Ex. Pratico 8 - IAI Não Supervisionada para Agrupamentos e Distâncias Multivariados, seleção de variáveis preditoras, equalização e detecção de outliers. ANOVA, RobustANOVA, Box and Wisker Plot. DL: 23/10

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

Sequencia:

I - Fazer tabela dinâmica e rodar no SAS Cluster Analysis sem preocupação com outliers, com equalização 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;

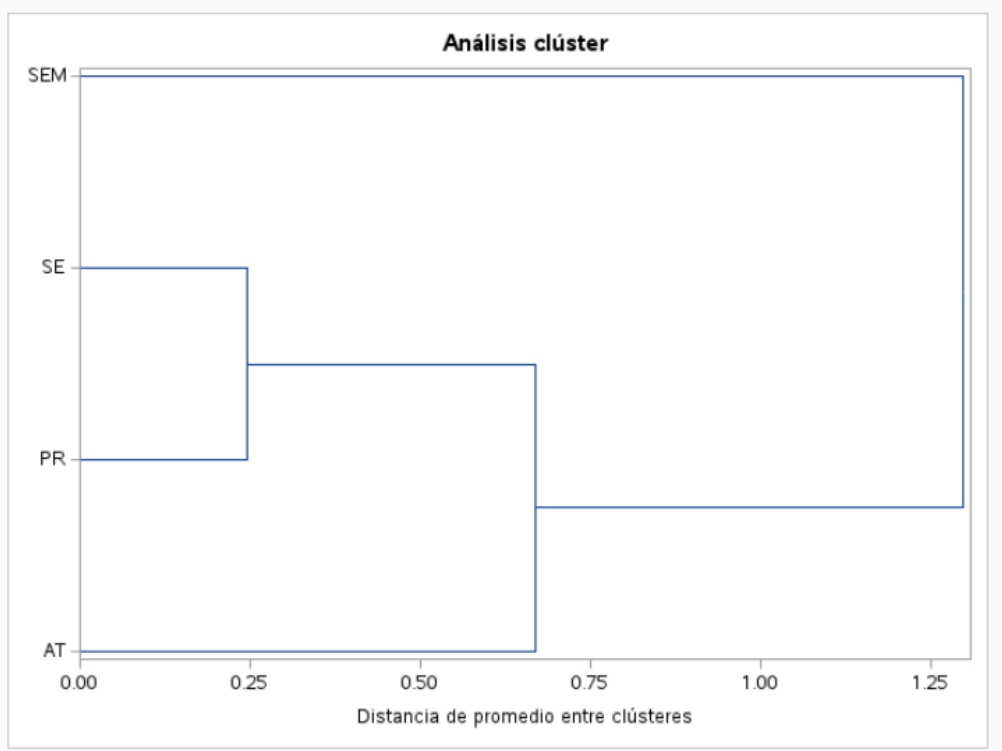
var IMC Movim Kcal;

id Cat;

run;

PROC TREE DATA = arvore;

RUN;



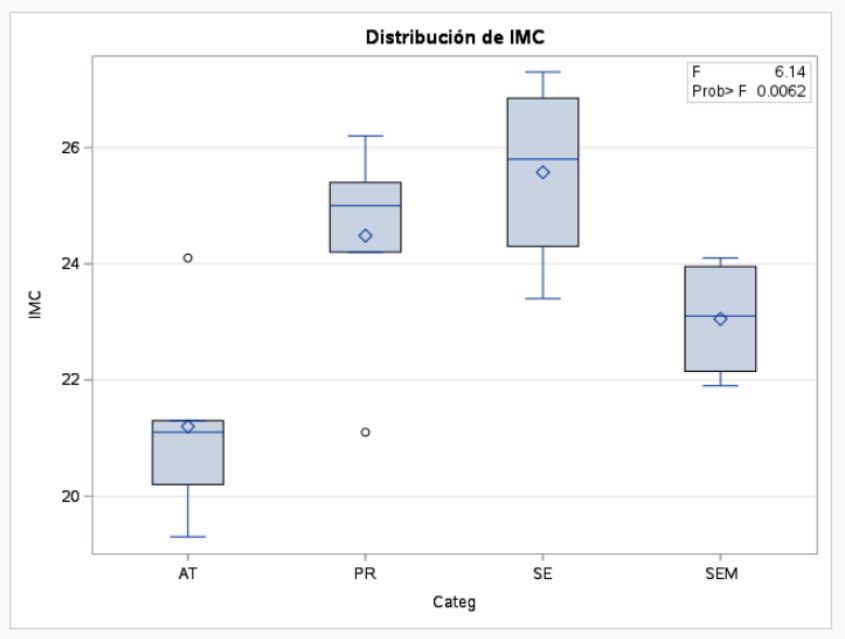
II - Eliminar os outlier, equalizar e excluir variáveis preditoras que não tem significância estatística.

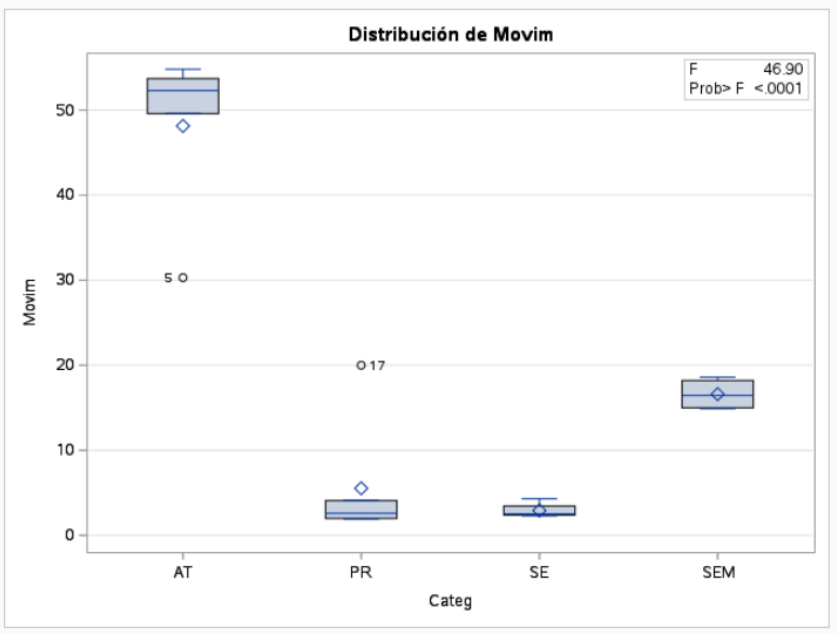
Rodar ANOVA, eliminar outliers. Programa. Eliminar variáveis preditoras não significativas (RobustANOVA). Equalizar dados, todas as variáveis preditoras na mesma escala.

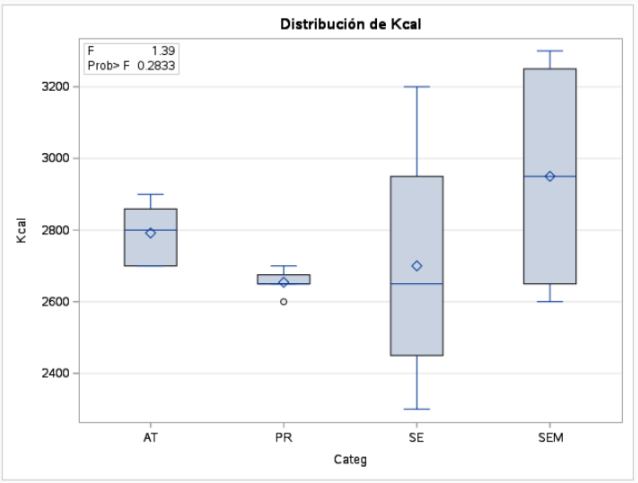
Programa para detectar outliers e primeira ideia de significância estatística das variáveis preditoras

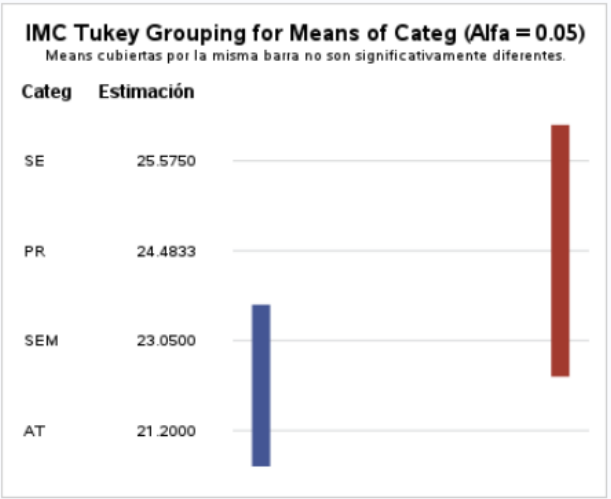
```
data outlier;

input Categ $ IMC Movim Kcal;

datalines;

AT 20.2 53.7 2859

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; */
```

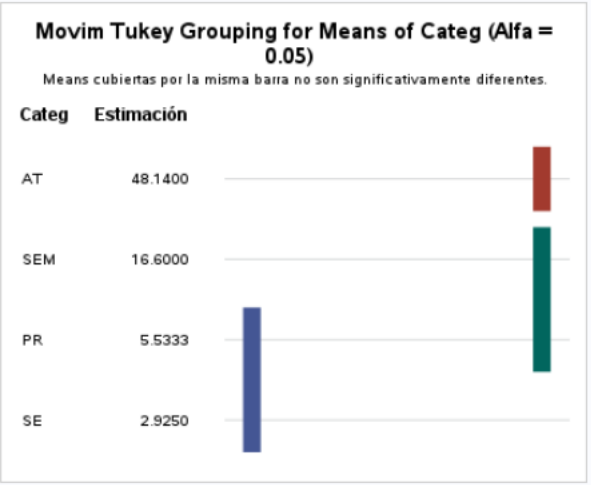
```
proc anova;
   class Categ;
   model IMC Movim Kcal = Categ;
   means Categ / tukey lines;
run;
```

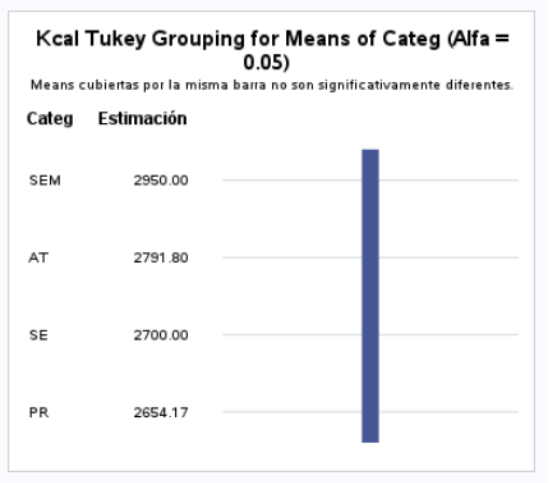





Distribución de Kcal
F 1.39
Prob> F 0.2833
3200
3000
2800
2600
2400
Kcal
AT
PR
SE
SEM
Categ


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



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 2859
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

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

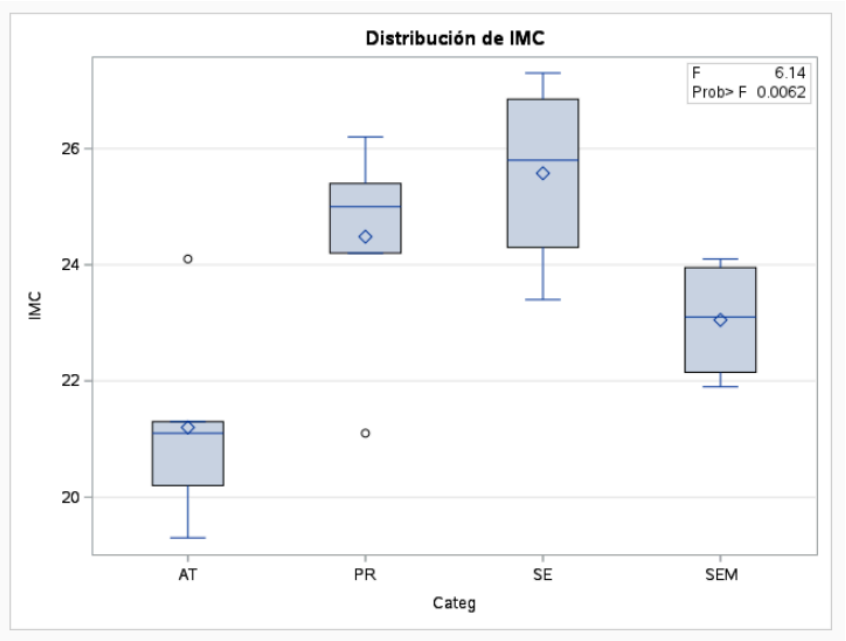
Test de Kruskal-Wallis		
Chi-cuadrado	DF	Pr > ChiSq
13.3316	3	0.0040

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

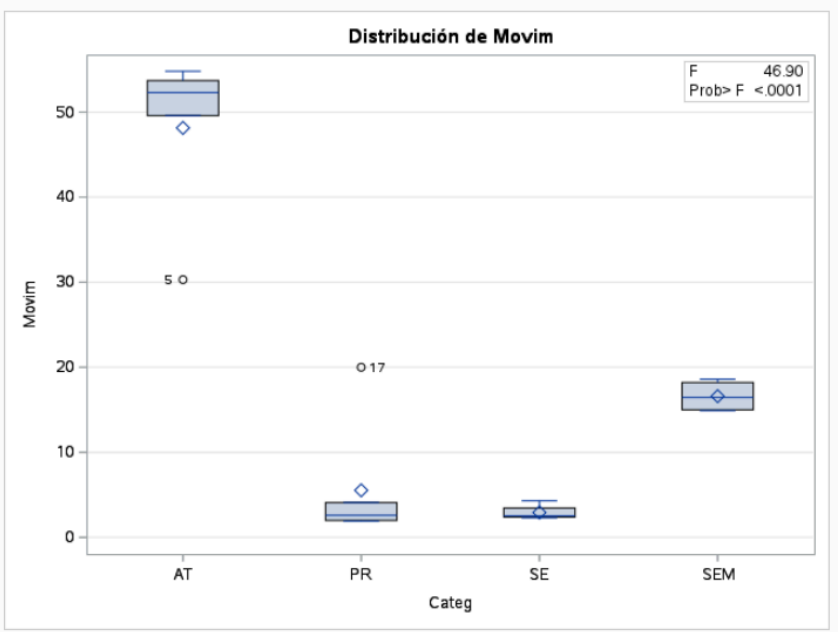


Atleta com IMC aproximadamente 24

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

Professor com IMC entre 21 e 22

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



Atleta com Movimento 30

AT	21,1	52,3	2900
AT	24,1	30,3	2700
SEM	22,4	14,9	2600

O mesmo atleta já eliminado;

Professor com movimento 20

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

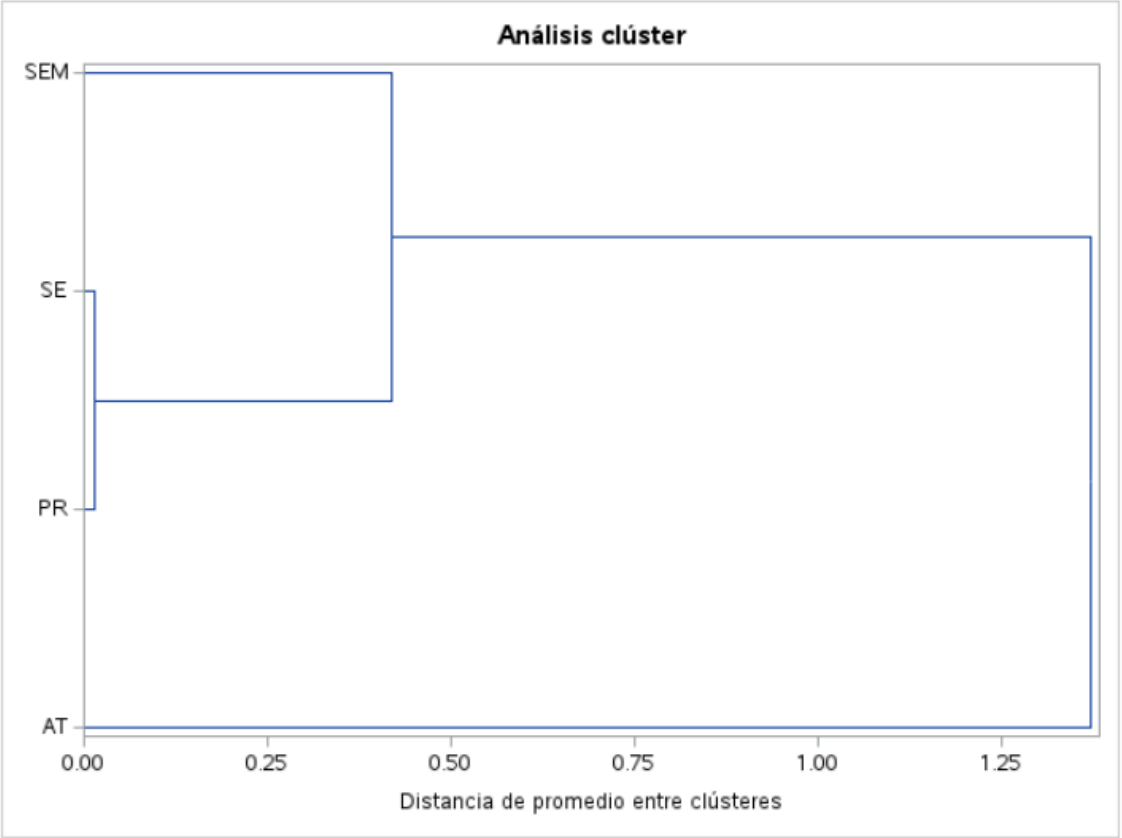
O mesmo professor já eliminado

Agora vamos rodar a IA, cluster analysis, com dados sem outliers e sem a variável Kcal.

Categ	IMC	Movim
AT	20,475	52,6
PR	25,16	2,64
SE	25,575	2,925
SEM	23,05	16,6

Programa cluster com esses dados

Cluster Correto – Sem outliers e sem Kcal



Podemos ver que os resultados do cluster são muito mais logicas

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

Gerar tabela do Excel sem outliers.

III Comparação dos dois Clusters

Programa de cluster com todos os pré-requisitos OK