TumoresSolidos

July 26, 2017

Rodrigo Toscano Ney
Relatório 1 - PESC - 19/07/2017

0.0.1 Objetivo :

O objetivo deste projeto é analisar a distribuição da base de dados de tumores sólidos pediátricos.

```
In [196]: from sklearn.decomposition import PCA
          from sklearn.cluster import KMeans
          import pandas as pd
          import numpy as np
          import matplotlib.pyplot as plt
          %matplotlib inline
```

0.1 Pré processamento dos dados de Tumores

São carregados dados dos marcadores:

```
In [197]: tumors = pd.read_csv('train/tumores_20_06_2017_no_comma.csv', encoding='latin1')
          tumors.sample(10)
```

Out[197]:

	CASE NUMBER	sample	DISEASE	GROUP	\
0	33	mass	Adrenal carcinoma	Adrenal carcinoma	
71	600	BM	neuroblastoma	disease free	
104	662	mass	neuroblastoma	Neuroblastic tumor	
113	1406	mass	neuroblastoma	Neuroblastic tumor	
148	1702	mass	renal carcinoma	renal carcinoma	
55	1513	ascitic fluid	germ cell tumor	germ cell tumor	
125	26	urine	rhabdomyosarcoma	rhabdomyosarcoma	
7	54	lymph node	Aplastic Lymphoma	no diagnosis	
65	3	mass formol	neuroblastoma	Neuroblastic tumor	
68	15	BM	neuroblastoma	disease free	

	Hematopoietico	Events:A	Visibility %:A	BV421:A	CD10:A	CD105:A	\
0	nao	29998.000	63.64	NaN	106.92	NaN	
71	livre de doenca	0.000	NaN	NaN	NaN	NaN	
104	nao	3086.000	21.81	NaN	93.64	NaN	
113	nao	1345.000	36.67	NaN	NaN	NaN	

```
148              nao   61011.000          76.57  8057.21  158.62     64.1
55               nao  194214.000          58.01   661.55     NaN      NaN
125              nao      17.192         100.00      NaN     NaN      NaN
7    sem diagnćstico       0.000            NaN      NaN     NaN      NaN
65               nao     681.000          55.77      NaN  240.61      NaN
68   livre de doenca       0.000            NaN      NaN     NaN      NaN

       ...       FSC-A:T  FSC-H:T  GD2:T  HLADR:T   MPO:T  nuMIOGENIN:T  \
0      ...      87120.90      NaN    NaN      NaN   97.69           NaN
71     ...      69413.40  55419.5    NaN      NaN  146.88           NaN
104    ...     101426.40  88832.0    NaN      NaN  154.09           NaN
113    ...      84809.27  73473.0    NaN      NaN     NaN           NaN
148    ...      40650.75  32704.5    NaN      NaN     NaN           NaN
55     ...      83919.29  71168.0  50.79      NaN     NaN          62.3
125    ...           NaN      NaN    NaN      NaN     NaN           NaN
7      ...     107319.59      NaN    NaN      NaN     NaN           NaN
65     ...      59264.21      NaN    NaN      NaN  132.74           NaN
68     ...           NaN      NaN    NaN      NaN     NaN           NaN

     nuMYOD:T  OCT3/4:T   SSC-A:T  SSC-H:T
0         NaN       NaN   8618.40      NaN
71        NaN       NaN   9876.29      NaN
104       NaN       NaN   6753.89      NaN
113       NaN       NaN  13599.36  10180.0
148       NaN       NaN  13425.55   8711.5
55        NaN       NaN   5901.12      NaN
125       NaN       NaN       NaN      NaN
7         NaN       NaN   8879.04      NaN
65        NaN       NaN   7586.60      NaN
68        NaN       NaN       NaN      NaN

[10 rows x 654 columns]
```

Pegamos do data frame apenas dados de tumores não hematopoieticos, inicialmente

```
In [198]: no_hematopoietic_tumors = tumors[tumors.Hematopoietico == 'nao']
          del no_hematopoietic_tumors['CASE NUMBER']
          no_hematopoietic_tumors.sample(10)
```

```
Out[198]:          sample                      DISEASE  \
          117        mass             neuroblastoma
          113        mass             neuroblastoma
          48         mass           germ cell tumor
          65   mass formol             neuroblastoma
          103  mass(posQT)             neuroblastoma
          41         mass      ganglioneuroblastoma
          144        mass               wilms tumor
          61         mass      mesoblastic nephroma
```

```
38          mass  Extraesqueletal Ewing Sarcoma
114           BM                   neuroblastoma

                            GROUP Hematopoietico   Events:A  Visibility %:A  \
117             Neuroblastic tumor            nao   410263.0           75.72
113             Neuroblastic tumor            nao     1345.0           36.67
48                 germ cell tumor            nao     1579.0           15.51
65              Neuroblastic tumor            nao      681.0           55.77
103             Neuroblastic tumor            nao     3029.0           21.68
41              Neuroblastic tumor            nao    44526.0           87.27
144                    wilms tumor            nao  1107266.0           95.98
61            mesoblastic nephroma            nao   846706.0           92.97
38   Extraesqueletal Ewing Sarcoma            nao    84801.0           86.78
114             Neuroblastic tumor            nao    26151.0            0.88

     BV421:A  CD10:A  CD105:A  CD117:A   ...       FSC-A:T  FSC-H:T   GD2:T  \
117   521.94     NaN  8964.97      NaN   ...      91791.16  78788.0  182.68
113      NaN     NaN      NaN      NaN   ...      84809.27  73473.0     NaN
48       NaN  104.77      NaN      NaN   ...      74623.83      NaN     NaN
65       NaN  240.61      NaN   180.28   ...      59264.21      NaN     NaN
103      NaN  163.56      NaN   685.33   ...     109197.00  95655.0     NaN
41       NaN -109.93      NaN   196.41   ...     116437.95      NaN     NaN
144  3368.45   95.16   414.91      NaN   ...      91755.30  76024.0   68.84
61       NaN   14.09      NaN    41.18   ...      95452.17  82698.0     NaN
38       NaN  125.86      NaN    51.56   ...      89457.91      NaN     NaN
114   491.91     NaN      NaN      NaN   ...      82067.40  72516.0   66.73

     HLADR:T     MPO:T  nuMIOGENIN:T  nuMYOD:T  OCT3/4:T   SSC-A:T  SSC-H:T
117      NaN       NaN        119.24       NaN       NaN   8004.00   5996.0
113      NaN       NaN           NaN       NaN       NaN  13599.36  10180.0
48       NaN       NaN           NaN       NaN       NaN   9463.68      NaN
65       NaN    132.74           NaN       NaN       NaN   7586.60      NaN
103      NaN    157.73           NaN       NaN       NaN   9049.04      NaN
41       NaN       NaN           NaN       NaN       NaN  14032.80      NaN
144      NaN       NaN         62.06       NaN       NaN   5850.72      NaN
61       NaN       NaN           NaN       NaN       NaN   9352.08      NaN
38       NaN  15731.41           NaN       NaN       NaN  10125.50      NaN
114      NaN       NaN        148.03       NaN       NaN   9820.80   7856.0

[10 rows x 653 columns]
```

In [199]:
```
no_hematopoietic_tumors.describe()
```

Out[199]:
```
           Events:A  Visibility %:A       BV421:A        CD10:A      CD105:A  \
count  7.900000e+01       79.000000     20.000000     56.000000    12.000000
mean   1.317935e+05       45.995570   7008.040000   1074.790536  1220.787500
std    2.788149e+05       36.649711  14664.139017   6027.646038  2512.201675
min    2.194000e+00        0.000000    122.900000   -849.260000    64.100000
```

```
25%    2.676000e+03        8.600000    572.145000     72.117500   117.245000
50%    1.886300e+04       36.670000   1079.565000    157.965000   422.600000
75%    1.284760e+05       86.105000   3449.897500    346.790000   710.630000
max    1.552006e+06      100.000000  58332.410000  45213.890000  8964.970000

           CD117:A       CD123:A       CD14:A        CD15:A       CD16:A  \
count    43.000000     20.000000    23.000000      4.000000    19.000000
mean    367.994186   2679.300000   820.174348   6141.592500   712.970526
std     723.874660   4392.700064  1087.148069   6979.418317   705.658442
min    -107.780000     91.640000  -462.970000    107.240000   171.430000
25%      74.590000    316.022500    16.960000    126.282500   274.250000
50%     156.740000   1079.370000   442.180000   5777.960000   353.160000
75%     301.635000   1894.635000  1171.180000  11793.270000   832.665000
max    3747.630000  16884.920000  3599.610000  12903.210000  2593.010000

           ...             FSC-A:T        FSC-H:T        GD2:T      HLADR:T  \
count      ...           67.000000      51.000000    27.000000    14.000000
mean       ...        83756.114179   74307.735294   273.694074   429.749286
std        ...        18359.825414   13937.102350   586.276938   367.661711
min        ...        32801.850000   32704.500000    50.790000   111.520000
25%        ...        72724.480000   66410.000000    91.735000   228.182500
50%        ...        86071.450000   75684.500000   174.550000   349.535000
75%        ...        92395.660000   81894.750000   217.850000   471.137500
max        ...       119606.310000  104300.000000  3181.200000  1583.560000

              MPO:T  nuMIOGENIN:T  nuMYOD:T  OCT3/4:T       SSC-A:T  \
count     25.000000     14.000000      1.00       0.0     65.000000
mean     757.073200    144.993571    206.16       NaN   9538.436462
std     3119.922326     98.064005       NaN       NaN   2828.509136
min       66.070000     43.470000    206.16       NaN   4304.300000
25%       98.940000     64.807500    206.16       NaN   7508.160000
50%      132.440000    135.120000    206.16       NaN   9340.900000
75%      163.410000    175.635000    206.16       NaN  10920.960000
max    15731.410000    405.970000    206.16       NaN  16264.800000

           SSC-H:T
count     18.000000
mean    7188.027778
std     1621.013340
min     4950.000000
25%     5617.625000
50%     7653.500000
75%     8347.500000
max    10180.000000

[8 rows x 649 columns]
```

Listamos quais são os grupos de tumores que temos dados

```
In [200]: for key in no_hematopoietic_tumors.GROUP.unique():
              print key
```

```
Adrenal carcinoma
chondrosarcoma
cystic nephroma
Extraesqueletal Ewing Sarcoma
Neuroblastic tumor
germ cell tumor
vascular tumor
mesoblastic nephroma
sopharyngeal carcinoma
normal kidney
rhabdomyosarcoma
thyroid-like follicular carcinoma of kidney
undiferenciated malignt neoplasm
wilms tumor
osteosarcoma
renal carcinoma
```

Fazemos então uma breve analise das colunas de marcadores, para descobrir quantos dados temos e quais são seus comportamentos estatísticos para cada grupo de tumor.

```
In [201]: string_columns = ['sample', 'DISEASE', 'GROUP', 'Hematopoietico']
          groups_statistics_list = []

          for key in no_hematopoietic_tumors.GROUP.unique():
              grouped_tumor = no_hematopoietic_tumors[no_hematopoietic_tumors.GROUP == key]
              df_group_result = pd.DataFrame()

              cur_index = 0
              for column in grouped_tumor:
                  if column not in string_columns and 'Visibility' not in column and 'Events' n
                      my_group_colum = pd.to_numeric(grouped_tumor[column])
                      qt_data = grouped_tumor[column].size - grouped_tumor[column].isnull().sum
                      qt_nulls = grouped_tumor[column].isnull().sum()
                      mean = 0
                      variance = 0
                      c_min = None
                      c_max = None
                      if qt_data > 0:
                          mean = my_group_colum.mean()
                          c_min = my_group_colum.min()
                          c_max = my_group_colum.max()
                          if qt_data > 1:
                              variance = my_group_colum.var()
```

```
                    data = pd.DataFrame({"grupo": key,
                                         "linhas": my_group_colum.size,
                                         "marcador": column,
                                         "dados": qt_data,
                                         "nulos": qt_nulls,
                                         "min": c_min,
                                         "max": c_max,
                                         "media": mean,
                                         "variancia": variance},
                                        index=[cur_index],
                                        columns=["grupo", "linhas", "marcador", "dados", "nul
                    df_group_result = df_group_result.append(data)
                    cur_index += 1
                df_group_result.to_csv('result/' + str(key).replace(" ", "_").lower() + '_statist
                groups_statistics_list.append(df_group_result)
                print(df_group_result.sample(10))
```

```
                grupo  linhas      marcador  dados  nulos      min      max  \
244  Adrenal carcinoma       1        CD22:L      0      1     None     None
425  Adrenal carcinoma       1        CD44:N      0      1     None     None
10   Adrenal carcinoma       1  CD19+TCRGD:A      0      1     None     None
393  Adrenal carcinoma       1        MPO:MO      1      0  7081.26  7081.26
570  Adrenal carcinoma       1       BV421:T      0      1     None     None
349  Adrenal carcinoma       1       CD16:MO      0      1     None     None
549  Adrenal carcinoma       1    CD8+IgL:RC      0      1     None     None
590  Adrenal carcinoma       1         CD3:T      1      0  20047.1  20047.1
77   Adrenal carcinoma       1         CD3:B      1      0     1.77     1.77
581  Adrenal carcinoma       1        CD1a:T      0      1     None     None

        media  variancia
244      0.00          0
425      0.00          0
10       0.00          0
393   7081.26          0
570      0.00          0
349      0.00          0
549      0.00          0
590  20047.11          0
77       1.77          0
581      0.00          0
              grupo  linhas       marcador  dados  nulos      min      max  \
352  chondrosarcoma       1  CD19+TCRGD:MO      0      1     None     None
480  chondrosarcoma       1        CD38:NK      1      0    64.31    64.31
99   chondrosarcoma       1        cyCD3:B      0      1     None     None
41   chondrosarcoma       1         CD99:A      1      0   369.04   369.04
228  chondrosarcoma       1        BV421:L      0      1     None     None
15   chondrosarcoma       1         CD21:A      0      1     None     None
```

```
327  chondrosarcoma       1      cyCD3:MSC      0      1     None     None
44   chondrosarcoma       1      cyCD79a:A      0      1     None     None
511  chondrosarcoma       1       SSC-A:NK      1      0  15578.4  15578.4
284  chondrosarcoma       1        SSC-H:L      0      1     None     None

        media  variancia
352      0.00          0
480     64.31          0
99       0.00          0
41     369.04          0
228      0.00          0
15       0.00          0
327      0.00          0
44       0.00          0
511  15578.42          0
284      0.00          0
               grupo  linhas  marcador  dados  nulos     min     max  \
51   cystic nephroma       1     MPO:A      0      1    None    None
70   cystic nephroma       1    CD20:B      0      1    None    None
601  cystic nephroma       1    CD57:T      0      1    None    None
226  cystic nephroma       1   SSC-A:E      1      0  170453  170453
278  cystic nephroma       1   HLADR:L      1      0  740.76  740.76
155  cystic nephroma       1  CD99:END      0      1    None    None
154  cystic nephroma       1  CD90:END      0      1    None    None
368  cystic nephroma       1   CD44:MO      1      0    2756    2756
432  cystic nephroma       1     CD7:N      0      1    None    None
405  cystic nephroma       1    CD15:N      0      1    None    None

         media  variancia
51        0.00          0
70        0.00          0
601       0.00          0
226  170452.81          0
278     740.76          0
155       0.00          0
154       0.00          0
368    2756.00          0
432       0.00          0
405       0.00          0
                            grupo  linhas      marcador  dados  nulos  \
7    Extraesqueletal Ewing Sarcoma       3        CD16:A      1      2
613  Extraesqueletal Ewing Sarcoma       3  cyCD3+CD19:T      0      3
324  Extraesqueletal Ewing Sarcoma       3       CD9:MSC      0      3
177  Extraesqueletal Ewing Sarcoma       3        CD15:E      0      3
616  Extraesqueletal Ewing Sarcoma       3   Epcam+CD4:T      0      3
612  Extraesqueletal Ewing Sarcoma       3       cyCD3:T      2      1
56   Extraesqueletal Ewing Sarcoma       3       SSC-H:A      1      2
497  Extraesqueletal Ewing Sarcoma       3       CD99:NK      3      0
```

```
493  Extraesqueletal Ewing Sarcoma       3       CD81:NK      3      0
354  Extraesqueletal Ewing Sarcoma       3        CD2:MO      0      3

         min      max         media      variancia
7     303.93   303.93    303.930000       0.000000
613     None     None      0.000000       0.000000
324     None     None      0.000000       0.000000
177     None     None      0.000000       0.000000
616     None     None      0.000000       0.000000
612  4951.74  5174.03   5062.885000   24706.422050
56     15986    15986  15986.000000       0.000000
497    95.04   553.89    333.066667   52857.796633
493 1014.71  2769.94   1745.053333  835274.919633
354     None     None      0.000000       0.000000
                 grupo  linhas      marcador  dados  nulos      min      max  \
528  Neuroblastic tumor      36       CD21:RC      0     36     None     None
402  Neuroblastic tumor      36       CD117:N     17     19   126.32  1498.69
323  Neuroblastic tumor      36  CD8-CD99:MSC      1     35  1874.08  1874.08
49   Neuroblastic tumor      36         GD2:A     32      4   403.16   121184
36   Neuroblastic tumor      36     CD8+IgL:A     20     16  -124.09  1294.38
165  Neuroblastic tumor      36       MPO:END      4     32   218.03   342.19
130  Neuroblastic tumor      36      CD22:END      0     36     None     None
325  Neuroblastic tumor      36      CD90:MSC      0     36     None     None
178  Neuroblastic tumor      36        CD16:E      6     30  1790.17   8598.9
287  Neuroblastic tumor      36     CD105:MSC      0     36     None     None

            media     variancia
528      0.000000  0.000000e+00
402    371.945294  1.247825e+05
323   1874.080000  0.000000e+00
49   19888.836250  8.268694e+08
36     375.546500  1.492373e+05
165    277.580000  4.473167e+03
130      0.000000  0.000000e+00
325      0.000000  0.000000e+00
178   4623.618333  7.179924e+06
287      0.000000  0.000000e+00
               grupo  linhas      marcador  dados  nulos     min     max  \
19   germ cell tumor      10        CD28:A      0     10    None    None
457  germ cell tumor      10       CD10:NK      2      8   13.81  130.76
172  germ cell tumor      10        CD10:E      0     10    None    None
538  germ cell tumor      10        CD4:RC      0     10    None    None
605  germ cell tumor      10         CD8:T      0     10    None    None
580  germ cell tumor      10  CD19+TCRGD:T      2      8   44.01   135.6
184  germ cell tumor      10        CD20:E      0     10    None    None
127  germ cell tumor      10      CD20:END      1      9  666.76  666.76
372  germ cell tumor      10   CD56+IgK:MO      3      7  123.89  786.51
115  germ cell tumor      10      CD10:END      2      8 -238.26  963.02
```

```
          media      variancia
19     0.000000       0.000000
457   72.285000    6838.651250
172    0.000000       0.000000
538    0.000000       0.000000
605    0.000000       0.000000
580   89.805000    4194.364050
184    0.000000       0.000000
127  666.760000       0.000000
372  470.506667  110469.198233
115  362.380000  721536.819200
                grupo  linhas    marcador  dados  nulos     min     max   media  \
365  vascular tumor       2     CD34:MO      1      1  413.65  413.65  413.65
578  vascular tumor       2      CD19:T      0      2    None    None    0.00
180  vascular tumor       2  CD19_CD4:E      0      2    None    None    0.00
156  vascular tumor       2   cyCD3:END      0      2    None    None    0.00
316  vascular tumor       2    CD57:MSC      0      2    None    None    0.00
554  vascular tumor       2     CD99:RC      0      2    None    None    0.00
484  vascular tumor       2      CD5:NK      0      2    None    None    0.00
54   vascular tumor       2    OCT3/4:A      0      2    None    None    0.00
279  vascular tumor       2       MPO:L      0      2    None    None    0.00
434  vascular tumor       2       CD8:N      0      2    None    None    0.00

     variancia
365        0.0
578        0.0
180        0.0
156        0.0
316        0.0
554        0.0
484        0.0
54         0.0
279        0.0
434        0.0
                    grupo  linhas     marcador  dados  nulos      min  \
411  mesoblastic nephroma       1        CD2:N      0      1     None
116  mesoblastic nephroma       1    CD105:END      0      1     None
418  mesoblastic nephroma       1       CD28:N      0      1     None
377  mesoblastic nephroma       1       CD8:MO      0      1     None
131  mesoblastic nephroma       1     CD24:END      0      1     None
480  mesoblastic nephroma       1      CD38:NK      1      0  2031.59
543  mesoblastic nephroma       1  CD56+IgK:RC      0      1     None
540  mesoblastic nephroma       1      CD45:RC      0      1     None
349  mesoblastic nephroma       1      CD16:MO      0      1     None
479  mesoblastic nephroma       1      CD34:NK      0      1     None

         max    media  variancia
```

```
411     None     0.00          0
116     None     0.00          0
418     None     0.00          0
377     None     0.00          0
131     None     0.00          0
480  2031.59  2031.59          0
543     None     0.00          0
540     None     0.00          0
349     None     0.00          0
479     None     0.00          0
                     grupo  linhas  marcador  dados  nulos      min      max  \
298  sopharyngeal carcinoma       1  CD20:MSC      0      1     None     None
412  sopharyngeal carcinoma       1    CD20:N      1      0   555.98   555.98
286  sopharyngeal carcinoma       1  CD10:MSC      0      1     None     None
228  sopharyngeal carcinoma       1   BV421:L      0      1     None     None
198  sopharyngeal carcinoma       1    CD45:E      1      0  6742.01  6742.01
23   sopharyngeal carcinoma       1    CD34:A      0      1     None     None
7    sopharyngeal carcinoma       1    CD16:A      0      1     None     None
70   sopharyngeal carcinoma       1    CD20:B      0      1     None     None
444  sopharyngeal carcinoma       1   EPCAM:N      1      0   224.61   224.61
131  sopharyngeal carcinoma       1  CD24:END      0      1     None     None

       media  variancia
298     0.00          0
412   555.98          0
286     0.00          0
228     0.00          0
198  6742.01          0
23      0.00          0
7       0.00          0
70      0.00          0
444   224.61          0
131     0.00          0
            grupo  linhas      marcador  dados  nulos      min      max  \
396  normal kidney       1     OCT3/4:MO      0      1     None     None
210  normal kidney       1         CD9:E      0      1     None     None
67   normal kidney       1  CD19+TCRGD:B      0      1     None     None
199  normal kidney       1         CD5:E      0      1     None     None
172  normal kidney       1        CD10:E      0      1     None     None
525  normal kidney       1        CD2:RC      0      1     None     None
161  normal kidney       1     FSC-A:END      0      1     None     None
590  normal kidney       1         CD3:T      1      0    23391    23391
541  normal kidney       1        CD5:RC      0      1     None     None
480  normal kidney       1       CD38:NK      1      0  5388.45  5388.45

       media  variancia
396     0.00          0
210     0.00          0
```

```
67       0.00          0
199      0.00          0
172      0.00          0
525      0.00          0
161      0.00          0
590  23391.03          0
541      0.00          0
480   5388.45          0
                grupo  linhas      marcador  dados  nulos      min      max  \
271  rhabdomyosarcoma       7  cyCD3+CD19:L      0      7     None     None
93   rhabdomyosarcoma       7     CD8+IgL:B      0      7     None     None
338  rhabdomyosarcoma       7    nuMYOD:MSC      0      7     None     None
430  rhabdomyosarcoma       7        CD57:N      3      4    93.49   175.72
424  rhabdomyosarcoma       7         CD4:N      1      6  1099.17  1099.17
582  rhabdomyosarcoma       7         CD2:T      0      7     None     None
206  rhabdomyosarcoma       7         CD8:E      1      6   335.42   335.42
245  rhabdomyosarcoma       7        CD24:L      0      7     None     None
487  rhabdomyosarcoma       7       CD57:NK      0      7     None     None
78   rhabdomyosarcoma       7        CD30:B      0      7     None     None

       media  variancia
271     0.00     0.0000
93      0.00     0.0000
338     0.00     0.0000
430   140.39  1790.8419
424  1099.17     0.0000
582     0.00     0.0000
206   335.42     0.0000
245     0.00     0.0000
487     0.00     0.0000
78      0.00     0.0000
                                          grupo  linhas        marcador  \
259  thyroid-like follicular carcinoma of kidney       1          CD57:L
337  thyroid-like follicular carcinoma of kidney       1  nuMIOGENIN:MSC
90   thyroid-like follicular carcinoma of kidney       1           CD7:B
521  thyroid-like follicular carcinoma of kidney       1         CD19:RC
4    thyroid-like follicular carcinoma of kidney       1         CD123:A
319  thyroid-like follicular carcinoma of kidney       1        CD71:MSC
274  thyroid-like follicular carcinoma of kidney       1     Epcam+CD4:L
404  thyroid-like follicular carcinoma of kidney       1          CD14:N
500  thyroid-like follicular carcinoma of kidney       1      cyCD79a:NK
103  thyroid-like follicular carcinoma of kidney       1     Epcam+CD4:B

     dados  nulos      min      max    media  variancia
259      0      1     None     None     0.00          0
337      0      1     None     None     0.00          0
90       0      1     None     None     0.00          0
521      0      1     None     None     0.00          0
```

```
4        0      1     None     None     0.00          0
319      0      1     None     None     0.00          0
274      1      0  2591.53  2591.53  2591.53          0
404      0      1     None     None     0.00          0
500      0      1     None     None     0.00          0
103      1      0   290.38   290.38   290.38          0
                                 grupo  linhas    marcador  dados  nulos  \
313  undiferenciated malignt neoplasm       2     CD5:MSC      0      2
320  undiferenciated malignt neoplasm       2     CD8:MSC      0      2
512  undiferenciated malignt neoplasm       2    SSC-H:NK      1      1
359  undiferenciated malignt neoplasm       2     CD24:MO      0      2
558  undiferenciated malignt neoplasm       2    EPCAM:RC      0      2
566  undiferenciated malignt neoplasm       2   nuMYOD:RC      0      2
567  undiferenciated malignt neoplasm       2   OCT3/4:RC      0      2
408  undiferenciated malignt neoplasm       2  CD19_CD4:N      0      2
5    undiferenciated malignt neoplasm       2      CD14:A      1      1
233  undiferenciated malignt neoplasm       2      CD14:L      1      1

        min     max    media  variancia
313    None    None     0.00        0.0
320    None    None     0.00        0.0
512    9528    9528  9528.00        0.0
359    None    None     0.00        0.0
558    None    None     0.00        0.0
566    None    None     0.00        0.0
567    None    None     0.00        0.0
408    None    None     0.00        0.0
5    116.92  116.92   116.92        0.0
233   -2.55   -2.55    -2.55        0.0
           grupo  linhas     marcador  dados  nulos      min      max  \
0    wilms tumor      10      BV421:A      1      9  3368.45  3368.45
281  wilms tumor      10     nuMYOD:L      4      6    68.93   159.51
370  wilms tumor      10       CD5:MO      2      8   286.44    764.1
325  wilms tumor      10     CD90:MSC      0     10     None     None
274  wilms tumor      10  Epcam+CD4:L      1      9   118.61   118.61
81   wilms tumor      10       CD38:B      2      8   420.73  1165.71
494  wilms tumor      10  CD8-CD99:NK      0     10     None     None
510  wilms tumor      10    OCT3/4:NK      0     10     None     None
578  wilms tumor      10       CD19:T      0     10     None     None
149  wilms tumor      10      CD8:END      0     10     None     None

         media      variancia
0    3368.4500       0.000000
281   113.6125    1413.378158
370   525.2700  114079.537800
325     0.0000       0.000000
274   118.6100       0.000000
81    793.2200  277497.600200
```

```
494     0.0000       0.000000
510     0.0000       0.000000
578     0.0000       0.000000
149     0.0000       0.000000
           grupo  linhas      marcador  dados  nulos    min    max  media  \
545  osteosarcoma       1       CD58:RC      0      1   None   None   0.00
531  osteosarcoma       1      CD271:RC      0      1   None   None   0.00
117  osteosarcoma       1     CD117:END      0      1   None   None   0.00
351  osteosarcoma       1   CD19_CD4:MO      0      1   None   None   0.00
505  osteosarcoma       1        GD2:NK      1      0  81.53  81.53  81.53
304  osteosarcoma       1      CD28:MSC      0      1   None   None   0.00
367  osteosarcoma       1        CD4:MO      0      1   None   None   0.00
68   osteosarcoma       1        CD1a:B      0      1   None   None   0.00
299  osteosarcoma       1  CD20+CD4:MSC      0      1   None   None   0.00
260  osteosarcoma       1        CD58:L      0      1   None   None   0.00

     variancia
545          0
531          0
117          0
351          0
505          0
304          0
367          0
68           0
299          0
260          0
              grupo  linhas     marcador  dados  nulos      min      max  \
356  renal carcinoma       1  CD20+CD4:MO      1      0  6872.91  6872.91
277  renal carcinoma       1        GD2:L      0      1     None     None
433  renal carcinoma       1       CD71:N      0      1     None     None
386  renal carcinoma       1   cyCD79a:MO      0      1     None     None
428  renal carcinoma       1       CD56:N      1      0    88.82    88.82
362  renal carcinoma       1       CD3:MO      1      0   439.14   439.14
516  renal carcinoma       1     CD117:RC      0      1     None     None
170  renal carcinoma       1    SSC-H:END      0      1     None     None
175  renal carcinoma       1      CD123:E      0      1     None     None
16   renal carcinoma       1       CD22:A      0      1     None     None

       media  variancia
356  6872.91          0
277     0.00          0
433     0.00          0
386     0.00          0
428    88.82          0
362   439.14          0
516     0.00          0
170     0.00          0
```

```
175      0.00          0
16       0.00          0
```

0.2 KMeans Clustering:

Roda-se então um algoritmo de clustering chamado KMeans, dividindo os dados em grupos.

Para isso, se cria um dataframe com todos os grupos e suas estatísticas.

```
In [202]: df_allgroup_result = pd.DataFrame(columns=["grupo", "linhas", "marcador", "dados", "
          for stat in groups_statistics_list:
              df_allgroup_result = df_allgroup_result.append(stat)

          df_allgroup_result.sample(10)
```

Out[202]:

	grupo	linhas	marcador	dados	nulos	min	\
287	vascular tumor	2	CD105:MSC	0	2	None	
147	renal carcinoma	1	CD7:END	0	1	None	
271	germ cell tumor	10	cyCD3+CD19:L	4	6	-133.8	
164	renal carcinoma	1	HLADR:END	0	1	None	
478	normal kidney	1	CD309:NK	1	0	41.63	
463	wilms tumor	10	CD16:NK	0	10	None	
452	sopharyngeal carcinoma	1	nuMYOD:N	1	0	1110.88	
90	normal kidney	1	CD7:B	0	1	None	
438	rhabdomyosarcoma	7	CD9:N	3	4	240.34	
104	vascular tumor	2	FSC-A:B	1	1	50436.9	

	max	media	variancia
287	None	0.000000	0
147	None	0.000000	0
271	3312.88	1794.937500	2.10646e+06
164	None	0.000000	0
478	41.63	41.630000	0
463	None	0.000000	0
452	1110.88	1110.880000	0
90	None	0.000000	0
438	1070.92	599.353333	181967
104	50436.9	50436.900000	0

Como teste inicial, focamos apenas no anticorpo CD10:A. Escolhemos apenas grupos que tenham pelo menos 1 linha com dados desse anticorpo.

```
In [203]: df_kmeans_1 = df_allgroup_result[df_allgroup_result.marcador == 'CD10:A']
          df_kmeans_1 = df_kmeans_1[df_kmeans_1.dados > 0]
          del df_kmeans_1['marcador']
          df_kmeans_1 = df_kmeans_1.set_index('grupo')
          df_kmeans_1.sample(10)
```

Out[203]:

grupo	linhas	dados	nulos	min	max	media	variancia
renal carcinoma	1	1	0	158.62	158.62	158.620000	0
osteosarcoma	1	1	0	305.83	305.83	305.830000	0
wilms tumor	10	9	1	69.95	962.87	254.547778	78415.3
mesoblastic nephroma	1	1	0	14.09	14.09	14.090000	0
cystic nephroma	1	1	0	1818.79	1818.79	1818.790000	0
Adrenal carcinoma	1	1	0	106.92	106.92	106.920000	0
Extraesqueletal Ewing Sarcoma	3	2	1	125.86	505.78	315.820000	72169.6
chondrosarcoma	1	1	0	493.96	493.96	493.960000	0
vascular tumor	2	2	0	432.54	3120.43	1776.485000	3.61238e+06
Neuroblastic tumor	36	21	15	-849.26	385.2	64.823810	58347.9

In [204]:
```
df_kmeans_1.shape
```

Out[204]:
```
(15, 7)
```

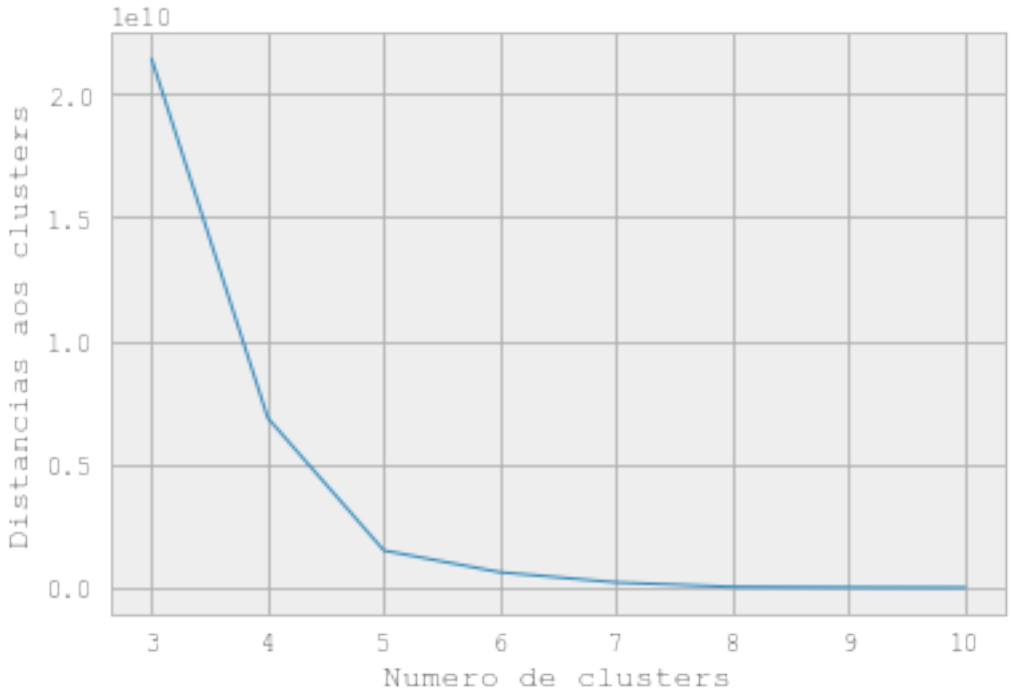
Treinamos e rodamos o kmeans, definindo numero de clusters desde 20% a 50% do número de grupos possiveis, para analisar sua accurácia.

In [211]:
```
len_groups = len(df_kmeans_1.index)

init_clusters = int(len_groups*0.2)
end_clusters = int(len_groups*0.75)
scores = []
for k in range(init_clusters, end_clusters):
    k_means = KMeans(n_clusters=k, n_jobs=1)
    k_means.fit(df_kmeans_1)
    scores.append(-k_means.score(df_kmeans_1))
```

In [212]:
```
plt.plot(range(init_clusters, end_clusters), scores)
plt.xlabel('Numero de clusters')
plt.ylabel('Distancias aos clusters')
```

Out[212]:
```
<matplotlib.text.Text at 0x27d74bd0>
```




```
In [213]: k_means = KMeans(n_clusters=3, n_jobs=-1, max_iter=2000)
          k_means.fit(df_kmeans_1)

          prediction = pd.Series(k_means.predict(df_kmeans_1), index=df_kmeans_1.index)
          print prediction
```

```
grupo
Adrenal carcinoma                   2
chondrosarcoma                      2
cystic nephroma                     2
Extraesqueletal Ewing Sarcoma       2
Neuroblastic tumor                  2
germ cell tumor                     2
vascular tumor                      1
mesoblastic nephroma                2
sopharyngeal carcinoma              2
normal kidney                       2
rhabdomyosarcoma                    2
undiferenciated malignt neoplasm    0
wilms tumor                         2
osteosarcoma                        2
renal carcinoma                     2
dtype: int32
```

```
In [214]: k_means = KMeans(n_clusters=4, n_jobs=-1, max_iter=2000)
          k_means.fit(df_kmeans_1)

          prediction = pd.Series(k_means.predict(df_kmeans_1), index=df_kmeans_1.index)
          print prediction
```

```
grupo
Adrenal carcinoma                   3
chondrosarcoma                      3
cystic nephroma                     3
Extraesqueletal Ewing Sarcoma       0
Neuroblastic tumor                  0
germ cell tumor                     3
vascular tumor                      1
mesoblastic nephroma                3
sopharyngeal carcinoma              3
normal kidney                       3
rhabdomyosarcoma                    0
undiferenciated malignt neoplasm    2
wilms tumor                         0
osteosarcoma                        3
renal carcinoma                     3
dtype: int32
```

```
In [215]: k_means = KMeans(n_clusters=5, n_jobs=-1, max_iter=2000)
          k_means.fit(df_kmeans_1)

          prediction = pd.Series(k_means.predict(df_kmeans_1), index=df_kmeans_1.index)
          print prediction
```

```
grupo
Adrenal carcinoma                   4
chondrosarcoma                      4
cystic nephroma                     4
Extraesqueletal Ewing Sarcoma       2
Neuroblastic tumor                  2
germ cell tumor                     4
vascular tumor                      1
mesoblastic nephroma                4
sopharyngeal carcinoma              4
normal kidney                       0
rhabdomyosarcoma                    2
undiferenciated malignt neoplasm    3
wilms tumor                         2
osteosarcoma                        4
renal carcinoma                     4
dtype: int32
```

```
In [218]: k_means = KMeans(n_clusters=6, n_jobs=-1, max_iter=2000)
          k_means.fit(df_kmeans_1)

          prediction = pd.Series(k_means.predict(df_kmeans_1), index=df_kmeans_1.index)
          print prediction
```

```
grupo
Adrenal carcinoma                   3
chondrosarcoma                      3
cystic nephroma                     3
Extraesqueletal Ewing Sarcoma       5
Neuroblastic tumor                  0
germ cell tumor                     0
vascular tumor                      1
mesoblastic nephroma                3
sopharyngeal carcinoma              3
normal kidney                       4
rhabdomyosarcoma                    5
undiferenciated malignt neoplasm    2
wilms tumor                         5
osteosarcoma                        3
renal carcinoma                     3
dtype: int32
```

```
In [217]: k_means = KMeans(n_clusters=7, n_jobs=-1, max_iter=2000)
          k_means.fit(df_kmeans_1)

          prediction = pd.Series(k_means.predict(df_kmeans_1), index=df_kmeans_1.index)
          print prediction
```

```
grupo
Adrenal carcinoma                   0
chondrosarcoma                      0
cystic nephroma                     0
Extraesqueletal Ewing Sarcoma       3
Neuroblastic tumor                  3
germ cell tumor                     6
vascular tumor                      1
mesoblastic nephroma                0
sopharyngeal carcinoma              0
normal kidney                       5
rhabdomyosarcoma                    4
undiferenciated malignt neoplasm    2
wilms tumor                         4
osteosarcoma                        0
renal carcinoma                     0
dtype: int32
```