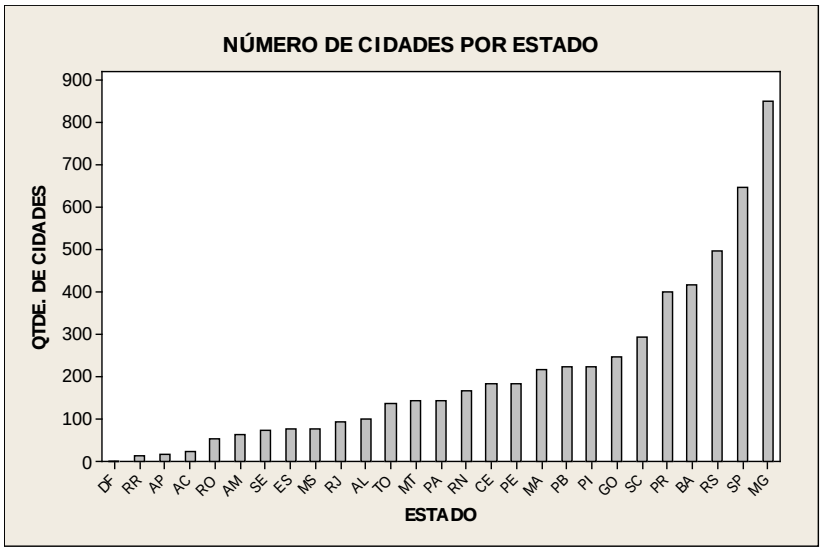
3.1.1 Variável: “ESTADO”

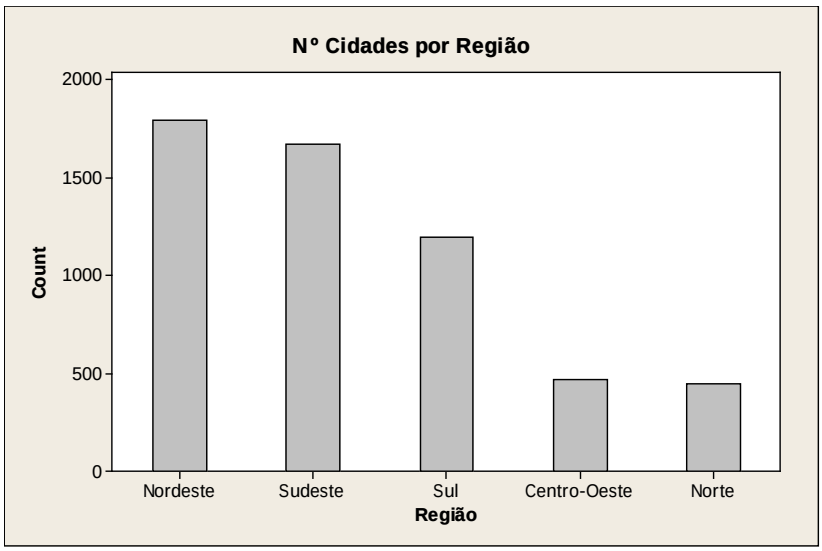
Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso é o estado que possui o maior número de cidades (852).

3.1.2 Variável: “REGIÃO”

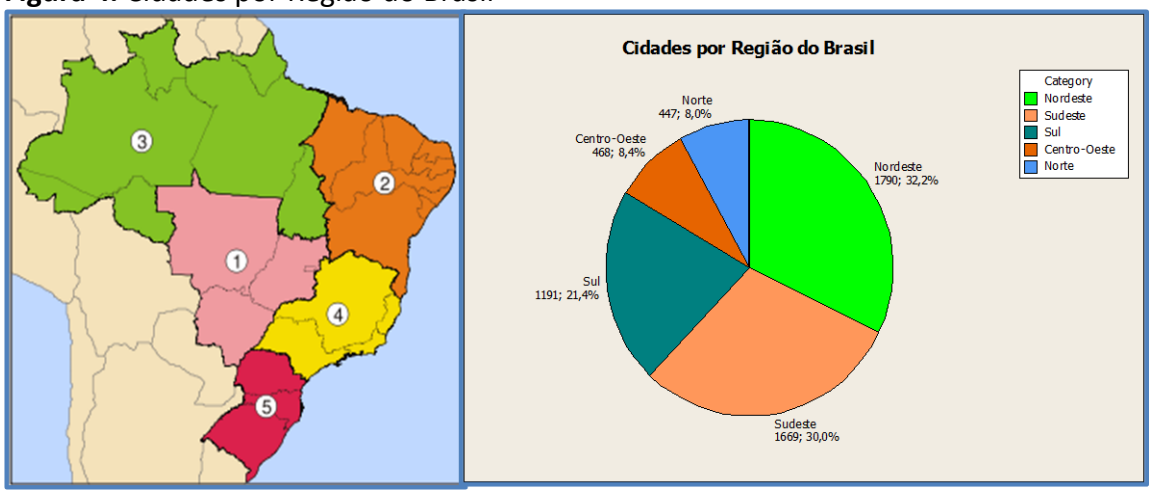
Figura 3. Número de Cidades por Estado e Região do Brasil





Nos gráficos ao lado podemos ter uma dimensão do número de cidades por estado e por região. A região do Brasil com o maior número de cidades é a Nordeste (1790), seguida pela região Sudeste (1669) e pela região Sul, com 1191 cidades. As regiões com menor número de cidades

Figura 4. Cidades por Região do Brasil



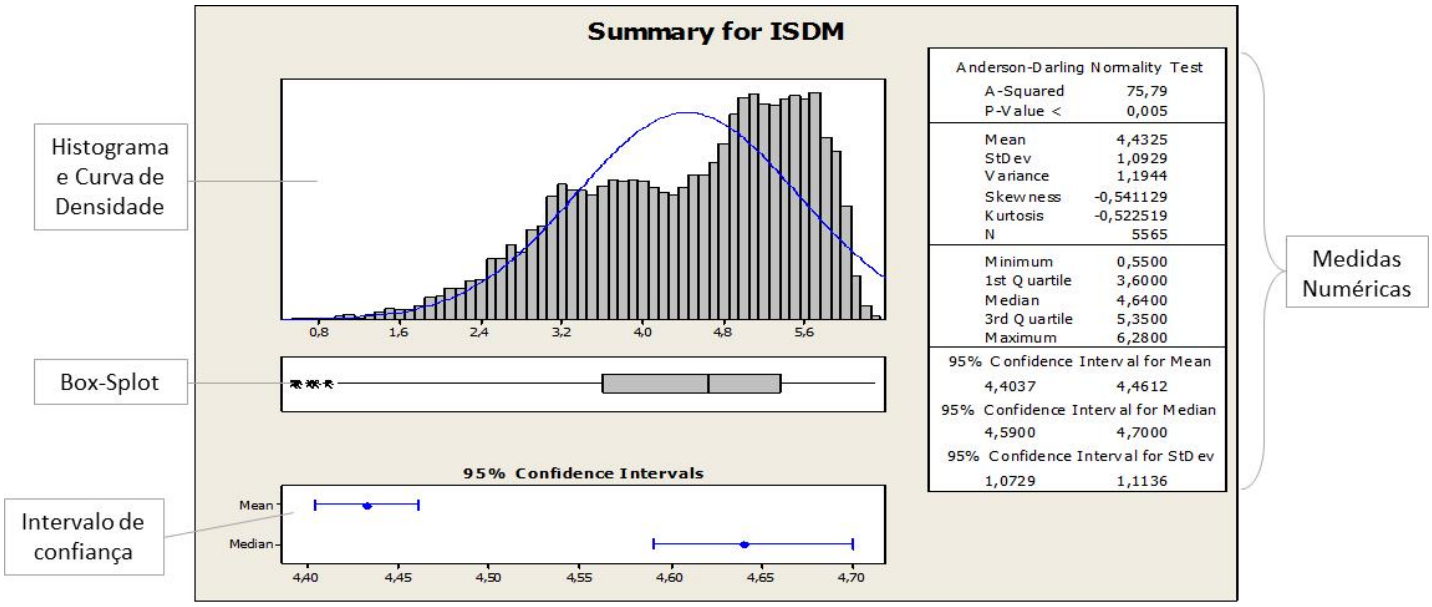
3.2 ANÁLISE EXPLORATÓRIA DAS VARIÁVEIS ANALÍTICAS

Serão analisadas as variáveis separadamente quanto a sua forma de distribuição, os valores atípicos, medidas de centro e dispersão. Para tanto contamos com o auxílio de gráficos (histogramas, gráficos de ramos, box-plot, dot-plot e curvas de densidade) e de medidas numéricas (média, mediana, quartis, desvio-padrão, variância, intervalo de confiança e teste de normalidade de Anderson-Darling).

3.2.1 VARIÁVEL ISDM

STAT >> BASIC STATISTICS >> GRAPHICAL SUMMARY

Segue abaixo quadro contendo Histograma, Curva de Densidade, Box-Plot, Intervalo de confiança da média e mediana, além das medidas numéricas como média, desvio-padrão, variância, quantidade de observações, valores mínimos, máximos, informações dos quartis e o teste de normalidade de Anderson-Darling (A-Squared e P-Value), para a variável ISDM.



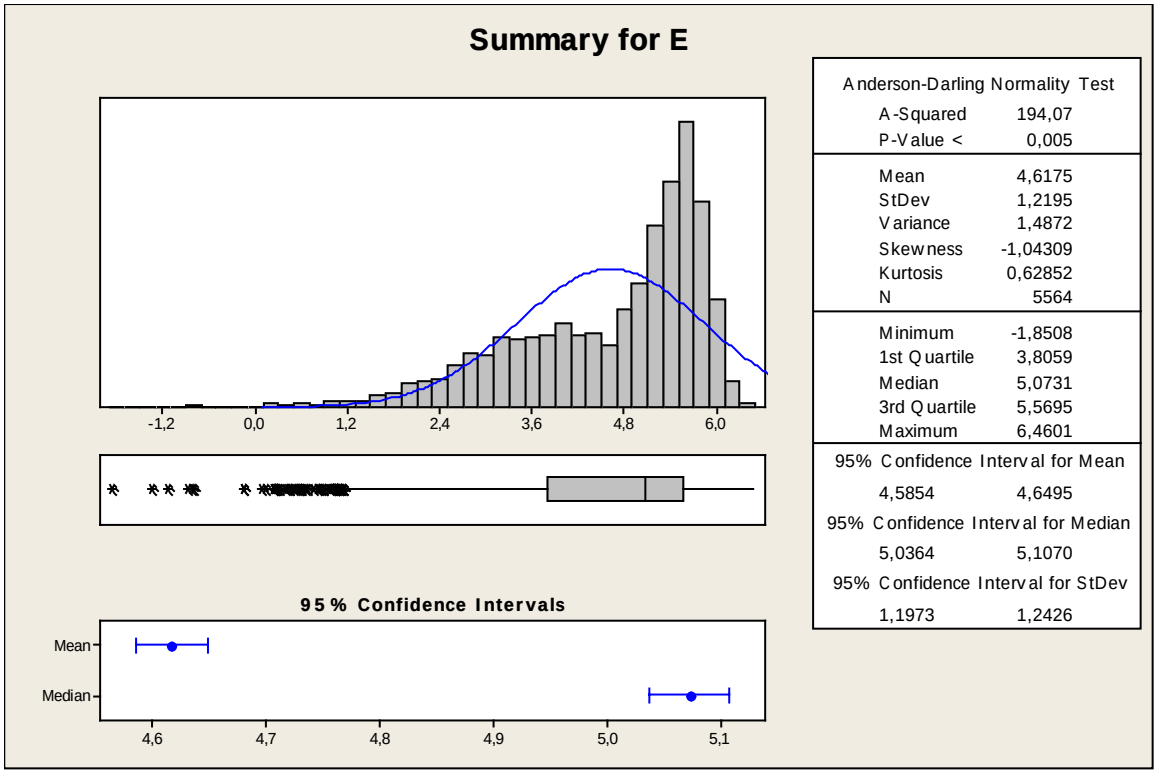
As principais observações que podemos fazer são:

- **Forma:** O Histograma nos permite verificar que trata-se de uma distribuição visivelmente assimétrica para a direita, o que é comum para variáveis que indiquem ganhos. Esta conclusão está comprovada pelo teste de normalidade de Anderson-Darling que indica que a distribuição não pode ser considerada uma Normal. Muitos municípios enfrentam problemas de ordem sustentável, enquanto poucos possuem uma situação mais plena. Fato

que se dá também pelo desequilíbrio econômico e social das mais variadas regiões do Brasil. Embora o ISDM de alguns municípios possuir valor alto, o que faz o gráfico se estender para a direita, a distribuição tem um único pico que representa os municípios com ISDM de 5,65 a 5,75. O Box-Plot nos deixa ainda mais clara esta assimetria da distribuição.

- **Centro e Dispersão:** A mediada do IFDM é de 4,64, ou seja, metade dos municípios possuem valores inferiores ou iguais à mediana e metade da população terá valores superiores ou iguais à este número.

3.2.2 VARIÁVEL EDUCAÇÃO

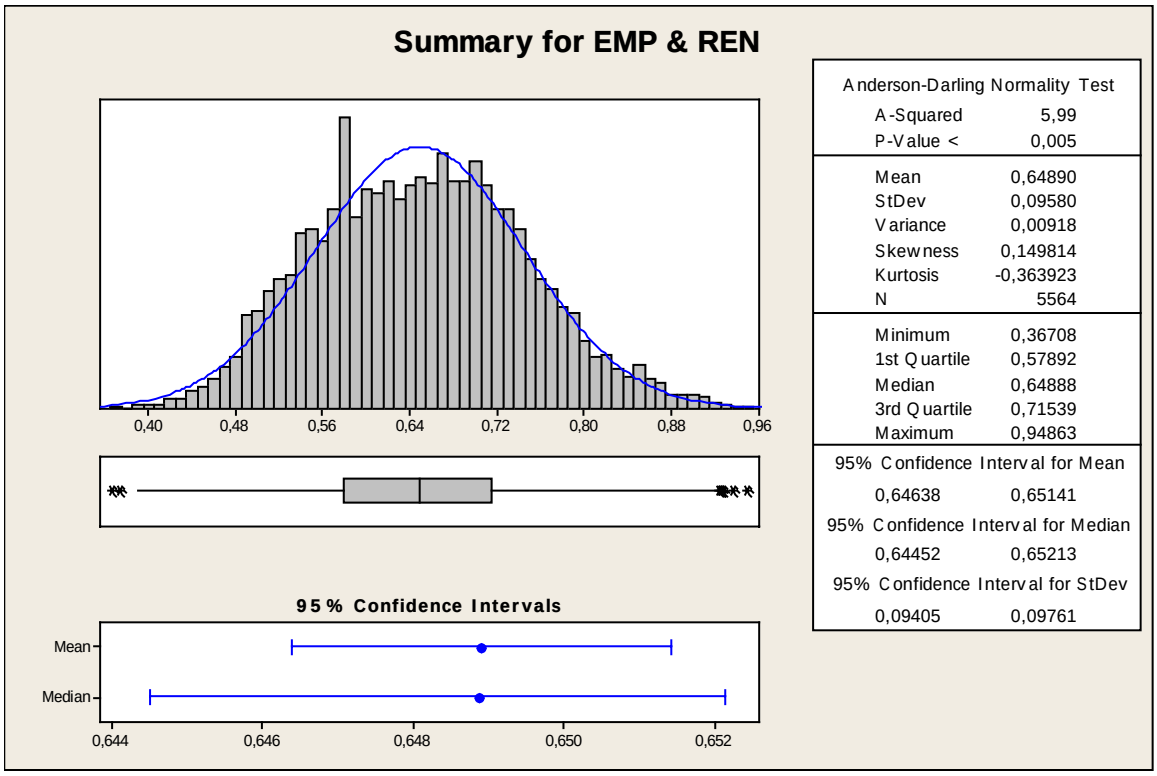


As principais observações que podemos fazer são:

- **Forma:** O Histograma nos permite verificar que trata-se de uma distribuição que tende a ser simétrica cujo pico concentra-se no centro, o que é comum para variáveis que indiquem desempenho regular. A curva apresenta várias corcovas, o que indica que temos diversas realidades sobre a questão da variabilidade sobre Educação nos municípios do Brasil. Os dados se dispersam muito, não existe um padrão na questão e pode-se concluir que existe muita diversidade entre os dados.

- Centro e Dispersão: A mediana nos indica que aproximadamente metade dos municípios tem Educação menor do que 5,0731. A Educação média é 4,6175 e o desvio-padrão (medida de dispersão) é de 1,2195, que implica em uma dispersão média para a questão.

3.2.3 VARIÁVEL EMPREGO E RENDA



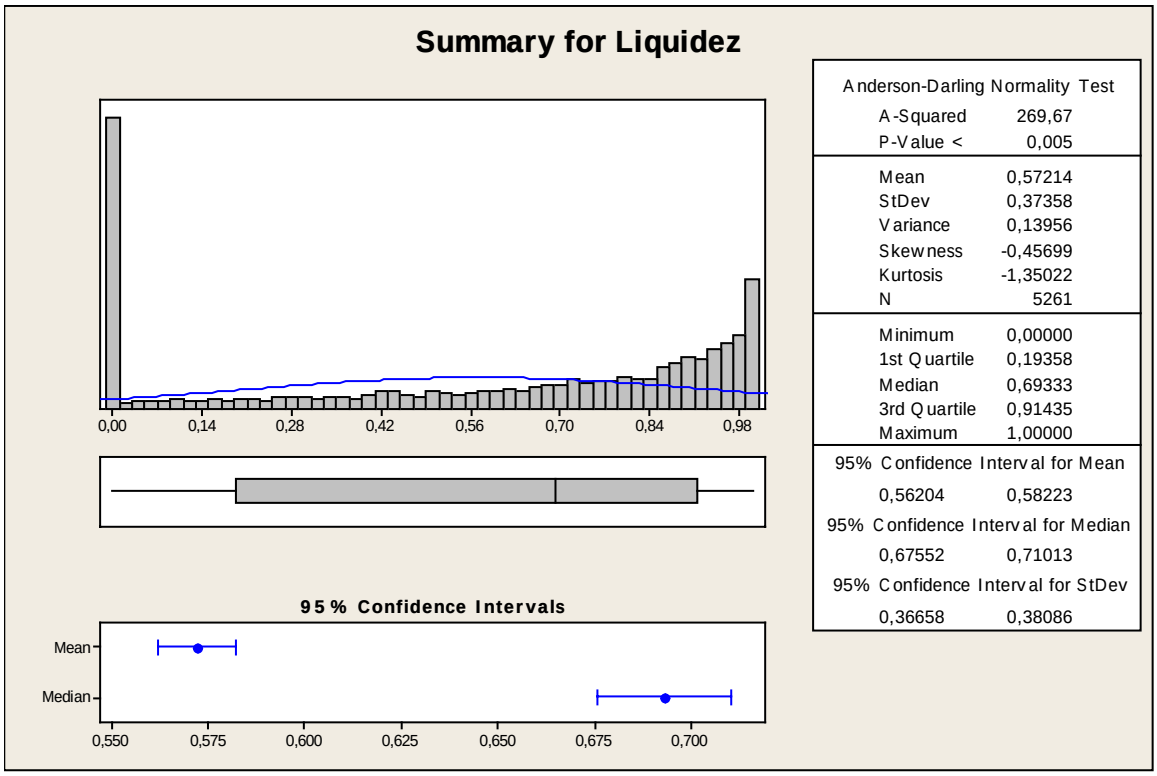
Anderson-Darling Normality Test	
A-Squared	5,99
P-Value <	0,005
Mean	0,64890
StDev	0,09580
Variance	0,00918
Skewness	0,149814
Kurtosis	-0,363923
N	5564
Minimum	0,36708
1st Quartile	0,57892
Median	0,64888
3rd Quartile	0,71539
Maximum	0,94863
95% Confidence Interval for Mean	
0,64638	0,65141
95% Confidence Interval for Median	
0,64452	0,65213
95% Confidence Interval for StDev	
0,09405	0,09761

- Forma: O Histograma nos permite verificar que trata-se de uma distribuição fortemente assimétrica tendendo para a esquerda, o que é comum para variáveis que indiquem desempenho baixo e menores números dentro de toda a distribuição dos dados. Esta conclusão está comprovada pelo teste de normalidade de Anderson-Darling que indica que a distribuição não pode ser considerada uma Normal. A maior parte das cidades possui valores baixos de emprego e renda. Muitas cidades possuem um nível médio de emprego e renda e poucas possuem um nível alto de emprego e renda. Existe apenas uma corcova no gráfico.

- Centro e Dispersão: A mediana nos indica que aproximadamente metade dos municípios tem emprego e renda menor do que 0,64888. A média de emprego e renda é 0,64890 e o desvio-padrão (medida de dispersão) é 0,09580, que implica em uma dispersão alta do índice de emprego e renda.

3.2.4 VARIÁVEL LIQUIDEZ

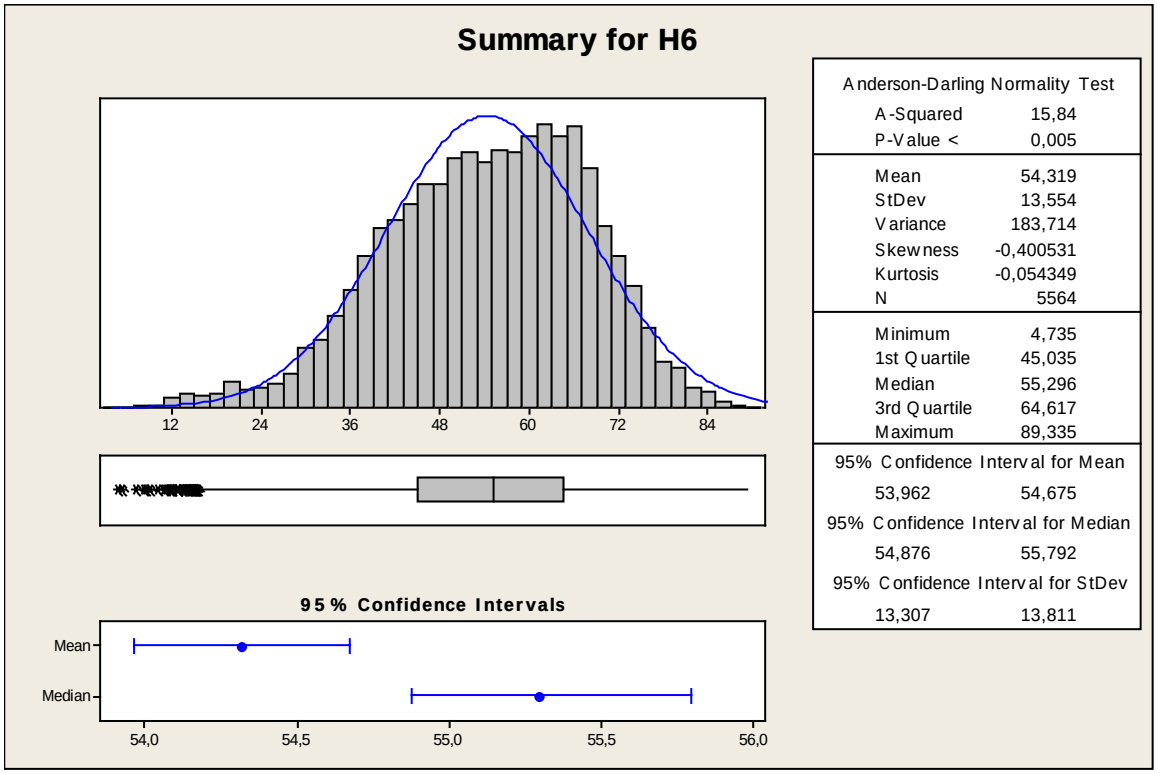
O indicador demonstra se o município possui recursos financeiros suficientes para fazer frente ao montante de restos a pagar. Se o município apresentar mais restos a pagar do que ativos financeiros disponíveis a pontuação será zero. Na leitura dos resultados, quanto mais próximo de 1,00, menos o município está postergando pagamentos para o exercício seguinte sem a devida cobertura.



- Forma: O Histograma nos permite verificar que trata-se de uma distribuição totalmente assimétrica tendendo levemente para a direita, o que é comum para variáveis que indiquem desempenho baixo e menores números dentro de toda a distribuição dos dados. Esta conclusão está comprovada pelo teste de normalidade de Anderson-Darling que indica que a distribuição não pode ser considerada uma Normal. Os valores de liquidez se espalham por todo o gráfico, não tendo um pico dos dados.

- Centro e Dispersão: A mediana nos indica que aproximadamente metade dos municípios tem liquidez menor do que 0,69333. A liquidez média é de 0,57214 e o desvio-padrão (medida de dispersão) é de 0,37358, que implica em uma dispersão absoluta do índice de liquidez.

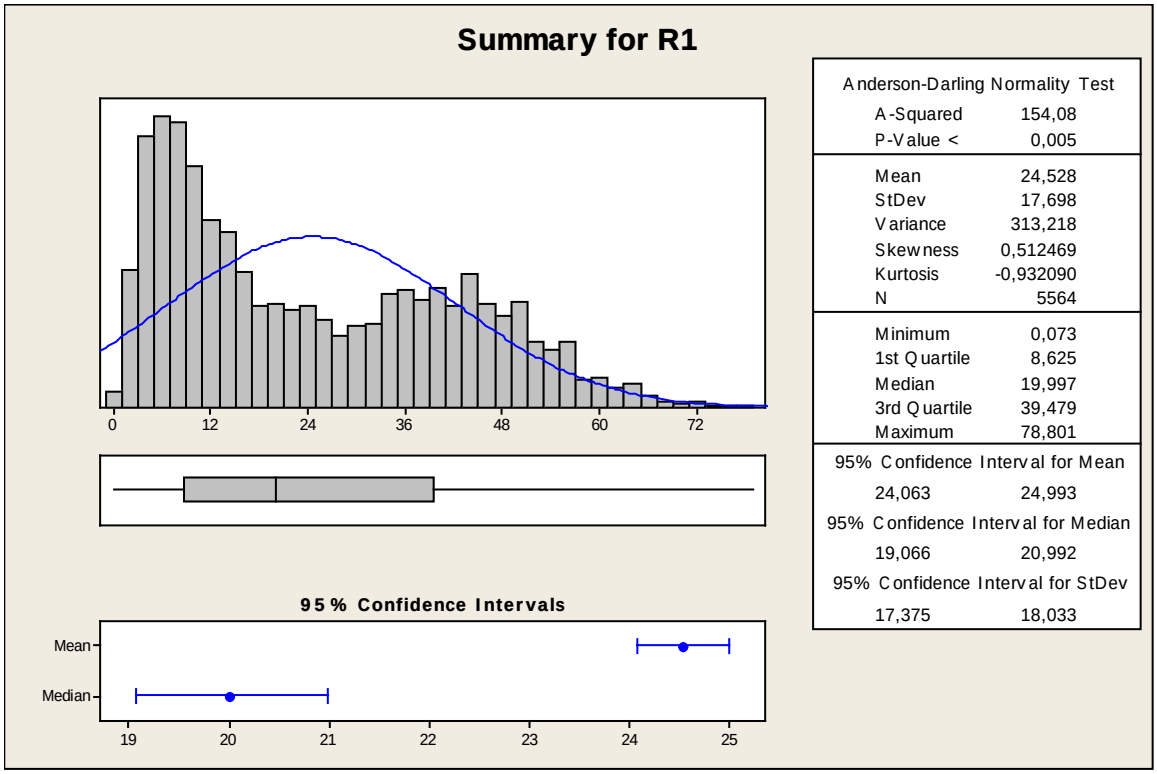
3.2.5 VARIÁVEL H6 (Pessoas que vivem em domicílio que tem densidade de moradores por dormitório inferior a 2)



- Forma: O Histograma nos permite verificar que trata-se de uma distribuição que tende a ser levemente assimétrica cujo pico concentra-se à direita, o que é comum para variáveis que indiquem desempenho médio para alto. A curva apresenta algumas corcovas, o que indica que temos um comportamento atípico da variabilidade sobre os dados de H6. Os dados se dispersam bastante, e podemos afirmar que a variável H6 tem alta dispersão em relação aos municípios do Brasil.

- Centro e Dispersão: A mediana nos indica que aproximadamente metade dos municípios tem H6 menor do que 55,296. O H6 médio é de 54,319 e o desvio-padrão (medida de dispersão) é de 13,554, que implica em uma dispersão média para H6.

3.2.6 VARIÁVEL R1 (Pessoas com renda domiciliar per capita abaixo da linha de pobreza (R$ 140,00))

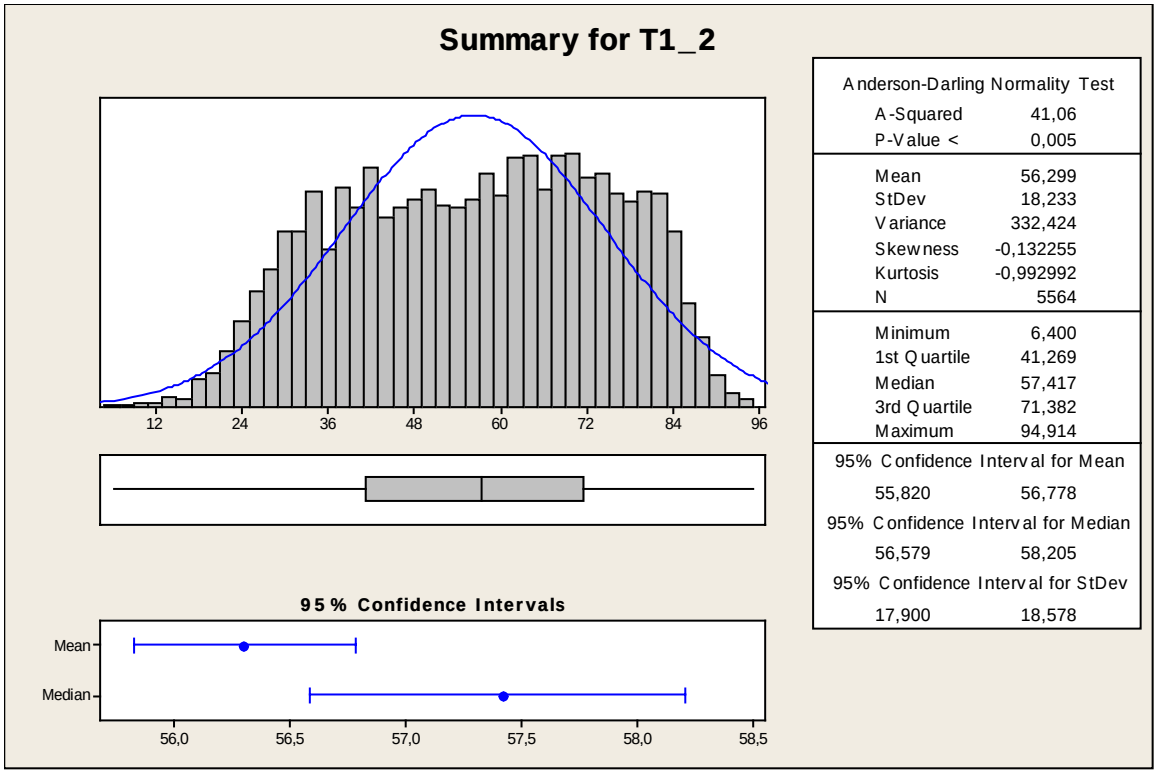


Anderson-Darling Normality Test	
A-Squared	154,08
P-Value <	0,005
Mean	24,528
StDev	17,698
Variance	313,218
Skewness	0,512469
Kurtosis	-0,932090
N	5564
Minimum	0,073
1st Quartile	8,625
Median	19,997
3rd Quartile	39,479
Maximum	78,801
95% Confidence Interval for Mean	
24,063	24,993
95% Confidence Interval for Median	
19,066	20,992
95% Confidence Interval for StDev	
17,375	18,033

- Forma: O Histograma nos permite verificar que trata-se de uma distribuição que tende a ser levemente assimétrica cujo pico concentra-se à esquerda, o que é comum para variáveis que indiquem desempenho baixo. A curva apresenta algumas corcovas, sendo duas altamente acentuadas, a primeira com maior pico e localizada fortemente à esquerda do gráfico. Indica que o comportamento atípico da variabilidade sobre os dados de R1. Os dados se dispersam bastante, e podemos afirmar que a variável R1 tem alta dispersão em relação aos municípios do Brasil.

- Centro e Dispersão: A mediana nos indica que aproximadamente metade dos municípios tem R1 menor do que 19,997. O R1 médio é de 24,528 e o desvio-padrão (medida de dispersão) é de 17,698, que implica em uma dispersão alta para R1.

3.2.7 VARIÁVEL T1_2 (Taxa de formalização entre os empregados)



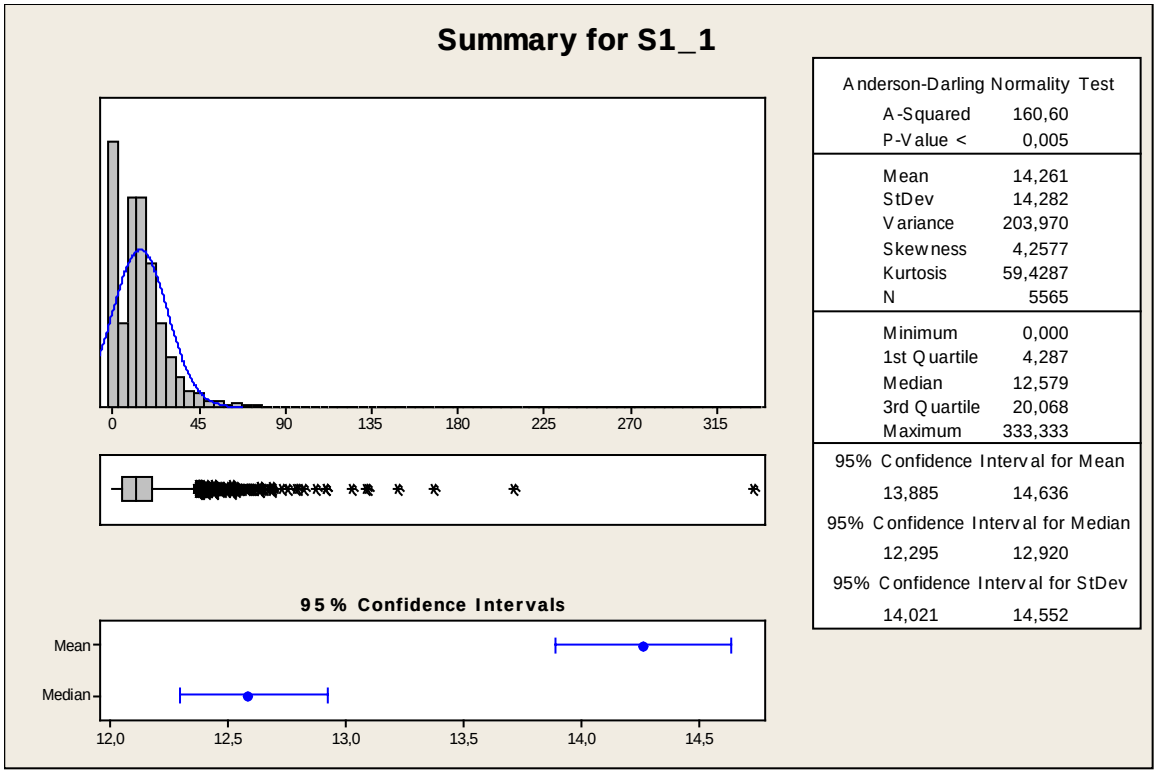
Anderson-Darling Normality Test	
A-Squared	41,06
P-Value <	0,005
Mean	56,299
StDev	18,233
Variance	332,424
Skewness	-0,132255
Kurtosis	-0,992992
N	5564
Minimum	6,400
1st Quartile	41,269
Median	57,417
3rd Quartile	71,382
Maximum	94,914
95% Confidence Interval for Mean	
55,820	56,778
95% Confidence Interval for Median	
56,579	58,205
95% Confidence Interval for StDev	
17,900	18,578

As principais observações que podemos fazer são:

- **Forma:** O Histograma nos permite verificar que trata-se de uma distribuição simétrica, embora o gráfico apresente várias corcovas na sua distribuiçõ. Indica que trata-se de um desempenho regular. Esta conclusão está comprovada pelo teste de normalidade de Anderson-Darling que indica que a distribuição pode ser considerada uma Normal. Muitas cidades possuem um baixo nível de desenvolvimento, muitas cidades possuem um nível médio de desenvolvimento e muitas possuem um nível alto de desenvolvimento. Existem várias corcovas no gráfico que nos mostra que existem N realidades nos dados analisados, ou seja, existem vários tipos de municípios dentro do Brasil em relação a formalização dos empregos.

- **Centro e Dispersão:** A mediana nos indica que aproximadamente metade dos municípios tem T1_2 menor do que 57,417. O T1_2 médio é de 56,299, e o desvio-padrão (medida de dispersão) é de 18,233, que implica em uma dispersão grande da população de T1_2.

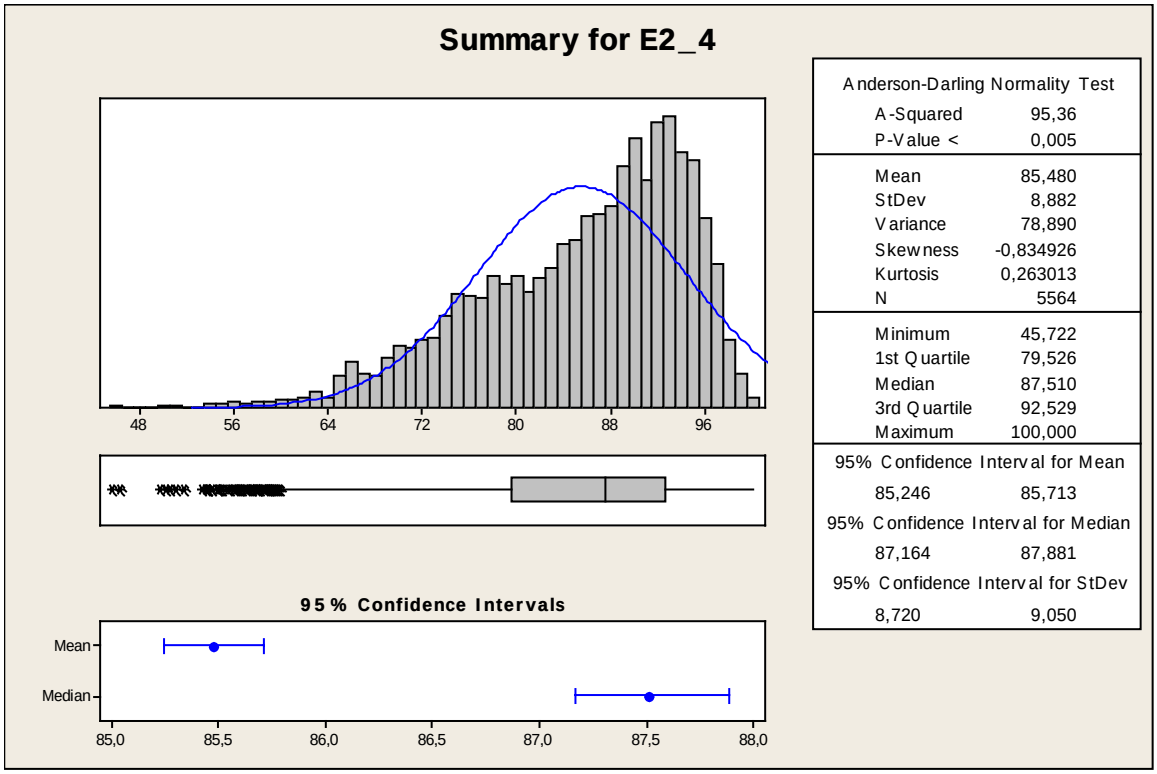
3.2.8 VARIÁVEL S1_1 (Taxa de mortalidade infantil, por mil nascidos vivos)



- **Forma:** O Histograma nos permite verificar que trata-se de uma distribuição fortemente assimétrica tendendo para a esquerda, o que é comum para variáveis que indiquem desempenho baixo e menores números dentro de toda a distribuição dos dados. Esta conclusão está comprovada pelo teste de normalidade de Anderson-Darling que indica que a distribuição não pode ser considerada uma Normal. A maior parte das cidades possui valores baixos de S1_1. Pouca cidades possuem um nível médio de S1_1 e quase nenhuma possuem um nível alto de S1_1. Existem duas corcovas visíveis no gráfico. Como trata-se de nascido vivos, o número baixo é bom porque a maioria dos nascidos vivos sobrevivem após um ano de vida.

- **Centro e Dispersão:** A mediana nos indica que aproximadamente metade dos municípios tem S1_1 menor do que 12,579. O S1_1 médio é de 14,261 e o desvio-padrão (medida de dispersão) é de 14,282, que implica em uma dispersão baixa do índice de S1_1.

3.2.9 VARIÁVEL E2_4 (Crianças entre 7 e 14 anos que estudam na série correta segundo sua idade)



Anderson-Darling Normality Test	
A-Squared	95,36
P-Value <	0,005
Mean	85,480
StDev	8,882
Variance	78,890
Skewness	-0,834926
Kurtosis	0,263013
N	5564
Minimum	45,722
1st Quartile	79,526
Median	87,510
3rd Quartile	92,529
Maximum	100,000
95% Confidence Interval for Mean	
85,246	85,713
95% Confidence Interval for Median	
87,164	87,881
95% Confidence Interval for StDev	
8,720	9,050

- **Forma:** O Histograma nos permite verificar que trata-se de uma distribuição fortemente assimétrica tendendo para a direita, o que é comum para variáveis que indiquem desempenho alto e taxas elevadas. Esta conclusão está comprovada pelo teste de normalidade de Anderson-Darling que indica que a distribuição não pode ser considerada uma Normal. A curva apresenta várias corcovas, o que indica que temos diversas realidades sobre a questão da série correta dos alunos. Os dados se dispersam muito, não existe um padrão na questão e pode-se concluir que existe muita diversidade entre a questão do grau correto de idade e escolaridade nos municípios.

- **Centro e Dispersão:** A mediana nos indica que aproximadamente metade dos municípios tem E2_4 menor do que 87,510. O E2_4 médio é de 85,480 e o desvio-padrão (medida de dispersão) é de 8,882, que implica em uma dispersão grande para a questão.

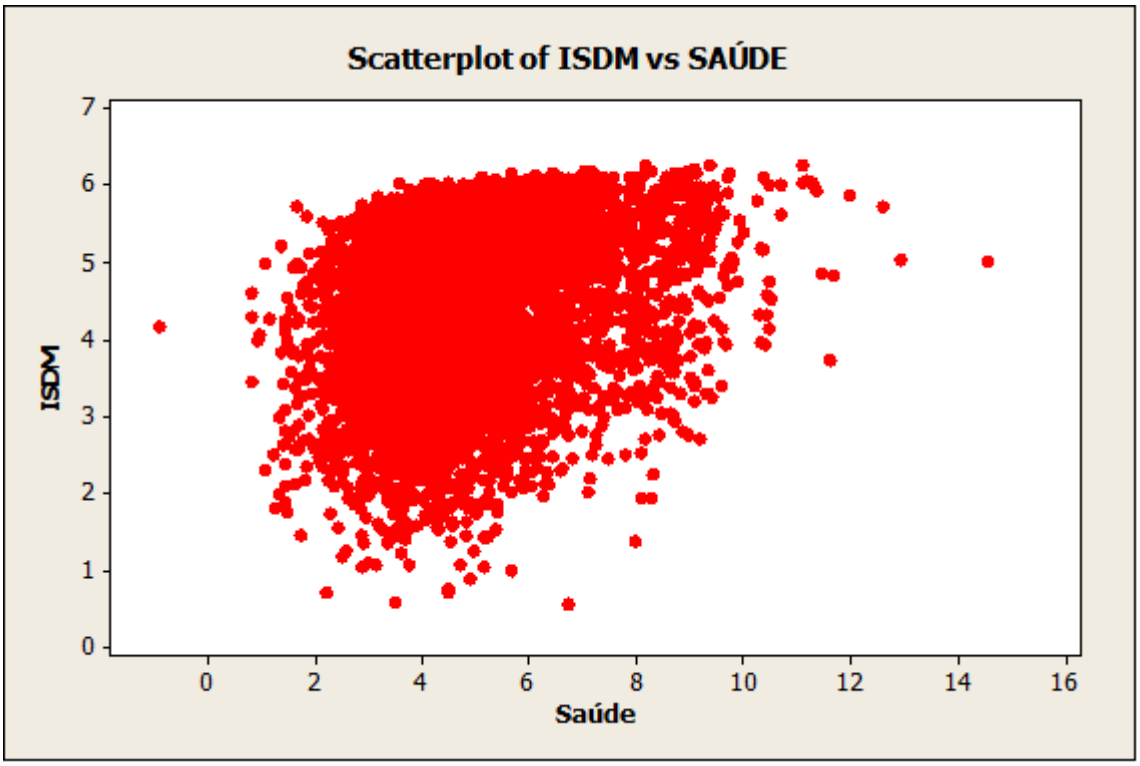
3.3 RELAÇÃO ENTRE VARIÁVEIS: CORRELAÇÃO, REGRESSÃO E TESTE QUI-QUADRADO

Gráficos de dispersão devem ser inicialmente analisados quanto a seu padrão geral e seus desvios relativos ao padrão. A descrição do padrão geral pode ser feita pela verificação de sua forma, direção e intensidade.

3.3.1 GRÁFICOS DE DISPERSÃO entre variáveis Saúde e ISDM

GRAPH >> SCATTERPLOT >> SIMPLE

A quantidade de dados analisados é muito grande, são 5565 municípios, o que causa uma "mancha" no gráfico e dificulta a visualização. Uma forma de contornar esta situação seria selecionar os dados por amostragem, mas neste caso não é aplicado, pois não existem critérios específicos que garantiriam a fidelidade da amostra em relação à população.



Gráficos de dispersão devem ser inicialmente analisados quanto a seu padrão geral e seus desvios relativos ao padrão. A descrição do padrão geral pode ser feita pela verificação de sua forma, direção e intensidade.

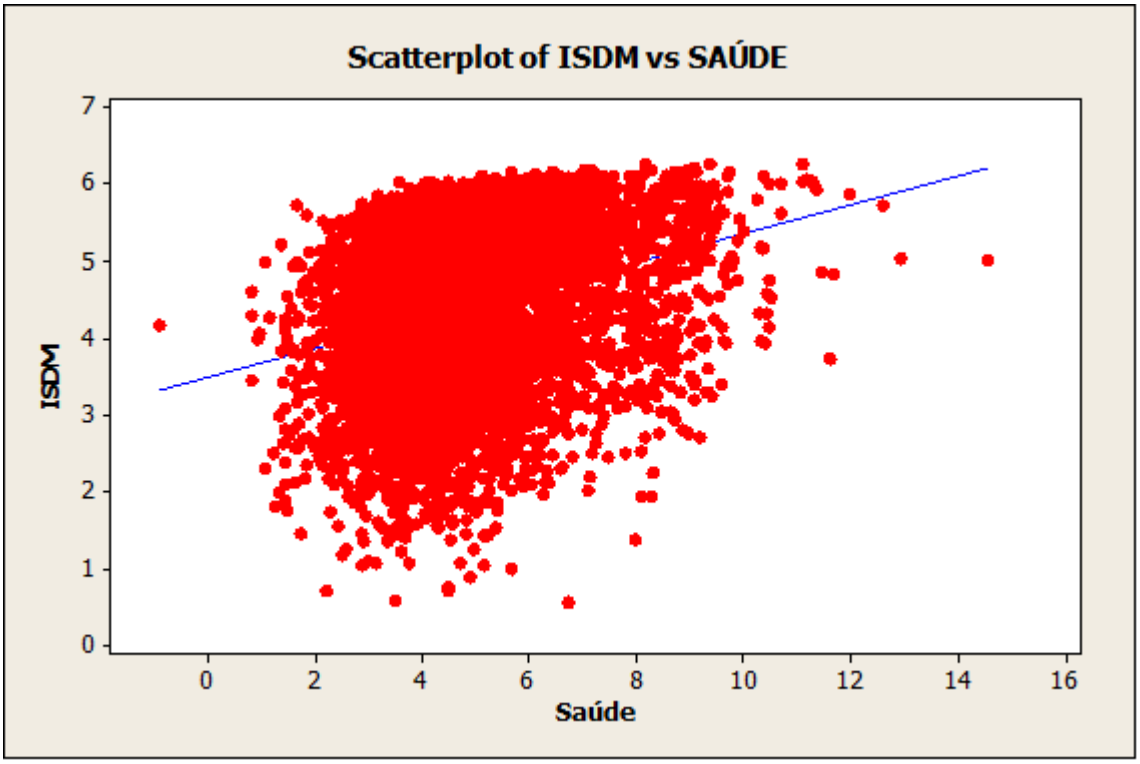
Direção: Da análise das correlações acima percebemos que quase todas possuem associações bastante neutras/lineares, ou seja, o crescimento de uma variável não é obrigatoriamente acompanhado do crescimento da outra. Contudo, parece que não há nenhuma associação negativa, ao menos de evidência visual.

Intensidade: O gráfico acima parece indicar a existência de relações lineares, embora a característica de uma reta seja constante na imagem.

Forma: O gráfico apresenta conglomerados que sugerem relações lineares, embora prejudicado pelo excesso de dados da população (5565 linhas).

3.3.2 LINHAS DE TENDÊNCIAS entre Educação e Emprego e Renda

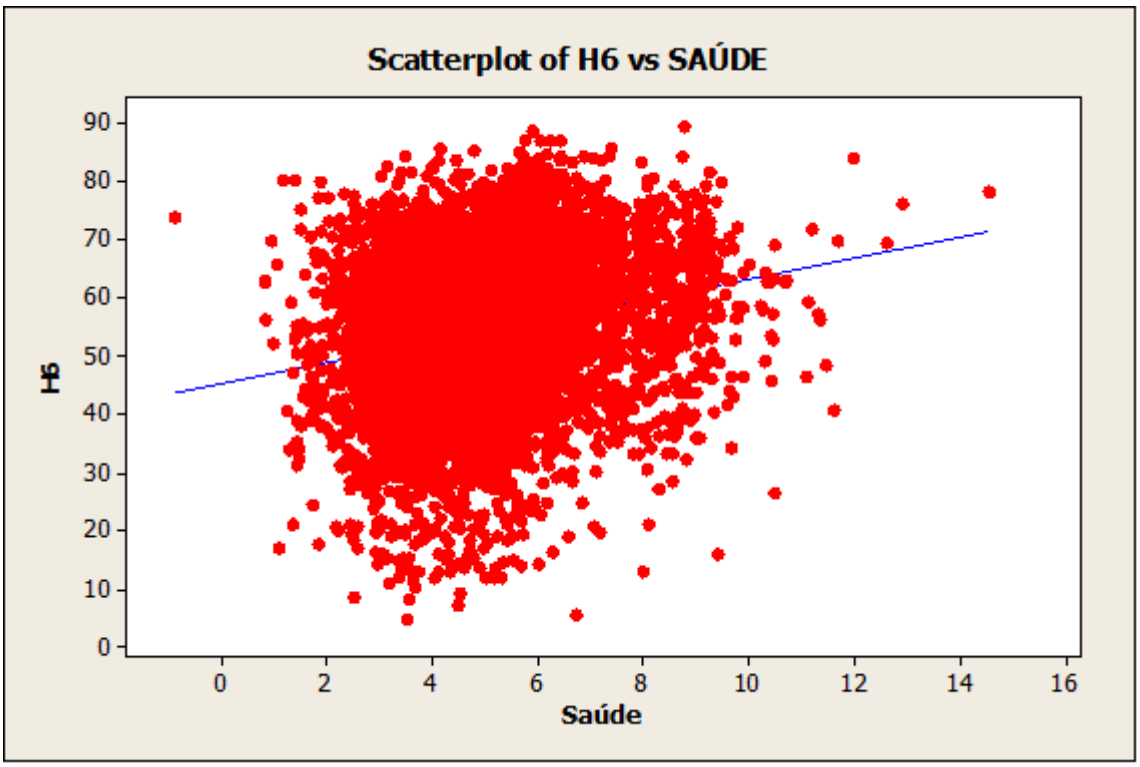
GRAPH >> SCATTERPLOT >> WITH REGRESSION



Para se verificar qual o tipo de relação (linear, quadrática, cúbica, exponencial, etc.) existente entre as variáveis, adicionamos em cada gráfico de dispersão uma linha de tendência.

O gráfico analisado neste caso contém a variável Saúde em relação ISDM. Podemos afirmar que os pontos estão muito próximos da linha e são ligeiramente lineares, o que nos aponta que o tipo de relação entre as variáveis é linear, embora existam valores atípicos distribuídos por toda a extensão da reta.

3.3.3 LINHAS DE TENDÊNCIAS entre Saúde e H6 (Proporção de pessoas que vivem em domicílio que tem densidade de moradores por dormitório inferior a 2)



Este gráfico compara a tendência entre as variáveis Saúde e H6. Se compararmos com o gráfico anterior, podemos constatar que a "nuvem de pontos" continua relativamente linear, apesar de demonstrar tendência crescente. Pode-se concluir que quando aumenta o índice de Saúde há um desempenho positivo do ISDM e da questão da habitação.

3.3.4 CORRELAÇÃO LINEAR

A matriz de correlação inclui o teste de significância p-value. Para a correlação foi utilizado o índice de Pearson. Vale ressaltar que o índice de correlação entre as variáveis não requer que exista uma relação de causa-efeito entre ambas.

Esta primeira visão exibe a correlação entre todas as variáveis utilizadas no trabalho.

STAT >> BASIC STATISTICS >> CORRELATION

Correlations: ISDM; EMP & REN; IFGF; Liquidez; H6; R1; T1_2; S; S1_1; E; E2_4

	ISDM	EMP & REN	IFGF	Liquidez	H6	R1
EMP & REN	0,815					
	0,000					
IFGF	0,420	0,446				
	0,000	0,000				
Liquidez	0,258	0,261	0,760			
	0,000	0,000	0,000			
H6	0,695	0,522	0,327	0,244		
	0,000	0,000	0,000	0,000		
R1	-0,951	-0,801	-0,455	-0,293	-0,709	
	0,000	0,000	0,000	0,000	0,000	
T1_2	0,806	0,737	0,430	0,291	0,449	-0,781
	0,000	0,000	0,000	0,000	0,000	0,000
S	0,286	0,205	0,106	0,069	0,220	-0,195
	0,000	0,000	0,000	0,000	0,000	0,000
S1_1	-0,147	-0,182	-0,066	-0,044	-0,115	0,140
	0,000	0,000	0,000	0,001	0,000	0,000
E	0,884	0,739	0,456	0,289	0,722	-0,868
	0,000	0,000	0,000	0,000	0,000	0,000
E2_4	0,764	0,705	0,419	0,244	0,613	-0,768
	0,000	0,000	0,000	0,000	0,000	0,000

	T1_2	S	S1_1	E
S	0,137			
	0,000			
S1_1	-0,112	-0,196		
	0,000	0,000		
E	0,664	0,215	-0,131	
	0,000	0,000	0,000	
E2_4	0,599	0,194	-0,128	0,811
	0,000	0,000	0,000	0,000

A correlação é sempre um número entre zero e um e mede a intensidade de relações lineares. A correlação entre as variáveis analisadas é positiva em alguns casos e negativa em outros, mas de fraca intensidade, com exceção da correlação entre Renda (R1) e ISDM. Os valores mais representativos estão marcado com verde quando positivos e vermelhos quando negativos. Indica que a correlação entre estas variáveis é mais intensa. Portanto, podemos afirmar que estas variáveis possuem relações lineares.

3.3.5 REGRESSÃO DE MÍNIMOS QUADRADOS

A correlação mede a direção e a intensidade da relação linear (linha reta) entre duas variáveis quantitativas. Se um diagrama de dispersão mostra uma relação linear, é interessante resumirmos esse padrão geral traçando uma reta no diagrama de dispersão. Uma reta de regressão resume a relação entre duas variáveis, mas somente em um contexto específico: quando uma das variáveis ajuda a explicar ou a predizer a outra, ou seja, a regressão descreve uma relação entre uma variável explanatória e uma variável resposta.

Abaixo, está o resultado da regressão entre as variáveis Saúde e ISDM.

Regression Analysis: S versus ISDM

The regression equation is
S = 3,08 + 0,437 ISDM

Predictor	Coef	SE Coef	T	P
Constant	3,07899	0,08960	34,37	0,000
ISDM	0,43741	0,01963	22,29	0,000

S = 1,59986 R-Sq = 8,2% R-Sq(adj) = 8,2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1271,3	1271,3	496,68	0,000
Residual Error	5562	14236,2	2,6		
Total	5563	15507,5			

A tabela acima exibe o resultado da fórmula entre as variáveis Saúde e ISDM. Se substituísse o valor de Saúde se chegaria ao valor do ISDM esperado. A é a expressão numérica da reta de tendência que vimos nos itens acima. Esta equação tem um poder explicativo de 89,6%, que é o R-Quadrado. O valor da constante 3,08 significa que, se o ISDM fosse zero, o valor da Saúde seria 3,08.

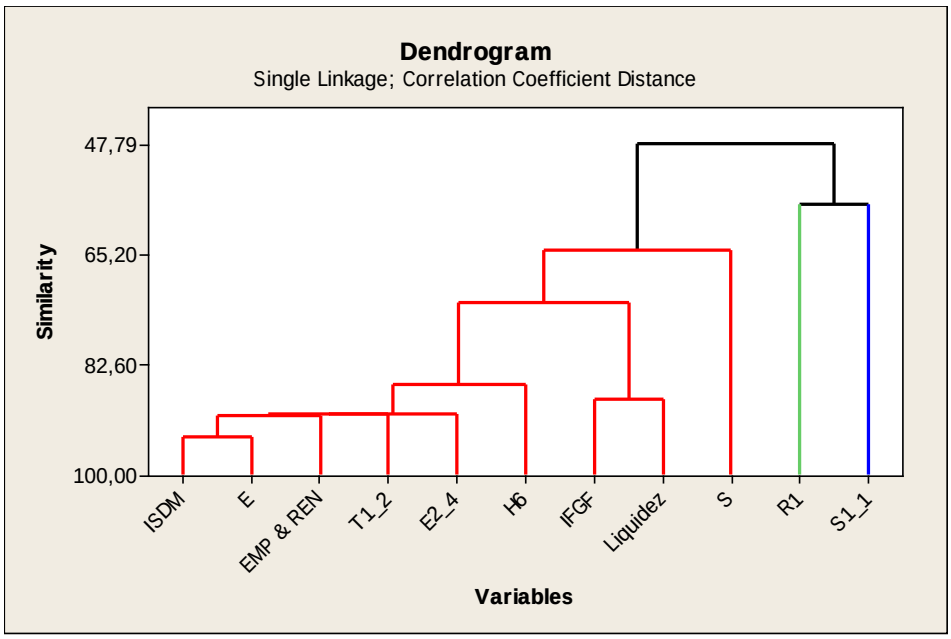
3.3.6 DENDROGRAMA

Um Dendrograma (dendr(o) = árvore) é um tipo específico de diagrama ou representação icônica que organiza determinados fatores e variáveis. É um diagrama de similaridade.

A interpretação de um dendrograma de similaridade entre amostras fundamenta-se na intuição: duas amostras próximas devem ter também valores semelhantes para as variáveis medidas. Ou seja, elas devem ser próximas matematicamente no espaço multidimensional. Portanto, quanto maior a proximidade entre as medidas relativas às amostras, maior a similaridade entre elas. O dendrograma hierarquiza esta similaridade de modo que podemos ter uma visão bidimensional da similaridade ou dissimilaridade de todo o conjunto de amostras utilizado no estudo.

Segue abaixo o Dendrograma das variáveis analisadas:

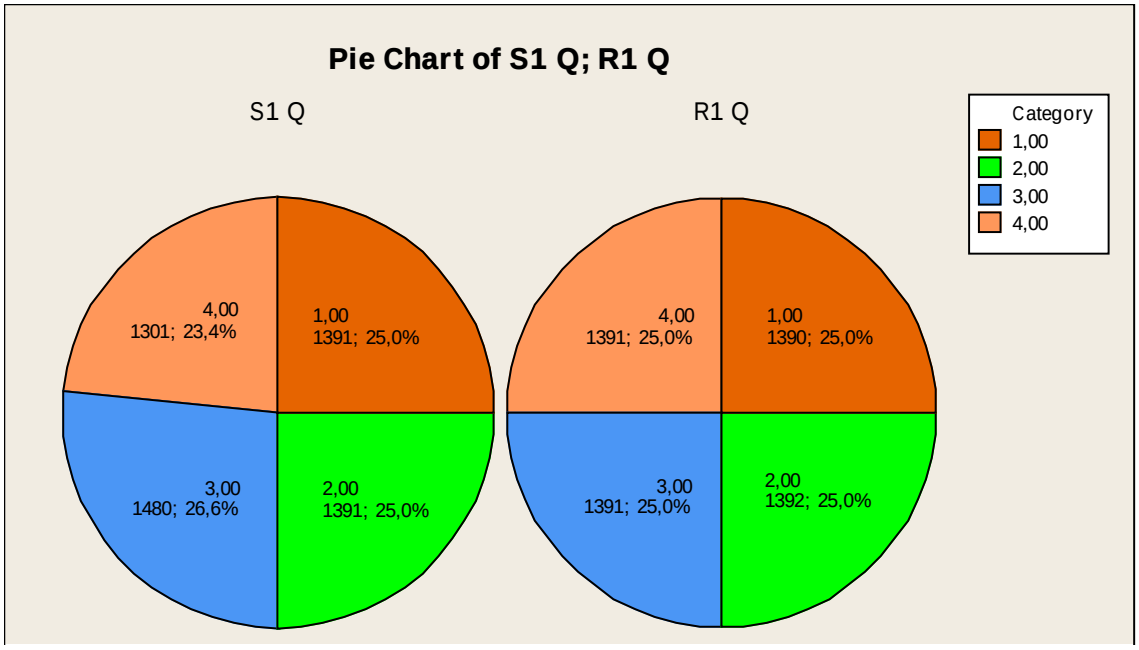
STAT >> MULTIVARIATE >> CLUSTER VARIABLE



As variáveis ISDM e E são as que possuem o maior nível de similaridade, por volta de 94%. As variáveis (Emprego & Renda, T1_2 e E2_4) também são muito similares, variando até 85%. Com menor nível de similaridade estão as variáveis H6, IFGF e Liquidez. Já as Saúde, R1 e S1_1 encontram-se com baixo nível de similaridade.

3.3.7 RELAÇÕES ENTRE AS VARIÁVEIS CATEGÓRICAS

Para correlacionarmos duas variáveis categóricas, criamos duas colunas categorizadas com informações sobre dados de S1_1 por quartil, tendo valores que variam de 1 a 4, e repetimos o processo para a variável R1, com os mesmos valores categóricos.



Os gráficos acima foram gerados a partir das informações dos quartis utilizando a função Data >> Code >> Numeric to Numeric e colocando os intervalos para geração das variáveis categóricas. Em seguida foi gerado um Pie Chart através da função Graph >> Pie Chart.

Para analisar a semelhança entre as variáveis categóricas será executada a tabulação cruzada entre elas.

STAT >> TABLES >> CROSS TABULATION AND CHI SQUARE

Tabulated statistics: R1 Q; S1 Q

Rows: R1 Q Columns: S1 Q

	1,00	2,00	3,00	4,00	333,33	All
1,00	465	428	287	210	0	1390
2,00	436	302	365	289	0	1392
3,00	313	307	400	370	1	1391
4,00	177	354	428	432	0	1391
All	1391	1391	1480	1301	1	5564

Cell Contents: Count

Pearson Chi-Square = 298,009; DF = 12
Likelihood Ratio Chi-Square = 313,706; DF = 12

As linhas são representadas por R1 e as colunas por S1_1. Os dados aparecem distribuídos entre cada quartil de uma variável.

3.4 MODELOS DE REGRESSÃO LINEAR MULTIPLOS

3. 4.1 CORRELAÇÃO LINEAR, ANÁLISE DE REGRAÇÃO E STEPWISE

Para o estudo em questão, queremos entender quais variáveis explicam melhor a variável específica. Para tanto utilizaremos o grupo das variáveis analíticas e sintética, comparando com a variável Saúde.

Inicialmente serão analisadas as correlações lineares entre a variável SAÚDE com as variáveis analíticas e sintéticas, relacionadas a este estudo, para verificar quais variáveis melhor explicam a SAÚDE.

Correlations: ISDM; EMP & REN; IFGF; Liquidez; H6; R1; T1_2; S; S1_1; E; E2_4

	ISDM	EMP & REN	IFGF	Liquidez	H6	R1
EMP & REN	**0,815**					
	0,000					
IFGF	0,420	0,446				
	0,000	0,000				
Liquidez	0,258	0,261	**0,760**			
	0,000	0,000	0,000			
H6	0,695	0,522	0,327	0,244		
	0,000	0,000	0,000	0,000		
R1	-0,951	-0,801	-0,455	-0,293	-0,709	
	0,000	0,000	0,000	0,000	0,000	
T1_2	**0,806**	**0,737**	0,430	0,291	0,449	-0,781
	0,000	0,000	0,000	0,000	0,000	0,000
S	0,286	0,205	0,106	0,069	0,220	-0,195
	0,000	0,000	0,000	0,000	0,000	0,000
S1_1	-0,147	-0,182	-0,066	-0,044	-0,115	0,140
	0,000	0,000	0,000	0,001	0,000	0,000
E	**0,884**	**0,739**	0,456	0,289	0,722	-0,868
	0,000	0,000	0,000	0,000	0,000	0,000
E2_4	**0,764**	**0,705**	0,419	0,244	0,613	-0,768
	0,000	0,000	0,000	0,000	0,000	0,000

	T1_2	S	S1_1	E
S	0,137			
	0,000			
S1_1	-0,112	-0,196		
	0,000	0,000		
E	0,664	0,215	-0,131	
	0,000	0,000	0,000	
E2_4	0,599	0,194	-0,128	0,811
	0,000	0,000	0,000	0,000

As correlações significativas de acordo com o P-Value, para este trabalho, será considerada significativa quando >= 0,70). No geral, Saúde em fraca correlação com todas as demais variáveis. Apresentam uma correlação satisfatoriamente forte entre si as variáveis: ISDM e Emprego & Renda com Trabalho (T1_2), Educação (E) e E2_4 (Proporção de crianças de 7 a 14 anos na série adequada para sua idade). ISM e Emprego & Renda também tem forte correlação, assim como Liquidez e IFGF.

3.4.2 REGRESSÃO: SAÚDE COM DEMAIS VARIÁVEIS DO ESTUDO

Regression Analysis: S versus ISDM; EMP & REN; ...

```
The regression equation is
S = - 4,73 + 2,00 ISDM - 0,069 EMP & REN + 0,722 IFGF + 0,0673 Liquidez
    + 0,0125 H6 + 0,0762 R1 - 0,0225 T1_2 - 0,0179 S1_1 - 0,330 E + 0,0115 E2_4


5261 cases used, 303 cases contain missing values


Predictor       Coef   SE Coef       T      P
Constant     -4,7318    0,5069   -9,34  0,000
ISDM         2,00326   0,07398   27,08  0,000
EMP & REN    -0,0695    0,4036   -0,17  0,863
IFGF          0,7219    0,2351    3,07  0,002
Liquidez     0,06729   0,08620    0,78  0,435
H6          0,012469  0,002383    5,23  0,000
R1          0,076160  0,004051   18,80  0,000
T1_2       -0,022524  0,002102  -10,71  0,000
S1_1       -0,017928  0,001466  -12,23  0,000
E           -0,32967   0,04362   -7,56  0,000
E2_4        0,011544  0,004223    2,73  0,006


S = 1,49408   R-Sq = 20,1%   R-Sq(adj) = 20,0%


Analysis of Variance

Source            DF        SS      MS       F      P
Regression        10   2949,04  294,90  132,11  0,000
Residual Error  5250  11719,37    2,23
Total           5260  14668,41
```

O R-Square é baixo = 20,1% e todos os valores Betas da equação apresentam valores próximos a zero, com exceção do ISDM que indica alto poder explicativo da variável Saúde. O P-value das variáveis possui valor baixo, sendo confiáveis para a explicação da variável Saúde. A exceção são as variáveis IFGF e Liquidez.

3.4.3 STEPWISE DA SAÚDE COM FILTRO DOS RESULTADOS OBTIDOS

A análise STEPWISE demonstra o percentual de composição das variáveis *Predictors* na equação da *Response.*

Stepwise Regression: S versus ISDM; EMP & REN; ...

Alpha-to-Enter: 0,15 Alpha-to-Remove: 0,15

Response is S on 10 predictors, with N = 5261
N(cases with missing observations) = 303 N(all cases) = 5564

Step	1	2	3	4	5	6
Constant	3,004	-3,932	-3,507	-3,300	-2,866	-3,395
ISDM	0,453	1,605	1,568	1,806	1,948	2,020
T-Value	22,07	25,36	25,11	27,55	27,15	27,77
P-Value	0,000	0,000	0,000	0,000	0,000	0,000
R1		0,0746	0,0746	0,0714	0,0672	0,0703
T-Value		19,17	19,45	18,75	17,24	17,91
P-Value		0,000	0,000	0,000	0,000	0,000
S1_1			-0,0184	-0,0181	-0,0181	-0,0181
T-Value			-12,41	-12,36	-12,38	-12,40
P-Value			0,000	0,000	0,000	0,000
T1_2				-0,0210	-0,0230	-0,0251
T-Value				-10,84	-11,61	-12,50
P-Value				0,000	0,000	0,000
E					-0,185	-0,231
T-Value					-4,84	-5,90
P-Value					0,000	0,000
IFGF						0,88
T-Value						5,49
P-Value						0,000
S	1,60	1,54	1,52	1,51	1,50	1,50
R-Sq	8,48	14,46	16,90	18,71	19,07	19,54
R-Sq(adj)	8,46	14,43	16,85	18,65	19,00	19,44
Mallows Cp	756,9	366,0	207,9	90,4	68,7	40,4

O Próximo passo é calcular a fórmula utilizando as variáveis analíticas e sintéticas demonstradas pela função Stepwise como sendo as que mais explicam a Saúde.

STAT >> REGRESSION >> REGRESSION

A fórmula resultante é:

```
S = - 4,73 + 2,00 ISDM - 0,069 EMP & REN + 0,722 IFGF + 0,0673 Liquidez +
0,0125 H6 + 0,0762 R1 - 0,0225 T1_2 - 0,0179 S1_1 - 0,330 E + 0,0115 E2_4
```

Nesta equação foram utilizadas as variáveis analíticas e sintéticas. Uma outra forma de se fazer este estudo seria isolar um primeiro grupo de cálculo utilizando apenas as variáveis analíticas e um segundo grupo com as variáveis sintéticas.

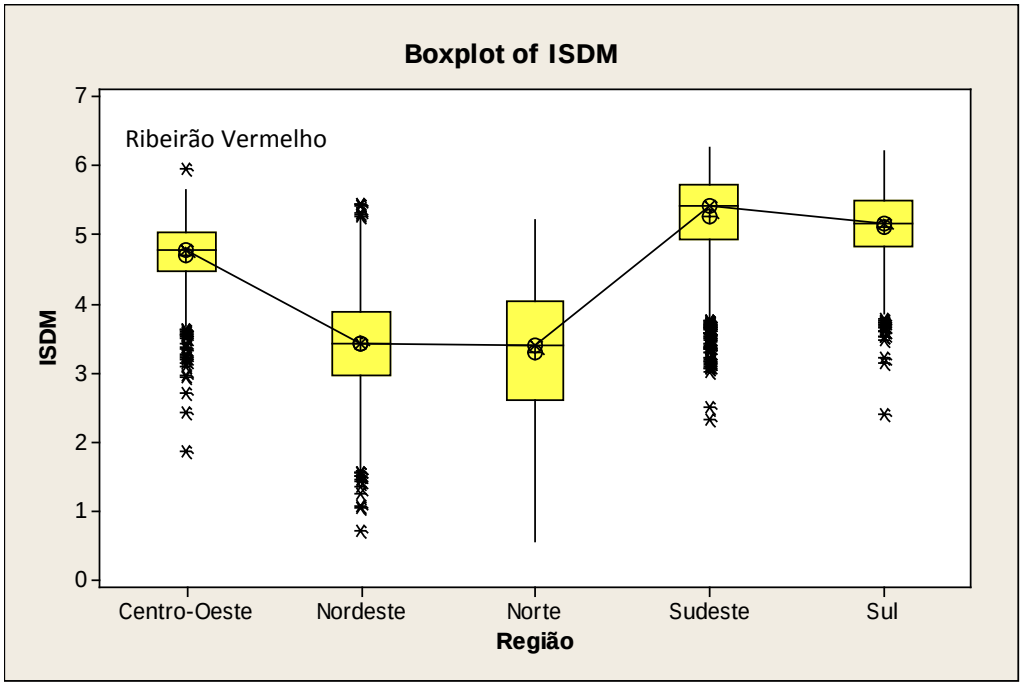
3.5 COMPARAÇÕES

O teste de hipótese nos permite comparar parâmetros de populações distintas de forma a fazermos inferências estatísticas sobre estas populações. Essencialmente as comparações realizadas nos testes de hipóteses se valem de testar uma hipótese nula (H0)e uma hipótese alternativa (H1) estabelecendo-se um grau de confiança em relação a se aceitar ou rejeitar as hipóteses estabelecidas.

Há dois tipos de abordagem para a realização dos testes de hipóteses: a do intervalo de confiança na qual se faz o teste objetivando verificar a pertinência de um parâmetro em um intervalo de valores com certa probabilidade de acerto e a do teste de significância que leva em consideração a probabilidade de cometer-se um erro do tipo I (rejeitar a hipótese nula quando ela é verdadeira).

Este trabalho propõe a comparação das médias entre as diversas regiões do Brasil, de acordo com as variáveis deste estudo. O objetivo é comparar a média dos indicadores e realizar testes de hipóteses das cidades com maiores índices de desenvolvimento.

3.5.1 – Variável ISDM por Região



A Região Sudeste possui o maior ISDM do país, o que indica que esta é a Região mais desenvolvida do Brasil, segundo a pesquisa. A região Sul encontra-se próxima a Região Sudeste, e ocupa o segundo lugar.

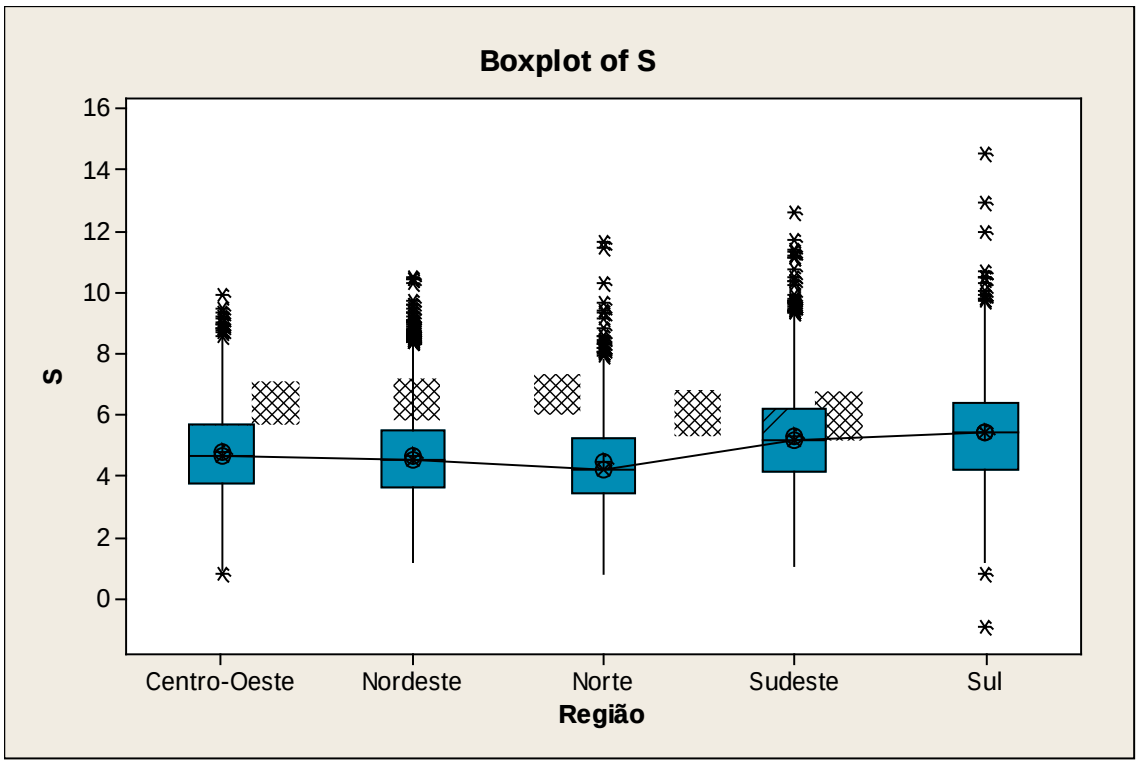
A Região que apresenta o ISDM médio mais baixo do País é a Norte, seguida bem próxima da Nordeste. Pelo tamanho da caixa do BloxPlot podemos visualizar a amplitude da variância. Podemos afirmar que os dados da Região Norte possuem maior variabilidade que os dados das demais regiões. As regiões que possuem menor variabilidade dos dados são Centro-Oeste e Sul.

One-way ANOVA: ISDM versus Região

```
Source    DF        SS        MS        F      P
Região     4  4187,835  1046,959  2369,23  0,000
Error   5559  2456,518     0,442
Total   5563  6644,353

S = 0,6648   R-Sq = 63,03%   R-Sq(adj) = 63,00%


                                    Individual 95% CIs For Mean Based on
                                    Pooled StDev
Level           N    Mean   StDev  ------+---------+---------+---------+---
Centro-Oeste  467  4,6934  0,5237                        (*)
Nordeste     1790  3,4112  0,7136    (*
Norte         447  3,3035  0,9794  (*)
Sudeste      1669  5,2606  0,6436                                  (*
Sul          1191  5,1271  0,5012                                *)
                                    ------+---------+---------+---------+---
                                        3,60      4,20      4,80      5,40

Pooled StDev = 0,6648
```

O grau de variação entre as Regiões é muito alto (2369), e o P-value nos indica que a informação é confiável e não existe chance deste valor ser diferente.

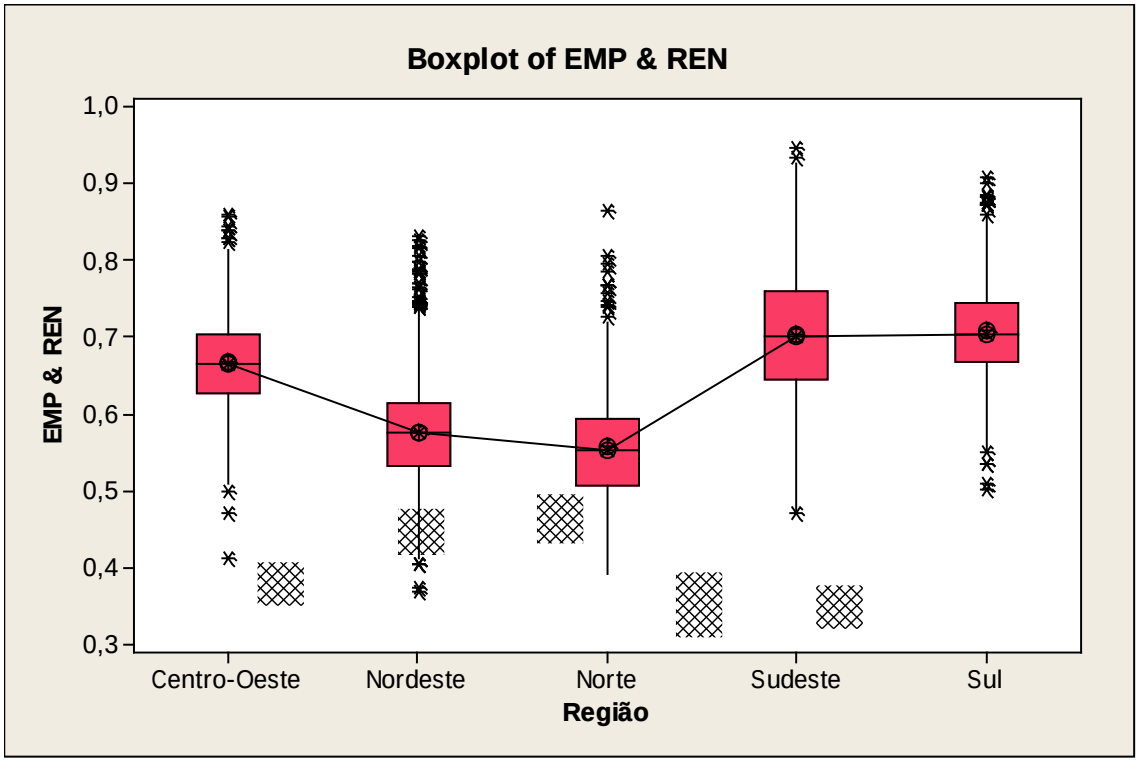
3.5.2 Variável SAÚDE por Região



One-way ANOVA: S versus Região

```
Source    DF        SS      MS      F      P
Região     4    754,22  188,56  71,05  0,000
Error   5559  14753,25    2,65
Total   5563  15507,47

S = 1,629   R-Sq = 4,86%   R-Sq(adj) = 4,80%


                                  Individual 95% CIs For Mean Based on
                                  Pooled StDev
Level            N   Mean  StDev  --------+---------+---------+---------+-
Centro-Oeste   467  4,781  1,557           (----*---)
Nordeste      1790  4,672  1,534         (-*--)
Norte          447  4,425  1,603  (---*----)
Sudeste       1669  5,309  1,671                              (--*-)
Sul           1191  5,444  1,741                                 (--*-)
                                  --------+---------+---------+---------+-
                                        4,55      4,90      5,25      5,60

Pooled StDev = 1,629
```

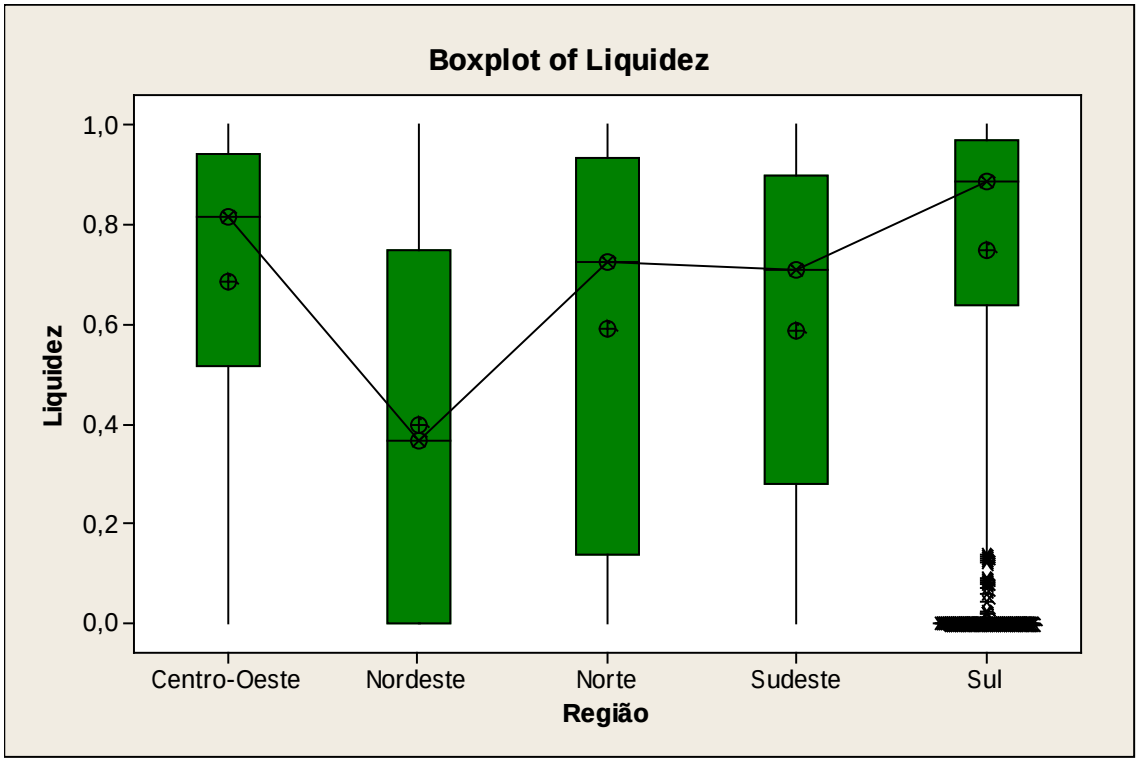
3.5.3 Variável EMPREGO E RENDA por Região



One-way ANOVA: EMP & REN versus Região

```
Source    DF        SS       MS        F      P
Região     4  22,72829  5,68207  1115,06  0,000
Error   5559  28,32730  0,00510
Total   5563  51,05559

S = 0,07138   R-Sq = 44,52%   R-Sq(adj) = 44,48%



Level            N     Mean    StDev
Centro-Oeste   467   0,66727  0,06610
Nordeste      1790   0,57598  0,06481
Norte          447   0,55717  0,07128
Sudeste       1669   0,70449  0,08502
Sul           1191   0,70781  0,06118

               Individual 95% CIs For Mean Based on Pooled StDev
Level            +---------+---------+---------+---------
Centro-Oeste                          (*-)
Nordeste              *)
Norte            (*-)
Sudeste                                        (*)
Sul                                             (*
                 +---------+---------+---------+---------
               0,550     0,600     0,650     0,700

Pooled StDev = 0,07138
```

3.5.4 Variável LIQUIDEZ por Região



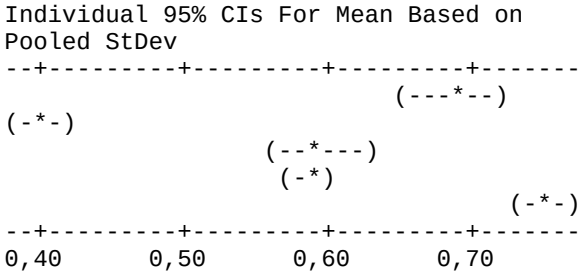
One-way ANOVA: Liquidez versus Região

Source	DF	SS	MS	F	P
Região	4	92,336	23,084	189,05	0,000
Error	5256	641,771	0,122		
Total	5260	734,107			

S = 0,3494 R-Sq = 12,58% R-Sq(adj) = 12,51%

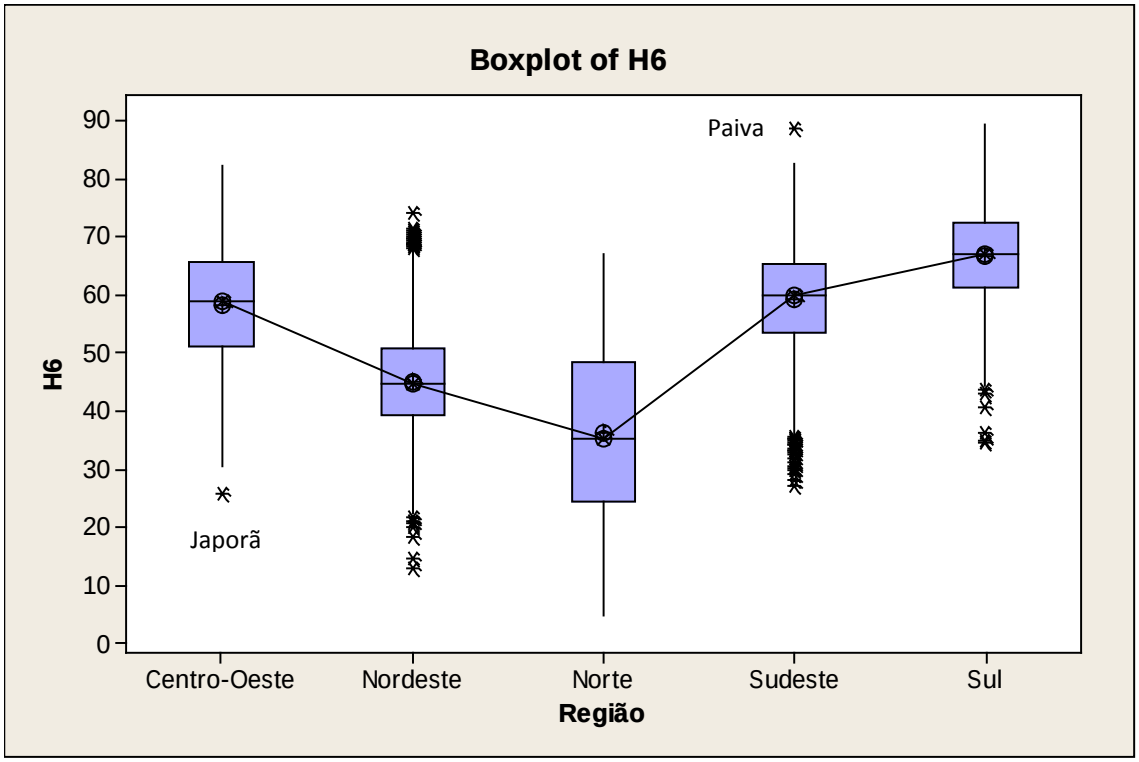
Level	N	Mean	StDev
Centro-Oeste	438	0,6856	0,3302
Nordeste	1650	**0,3985**	0,3696
Norte	391	0,5909	0,3864
Sudeste	1609	0,5861	0,3552
Sul	1173	**0,7486**	0,3031

Individual 95% CIs For Mean Based on Pooled StDev



Pooled StDev = 0,3494

3.5.5 Variável H6 por Região



One-way ANOVA: H6 versus Região

Source	DF	SS	MS	F	P
Região	4	525704	131426	1472,10	0,000
Error	5559	496296	89		
Total	5563	1022001			

S = 9,449 R-Sq = 51,44% R-Sq(adj) = 51,40%

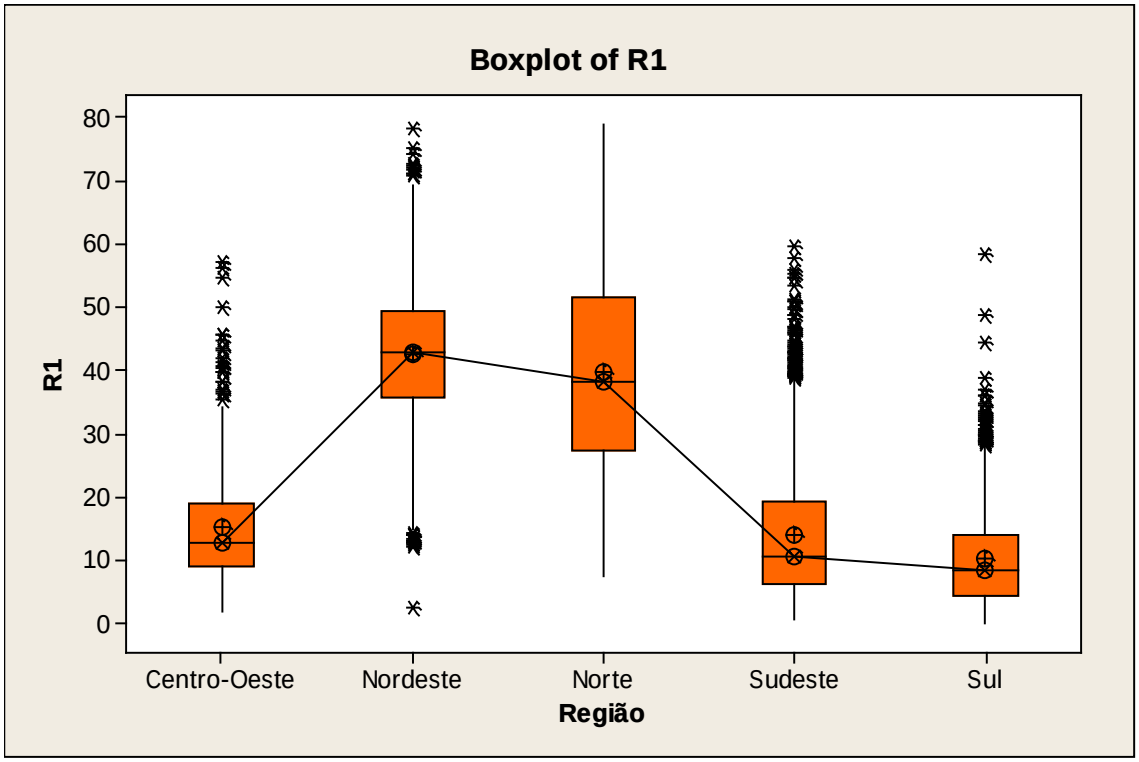
Level	N	Mean	StDev
Centro-Oeste	467	58,217	9,709
Nordeste	1790	45,127	8,829
Norte	447	36,253	14,618
Sudeste	1669	59,053	9,008
Sul	1191	66,750	8,314

Individual 95% CIs For Mean Based on Pooled StDev

```
------+---------+---------+---------+---
                                (*)
              *)
(*)
                                 (*
                                        *)
------+---------+---------+---------+---
   40,0      48,0      56,0      64,0
```

Pooled StDev = 9,449

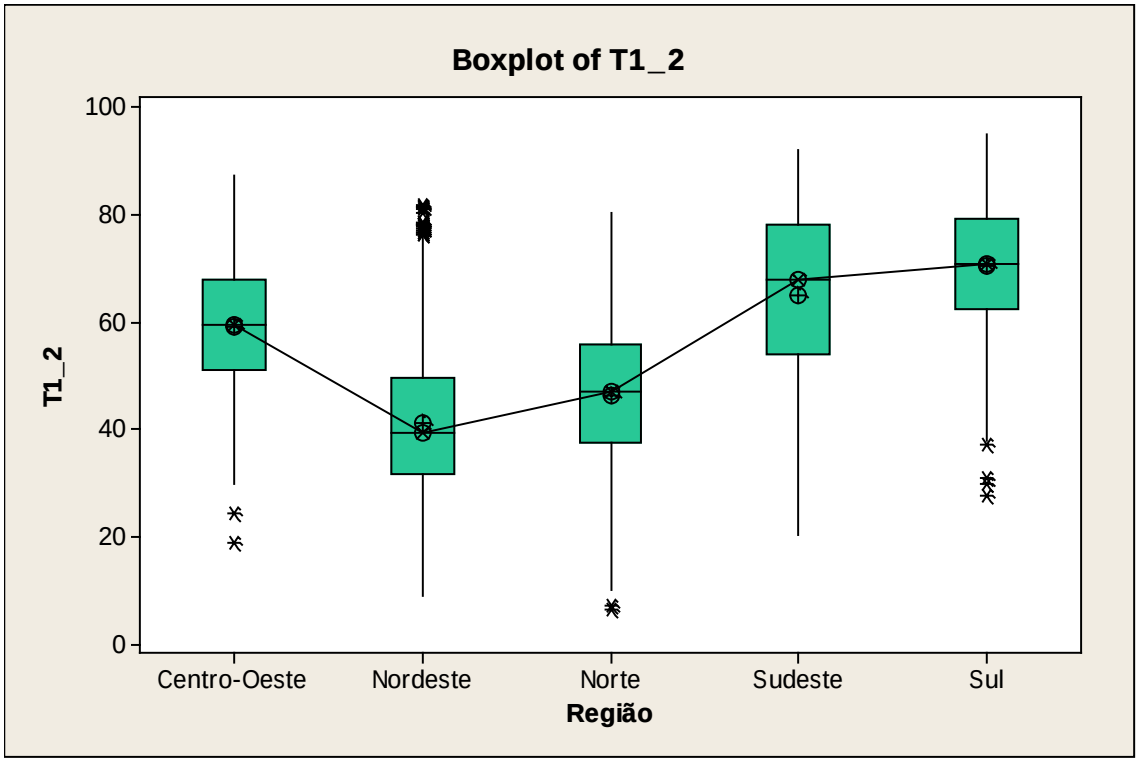
3.5.6 Variável R1 por Região



One-way ANOVA: R1 versus Região

```
Source    DF       SS      MS        F      P
Região     4  1148531  287133  2687,59  0,000
Error   5559   593903     107
Total   5563  1742434

S = 10,34   R-Sq = 65,92%   R-Sq(adj) = 65,89%


                                  Individual 95% CIs For Mean Based on
                                  Pooled StDev
Level           N   Mean  StDev   +---------+---------+---------+---------
Centro-Oeste  467  15,25   9,23       (*)
Nordeste     1790  42,49  10,60                                        *)
Norte         447  39,65  15,18                                   (*)
Sudeste      1669  14,07  10,54       *)
Sul          1191  10,16   7,40   *)
                                   +---------+---------+---------+---------
                                  10        20        30        40

Pooled StDev = 10,34
```

3.5.7 Variável T1_2 por Região



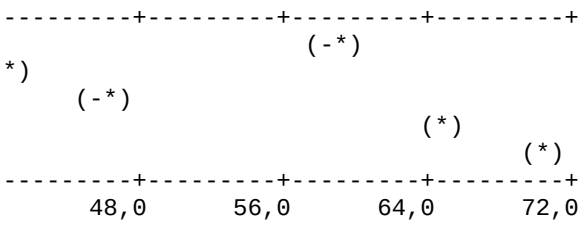
One-way ANOVA: T1_2 versus Região

Source	DF	SS	MS	F	P
Região	4	804370	201092	1069,83	0,000
Error	5559	1044907	188		
Total	5563	1849277			

S = 13,71 R-Sq = 43,50% R-Sq(adj) = 43,46%

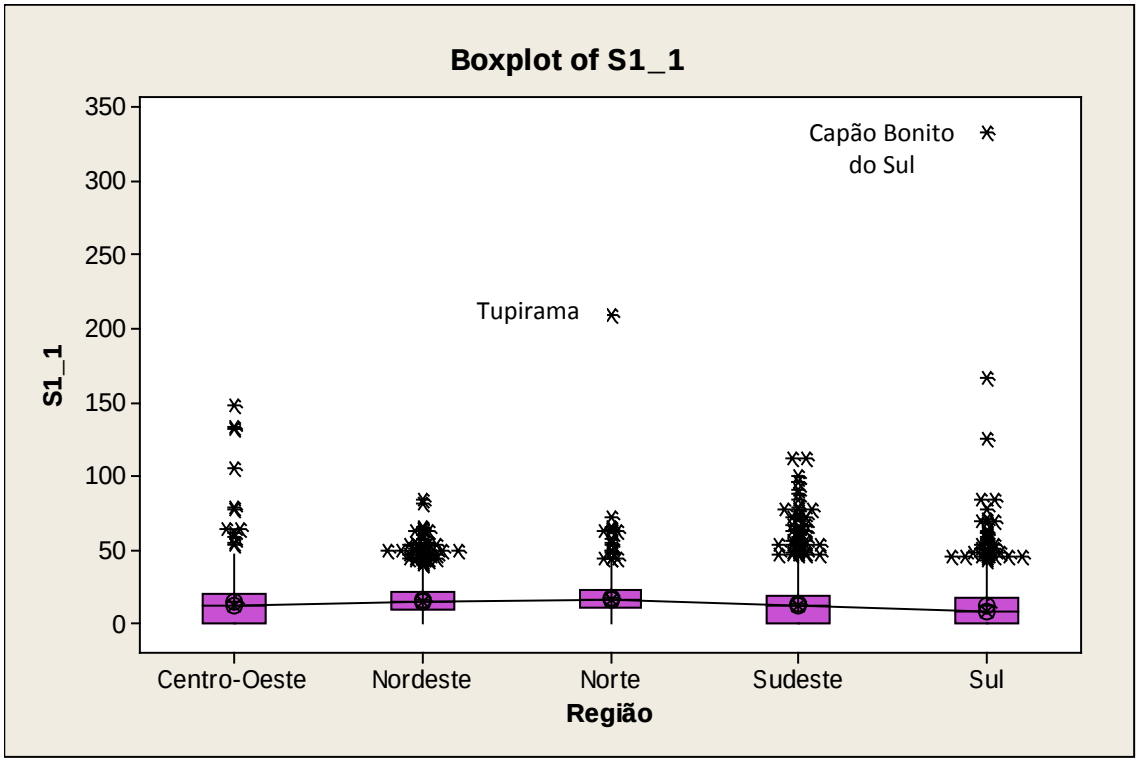
Level	N	Mean	StDev
Centro-Oeste	467	58,85	11,80
Nordeste	1790	41,17	13,36
Norte	447	46,10	13,65
Sudeste	1669	64,65	15,77
Sul	1191	70,16	11,69

Individual 95% CIs For Mean Based on Pooled StDev



Pooled StDev = 13,71

3.5.8 Variável S1_1 por Região



One-way ANOVA: S1_1 versus Região

Source	DF	SS	MS	F	P
Região	4	19369	4842	24,13	0,000
Error	5559	1115522	201		
Total	5563	1134891			

S = 14,17 R-Sq = 1,71% R-Sq(adj) = 1,64%

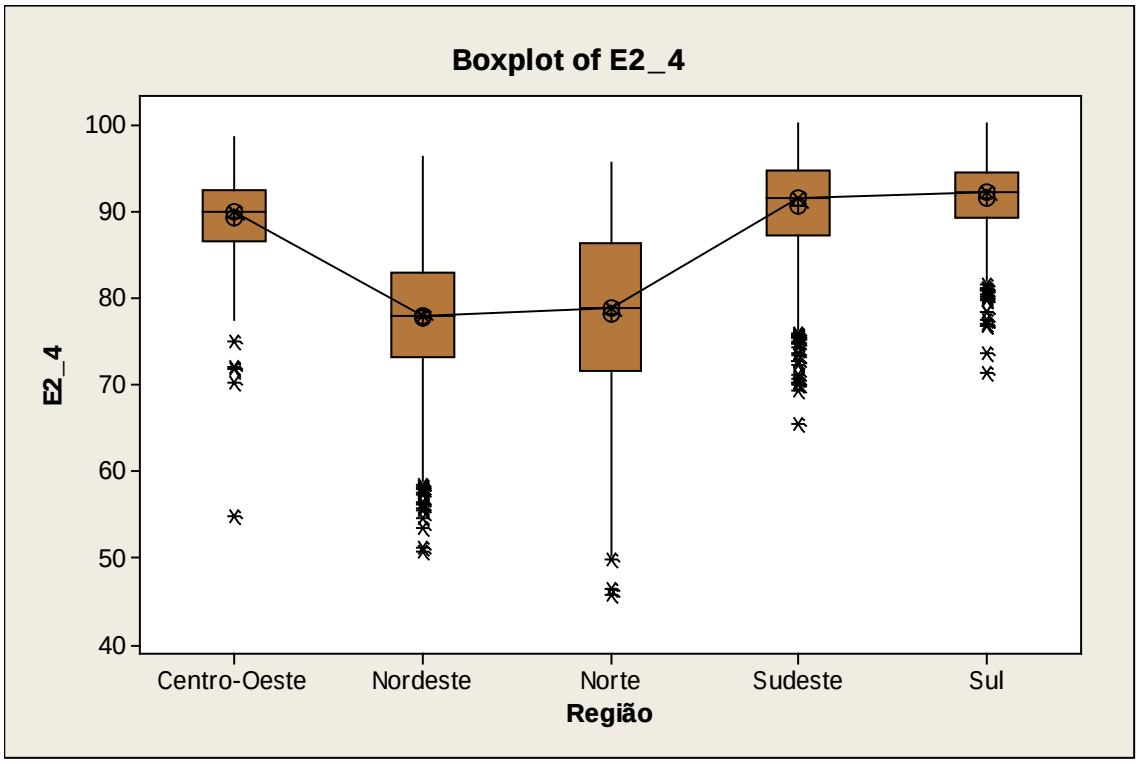
Level	N	Mean	StDev
Centro-Oeste	467	14,20	16,93
Nordeste	1790	15,80	10,90
Norte	447	17,82	15,13
Sudeste	1669	13,60	13,91
Sul	1191	11,56	17,01

Individual 95% CIs For Mean Based on Pooled StDev

```
-------+---------+---------+---------+--
         (----*----)
                   (-*--)
                        (----*-----)
         (-*--)
(--*--)
-------+---------+---------+---------+--
    12,5      15,0      17,5      20,0
```

Pooled StDev = 14,17

3.5.9 Variável E2_4 por Região



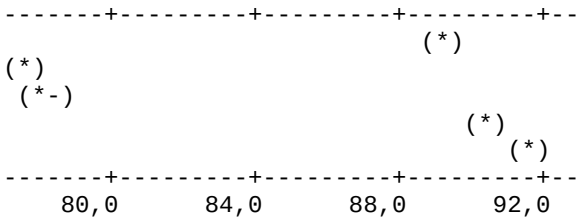
One-way ANOVA: E2_4 versus Região

Source	DF	SS	MS	F	P
Região	4	226215,9	56554,0	1478,42	0,000
Error	5559	212648,5	38,3		
Total	5563	438864,4			

S = 6,185 R-Sq = 51,55% R-Sq(adj) = 51,51%

Level	N	Mean	StDev
Centro-Oeste	467	89,202	4,818
Nordeste	1790	77,645	7,045
Norte	447	78,120	9,991
Sudeste	1669	90,476	5,518
Sul	1191	91,557	3,860

Individual 95% CIs For Mean Based on Pooled StDev



Pooled StDev = 6,185

3.8 ANÁLISE MULTIVARIADA – COMPONENTES PRINCIPAIS

Nesta parte, o objetivo é efetuar uma análise das correlações e dos componentes principais (análise multivariada) de dados quantitativos sobre os dados de desenvolvimento dos Municípios do Brasil. Iniciamos com a análise da estatística descritiva. Em seguida, passamos para a análise das correlações e dendrogramas. E por fim, utilizamos a análise dos componentes principais.

3.8.1 CORRELAÇÃO LINEAR

Segue abaixo a matriz de correlação incluindo o teste de significância p-value. Para a correlação foi utilizado o índice de Pearson. Vale ressaltar que o índice de correlação entre as variáveis não requer que exista uma relação de causa-efeito entre ambas.

Correlations: ISDM; EMP & REN; IFGF; Liquidez; H6; R1; T1_2; S; S1_1; E; E2_4

	ISDM	EMP & REN	IFGF	Liquidez	H6	R1
EMP & REN	**0,815**					
	0,000					
IFGF	0,420	0,446				
	0,000	0,000				
Liquidez	0,258	0,261	**0,760**			
	0,000	0,000	0,000			
H6	0,695	0,522	0,327	0,244		
	0,000	0,000	0,000	0,000		
R1	-0,951	-0,801	-0,455	-0,293	-0,709	
	0,000	0,000	0,000	0,000	0,000	
T1_2	**0,806**	**0,737**	0,430	0,291	0,449	-0,781
	0,000	0,000	0,000	0,000	0,000	0,000
S	0,286	0,205	0,106	0,069	0,220	-0,195
	0,000	0,000	0,000	0,000	0,000	0,000
S1_1	-0,147	-0,182	-0,066	-0,044	-0,115	0,140
	0,000	0,000	0,000	0,001	0,000	0,000
E	**0,884**	**0,739**	0,456	0,289	0,722	-0,868
	0,000	0,000	0,000	0,000	0,000	0,000
E2_4	**0,764**	**0,705**	0,419	0,244	0,613	-0,768
	0,000	0,000	0,000	0,000	0,000	0,000

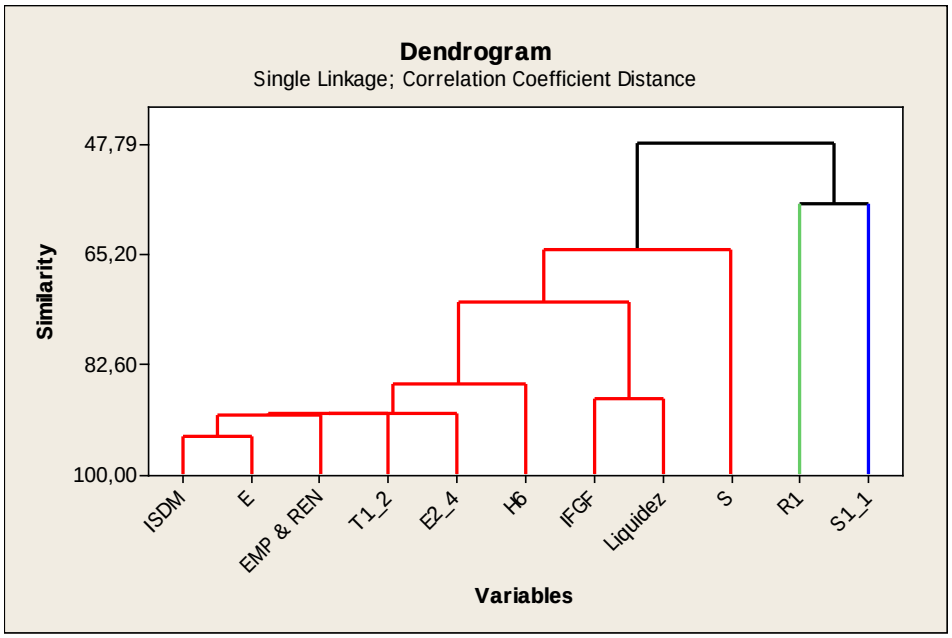
	T1_2	S	S1_1	E
S	0,137			
	0,000			
S1_1	-0,112	-0,196		
	0,000	0,000		
E	0,664	0,215	-0,131	
	0,000	0,000	0,000	
E2_4	0,599	0,194	-0,128	0,811
	0,000	0,000	0,000	0,000

As correlações significativas de acordo com o P-Value, para este trabalho, será considerada significativa quando >= 0,70). No geral, Saúde em fraca correlação com todas as demais variáveis. Apresentam uma correlação satisfatoriamente forte entre si as variáveis: ISDM e Emprego & Renda com Trabalho (T1_2), Educação (E) e E2_4 (Proporção de crianças de 7 a 14 anos na série adequada para sua idade). ISM e Emprego & Renda também tem forte correlação, assim como Liquidez e IFGF.

3.8.2 DENDROGRAMA

A interpretação de um dendrograma de similaridade entre amostras fundamenta-se na intuição: duas amostras próximas devem ter também valores semelhantes para as variáveis medidas. Ou seja, elas devem ser próximas matematicamente no espaço multidimensional. Portanto, quanto maior a proximidade entre as medidas relativas às amostras, maior a similaridade entre elas.

STAT >> MULTIVARIATE >> CLUSTER VARIABLE



As variáveis ISDM e E são as que possuem o maior nível de similaridade, por volta de 94%. As variáveis (Emprego & Renda, T1_2 e E2_4) também são muito similares, variando até 85%. Com menor nível de similaridade estão as variáveis H6, IFGF e Liquidez. Já as Saúde, R1 e S1_1 encontram-se com baixo nível de similaridade.

3.8.3. PRINCIPAIS COMPONENTES

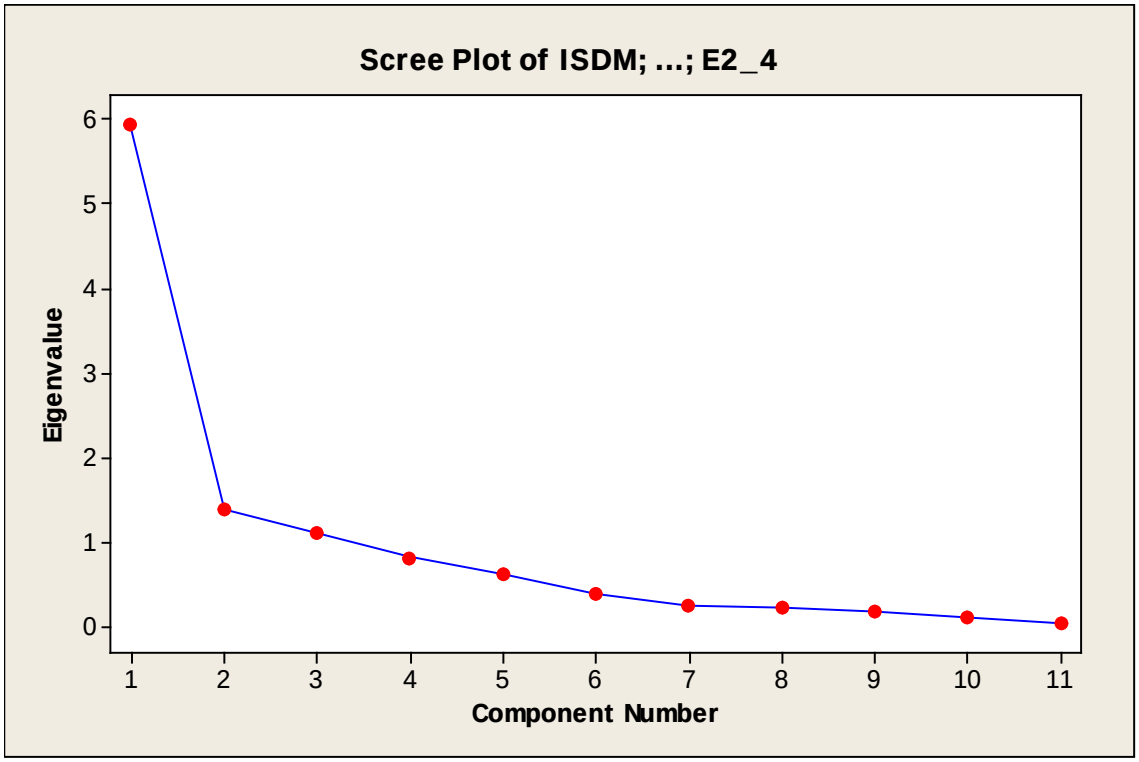
>> STAT >> MULTIVARIATE >> Principal Components

Principal Component Analysis: ISDM; EMP & REN; IFGF; Liquidez; H6; R1; T1_2; S;

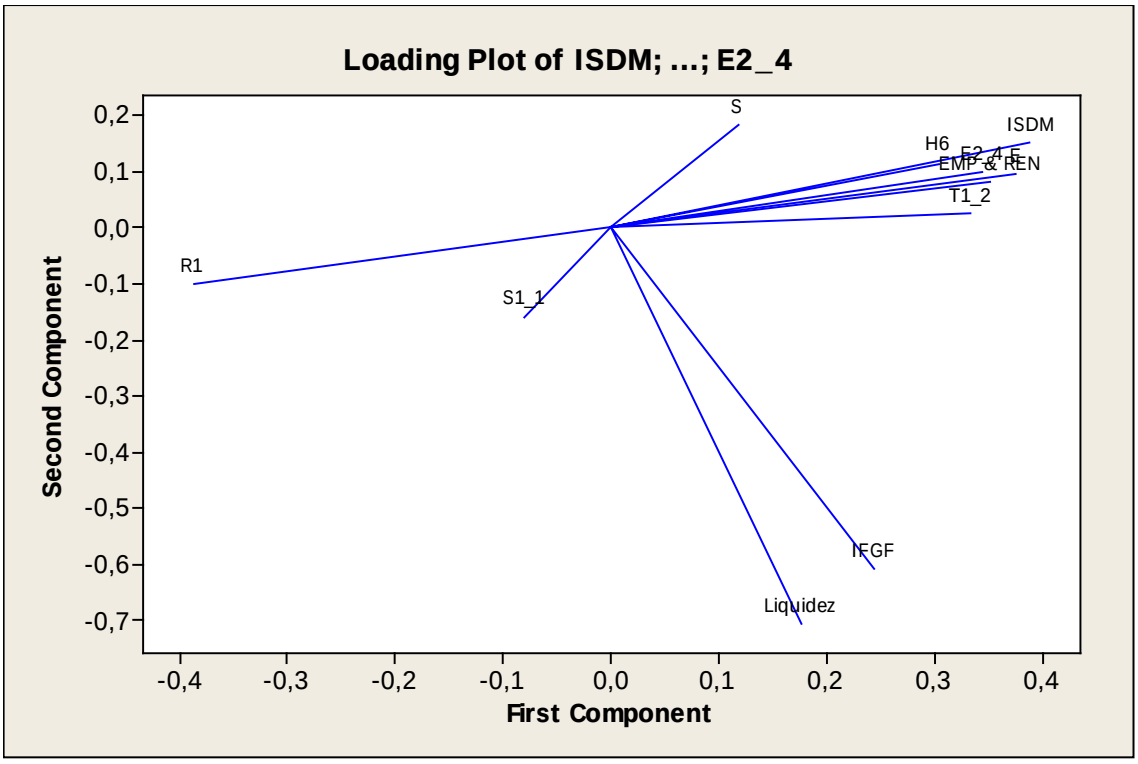
```
Eigenanalysis of the Correlation Matrix
5261 cases used, 303 cases contain missing values

Eigenvalue  5,9312  1,3778  1,0955  0,8126  0,6252  0,3780  0,2370  0,2089
Proportion   0,539   0,125   0,100   0,074   0,057   0,034   0,022   0,019
Cumulative   0,539   0,664   0,764   0,838   0,895   0,929   0,951   0,970

Eigenvalue  0,1835  0,1111  0,0392
Proportion   0,017   0,010   0,004
Cumulative   0,986   0,996   1,000


Variable       PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
ISDM         0,389   0,151  -0,061  -0,024   0,095  -0,160  -0,082  -0,038
EMP & REN    0,352   0,081  -0,041   0,130   0,324   0,151   0,782   0,302
IFGF         0,244  -0,609   0,111  -0,010   0,027   0,172   0,182  -0,689
Liquidez     0,176  -0,706   0,164  -0,036  -0,076  -0,145  -0,156   0,611
H6           0,303   0,114  -0,025  -0,171  -0,705  -0,403   0,203  -0,007
R1          -0,388  -0,102   0,108  -0,040  -0,007   0,149   0,052   0,057
T1_2         0,334   0,025  -0,124   0,151   0,509  -0,411  -0,379  -0,068
S            0,118   0,182   0,630  -0,705   0,226   0,014  -0,025  -0,004
S1_1        -0,080  -0,159  -0,719  -0,656   0,131   0,019   0,053   0,029
E            0,376   0,094  -0,084  -0,027  -0,191   0,160  -0,136  -0,123
E2_4         0,345   0,098  -0,079  -0,001  -0,137   0,723  -0,339   0,187

Variable       PC9    PC10    PC11
ISDM         0,307  -0,267  -0,783
EMP & REN   -0,077   0,118   0,031
IFGF        -0,114  -0,058  -0,047
Liquidez     0,130   0,010  -0,005
H6          -0,399   0,067  -0,024
R1          -0,288   0,614  -0,581
T1_2        -0,431   0,281   0,099
S           -0,023   0,014   0,080
S1_1        -0,006   0,001   0,008
E            0,538   0,656   0,164
E2_4        -0,389  -0,143  -0,045
```

Existe um peso muito grande da primeira variável e as demais estão mais distantes. As variável 2 e 3 possuem peso maior que 1, e as demais variáveis possuem um peso ABAIXO DE 0.8.

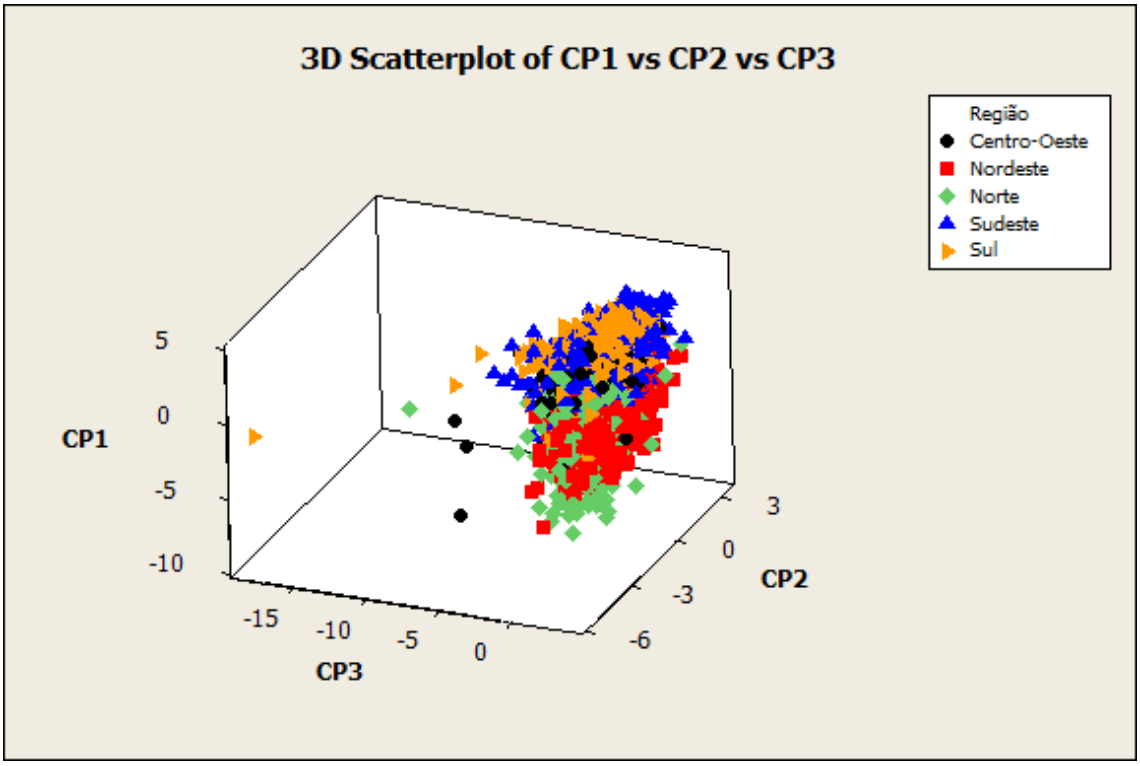
A conclusão é que podemos resumir as 11 variáveis em 3 principais variáveis para efeito de simplificação do trabalho com dados contendo muitas colunas.



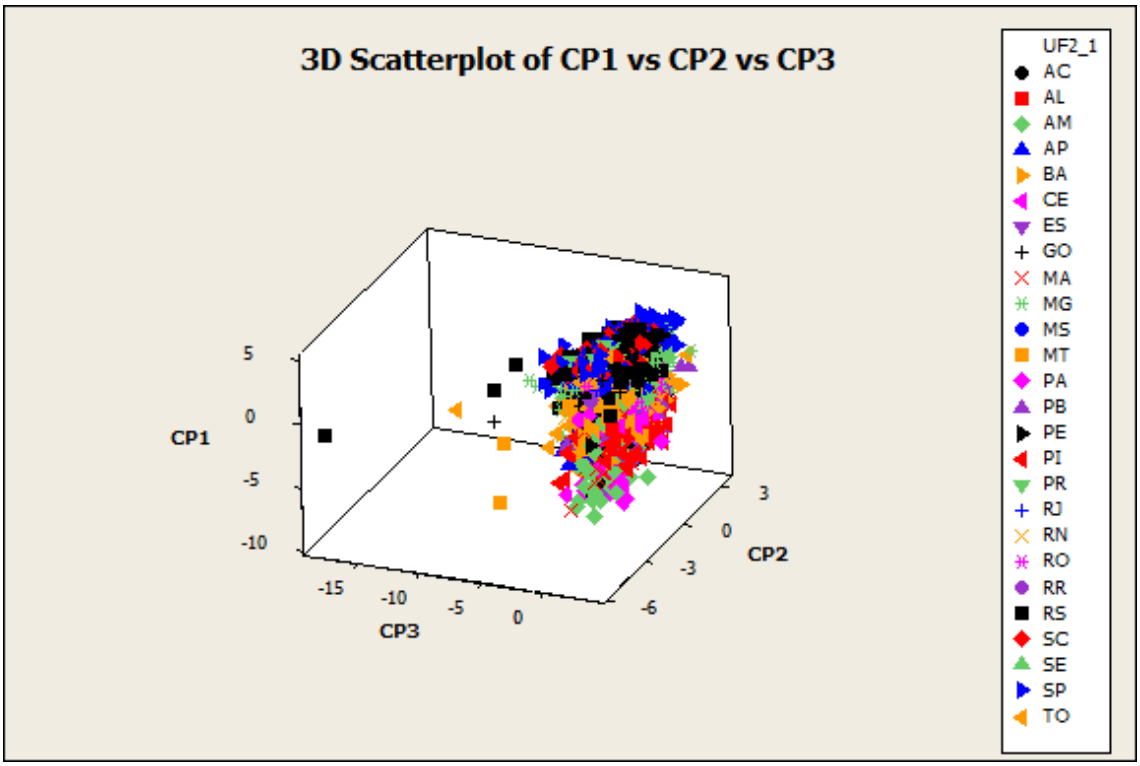
O gráfico acima demonstra a representatividade das variáveis para o componente, ou o grau de equivalência. Existe um peso muito grande da primeira variável e as demais estão bem distantes. As variáveis 2 e 3 possuem peso próximo de 1, as demais possuem um baixo peso.



Podemos observar que as variáveis R1, S e S1_1 encontram-se isoladas. Liquidez e IFGF estão próximas e, juntas, se isolam das restantes. As demais variáveis: ISDM, H6, T1_2, E2_4, Emprego & Renda e Educação formam o grupo mais próximo.



O gráfico acima é uma visão multidimensional das variáveis CP1, CP2 e CP3 agrupadas por região.



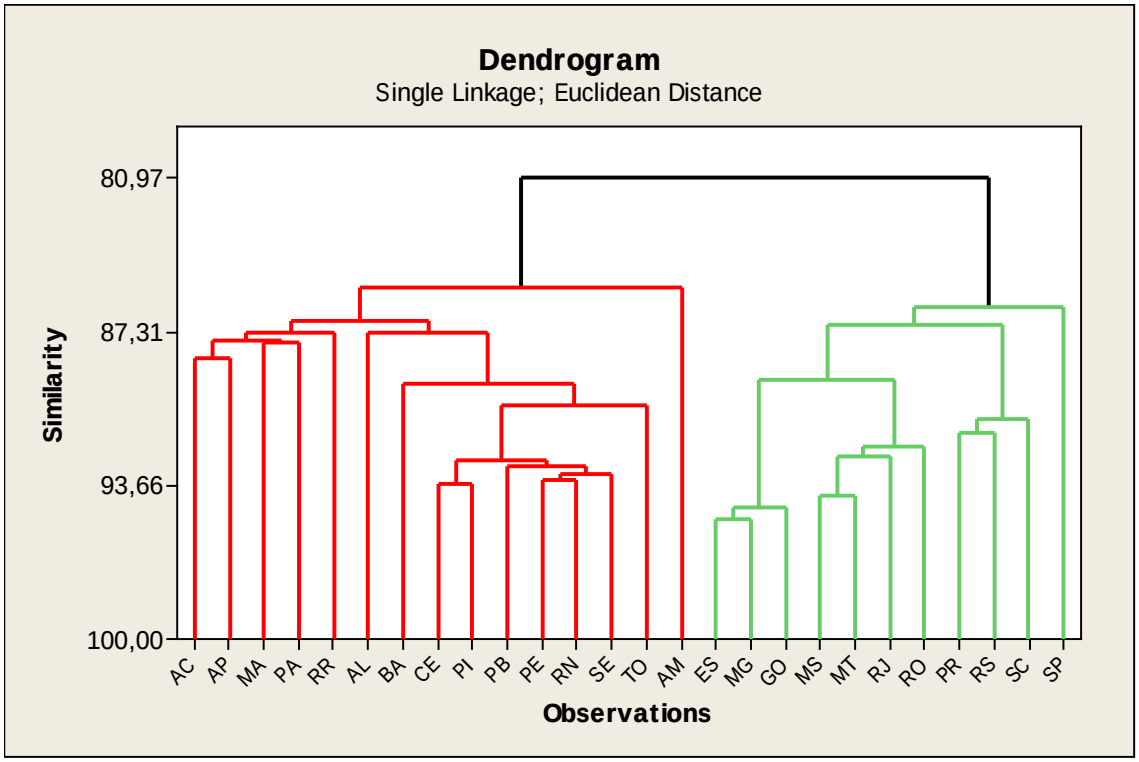
O gráfico acima repete a visão multidimensional das variáveis CP1, CP2 e CP3, agora agrupadas por estado. Nos dois gráficos a dificuldade de visualização dos dados ocorre pelo número elevado de indivíduos que compõem a população (5565 municípios).

Pelo resultado das análises da correlação linear, dendograma e principais componentes, os dados podem ser reduzidos para três variáveis, o que torna o trabalho com os números mais fáceis e de prático manuseamento.

3.9 ANÁLISE DE CONGLOMERADOS

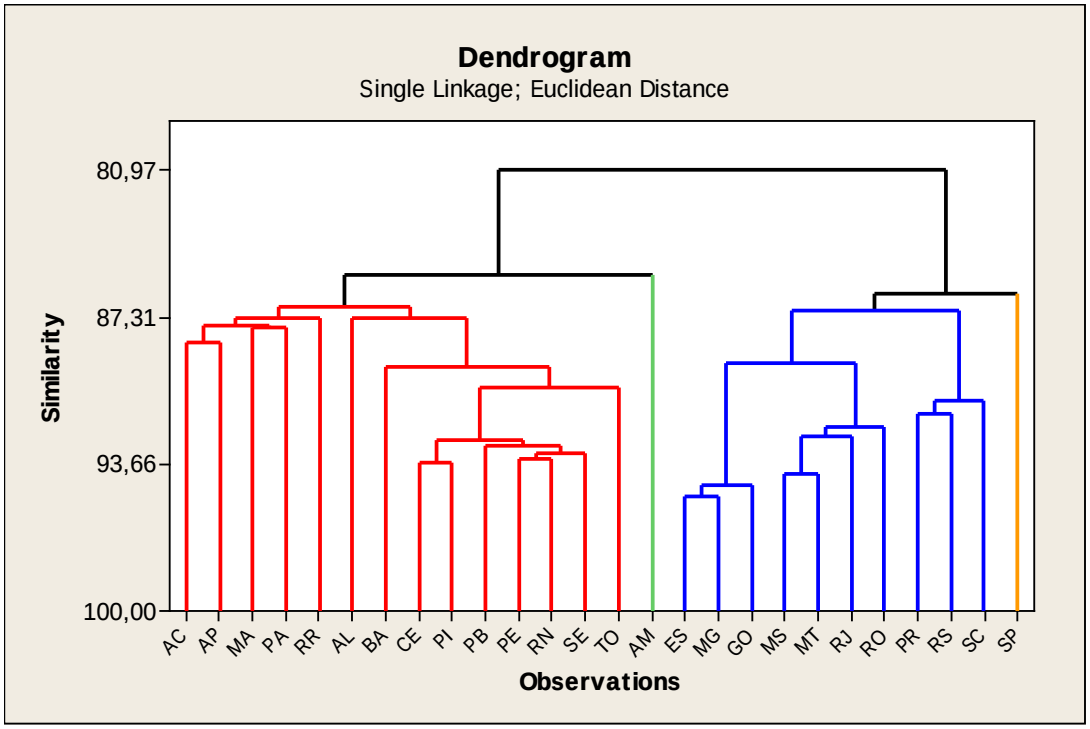
3.9.1. DENDROGRAMA DA MÉDIA DE DESENVOLVIMENTO POR ESTADO (-DF)

O Dendrograma permite uma análise do grau de similaridade dos dados para uma determinada variável. Abaixo, geramos o Dendrograma da **média** de desenvolvimento dos municípios, agrupado por Estado.



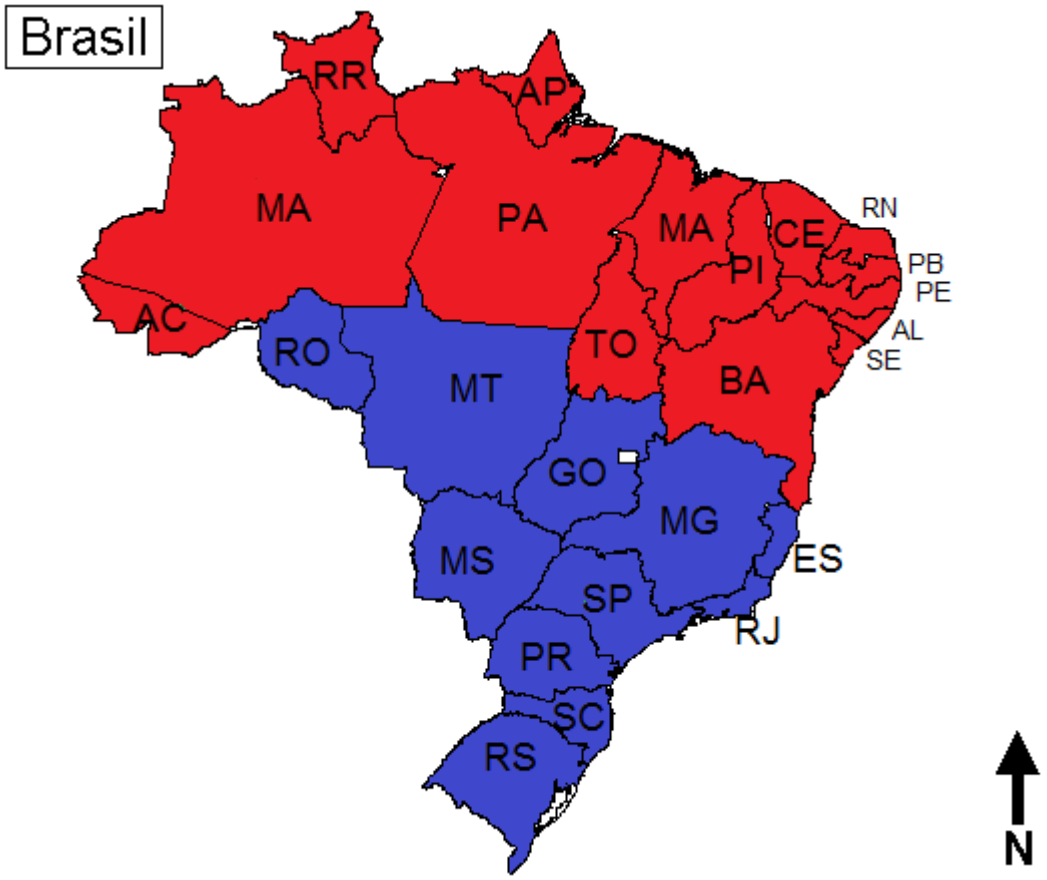
Podemos observar no gráfico acima que existem 2 grandes grupos por similaridade, e também alguns estados com baixo grau de similaridade.

É possível gerar o gráfico solicitando um número específico de cluster, no caso abaixo foi solicitado que se gerasse 4 clusters.



Os destaques deste dendograma ficaram para os estados AM e SP que possuem baixo nível de similaridade com os demais estados. Podemos concluir que o nível de desenvolvimento do Brasil pode ser dividido em 2,5 Brasis, sendo o primeiro grupo composto pelos estados em vermelho e o segundo grupo pelos estados em azul e o terceiro pelos estados com baixa similaridade sobre as médias de desenvolvimento dos municípios.

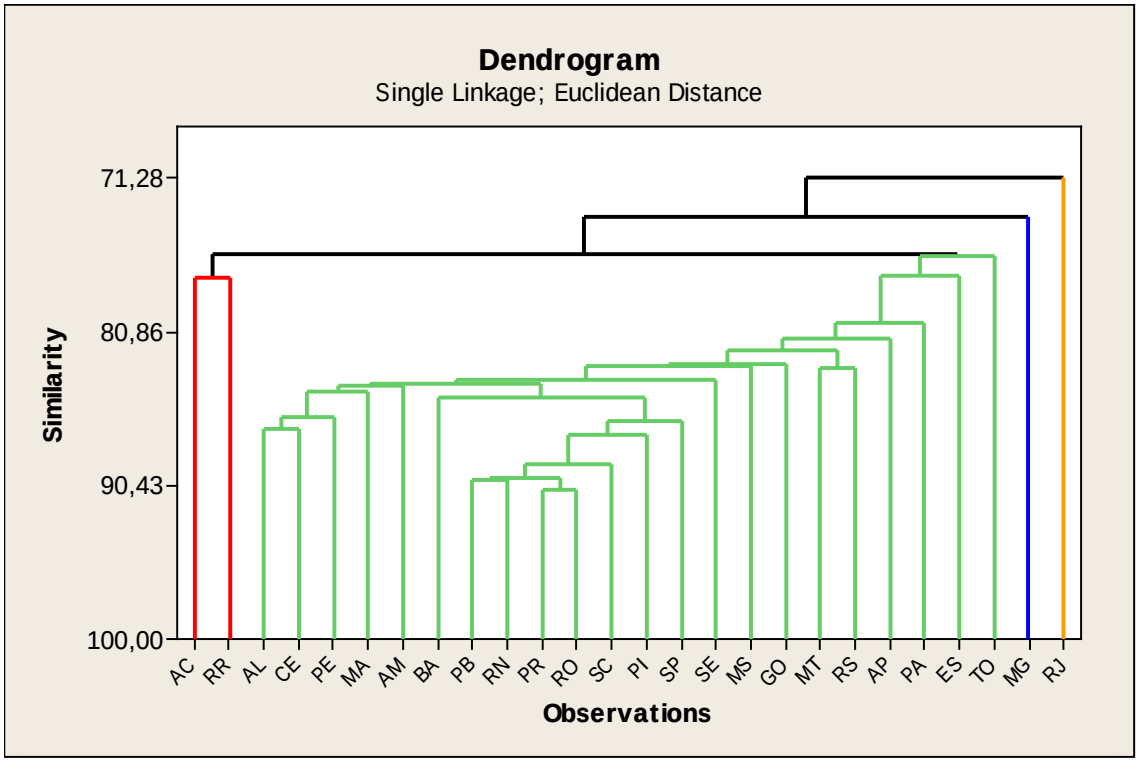
Mais prático então seria permanecer com o agrupamento em 2 Brasis.



Brasil Político – Representação dos 2 Brasis, segundo o índice médio de desenvolvimento dos municípios.

3.9.2. DENDROGRAMA DA DESIGUALDADE DE DESENVOLVIMENTO POR ESTADO (-DF)

Neste exemplo será demonstrado o índice de desigualdade de desenvolvimento dos municípios do Brasil agrupados por estados. Utilizaremos para isso o **desvio padrão** dos índices de desenvolvimento.

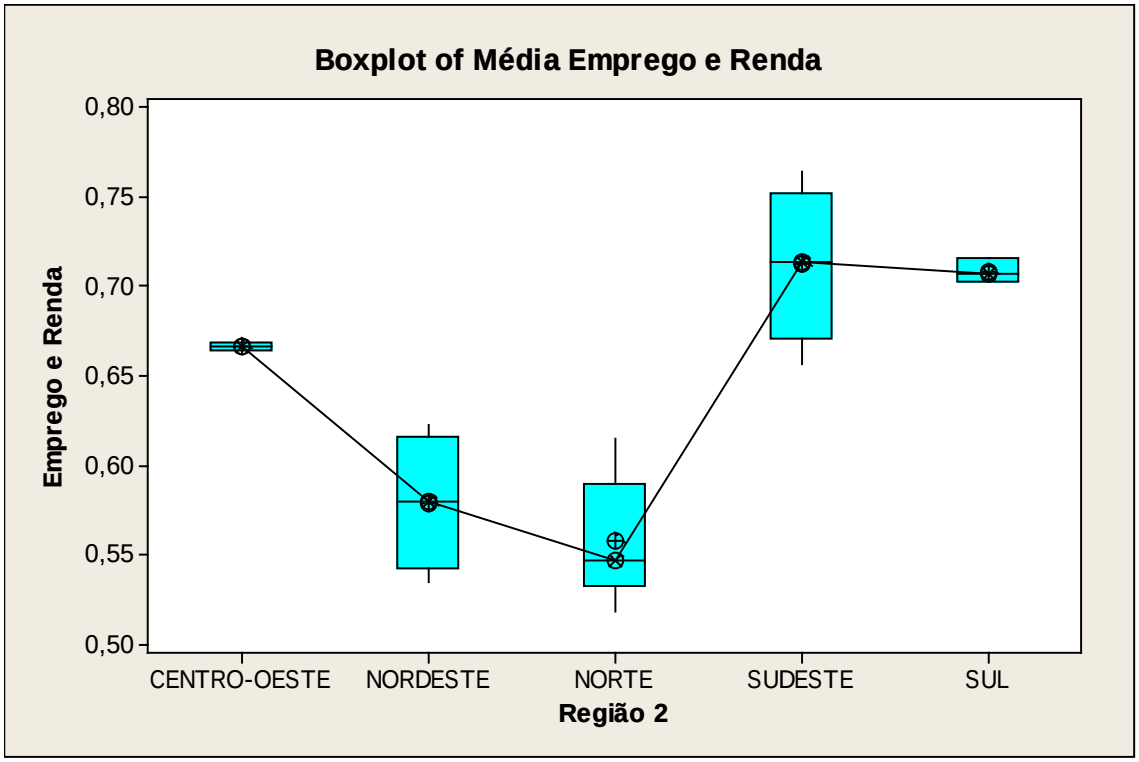


O grau de similaridade dos índices de desigualdade dos estados é muito variado. Foram considerados 4 cluster neste primeiro agrupamento, sendo o primeiro composto pelos estados do AC e RR, o segundo composto isoladamente por MG e o terceiro por RJ, também isolada. O grande grupo é composto pelos estados desde AL até TO.

3.9.3. ANÁLISE DAS VARIANCIAS DOS ÍNDICES DE DESENVOLVIMENTO POR ESTADO (– DF)

A análise das variâncias permite a verificação e visualização das médias e desvios padrões da variável a ser analisada. O gráfico *BOXPLOT* ilustra os agrupamentos, o seu tamanho varia de acordo com a quantidade de dados de cada grupo, e também é possível visualizar as ocorrências de *outliers* dentro de um grupo de dados.

A primeira análise é do índice médio de Emprego e Renda dos municípios do Brasil.



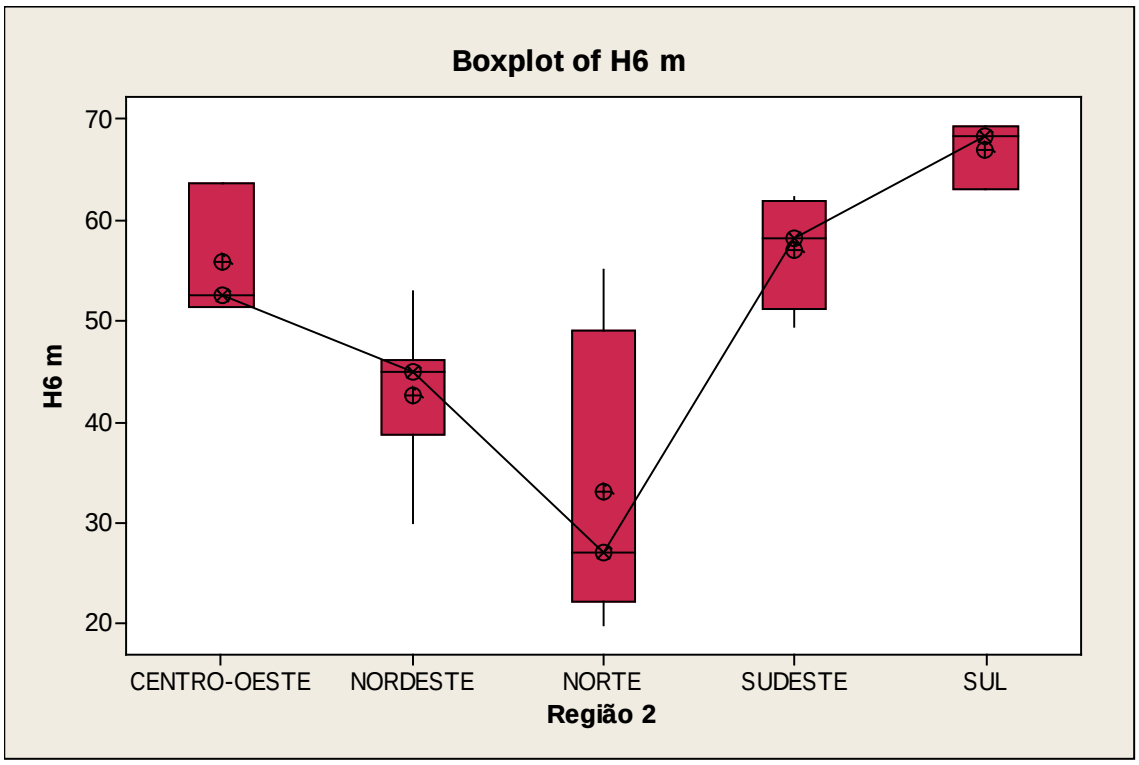
Este gráfico exibe os resultados das médias dos estados, agrupados por região. Podemos ver que a região Sudeste é a que possui maior índice médio de desenvolvimento, quase empatada com a região Sul. A região que possui o pior desempenho médio de desenvolvimento é a Norte seguido pela Nordeste.

No resumo descritivo dos dados podemos visualizar os desvios padrões de cada região, e as médias.

One-way ANOVA: ER m versus Região 2

```
Source     DF       SS       MS      F      P
Região 2    4  0,10495  0,02624  23,72  0,000
Error      21  0,02323  0,00111
Total      25  0,12818

S = 0,03326   R-Sq = 81,88%   R-Sq(adj) = 78,43%

                                      Individual 95% CIs For Mean Based on
                                      Pooled StDev
Level          N     Mean    StDev  --+---------+---------+---------+-------
CENTRO-OESTE   3  0,66637  0,00249                  (------*------)
NORDESTE      10  0,57832  0,03511       (--*---)
NORTE          6  0,55731  0,03528  (----*----)
SUDESTE        4  0,71223  0,04401                          (-----*----)
SUL            3  0,70858  0,00649                        (------*------)
                                      --+---------+---------+---------+-------
                                      0,540     0,600     0,660     0,720
Pooled StDev = 0,03326
```

O segundo gráfico mostra o resultado do índice médio H6, e mostra que a região mais adiantada em relação à Habitação é a Sul, seguida pela Sudeste, e a pior região é a Norte.



Podemos observar que os desvios padrões são altos, existe uma grande variação nos dados de habitação entre os municípios das regiões do Brasil.

One-way ANOVA: H6 m versus Região 2

Source	DF	SS	MS	F	P
Região 2	4	3159,2	789,8	10,03	0,000
Error	21	1654,3	78,8		
Total	25	4813,5			

S = 8,876 R-Sq = 65,63% R-Sq(adj) = 59,09%

Level	N	Mean	StDev
CENTRO-OESTE	3	55,863	6,752
NORDESTE	10	42,672	6,549
NORTE	6	**33,112**	14,541
SUDESTE	4	56,980	5,687
SUL	3	**66,837**	3,383

Individual 95% CIs For Mean Based on Pooled StDev

```
---+---------+---------+---------+------
                (------*------)
          (--*---)
(----*----)
                   (-----*-----)
                        (-------*------)
---+---------+---------+---------+------
  30        45        60        75
```

Pooled StDev = 8,876

3.10 ANÁLISE DISCRIMINANTE

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos. É uma técnica da estatística multivariada que estuda a separação de objetos de uma população em duas ou mais classes. A discriminação ou separação é a primeira etapa, sendo a parte exploratória da análise e consiste em se procurar características capazes de serem utilizadas para alocar objetos em diferentes grupos previamente definidos. A classificação ou alocação pode ser definida como um conjunto de regras que serão usadas para alocar novos objetos.

Este trabalho tem por objetivo efetuar uma análise comparativa de médias, intervalos de confiança e regressões de dados de indicadores relacionados ao desenvolvimento humano dos municípios do Brasil. Utilizamos a análise discriminante para tentar predizer ou explicar os indicadores relacionados ao desenvolvimento da educação dos municípios do Brasil.

3.10.1. ANÁLISE DISCRIMINANTE LINEAR POR REGIÃO

Nesta análise iremos discriminar os indicadores de desenvolvimento dos municípios do Brasil, e utilizaremos inicialmente a variável categórica Região.

STAT >> MULTIVARIATE >> DISCRIMINANT ANALISYS

Discriminant Analysis: Região versus ISDM; EMP & REN; ...

```
Linear Method for Response: Região


Predictors: ISDM; EMP & REN; IFGF; Liquidez; H6; R1; T1_2; S; S1_1; E; E2_4


Group  Centro-Oeste  Nordeste  Norte  Sudeste   Sul
Count           467      1790    447     1669  1191

Summary of classification
                                 True Group
Put into Group  Centro-Oeste  Nordeste  Norte  Sudeste    Sul
Centro-Oeste             304        19     75      135    250
Nordeste                   9      1454     52       86      4
Norte                     35       217    282       35      8
Sudeste                   39        97     26     1339    168
Sul                       80         3     12       74    761
Total N                  467      1790    447     1669   1191
N correct                304      1454    282     1339    761
Proportion             0,651     0,812  0,631    0,802  0,639

N = 5564           N Correct = 4140           Proportion Correct = 0,744
```

A região que acertou mais é Nordeste (0,812) e a que errou mais é o Norte (0,631). O gráfico exibe o cruzamento de dados entre as regiões. Por exemplo, a região Sudeste possui 1669 municípios e apenas 1339 correspondem a região, sendo que 135 são semelhantes aos dados da região Centro-Oeste. Podemos concluir que o agrupamento por região não é uma boa escolha segundo esta avaliação. O percentual correto **= 0,744.**

3.10.2. ANÁLISE DISCRIMINANTE LINEAR POR "3 BRASIS"

Aqui, iremos discriminar os indicadores de desenvolvimento dos municípios do Brasil, e utilizaremos a variável categórica 3 Brasis, que representa os agrupamentos segundo a análise anterior do Dendrograma por similaridade dos dados.

Discriminant Analysis: 3 BRA TOTAL versus ISDM; EMP & REN; ...

```
Linear Method for Response: 3 BRA TOTAL

Predictors: ISDM; EMP & REN; IFGF; Liquidez; H6; R1; T1_2; S; S1_1; E; E2_4

Group         B1         B2         B3
Count       2123       2732        709

Summary of classification

                        True Group
Put into Group      B1      B2      B3
B1                1887     156      59
B2                 118    2168      80
B3                 118     408     570
Total N           2123    2732     709
N correct         1887    2168     570
Proportion       0,889   0,794   0,804

N = 5564            N Correct = 4625           Proportion Correct = 0,831
```

O grupo que acertou mais é B1 (0,889) e a que errou mais é o B2 (0,794). O gráfico exibe o cruzamento de dados entre as classificações de 3 Brasis. Por exemplo, o B1 possui 2123 municípios e apenas 1887 correspondem a região, sendo que 118 são semelhantes aos dados de B2 e B3. O nome desta matriz é *confusion matrix* ou matriz de confusão. O percentual correto **= 0,831.** O percentual de acerto para esta análise foi maior que para o cruzamento dos dados de Regiões do Brasil.

3.10.3. ANÁLISE DISCRIMINANTE QUADRÁTICA POR “3 BRASIS”

Uma boa classificação deve resultar em pequenos erros, isto é, deve haver pouca probabilidade de má classificação, e para que isso ocorra a regra de classificação deve considerar as probabilidades a priori e os custos de má classificação. Outro fator que uma regra de classificação deve considerar é se as variâncias das populações são iguais ou não. Quando a regra de classificação assume que as variâncias das populações são iguais, as funções discriminantes são ditas lineares e quando não são funções discriminantes quadráticas. Vamos agora verificar a função quadrática para 3 Brasis.

Discriminant Analysis: 3 BRA TOTAL versus ISDM; EMP & REN; ...

Quadratic Method for Response: 3 BRA TOTAL

Predictors: ISDM; EMP & REN; IFGF; Liquidez; H6; R1; T1_2; S; S1_1; E; E2_4

Group	B1	B2	B3
Count	2123	2732	709

Summary of classification

	True Group		
Put into Group	B1	B2	B3
B1	1878	198	50
B2	103	2235	67
B3	142	299	592
Total N	2123	2732	709
N correct	1878	2235	592
Proportion	0,885	0,818	0,835

N = 5564 N Correct = 4705 Proportion Correct = 0,846

No modelo quadrático a proporção foi alterada em menos de 1,5% (de 0,831 para 0,846). Seguindo o pensamento da simplicidade, vamos escolher o método linear por ser o mais simples.

A parcimônia é a preferência pela explicação mais simples para uma observação. Esta geralmente é considerada a melhor maneira de julgar as hipóteses. Parcimônia também é um conceito utilizado na sistemática moderna que estabelece que ao construir e selecionar árvores filogenéticas, ou seja, os dados, o melhor critério é baseado em seus princípios. Normalmente é correto o relacionamento mais simples encontrado entre dois indivíduos,

aquele que apresente o menor número de passos intermediários ou mudanças evolucionárias. Portanto a diferença entre o método linear e o quadrático é pequena e não justifica a utilização do método quadrático.

3.10.4. ANÁLISE DISCRIMINANTE LINEAR PARA DADOS AGRUPADOS

Neste exemplo abaixo vamos através do dendrograma pesquisar o grau de similaridade das variáveis das médias do desenvolvimento dos municípios do Brasil. Com base na similaridade poderemos definimos agrupamento de dados e após utilizamos a análise discriminante para verificar a proporção correta dos agrupamentos.

Discriminant Analysis: 3 BRA versus ISDM m; ER m; ...

```
Linear Method for Response: 3 BRA


Predictors: ISDM m; ER m; IFGF m; LIQ m; H6 m; R1 m; T1_2 m; S m; S1_1 m; E m;
            E2_4 m


Group         1         2         3
Count        14        10         2


Summary of classification

                     True Group
Put into Group      B1      B2      B3
B1                  14       0       0
B2                   0      10       0
B3                   0       0       2
Total N             14      10       2
N correct           14      10       2
Proportion       1,000   1,000   1,000

N = 26           N Correct = 26           Proportion Correct = 1,000
```

Neste caso a proporção correta é de 100%, ou seja, os agrupamentos gerados anteriormente pelo agrupamento em 3 Brasis gerou a mesma proporção do método linear utilizado na análise discriminante.

3.11 REGRESSÃO LOGÍSTICA

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias1 2.

O êxito da regressão logística assenta sobretudo nas numerosas ferramentas que permitem interpretar de modo aprofundado os resultados obtidos. Em comparação com as técnicas conhecidas em regressão, em especial a regressão linear, a regressão logística distingue-se essencialmente pelo fato de a variável resposta ser categórica.

Trata-se de um modelo de regressão para variáveis dependentes ou de resposta binominalmente distribuídas. É útil para modelar a probabilidade de um evento ocorrer como função de outros fatores.

Stat >> Regression >> Ordinal Logistical Regression

3.11.1 REGRESSÃO LOGÍSTICA AGRUPADA POR REGIÃO

Ordinal Logistic Regression: Região versus ISDM; EMP & REN; ...

Link Function: Logit

Response Information

Variable	Value	Count
Região	Centro-Oeste	467
	Nordeste	1790
	Norte	447
	Sudeste	1669
	Sul	1191
	Total	5564

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Const(1)	-1,86903	0,620952	-3,01	0,003			
Const(2)	0,632157	0,621136	1,02	0,309			
Const(3)	1,22819	0,621302	1,98	0,048			
Const(4)	3,43042	0,622352	5,51	0,000			
ISDM	1,47236	0,101164	14,55	0,000	4,36	3,58	5,32
EMP & REN	1,05704	0,498255	2,12	0,034	2,88	1,08	7,64
IFGF	-2,09009	0,296313	-7,05	0,000	0,12	0,07	0,22
Liquidez	0,308219	0,106848	2,88	0,004	1,36	1,10	1,68
H6	-0,0425569	0,0029926	-14,22	0,000	0,96	0,95	0,96
R1	0,0602976	0,0052157	11,56	0,000	1,06	1,05	1,07
T1_2	-0,0483010	0,0026955	-17,92	0,000	0,95	0,95	0,96
S	-0,155589	0,0172922	-9,00	0,000	0,86	0,83	0,89
S1_1	0,0014347	0,0018411	0,78	0,436	1,00	1,00	1,01
E	-0,588083	0,0540832	-10,87	0,000	0,56	0,50	0,62
E2_4	-0,0097021	0,0051114	-1,90	0,058	0,99	0,98	1,00

Log-Likelihood = -6702,829
Test that all slopes are zero: G = 2914,043, DF = 11, P-Value = 0,000

3.11.1 REGRESSÃO LOGÍSTICA AGRUPADA POR REGIÃO

Ordinal Logistic Regression: Região versus ISDM; EMP & REN; ...

```
Link Function: Logit


Response Information

Variable  Value         Count
Região    Centro-Oeste    467
          Nordeste       1790
          Norte           447
          Sudeste        1669
          Sul            1191
          Total          5564


Logistic Regression Table

                                                          Odds     95% CI
Predictor         Coef    SE Coef       Z      P  Ratio  Lower  Upper
Const(1)     -1,86903   0,620952   -3,01  0,003
Const(2)     0,632157   0,621136    1,02  0,309
Const(3)      1,22819   0,621302    1,98  0,048
Const(4)      3,43042   0,622352    5,51  0,000
ISDM          1,47236   0,101164   14,55  0,000   4,36   3,58   5,32
EMP & REN     1,05704   0,498255    2,12  0,034   2,88   1,08   7,64
IFGF         -2,09009   0,296313   -7,05  0,000   0,12   0,07   0,22
Liquidez     0,308219   0,106848    2,88  0,004   1,36   1,10   1,68
H6         -0,0425569  0,0029926  -14,22  0,000   0,96   0,95   0,96
R1          0,0602976  0,0052157   11,56  0,000   1,06   1,05   1,07
T1_2       -0,0483010  0,0026955  -17,92  0,000   0,95   0,95   0,96
S           -0,155589  0,0172922   -9,00  0,000   0,86   0,83   0,89
S1_1        0,0014347  0,0018411    0,78  0,436   1,00   1,00   1,01
E           -0,588083  0,0540832  -10,87  0,000   0,56   0,50   0,62
E2_4       -0,0097021  0,0051114   -1,90  0,058   0,99   0,98   1,00


Log-Likelihood = -6702,829
Test that all slopes are zero: G = 2914,043, DF = 11, P-Value = 0,000
```

Enquanto método de predição para variáveis categóricas, a regressão logística é comparável às técnicas supervisionadas propostas em aprendizagem automática (árvores de decisão, redes neuronais, etc.), ou ainda a análise discriminante preditiva em estatística exploratória. É possível de colocá-la em concorrência para escolha do modelo mais adaptado para um certo problema preditivo a resolver.

3.11.2 REGRESSÃO LOGÍSTICA AGRUPADA POR “3 BRASIS”

Ordinal Logistic Regression: 3 BRA TOTAL versus ISDM; EMP & REN; ...

```
Link Function: Logit


Response Information

Variable     Value  Count
3 BRA TOTAL  B1      2123
             B2      2732
             B3       709
             Total   5564


Logistic Regression Table

                                                        Odds     95% CI
Predictor         Coef    SE Coef       Z      P  Ratio  Lower  Upper
Const(1)       18,0394   0,984262   18,33  0,000
Const(2)       23,0034    1,01870   22,58  0,000
ISDM          -1,37164   0,139888   -9,81  0,000   0,25   0,19   0,33
EMP & REN      0,372517   0,683848    0,54  0,586   1,45   0,38   5,54
IFGF          -1,66907   0,422054   -3,95  0,000   0,19   0,08   0,43
Liquidez     0,0601507   0,152355    0,39  0,693   1,06   0,79   1,43
H6           0,0617640  0,0042414   14,56  0,000   1,06   1,05   1,07
R1           0,0246666  0,0073452    3,36  0,001   1,02   1,01   1,04
T1_2         0,0059919  0,0037571    1,59  0,111   1,01   1,00   1,01
S            0,0549223  0,0233348    2,35  0,019   1,06   1,01   1,11
S1_1        -0,0027411  0,0025170   -1,09  0,276   1,00   0,99   1,00
E            -0,969315  0,0935021  -10,37  0,000   0,38   0,32   0,46
E2_4         -0,141731  0,0085344  -16,61  0,000   0,87   0,85   0,88


Log-Likelihood = -2799,238
Test that all slopes are zero: G = 5300,331, DF = 11, P-Value = 0,000
```

Comparando os dois exemplos, no primeiro ele executou 4 interações enquanto que para os 3 Brasis apenas duas interações. O valor de G foi aumentado de 2914 para 5300.

3.12 ANÁLISE DE CORRESPONDÊNCIA

Análise de correspondência é uma técnica de análise exploratória de dados adequada para analisar tabelas de duas entradas ou tabelas de múltiplas entradas, levando em conta algumas medidas de correspondência entre linhas e colunas. Consiste na conversão de uma matriz de dados não negativos em um tipo particular de representação gráfica em que as linhas e colunas da matriz são simultaneamente representadas em dimensão reduzida, isto é, por pontos no gráfico. Este método permite estudar as relações e semelhanças existentes

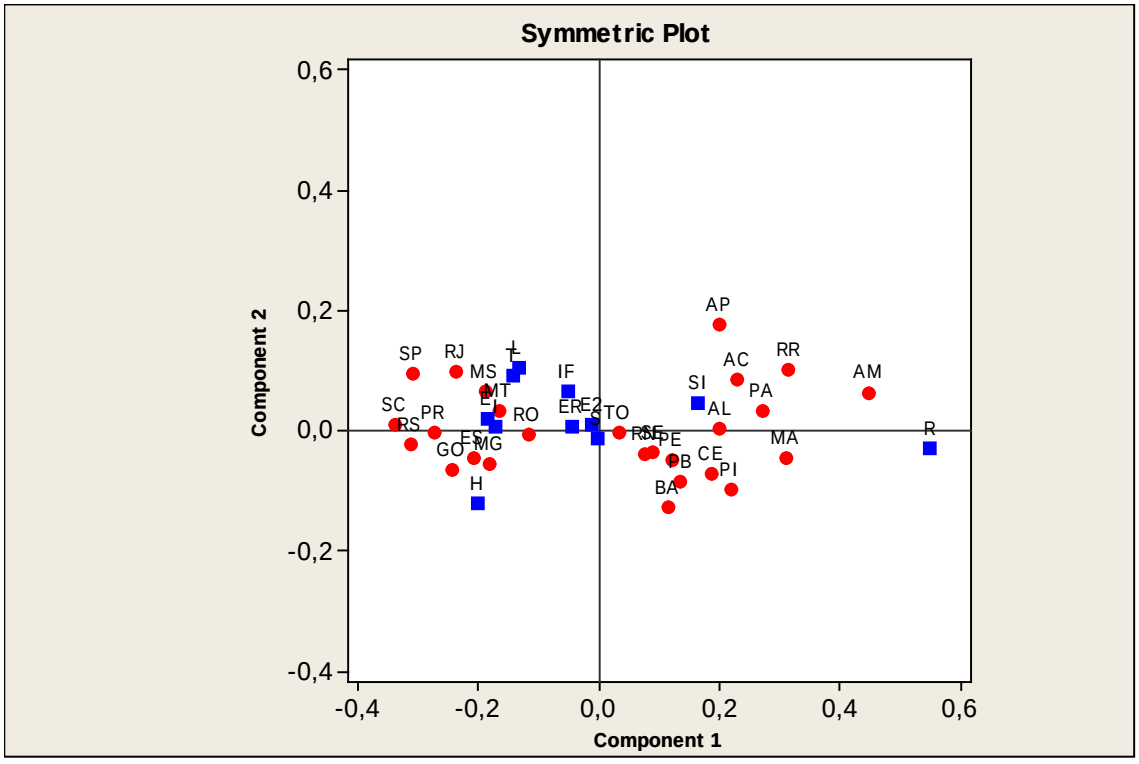
entre as categorias de linhas e entre as categorias de colunas de uma tabela de contingência ou o conjunto de categorias de linhas e o conjunto categorias de colunas.

A análise de correspondência mostra como as variáveis dispostas em linhas e colunas estão relacionadas e não somente se a relação existe. Embora seja considerada uma técnica descritiva e exploratória, esta análise simplifica dados complexos e produz análises exaustivas de informações que suportam conclusões a respeito das mesmas.

3.12.1. ANÁLISE DE CORRESPONDÊNCIA DOS ÍNDICES DE DESENVOLVIMENTO

Nesta análise serão trabalhados os estados e as médias de desenvolvimento por estado. Na análise de correspondência será gerado um mapa contendo quais estados estão mais próximos e quais variáveis tem a ver entre si. O comando para gerar o gráfico é:

STAT >> MULTIVARIATE >> SIMPLE CORRESPONDENCE ANALISYS

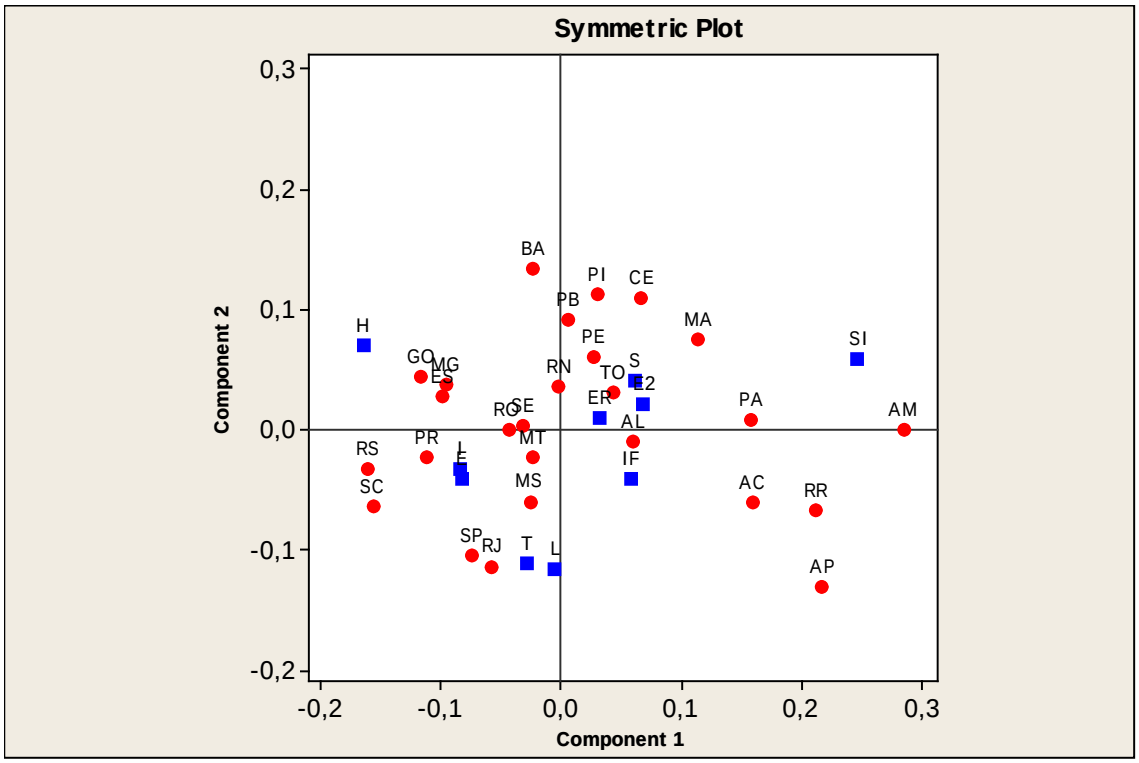


O gráfico acima é o resultado da análise de correspondência das médias de desenvolvimento dos municípios do Brasil, já agrupadas por estado. Os pontos azuis representam as variáveis ISDM (I), Emprego & Renda (ER), IFGF (IF), Liquidez (L), Habitação (H), Renda (R), Trabalho

(T), Saúde (S), Educação (E) e Percentual de crianças de 7 a 14 anos que estão na série correta segundo a idade (E2_4)-(E2). Os pontos em vermelho representam os estados do Brasil.

Todas as variáveis se encontram próximas ao agrupamento, porém a mais distante é Renda (R).

Eliminando a variável Renda, obtemos este resultado.



A análise de correspondência pode ser considerada como um caso especial da análise de componentes principais (TRABALHO 7), porém dirigida a dados categóricos organizados em tabelas de contingência e não a dados contínuos.

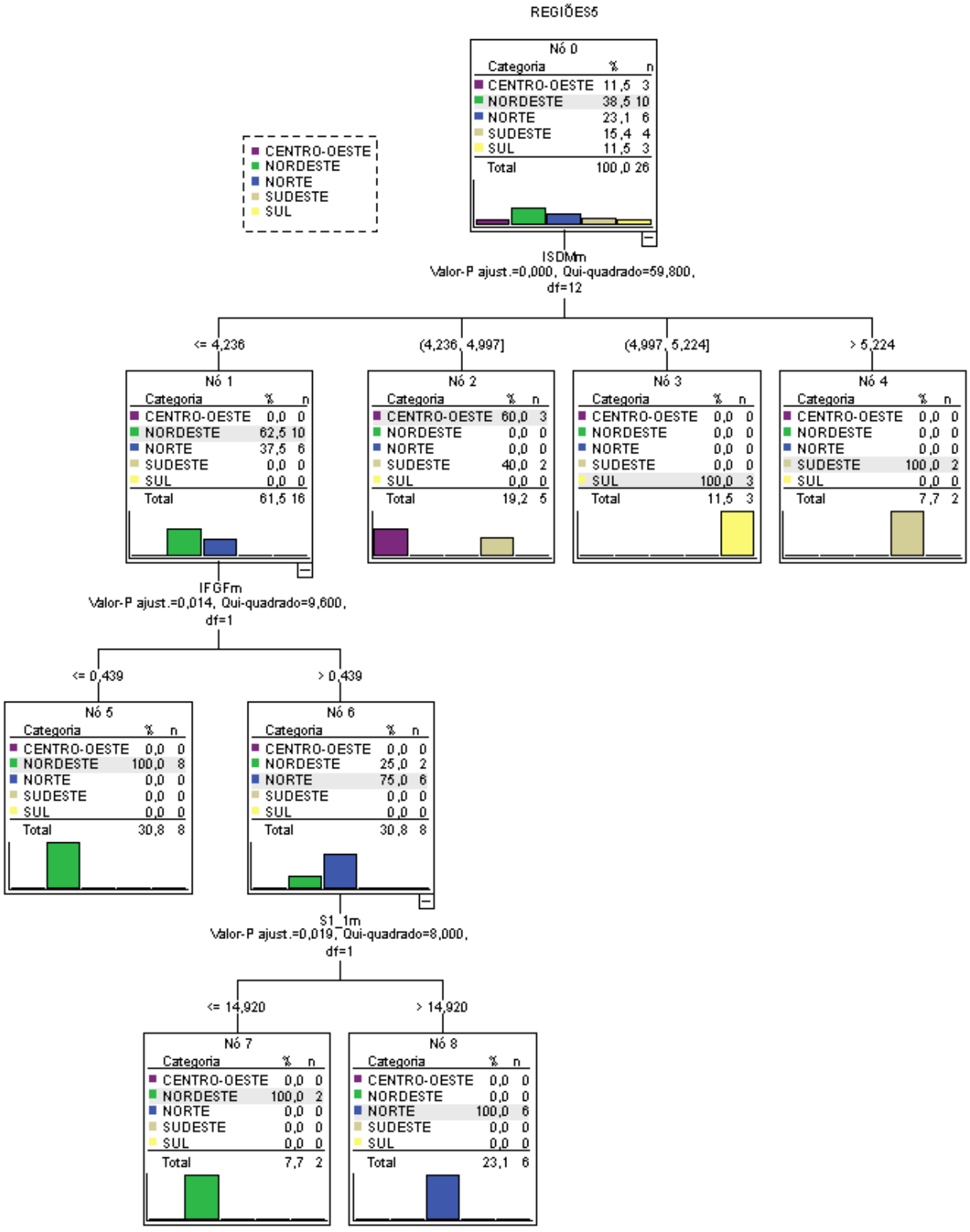
3.13 ÁRVORES DE CLASSIFICAÇÃO

3.13.1 ÁRVORE AGRUPADA POR REGIÃO COM AS MÉDIAS DE DESENVOLVIMENTO DOS MUNICÍPIOS

Resumo do modelo

Especificações	Método de crescimento	CHAID
	Variável dependente	REGIÕES5
	Variáveis independentes	ISDMm, ERm, IFGFm, LIQm, H6m, R1m, T1_2m, Sm, S1_1m, Em, E2_4m
	Validação	Nenhum
	Profundidade de árvore máxima	3
	Casos mínimos em nó pai	2
	Casos mínimos em nó filho	1
	Variáveis independentes incluídas	ISDMm, IFGFm, S1_1m
Resultados	Número de nós	9
	Número de nós de terminal	6
	Profundidade	3

Neste primeiro estudo a árvore de classificação será constituída pelo agrupamento das variáveis de desenvolvimento utilizadas neste estudo, por região.


REGIÕES5
CENTRO-OESTE
NORDESTE
NORTE
SUDESTE
SUL
Nó 0
Categoria % n
CENTRO-OESTE 11,5 3
NORDESTE 38,5 10
NORTE 23,1 6
SUDESTE 15,4 4
SUL 11,5 3
Total 100,0 26
ISDMm
Valor-P ajust.=0,000, Qui-quadrado=59,800, df=12
<= 4,236
(4,236, 4,997]
(4,997, 5,224]
> 5,224
Nó 1
Categoria % n
CENTRO-OESTE 0,0 0
NORDESTE 62,5 10
NORTE 37,5 6
SUDESTE 0,0 0
SUL 0,0 0
Total 61,5 16
Nó 2
Categoria % n
CENTRO-OESTE 60,0 3
NORDESTE 0,0 0
NORTE 0,0 0
SUDESTE 40,0 2
SUL 0,0 0
Total 19,2 5
Nó 3
Categoria % n
CENTRO-OESTE 0,0 0
NORDESTE 0,0 0
NORTE 0,0 0
SUDESTE 0,0 0
SUL 100,0 3
Total 11,5 3
Nó 4
Categoria % n
CENTRO-OESTE 0,0 0
NORDESTE 0,0 0
NORTE 0,0 0
SUDESTE 100,0 2
SUL 0,0 0
Total 7,7 2
IFGFm
Valor-P ajust.=0,014, Qui-quadrado=9,600, df=1
<= 0,439
> 0,439
Nó 5
Categoria % n
CENTRO-OESTE 0,0 0
NORDESTE 100,0 8
NORTE 0,0 0
SUDESTE 0,0 0
SUL 0,0 0
Total 30,8 8
Nó 6
Categoria % n
CENTRO-OESTE 0,0 0
NORDESTE 25,0 2
NORTE 75,0 6
SUDESTE 0,0 0
SUL 0,0 0
Total 30,8 8
S1_1m
Valor-P ajust.=0,019, Qui-quadrado=8,000, df=1
<= 14,920
> 14,920
Nó 7
Categoria % n
CENTRO-OESTE 0,0 0
NORDESTE 100,0 2
NORTE 0,0 0
SUDESTE 0,0 0
SUL 0,0 0
Total 7,7 2
Nó 8
Categoria % n
CENTRO-OESTE 0,0 0
NORDESTE 0,0 0
NORTE 100,0 6
SUDESTE 0,0 0
SUL 0,0 0
Total 23,1 6

A árvore indica que existem 6 nós a partir de ISDMm, IFGFm, e S1_1m. O primeiro nó representa os dados menores que 4,236, o segundo entre 4,236 a 4,997, o terceiro entre 4,997 e 5,224 e assim por diante.

Risco

Estimativas	Modelo padrão
,077	,052

Método de crescimento: CHAID

Variável dependente: REGIÕES5

Posto

Observado	Previsto					
	CENTRO-OESTE	NORDESTE	NORTE	SUDESTE	SUL	Porcentagem Correta
CENTRO-OESTE	3	0	0	0	0	100,0%
NORDESTE	0	10	0	0	0	100,0%
NORTE	0	0	6	0	0	100,0%
SUDESTE	2	0	0	2	0	50,0%
SUL	0	0	0	0	3	100,0%
Porcentagem global	19,2%	38,5%	23,1%	7,7%	11,5%	92,3%

Método de crescimento: CHAID

Variável dependente: REGIÕES5

O percentual de acerto é 92,3%.

3.13.2 ÁRVORE AGRUPADA POR "3BRASIS" COM OS ÍNDICES DE DESIGUALDADE (DESVIO PADRÃO)

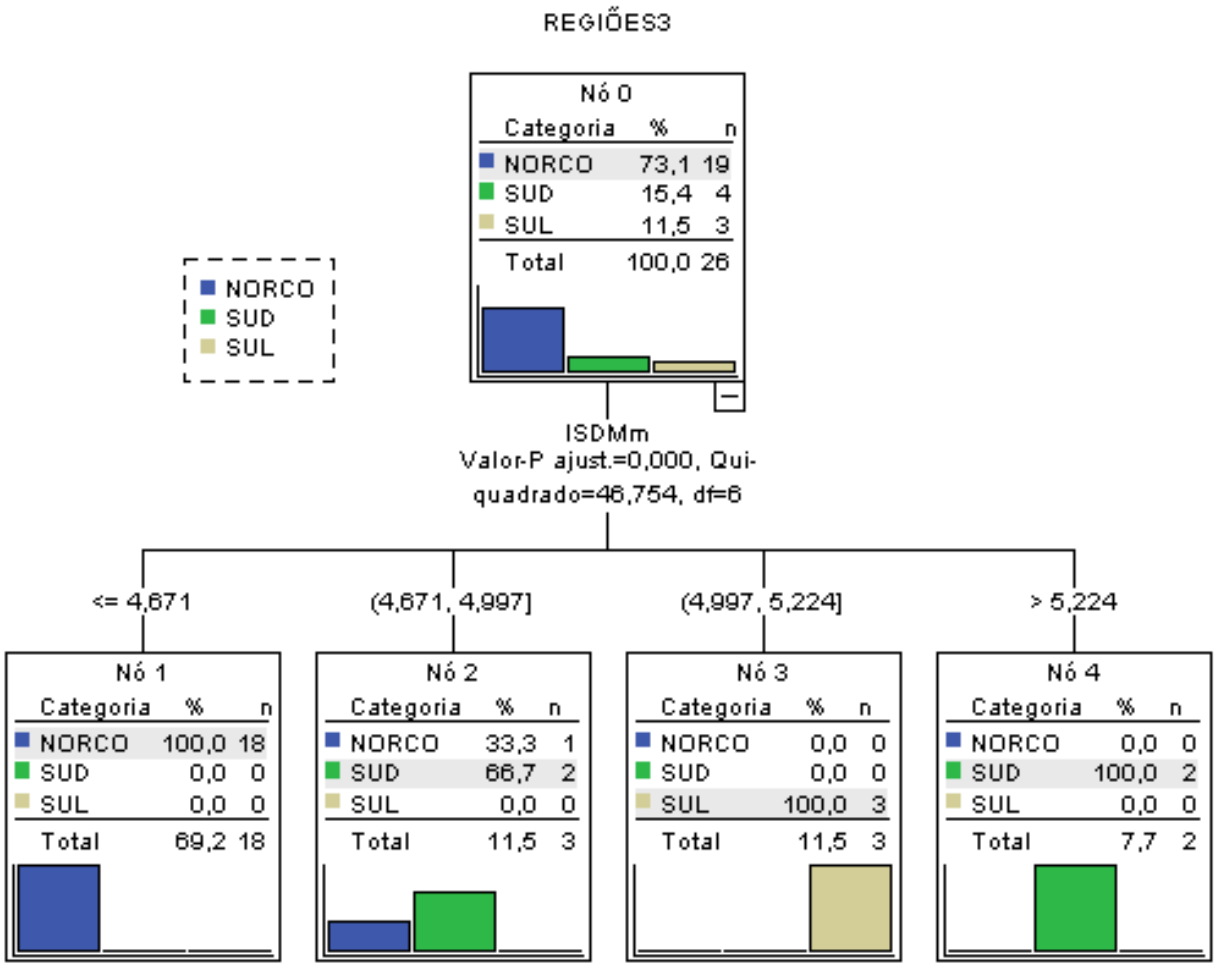
Resumo do modelo

Especificações	Método de crescimento	CHAID
	Variável dependente	REGIÕES3
	Variáveis independentes	ISDMsdn, ERsdn, IFGFsdn, LIQsdn, H6sdn, R1sdn, T1_1sdn, Ssdn, S1_1sdn, Esdn, E2_4sdn
	Validação	Nenhum
	Profundidade de árvore máxima	3
	Casos mínimos em nó pai	2
	Casos mínimos em nó filho	1
	Variáveis independentes incluídas	ISDMsdn, LIQsdn
Resultados	Número de nós	7
	Número de nós de terminal	5
	Profundidade	2

Esta árvore de classificação é um agrupamento dos 3 Brasis com as variáveis que mais se assemelham segundo os outros estudos. São elas: ISDMm, Sm, S1_1m.

Resumo do modelo

Especificações	Método de crescimento	CHAID
	Variável dependente	REGIÕES3
	Variáveis independentes	ISDMm, Sm, S1_1m
	Validação	Nenhum
	Profundidade de árvore máxima	3
	Casos mínimos em nó pai	2
	Casos mínimos em nó filho	1
	Variáveis independentes incluídas	ISDMm
Resultados	Número de nós	5
	Número de nós de terminal	4
	Profundidade	1



A árvore indica que existem 4 nós a partir de ISDMm. O primeiro nó representa os dados menores que 4,671, o segundo entre 4,671 a 4,997, o terceiro entre 4,997 e 5,224 e o último nó cujos valores são maiores que 5,224. O percentual de acerto é 96,2%.

Posto

Observado	Previsto			
	NORCO	SUD	SUL	Porcentagem Correta
NORCO	18	1	0	94,7%
SUD	0	4	0	100,0%
SUL	0	0	3	100,0%
Porcentagem global	69,2%	19,2%	11,5%	96,2%