Ex. Prático 10 para LCE105 e LCE1270
Ex. Prático 8 para LCE 136

IAI Não Supervisionada para Agrupamentos e Distâncias Multivariados – Cluster Analysis, seleção de variáveis preditoras, equalização e detecção de outliers. ANOVA, RobustANOVA, Box and Wisker Plot. DL: 13/10

Dados: Qualidade de Vida de Diferentes Categorias.

Categ	IMC	Movim	Kcal
AT	20,2	53,7	28??
AT	21,3	54,8	2700
AT	19,3	49,6	2800
AT	21,1	52,3	2900
AT	24,1	30,3	2700
SEM	22,4	14,9	2600
SEM	21,9	17,8	2700
SEM	23,8	18,6	3200
SEM	24,1	15,1	3300
SE	27,3	2,5	2700
SE	23,4	4,3	2300
SE	25,2	2,3	2600
SE	26,4	2,6	3200
PR	26,2	4,1	2600

PR	24,2	2,1	2700
PR	25,4	1,9	2650
PR	21,1	20	2650
PR	25,2	3,1	2650
PR	24,8	2	2675

Secuencia:

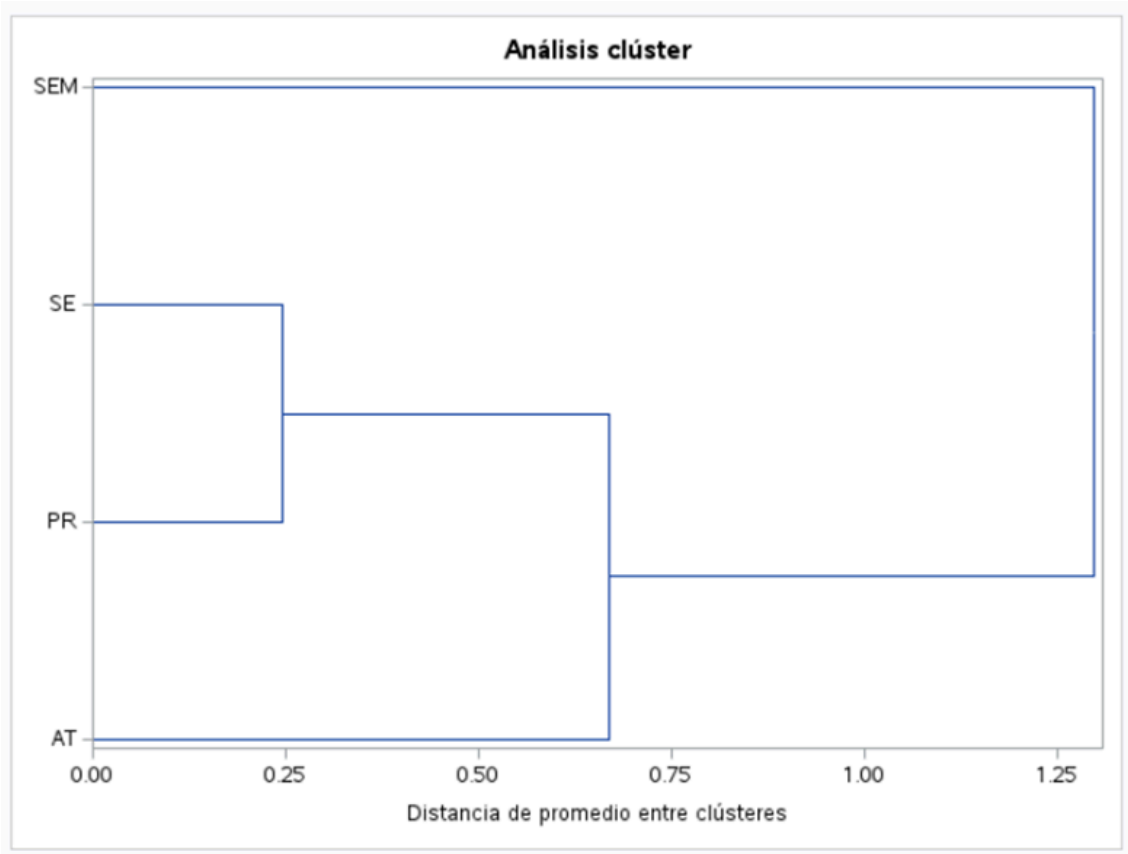
I - Fazer tabela dinâmica e rodar no SAS Cluster Analysis sem preocupação (aplicando ERADAMENTE) com outliers, com equalização (ordem de grandeza ou escala, todas as variáveis preditoras devem der o mesmo número de dígitos à esquerda da virgula, IMC tem 2 dígitos, Movimento 2 dígitos e quilocalorias 4 dígitos, deveríamos dividir essa última variável por 100) e significância das variáveis preditoras.

Programa de Cluster para dados sem tratamento

```
data qv;
input Cat $ IMC Movim Kcal;
datalines;
AT 21.2 48.14 2791.8
PR 24.48333333 5.533333333 2654.166667
SE 25.575 2.925 2700
SEM 23.05 16.6 2950
;
/* Fim do Data Step */
proc print;
run;
/* input Cat $ IMC Movim Kcal; */
proc cluster outtree = arvore method = average;
```

```
var IMC Movim Kcal;
id Cat;
run;
PROC TREE DATA = arvore;
RUN;
```

Dendrograma de Cluster Analysis Errado



A menor distancia aconteceu entre as categorias Professor e Sedentário, a maior entre Semi- atletas e as outras categorias. Resultado muito suspeito, que deveria se parecer com sedentários e professores deveria se a categoria Semialertas e não Atletas (IMC baixa e Movimento Alto).

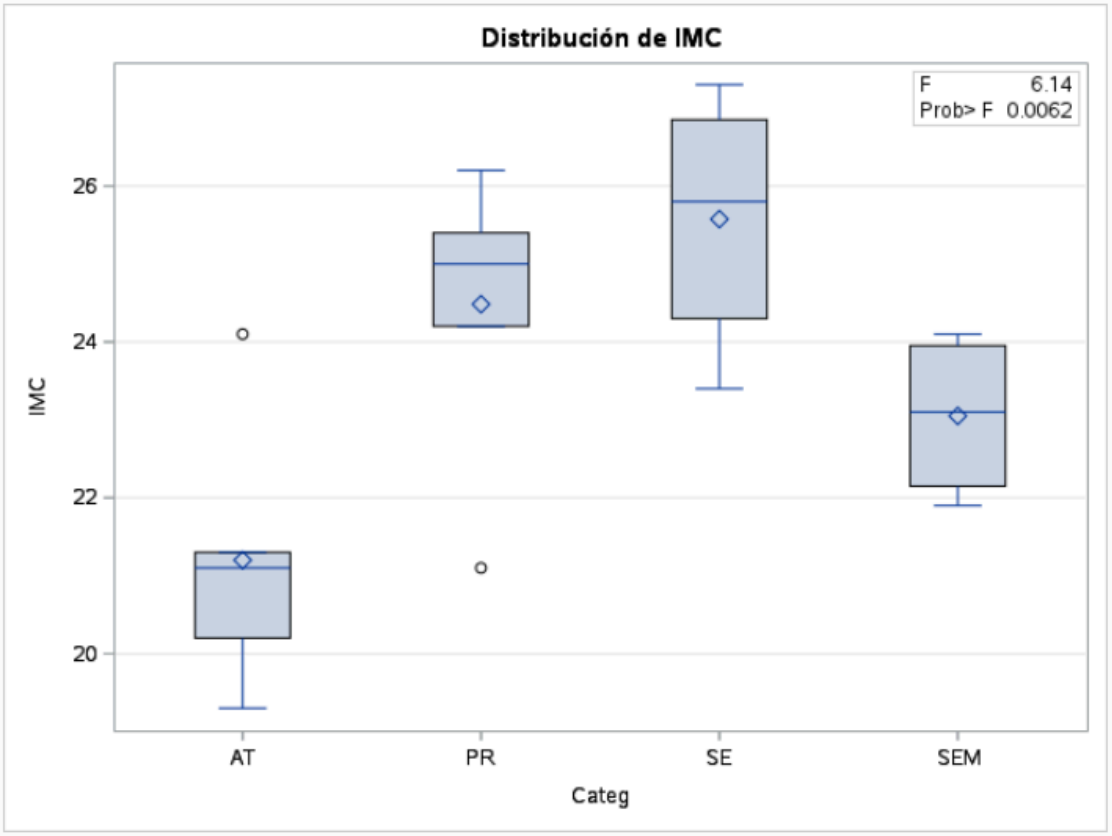
II - Eliminar os outliers, equalizar e excluir variáveis preditoras que não tem significância estatística.

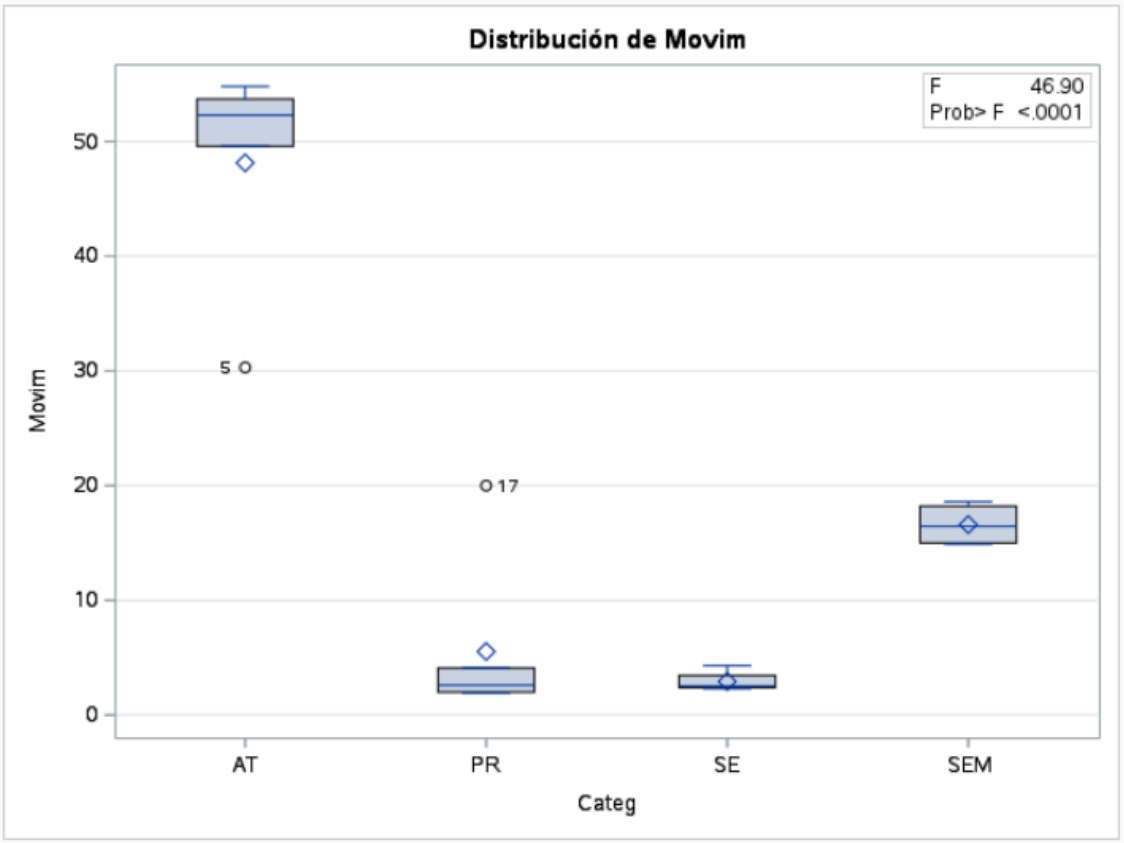
Rodar ANOVA, eliminar outliers. Programa. Eliminar variáveis preditoras não significativas (RobustANOVA). Equalizar dados, todas as variáveis preditoras na mesma escala.

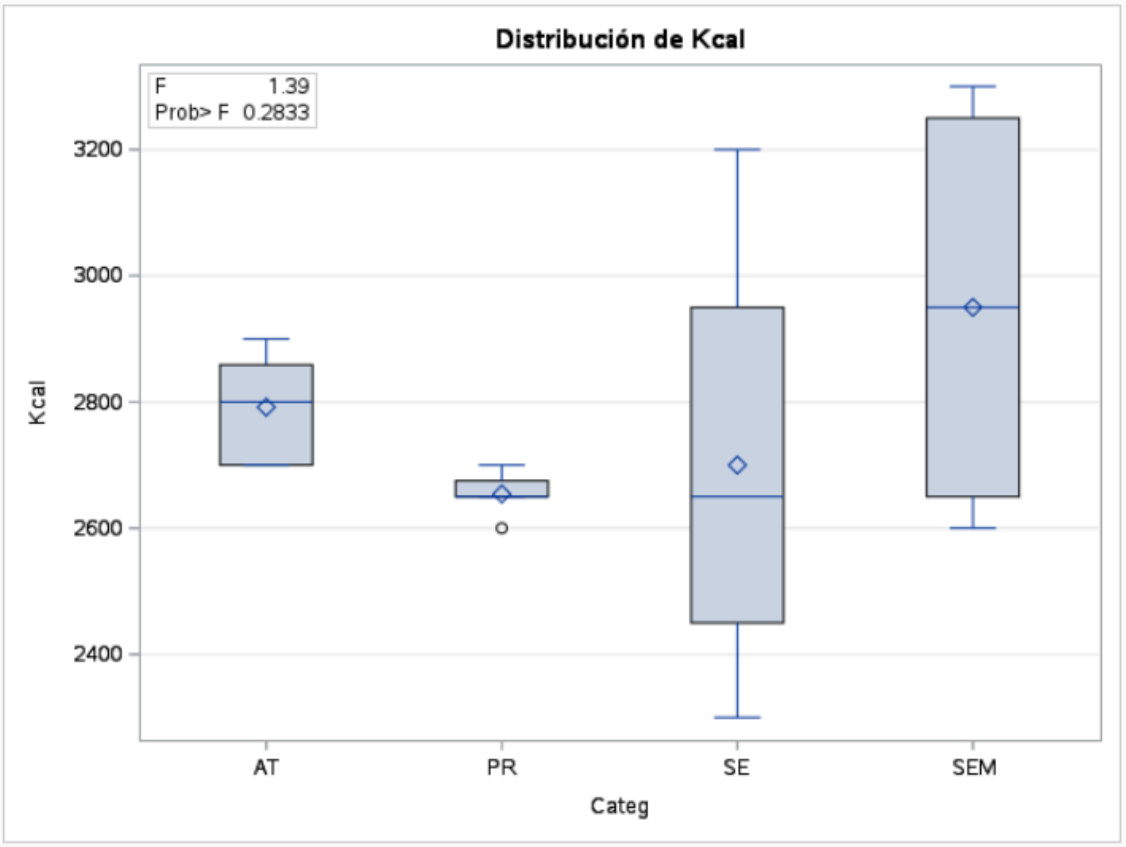
Programa para detectar outliers e primeira ideia de significância estatística das variáveis preditoras

```
data outlier;
nput Categ $ IMC Movim Kcal;
datalines;
AT 20.2 53.7 28 ??
AT 21.3 54.8 2700
AT 19.3 49.6 2800
AT 21.1 52.3 2900
AT 24.1 30.3 2700
SEM 22.4 14.9 2600
SEM 21.9 17.8 2700
SEM 23.8 18.6 3200
SEM 24.1 15.1 3300
SE 27.3 2.5 2700
SE 23.4 4.3 2300
SE 25.2 2.3 2600
SE 26.4 2.6 3200
PR 26.2 4.1 2600
PR 24.2 2.1 2700
PR 25.4 1.9 2650
PR 21.1 20 2650
PR 25.2 3.1 2650
PR 24.8 2 2675

proc print; run;
/* input Categ $ IMC Movim Kcal; */
proc anova;
  class Categ;
  model IMC Movim Kcal = Categ;
  means Categ / tukey lines;
run;
```


Distribución de IMC
F 6.14
Prob> F 0.0062
IMC
26
24
22
20
AT
PR
SE
SEM
Categ


Distribución de Movim
F 46.90
Prob> F <.0001
Movim
50
40
30
20
10
0
5
17
AT
PR
SE
SEM
Categ

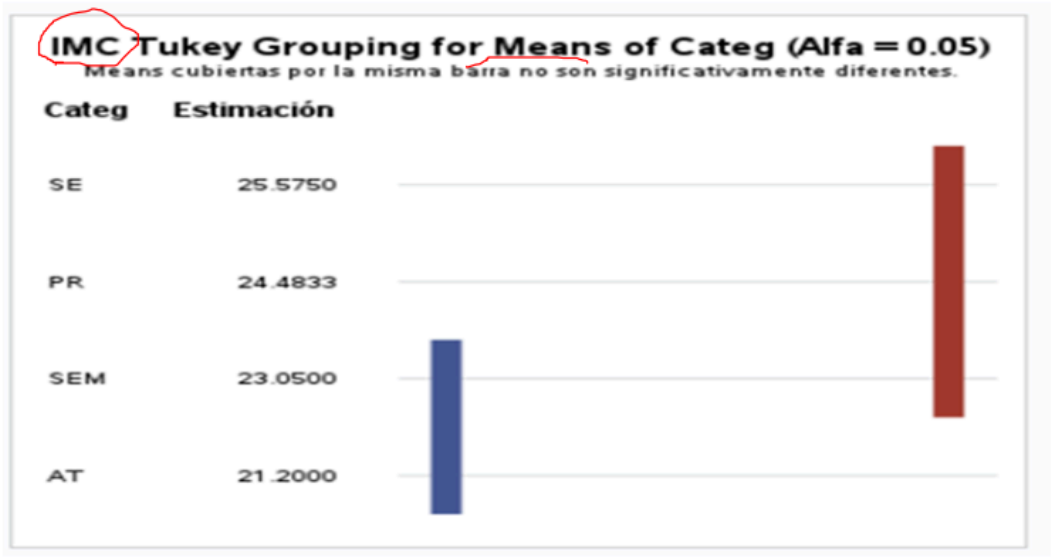


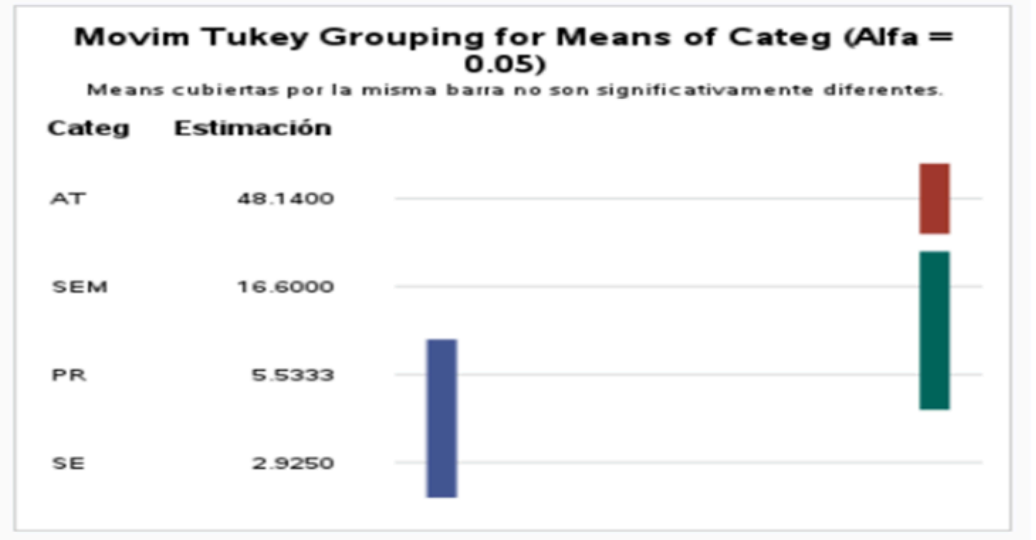
Podemos observar que aparecem duas bolinhas fora das caixas do Box and Wisker Plot (Gráfico de Caixa e Bigode), assim temos 2 outliers, que devemos eliminar para rodar a IA.

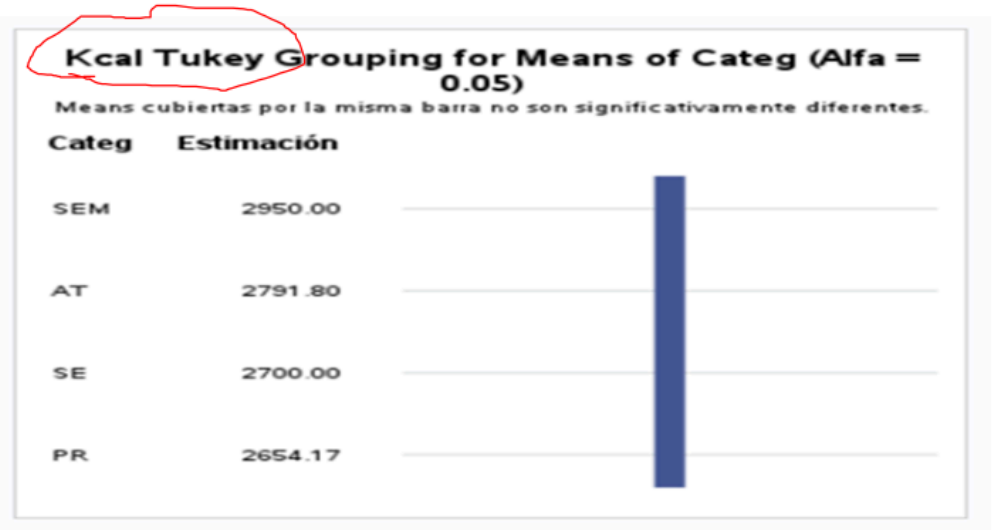
Temos que eliminar o ultimo atleta do banco de dados (Excel ou LO Calc)
Temos que **eliminar um professor com IMC menor do que 24, aproximadamente 21**.

Agora veremos os outliers do Movimento
Temos que eliminar um atleta com Movimento menor que 50, aproximadamente 30.
Podemos observar que esse atleta com Movimento 30 é o mesmo atleta com IMC 24,1, que já tínhamos eliminado na análise do IMC.
Agora temos que eliminar um professor com Movimento aproximadamente 20.
Vemos que já tínhamos eliminado esse professor, que 21,1 de IMC.

O Teste de Tukey dá uma ideia Visual de quais variáveis preditoras irão para a IA


IMC Tukey Grouping for Means of Categ (Alfa = 0.05)
Means cubiertas por la misma barra no son significativamente diferentes.
Categ
Estimación
SE
25.5750
PR
24.4833
SEM
23.0500
AT
21.2000


Movim Tukey Grouping for Means of Categ (Alfa = 0.05)
Means cubiertas por la misma barra no son significativamente diferentes.
Categ
Estimación
AT
48.1400
SEM
16.6000
PR
5.5333
SE
2.9250



O banco de dados ficará assim:

Reg	Categ	IMC	Movim	Kcal
1	AT	20,2	53,7	2859
2	AT	21,3	54,8	2700
3	AT	19,3	49,6	2800
4	AT	21,1	52,3	2900
5	AT	24,1	30,3	2700
6	SEM	22,4	14,9	2600
7	SEM	21,9	17,8	2700
8	SEM	23,8	18,6	3200
9	SEM	24,1	15,1	3300
10	SE	27,3	2,5	2700
11	SE	23,4	4,3	2300
12	SE	25,2	2,3	2600
13	SE	26,4	2,6	3200

14	PR	26,2	4,1	2600
15	PR	24,2	2,1	2700
16	PR	25,4	1,9	2650
17	PR	21,1	20	2650
18	PR	25,2	3,1	2650
19	PR	24,8	2	2675

Agora vamos testar quais variáveis preditoras devem ir para a IA

Utilizaremos RobustANOVA
Programa

```
data outlier;
input Categ $ IMC Movim Kcal;
datalines;
AT 20.2 53.7 28??
AT 21.3 54.8 2700
AT 19.3 49.6 2800
AT 21.1 52.3 2900
AT 24.1 30.3 2700
SEM 22.4 14.9 2600
SEM 21.9 17.8 2700
SEM 23.8 18.6 3200
SEM 24.1 15.1 3300
SE 27.3 2.5 2700
SE 23.4 4.3 2300
SE 25.2 2.3 2600
SE 26.4 2.6 3200
PR 26.2 4.1 2600
PR 24.2 2.1 2700
```

```
PR 25.4 1.9 2650
PR 21.1 20 2650
PR 25.2 3.1 2650
PR 24.8 2 2675
;
proc print; run;

/*
input Categ $ IMC Movim Kcal;
*/
Title "Robust ANOVA  Kuskal Wallis (um fator)";
proc npar1way wilcoxon dscf;
 class Categ;
 var IMC Movim Kcal;
run;
```

Procedimineto NPAR1WAY

Puntuaciones de Wilcoxon (Sumas de rango) para variable IMC
Clasificado por variable Categ

Categ	N	Suma de puntuaciones	Esperado debajo de H0	Desv. est. debajo de H0	Puntuación media
AT	5	22.00	50.0	10.787013	4.400000
SEM	4	32.50	40.0	9.986833	8.125000
SE	4	59.50	40.0	9.986833	14.875000
PR	6	76.00	60.0	11.386742	12.666667

Se utilizaron puntuaciones media para valores repetidos.

Test de Kruskal-Wallis		
Chi-cuadrado	DF	Pr > ChiSq
9.7707	3	0.0206

< 0,05

Vai para a IA

A margem de erro < 0,05, equivale a dizer que a margem de confiança é maior que 95%.

Neste caso temos 98% de confiança para falar que as categorias são diferentes pelo RobustANOVA, a melhor técnica do mundo para detectar diferenças entre categorias, sempre está certa, ANOVA em geral está sempre errada.

Puntuaciones de Wilcoxon (Sumas de rango) para variable Movim Clasificado por variable Categ					
Categ	N	Suma de puntuaciones	Esperado debajo de H0	Desv. est. debajo de H0	Puntuación media
AT	5	85.0	50.0	10.801234	17.000000
SEM	4	46.0	40.0	10.000000	11.500000
SE	4	24.0	40.0	10.000000	6.000000
PR	6	35.0	60.0	11.401754	5.833333

Test de Kruskal-Wallis		
Chi-cuadrado	DF	Pr > ChiSq
13.3316	3	0.0040

Puntuaciones de Wilcoxon (Sumas de rango) para variable Kcal Clasificado por variable Categ					
Categ	N	Suma de puntuaciones	Esperado debajo de H0	Desv. est. debajo de H0	Puntuación media
AT	5	67.00	50.0	10.662965	13.400000
SEM	4	50.50	40.0	9.871988	12.625000
SE	4	32.50	40.0	9.871988	8.125000
PR	6	40.00	60.0	11.255798	6.666667
Se utilizaron puntuaciones media para valores repetidos.					

Test de Kruskal-Wallis		
Chi-cuadrado	DF	Pr > ChiSq
5.3819	3	0.1459

Agora vamos eliminar os outliers do arquivo Excel ou L Office
Vejamos os box-plot da anova

Atleta com IMC aproximadamente 24
Professor com IMC entre 21 e 22

Seleção de Variáveis Preditoras

Devemos observar o p valor da Robust ANOVA

Como o p valor = 0,0206 é menor que 0,05, então a variável IMC vai para a IA.

Com 98% de confiança existe diferença estatisticamente significativa entre as categorias (AT SEM SE PR) para a variável IMC, assim a variável preditora IMC vai para a IA.

Com mais do que 99% de confiança a variável preditora Movimento vai para a IA. P valor < 0,004

Como a Margem de Erro (p valor = 0,1459) não foi menor que 0,05 então a variável preditora Kcal não vai para a IA.

Reflexão

Agora temos que rodar a IA, sem outliers, sem a variável Kcal e não faremos a equalização da variável Kcal por que essa variável será eliminada da IA.

Banco de Dados sem Outliers e sem Kcal

Registros	Categ	IMC	Movim
1	AT	20,2	53,??
2	AT	21,3	54,8
3	AT	19,3	49,6
4	AT	21,1	52,3
6	SEM	22,4	14,9
7	SEM	21,9	17,8
8	SEM	23,8	18,6
9	SEM	24,1	15,1
10	SE	27,3	2,5
11	SE	23,4	4,3
12	SE	25,2	2,3
13	SE	26,4	2,6
14	PR	26,2	4,1
15	PR	24,2	2,1

16	PR	25,4	1,9
18	PR	25,2	3,1
19	PR	24,8	2

Agora temos que calcular as medias aritméticas das categorias por tabela dinâmica

III Comparação dos dois Clusters

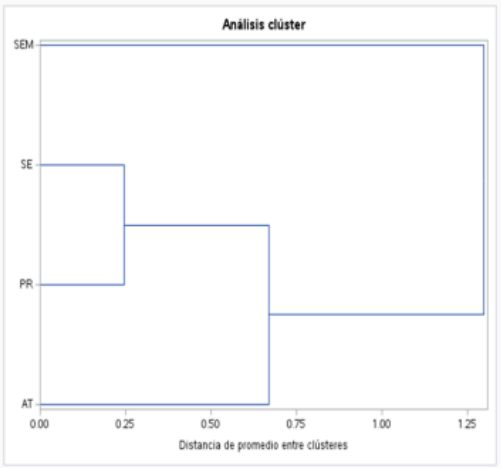
Programa de cluster com todos os pré-requisitos OK

```
data QV_OK;
input Categ $ IMC Movim;
datalines;
AT 20.475 52.6
PR 25.16 2.64
SE 25.575 2.925
SEM 23.05 16.6
;
/* fim do data step
  inicio do procedure step
*/
proc print;
run;
/* input Regis Categ $ IMC Movim; */
title "IAI Não Superv. para Agrupamentos - Cluster";
proc cluster outtree = arvore method = average;
var IMC Movim;
```

```
id Categ;
run;

PROC TREE DATA = arvore;
RUN;
```

Errado



Certo