Code challenge easynvest

(work in progress)

Description

This is the complete document of the challenge proposed by Easynvest. Disclosure of company name and publishing of the results were explicitly authorized by their recruiting team.

Easynvest is a fintech company, more specifically a digital broker-dealer which helps thousands of clients to invest their money easily and quickly. They are known for their online platform and strong digital presence.

The complete description of the challenge may be found in the file `challenge_description/challenge_description.pdf` (in Portuguese).

Introduction

The data set

The data set received was in the form of an Excel spreadsheet with two tabs. The first tab contained 4973 entries (N=4973), one unique ID and 10 characteristics (11 columns total).
The second tab has entries which are not described elsewhere. The lack of a formal description casts unecessary uncertainty into the data at hand. Lack of proper definition is a discouraged practice in data creation (e.g.: absence of research method and methodology). Data definition must not be open for interpretation.

A remarkable fact of this data set is that it does not contain any null values. Such high quality data sets are rare to find and may indicate that its source is very thoughtful of its data management.

A final remark is that the characteristics' names (column names) should not be considered self explanatory. A codebook is often use to describe published data. To illustrate the critique above consider the variable 'VALOR_01' (value_01; there are 4 of these variables). To what value does it refer to? Is it the amount already invested in the investment platform? Is it the income enumerated by different sources of income? Is it profit? If it is income, is it yearly or monthly? Another illustration is the 'GEO_REFERENCIA' (georeference) variable. It has values ranging from 10 to 999 but it is not explained elsewhere. Usual geolocation information are comprised of x and y coordinates or other better known formats. Also, it cannot refer to Brazilian municipalities because there

exists ~5500 of them.
Consequently this variable has been neglected in the present analysis.

As one can see, this seemingly unimportant differences may yield different interpretations later on the data analysis and render some conclusions useless or even worse: wrong.

Approach

As stated in the challenge description my work should:

1. **Group users finding well defined groups with common characteristics.**
    - In order to do that I have clustered the data set using the *K-Means* clustering algorithm.
2. **Justify the chosen clustering algorithm.**
    - This algorithm is one of the most *commonly used* algorithm in Data Sciences. As such one can easily find support, implementations, discussions and suggestions on various references. Such vast amount of information is not something to be neglected.
      The algorithm also *allows the specification of the number of clusters* to be found. This is seen as drawback according to some sources. Yet I think that it can be overcome with successively running the algorithm with a different cluster number. Specifying the number of clusters also impedes the algorithm to come up with a number of clusters which may be uninterpretable (too few, e. g. 2 or too many 10+).
      The algorithm tends to yield *clusters with similar size*. This may be a desired characteristic in a business setting for example, where investment of resources (time and capital) may be applied to a cluster of clients. In such cases one does not want to invest those in a cluster just to find out that it aggregates to just a few individuals of their clientele.
3. **Present metrics of perfomance for the chosen algorithm.**
    - In this case the *silhouette analysis* was performed to assess the effectivenss of the clustering algorithm.
      Also the *intra-group and inter-group standard deviation* and means were taken in consideration to interpret the results of this clustering algorithm.
4. **Discuss the metrics of performance to assess the clusters.**
    - See discussion of the clustering for a detailed assessment of the clustering algorithm.
5. **Explain the results.**
    - See the results and summary sections for a precise answer to this question.

Results

Preprocessing

Variable scaling

The received data needed preprocessing before applying te clustering method. That is because the K-Means clustering method is sensitive to variable scaling (more precisely to variance). Without scaling, variables tend to have a variances of different orders of magnitude (standard deviation for the data set before preprocessing):

variable	std
valor_01	6098.823
valor_02	89180.835
valor_03	37645.943
valor_04	23246.037
age	10.792
estado_civil_solteiro	0.500
estado_civil_casado	0.486
estado_civil_outro	0.297
genero_m	0.416
genero_f	0.416
perfil_a	0.458
perfil_b	0.416
perfil_c	0.216
perfil_d	0.161

Standard deviation for the data set after preprocessing (abbreviated):

variable	std
valor_01	1.0
valor_02	1.0
(...)	1.0
perfil_c	1.0
perfil_d	1.0

That means that without scaling the four variables of 'valor' would dominate the clustering sensitivity, rendering the presence of the other variables useless.

Nominal variables processing

Some presented variables are categorical and do not meaningfully present any interpretation from a numerical standpoint. For example, height may be compared so that a person who is 170 cm high is higher than someone who is 165 cm.
There is no parallel to variables which represent 'non rankable' variables such as gender and ethnicity. Assigning a value of 1 for male, 0 for female and 2 for non identified gender does not mean that in this scenario that male > female.

In order to overcome this problem categorical variables with N categories are transformed to new binary characteristics (then scaled as commented above). To illustrate suppose that we begin only with `col1` and `col_a`, `col_b` and `col_c` are generated from them:

col1	col_a	col_b	col_c
a	1	0	0
b	0	1	0
c	0	0	1
a	1	0	0

This allow them to be included in the K-Means clustering algorithm.

Clustering

Choice of the number of clusters

I have chosen the numbers of clusters to be six. See the discussion below for details.

Before diving in the details of my choice, one cannot overstress the importance of the choice of the number of clusters. This is arguably the most tricky decision in this challenge as it deals with a great mix of technical as well as non-technical details.

Silhouette analysis (technical analysis)

Silhouette analysis is a technique used to compare how well your data is sorted into clusters. It can be calculated to all data points and then averaged to provide a summary statistic. It ranges from -1 to 1:

- Values near to -1: the data point was incorrectly clustered and should belong to a different cluster
- Values near to zero: the point lies between two clusters and lack a sharp 'belonging attribute' (it could thus belong to both clusters)

- Values near to one: the data point was correctly classified and lies near to other data points in the same cluster. Its cluster is adequately away from other clusters

From a pure technical standpoint choosing the number of clusters such that the average value for silhouette is maximum is the best option. On the other hand, working with such a large number of clusters may hinder the interpretability of the results as clusters probably would not have a sharp distinction between them (consider that our data set has 10 dimensions originally). Probably the communication of such results for a multidiscipliniary team of mixed background would be noisy as well.

Real world analysis (non-technical analysis)

I have chosen the number of cluster to be 6 for a couple of different reasons. First of all, analyzing the average value for silhouette we can see that the average value for silhouette reaches a maximum at around 18 clusters.
Thus we naively could choose the number of clusters to be 18.

However in the context of the **interpretability and communications** of the results one would limit the number of clusters to a maximum of ~10.

Back to the average silhouettes, we can see that it is an increasing function between 2 and 6 clusters, almost doubling its value in this interval. This means that the samples are on average better defined in their own cluster and far away from other clusters. Another fact that indicates that 6 is a good number for clusters is that in this case just a few data points show a silhouette smaller than zero. In other words, just a few data points are incorrectly labeled in their cluster (those data points are unfrequent and are concentrated on cluster 2) (see below). Using the same argumentation the cluster that is best defined is cluster 1 because of the high incidence of data points at near 0.75 silhouette value.

See images for silhouette for all images.

Cluster interpretation

For cluster interpretation two resources are available:

1. Tables output to stdout during program execution.
2. Plots.

Please notice that all tables, plots and results are exhaustively detailed below. They are repeated here for convenience.

From a general standpoint clusters should have low intra-cluster variance and high inter-cluster variance for each variable.

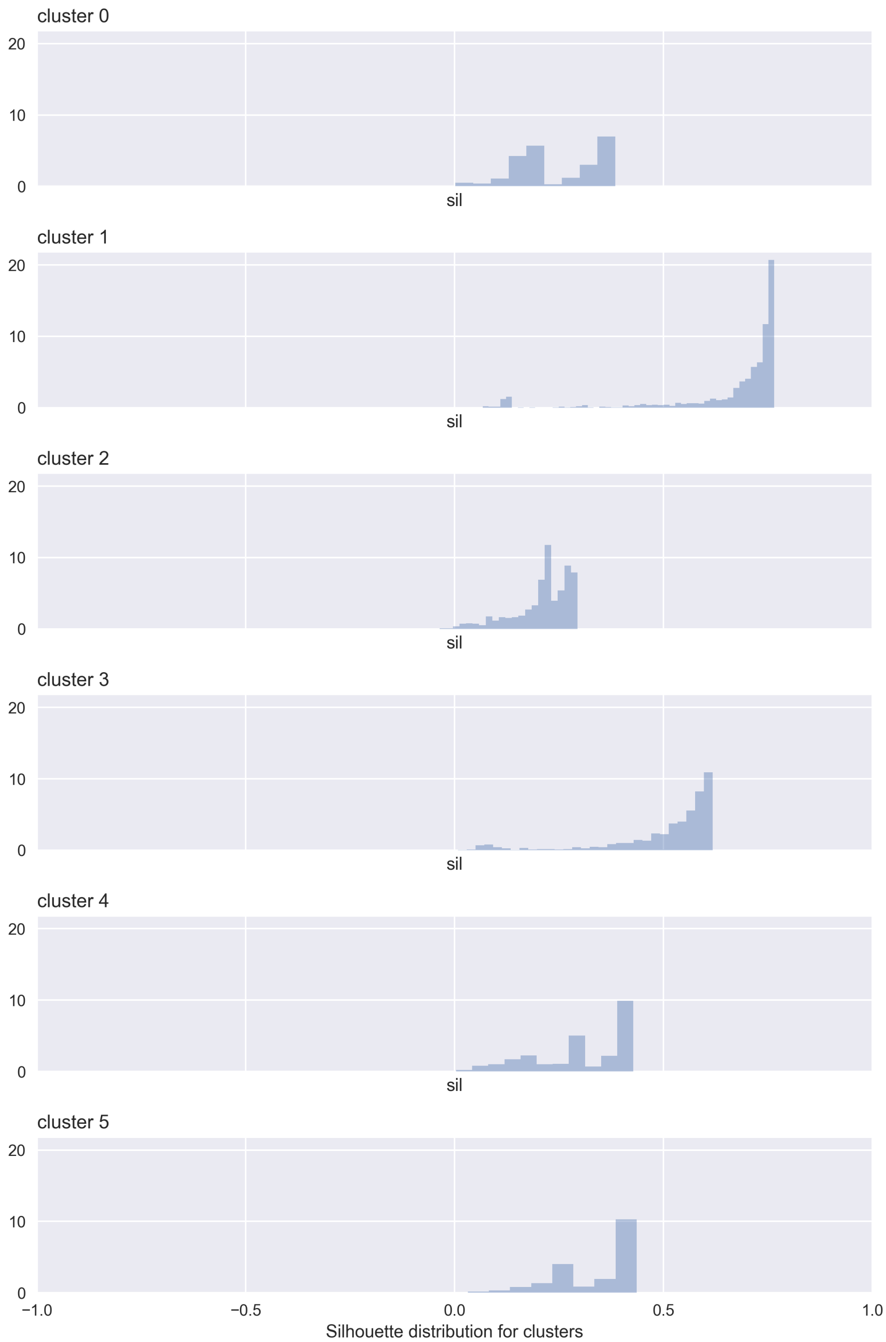
cluster 0
cluster 1
cluster 2
cluster 3
cluster 4
cluster 5
20
10
0
sil
−1.0
−0.5
0.0
0.5
1.0
Silhouette distribution for clusters

Figure 1:

Cluster 0

Distinctive features:

1. Has all individuals with ‘other’ marital status (that is, it is neither married nor single).

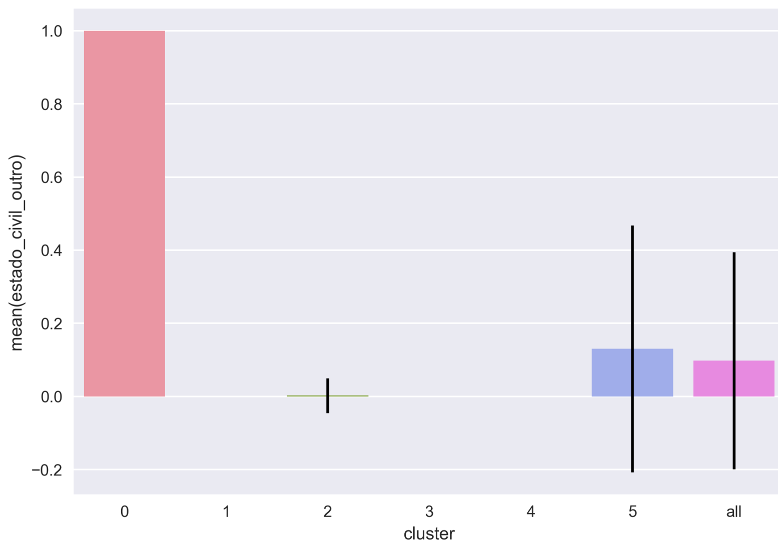

Figure 2:

1. Has the highest age mean of all groups even though there is a high dispersion both intra and inter cluster for this variable.

Cluster 1

Distinctive features:

1. Has the most concentration of single persons (‘solteiro’).

1. It is solely composed of male individuals (absence of ‘genenro_f’). This also happens to cluster 2 and cluster 3.
2. Has the most concentration of profile D (‘perfil_d’). Also contains a lot of profile A individuals.
3. Has the most concentration young people.
4. Has the highest average value of silhouette (see above).
5. It is the cluster which aggregates most individuals (~1400).

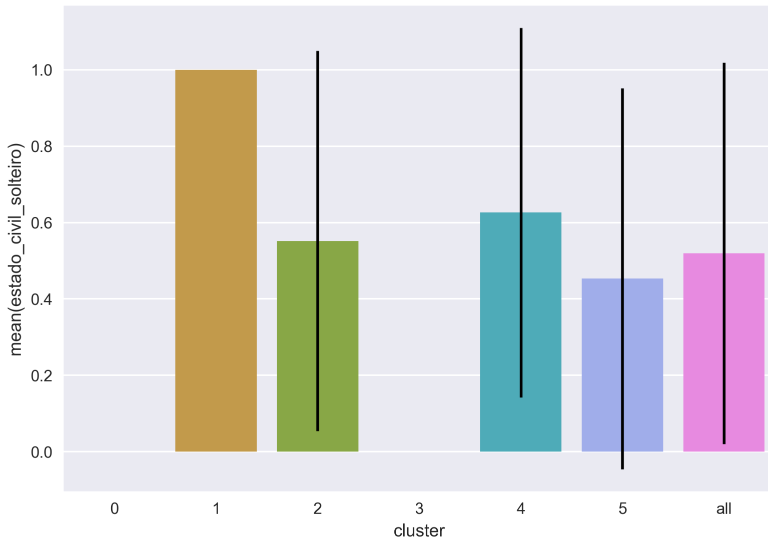

Figure 3:

Cluster 2

Distinctive features:

1. Has the highest averages for 'valor__02', 'valor__03' and 'valor__04'. In respect to these 3 variables all the other groups have lower averages.

1. Includes almost solely profile B people.
2. Contains almost solely males.

Cluster 3

Distinctive features:

1. It is the group with the highest proportion of married individuals ('estado_civil_casado'). And it is solely comprised of married individuals.

1. The cluster is entirely comprised of male individuals.
2. The cluster contains only individuals from profile A and profile D.

Cluster 4

Distinctive features:

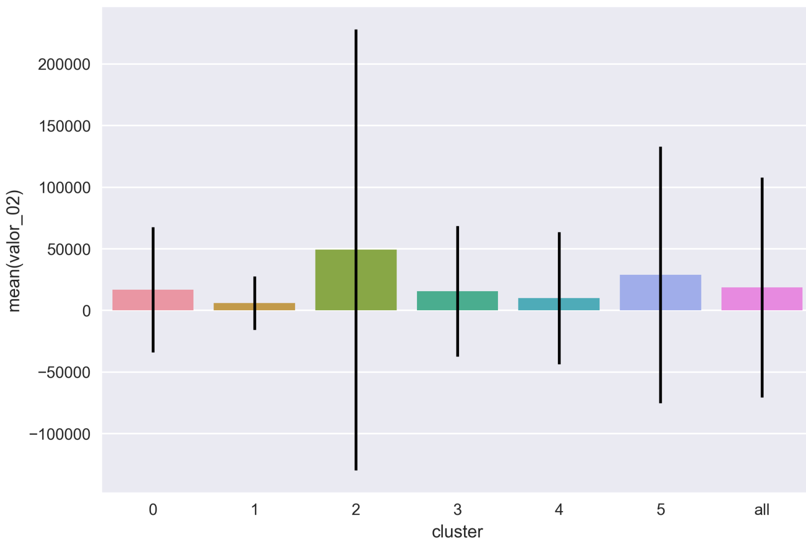

Figure 4:

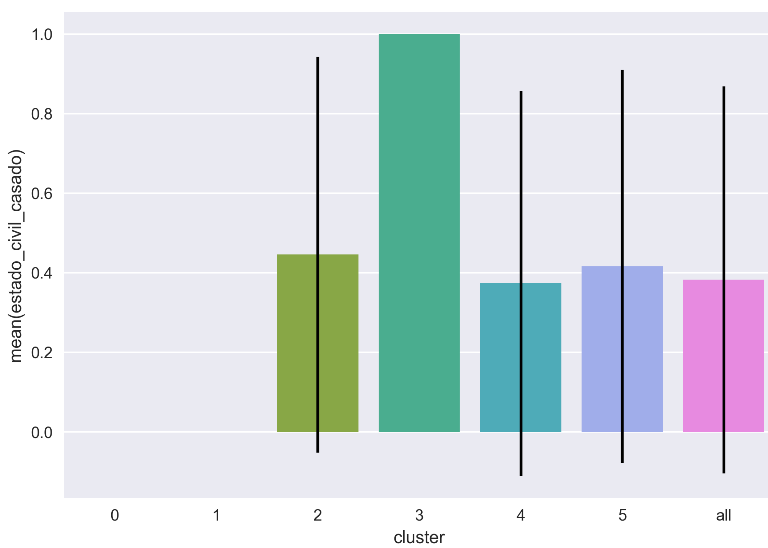

Figure 5:

1. Comprised solely of female subjects:

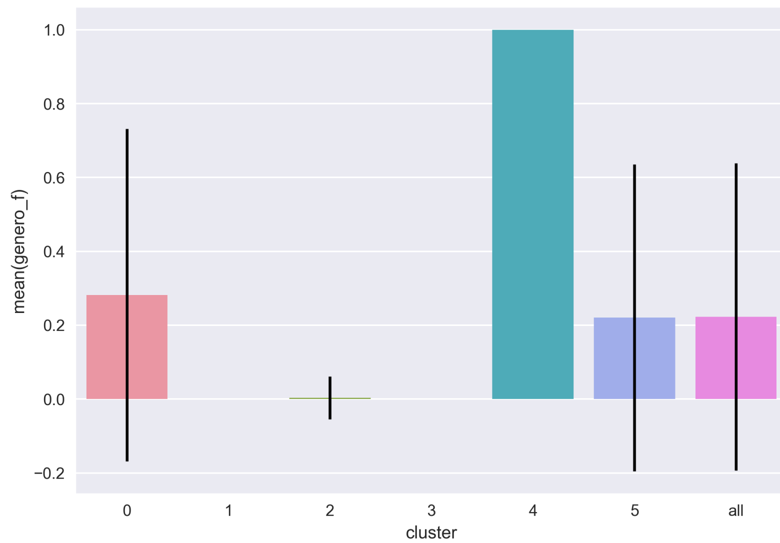

Figure 6:

1. Contains a fair distribution of different profiles ('perfil'), marital status, and values ('valor')

Cluster 5

Distinctive features:

1. Comprised solely of profile C and it contains this group entirely:

(the absence of the errorbar indicates that there in only one value for this variable in for this cluster).

1. It is the smallest of all clusters: 245 individuals.

Summary

The clusterization was conducted properly and yielded significant results. This is evidenced by:

- A satisfactory value of silhouette indicated good clustering (0.433)

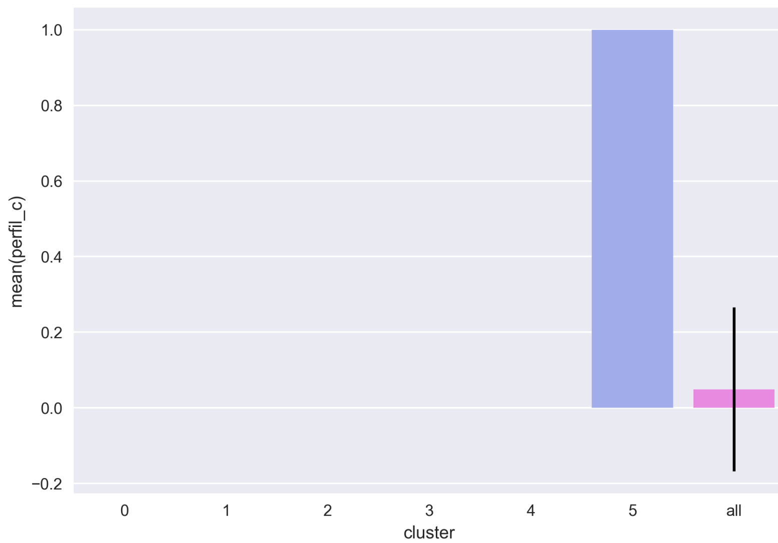

Figure 7:

- A reasonable amount of clusters, facilitating the communication and interpretation of the results (one of the strenghts of the algorithm)
- Some very sharp separations, some of which are coupled and yield easily interpretable results:
  - Cluster 0:
    - it contains all individuals with 'other' marital status
    - has the highest age mean
    - (notice that the two variables are correlated)
  - Cluster 1:
    - all individuals are single males
    - lowest age mean of all clusters
    - it is the largest cluster
  - Cluster 2:
    - concentrates individuals with high value variables ('valor') from 2 to 4
  - Cluster 3:
    - group entirely comprised of male individuals
    - group entirely comprised of married individuals
  - Cluster 4:
    - group entirely composed of female subjects
  - Cluster 5:

    * it contains all individuals with profile C (‘perfil_c’)

Additional information & Reproducibility

Reproducibility

Reproducibility is going to be assessed in this task. In order to comply with it the software versions needed to replicate the experiement are specified below.

Also non deterministic part of the algorithms are fixed using a defined random seed at `code/control.py` and invoked properly during code execution.

Finally the code is hosted on github to allow any team to replicate and judge the results themselves.

Tools

- Vim

  ```
  vim --version
  VIM - Vi IMproved 8.0 (2016 Sep 12, compiled Apr  4 2017 13:41:19)
  Included patches: 1-542
  Modified by <cygwin@cygwin.com>
  Compiled by <cygwin@cygwin.com>
  Huge version without GUI.
  ```

- python

  ```
  python3 --version
  Python 3.6.1
  ```

- python modules:

  ```
  data-utilities==1.2.6
  matplotlib==2.0.0
  numpy==1.12.1
  pandas==0.19.2
  scikit-learn==0.18.1
  scipy==0.19.0
  seaborn==0.7.1
  ```

- pandoc

  ```
  pandoc --version
  pandoc.exe 1.19.2.1
  Compiled with pandoc-types 1.17.0.4, texmath 0.9, skylighting 0.1.1.4
  Default user data directory: C:\Users\e061568\AppData\Roaming\pandoc
  ```

```
Copyright (C) 2006-2016 John MacFarlane
Web:  http://pandoc.org
This is free software; see the source for copying conditions.
There is no warranty, not even for merchantability or fitness
for a particular purpose.
```

Other remarks

Here are comments which would not fit elsewhere in the discussion of the document. Despite there are discussions on some of these topics I thought they would not fit well in the flow of the assignment.

- As suggested in the introduction, a more precise definition of the data set could improve the conclusions that could be drawn from it. When presenting a data set for someone who is not acquainted with how it was generated extreme care should be taken in order to communicate the variables, their origin and their precise meaning carefully.
- K-Means clusterization algorithm:
  - Randomization of the initial cluster points may yield very different clusters for the given data. This is considered a weakness. To ovecome this the algorithm may use the *k-means++* initial seeding which improve the initial assignment of the algorithm.
- Approximate running time for the `main.py` is 7 minutes on commodity hardware:

  ```
  time python3 main.py
      7:06.02
  ```

  It is dominated by the plotting of different k-means silhouettes. The rest of the code runs in less than one minute (full running time is kept for reproducibility).

Next steps

- XXX TODO: close and comment all open 'XXX TODO'.
- XXX TODO: coding conventions and style will also be assessed.
  - Comment that it is PEP8 compliant
    - Comment on python-mode and contributions
  - Comment on docstrings style
    - Comment on sphinx documentation

All output from python code

All the images

Silhouette

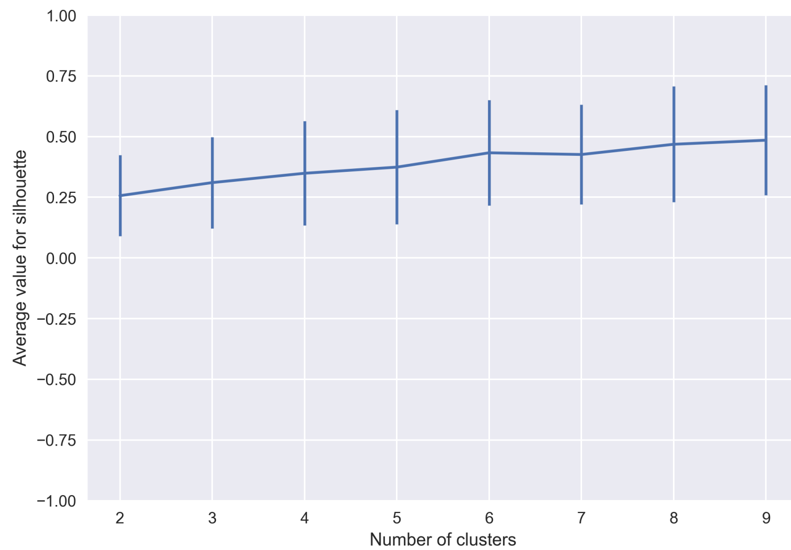

Figure 8: file silhouette.png

Clusters

Notice that the error bars represented here are +- 1 standard deviation.

Code output (stdout)

```
----------------------------------------------------------------------
The effect of preprocessing on standard deviation   ------------------
----------------------------------------------------------------------
    before:
valor_01                  6098.823
valor_02                 89180.835
valor_03                 37645.943
valor_04                 23246.037
```

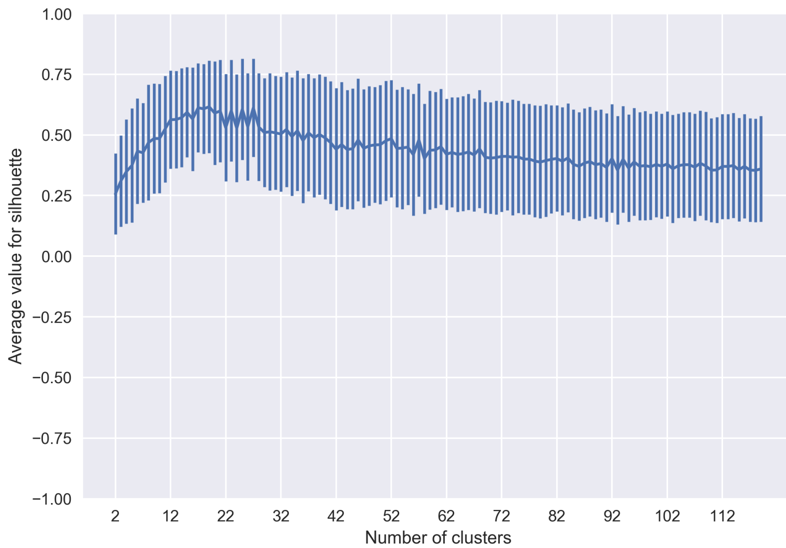

Figure 9: file silhouette_120.png

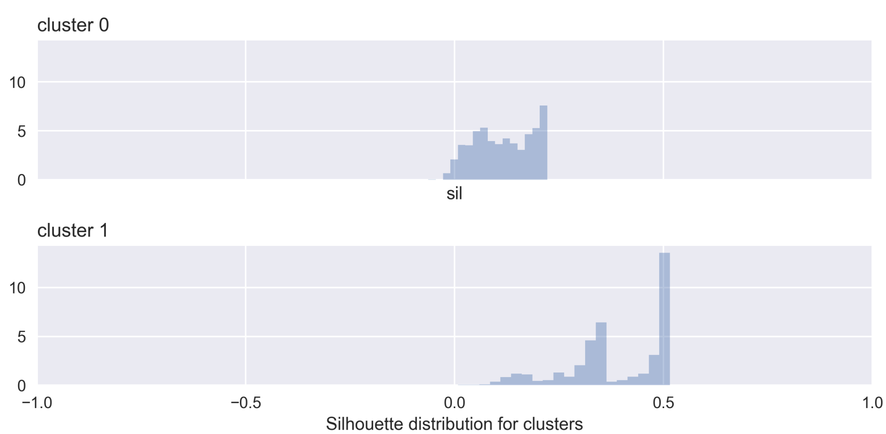

Figure 10: file silhouette_distribution_for_n=2_clusters.png

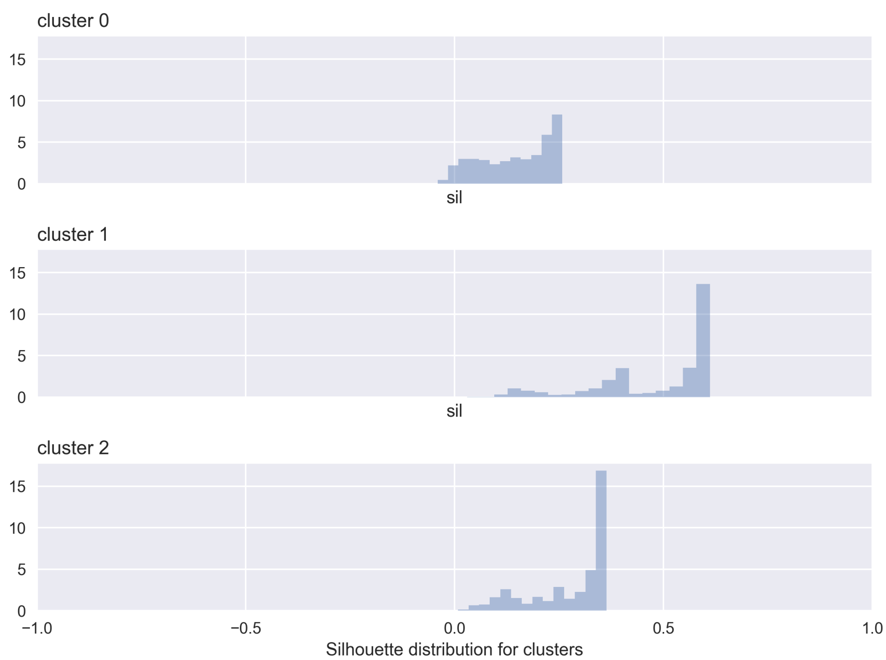

Figure 11: file silhouette_distribution_for_n=3_clusters.png

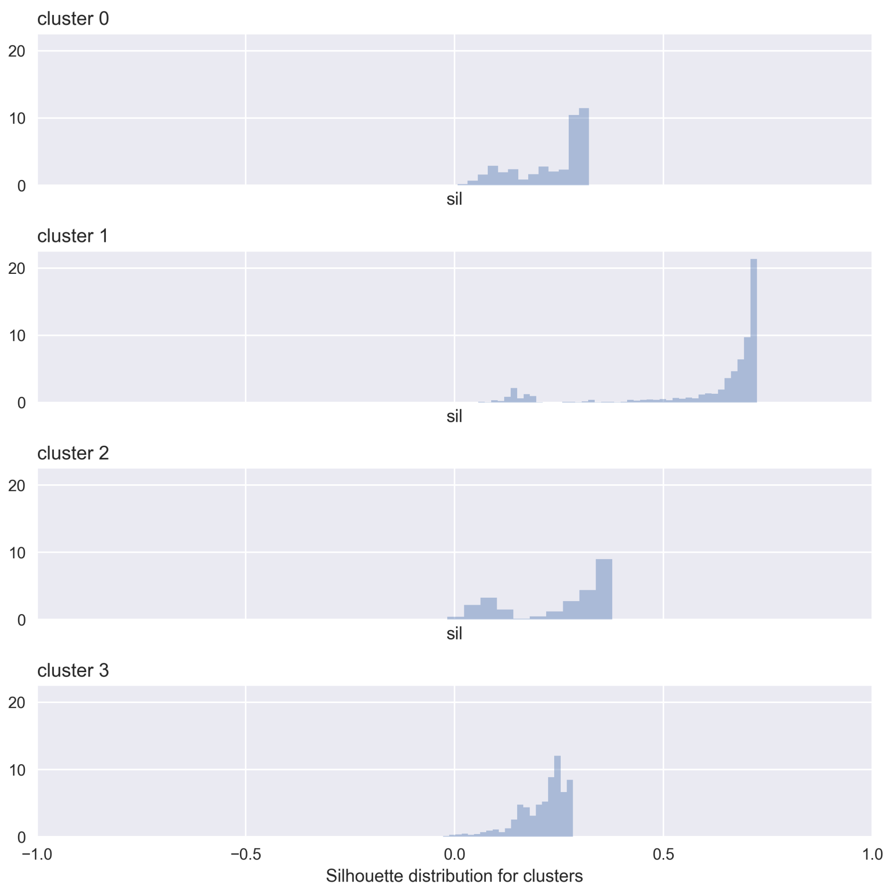

Figure 12: file silhouette_distribution_for_n=4_clusters.png

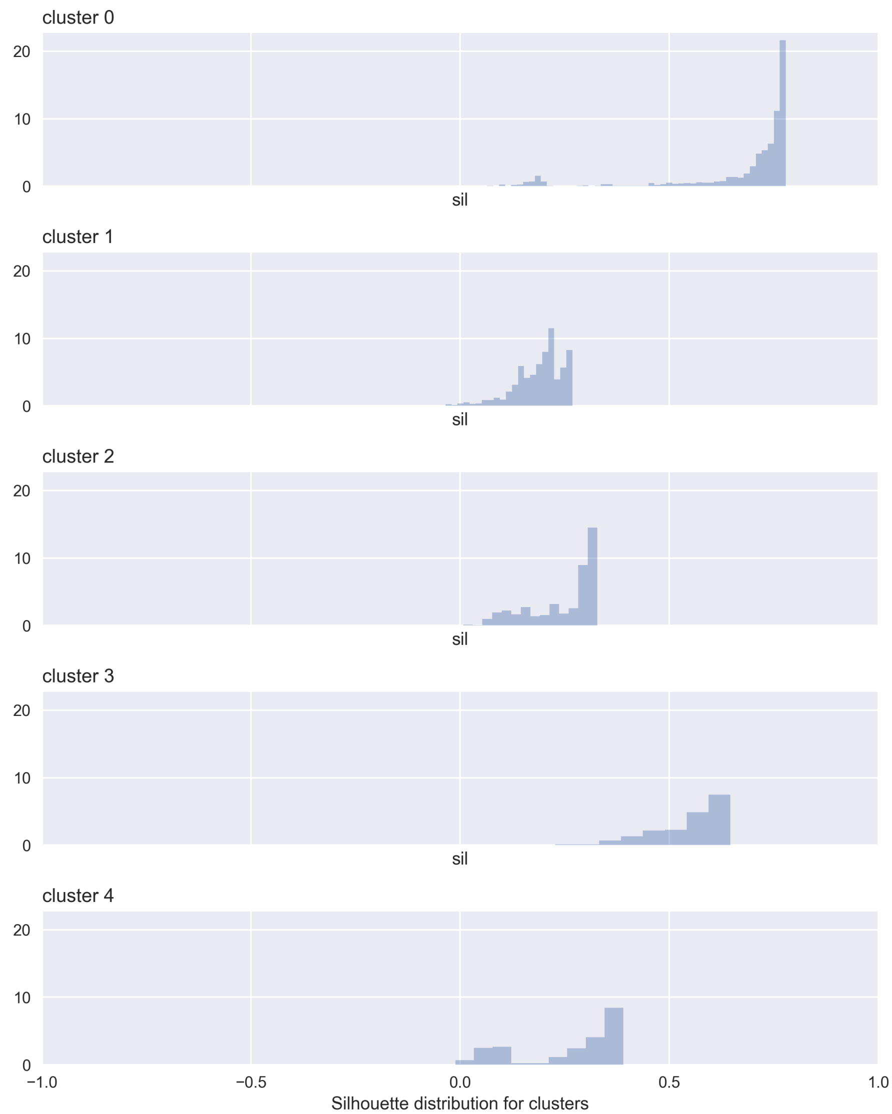

Figure 13: file silhouette_distribution_for_n=5_clusters.png

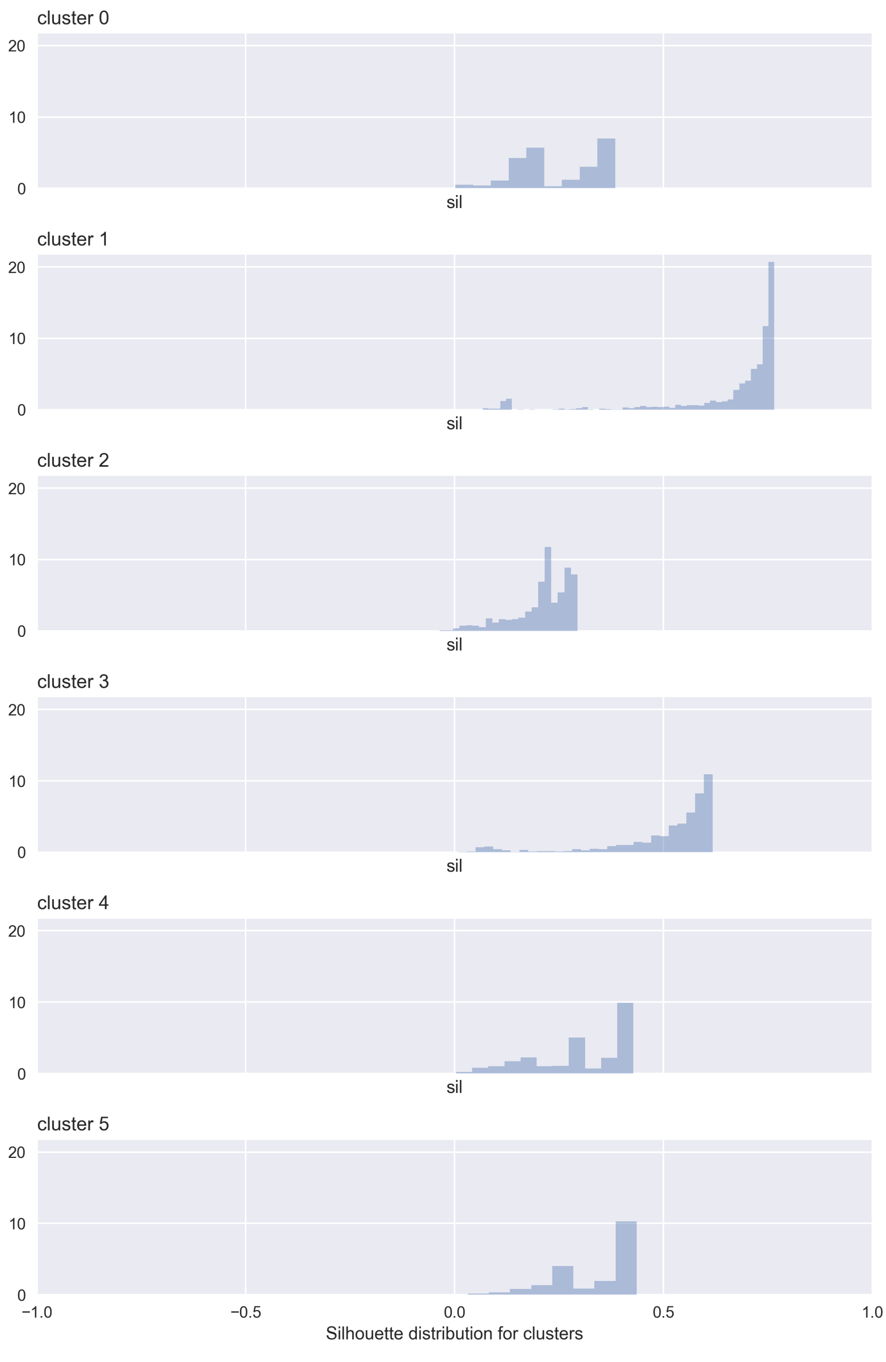

Figure 14: file silhouette_distribution_for_n=6_clusters.png

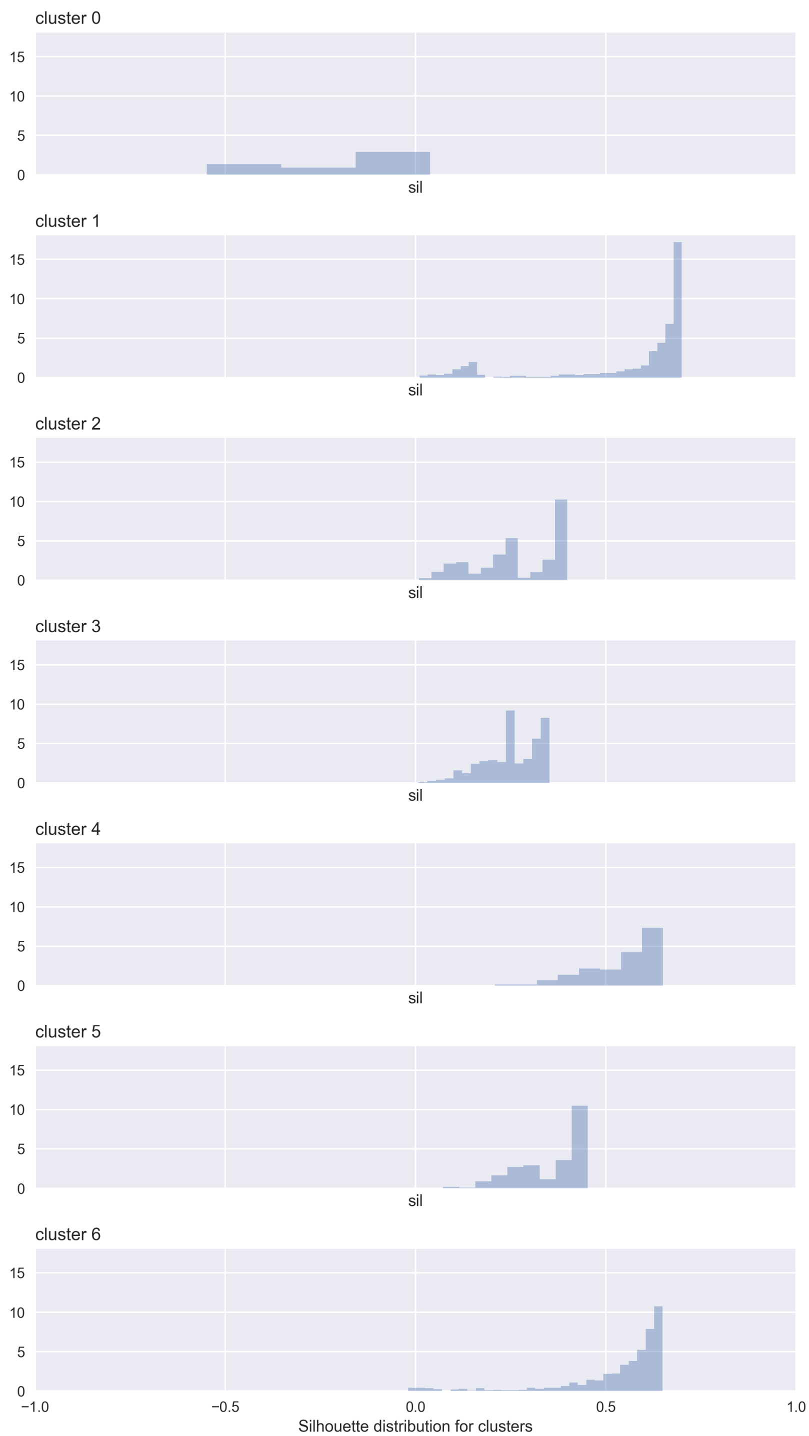

Figure 15: file silhouette_distribution_for_n=7_clusters.png

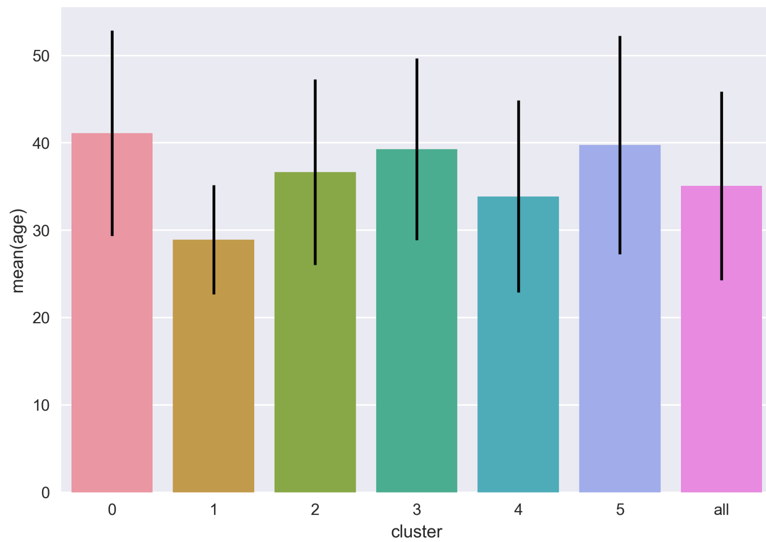

Figure 16: file cluster_analysis_cluster_mean_of_variables_age.png

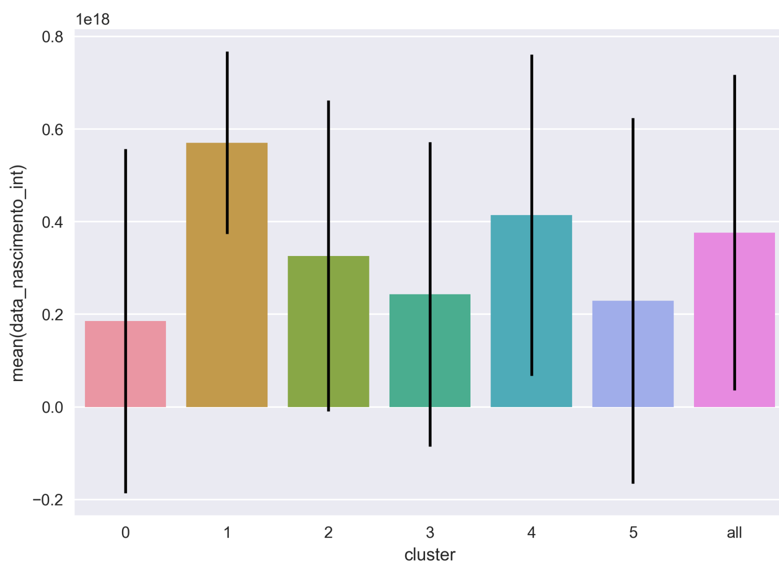

Figure 17: file cluster_analysis_cluster_mean_of_variables_data_nascimento_int.png

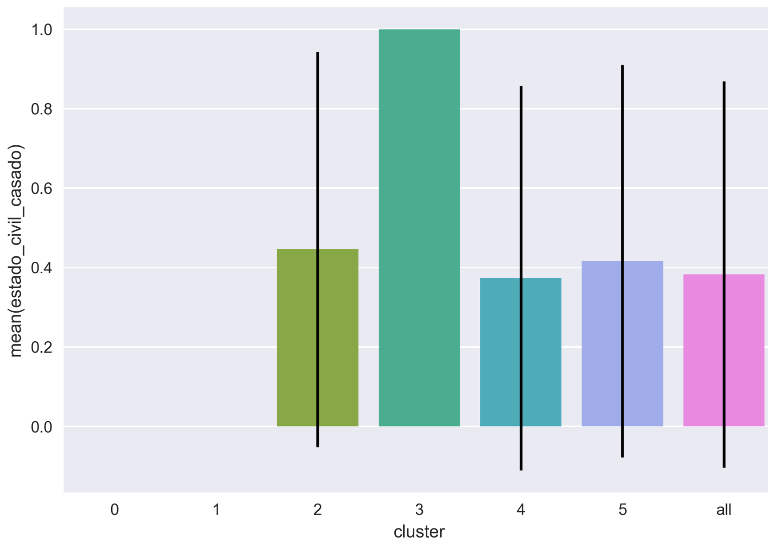

Figure 18: file cluster_analysis_cluster_mean_of_variables_estado_civil_casado.png

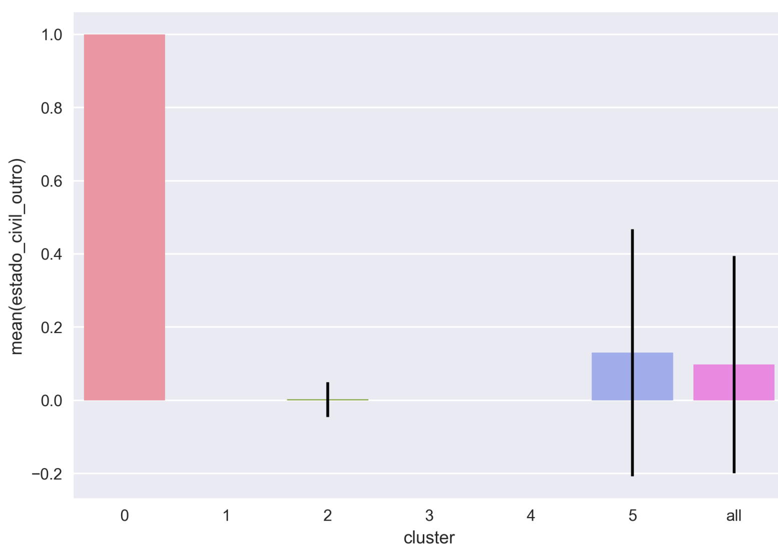

Figure 19: file cluster_analysis_cluster_mean_of_variables_estado_civil_outro.png

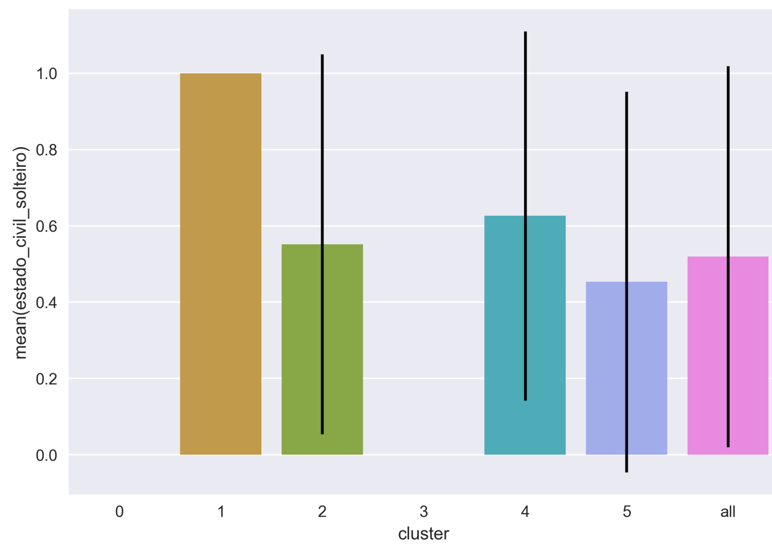

Figure 20: file cluster_analysis_cluster_mean_of_variables_estado_civil_solteiro.png

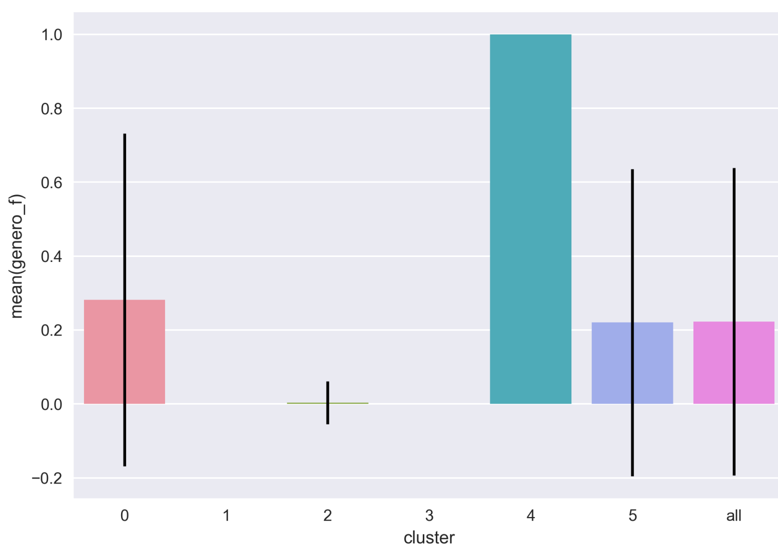

Figure 21: file cluster_analysis_cluster_mean_of_variables_genero_f.png

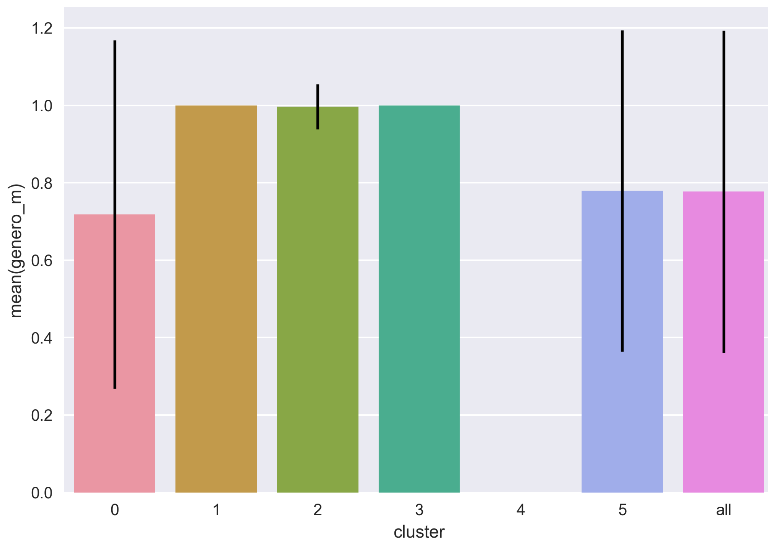

Figure 22: file cluster_analysis_cluster_mean_of_variables_genero_m.png

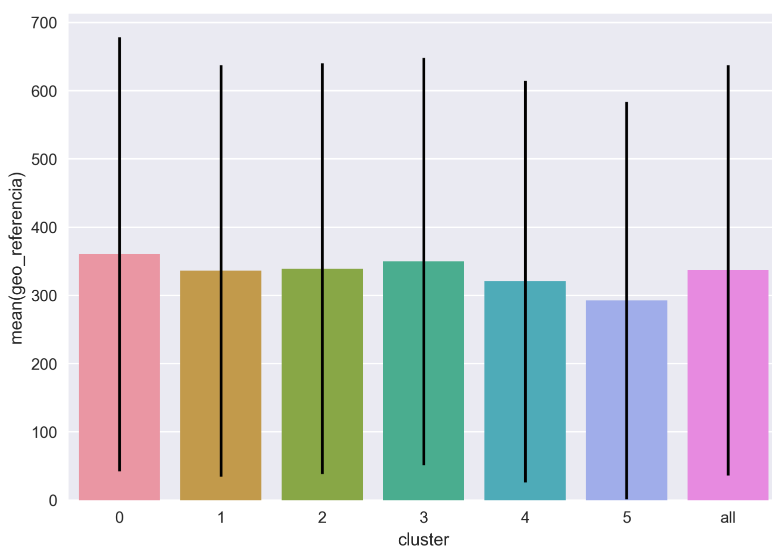

Figure 23: file cluster_analysis_cluster_mean_of_variables_geo_referencia.png

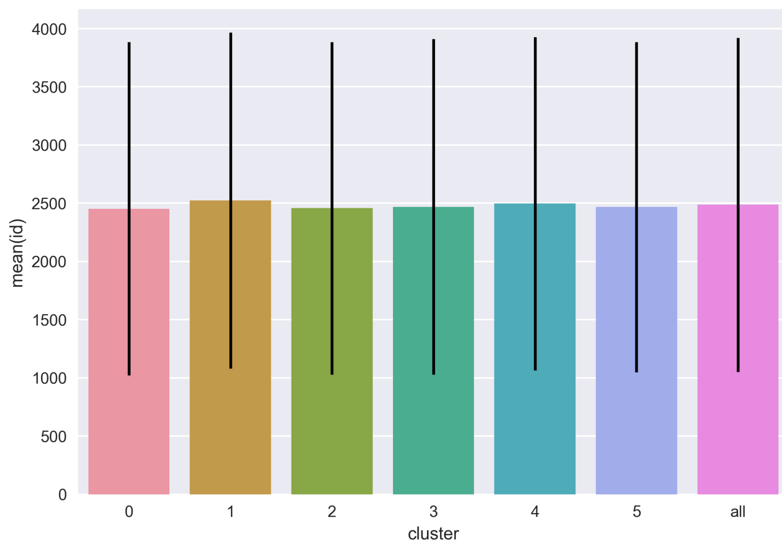

Figure 24: file cluster_analysis_cluster_mean_of_variables_id.png

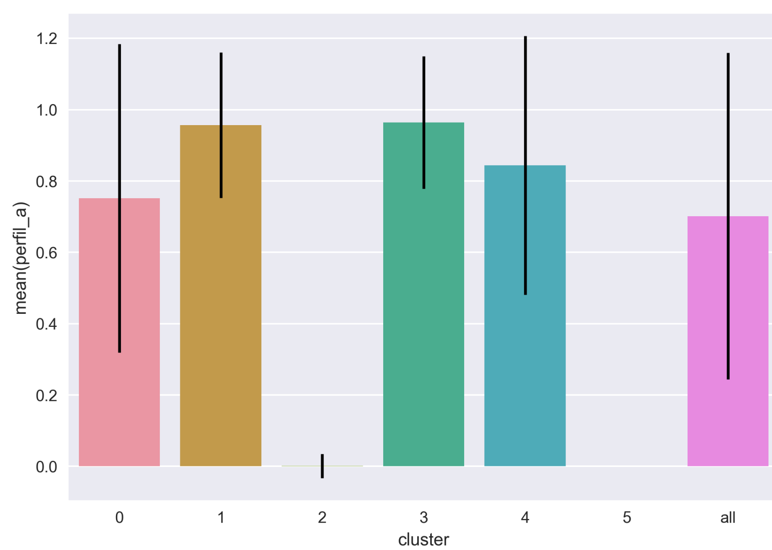

Figure 25: file cluster_analysis_cluster_mean_of_variables_perfil_a.png

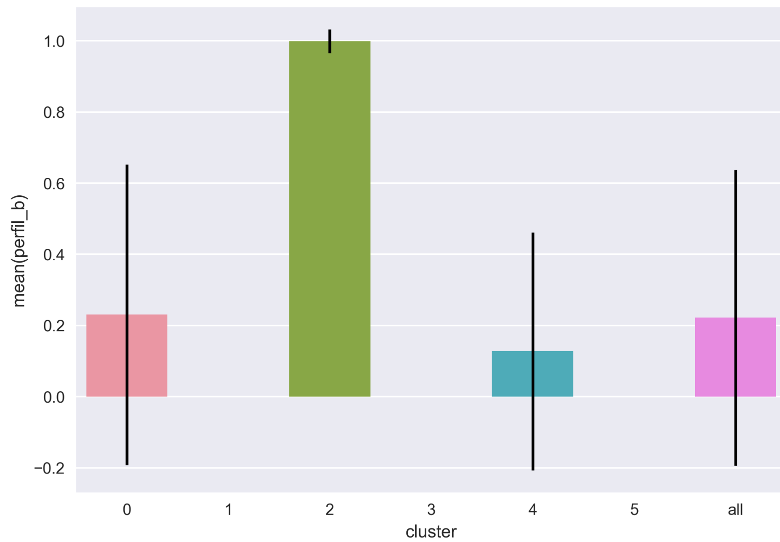

Figure 26: file cluster_analysis_cluster_mean_of_variables_perfil_b.png

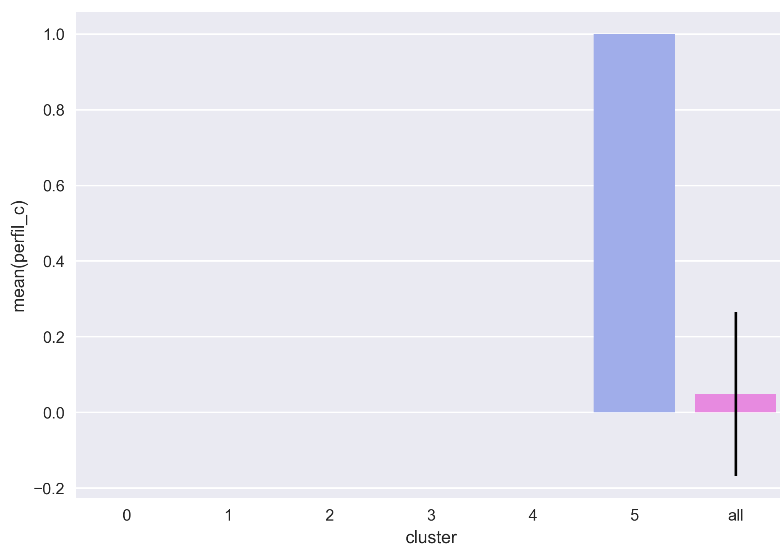

Figure 27: file cluster_analysis_cluster_mean_of_variables_perfil_c.png

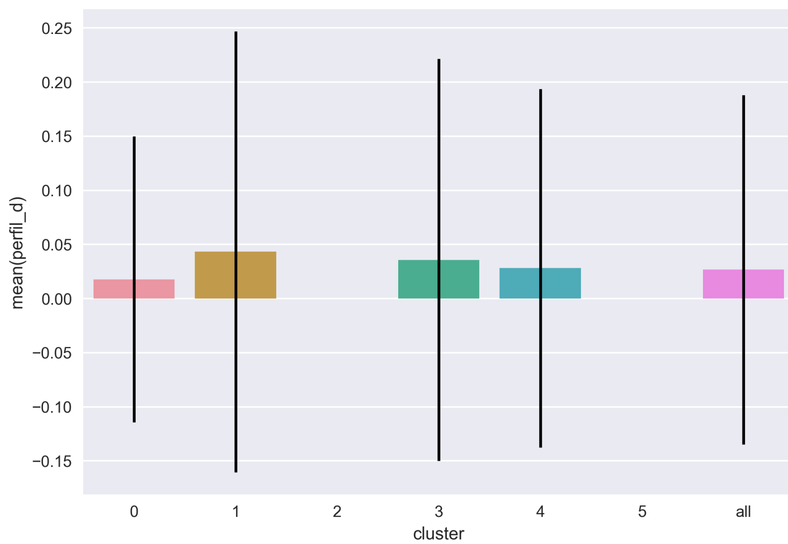

Figure 28: file cluster_analysis_cluster_mean_of_variables_perfil_d.png

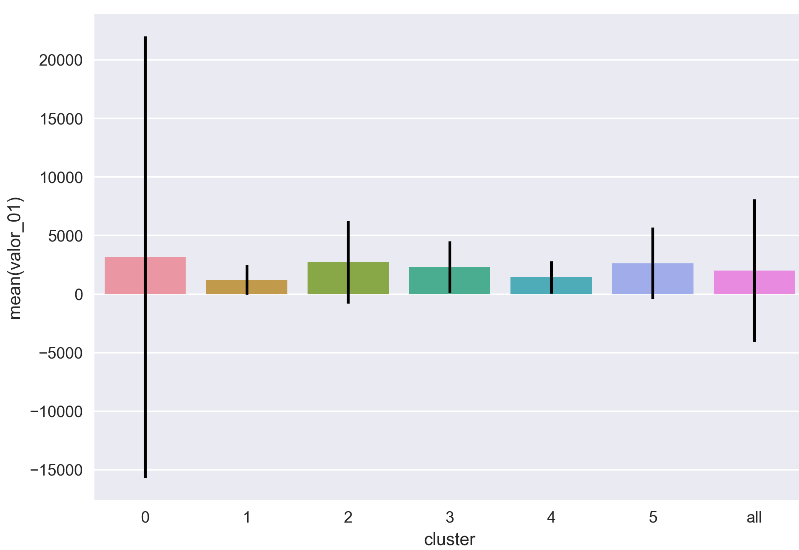

Figure 29: file cluster_analysis_cluster_mean_of_variables_valor_01.png

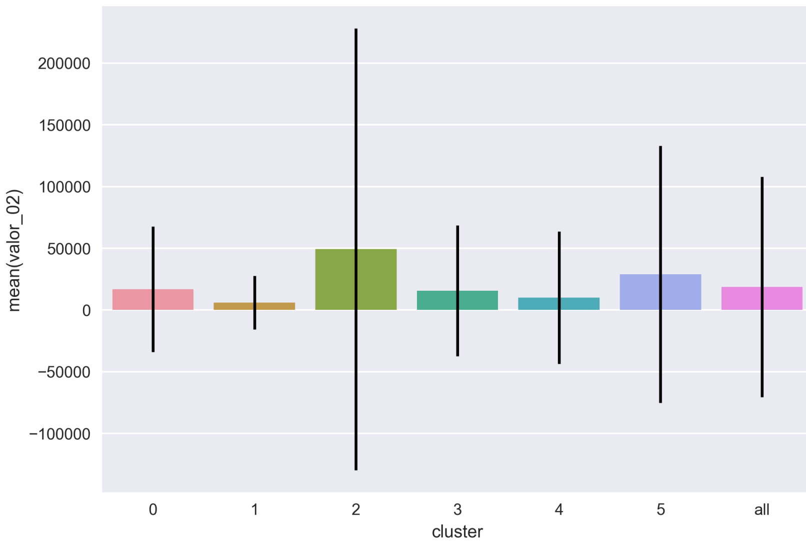

Figure 30: file cluster_analysis_cluster_mean_of_variables_valor_02.png

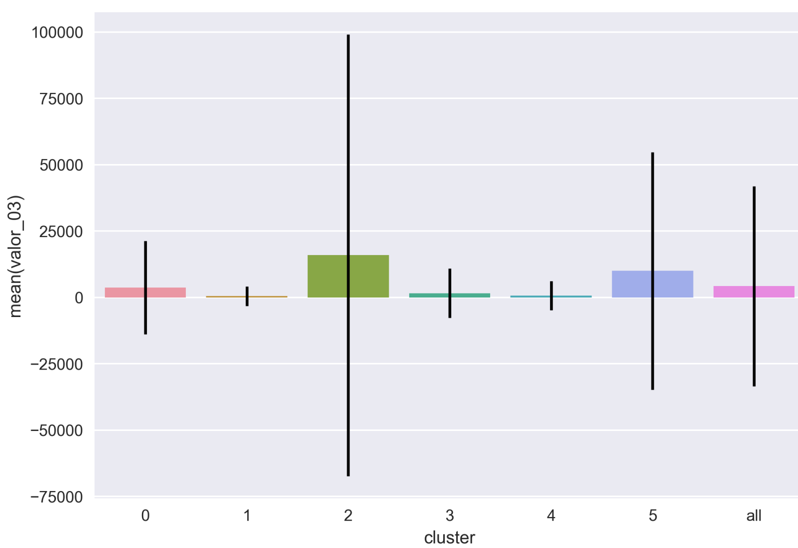

Figure 31: file cluster_analysis_cluster_mean_of_variables_valor_03.png

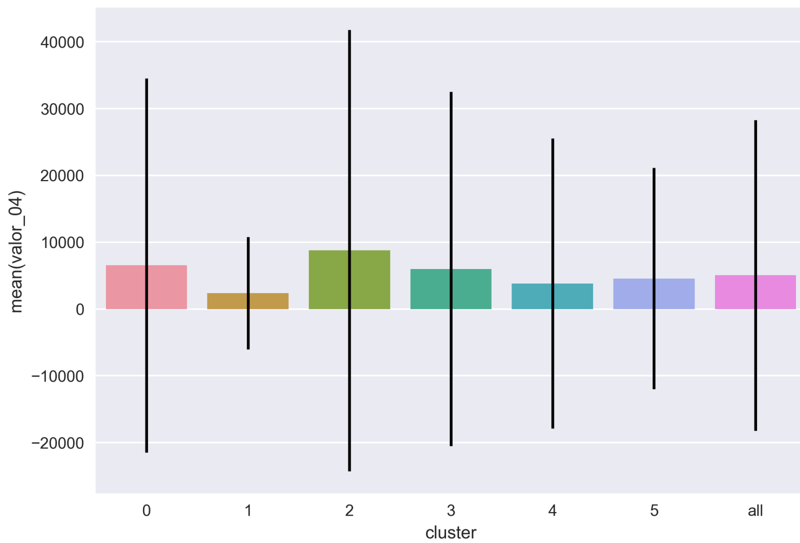

Figure 32: file cluster_analysis_cluster_mean_of_variables_valor_04.png

```
age                          10.792
estado_civil_solteiro         0.500
estado_civil_casado           0.486
estado_civil_outro            0.297
genero_m                      0.416
genero_f                      0.416
perfil_a                      0.458
perfil_b                      0.416
perfil_c                      0.216
perfil_d                      0.161
dtype: float64
    after:
valor_01                 1.0
valor_02                 1.0
valor_03                 1.0
valor_04                 1.0
age                      1.0
estado_civil_solteiro    1.0
estado_civil_casado      1.0
estado_civil_outro       1.0
genero_m                 1.0
genero_f                 1.0
```

```
perfil_a                 1.0
perfil_b                 1.0
perfil_c                 1.0
perfil_d                 1.0
dtype: float64

----------------------------------------------------------------------
Features for different number of clusters   --------------------------
----------------------------------------------------------------------
Silhouette average for 2 clusters 0.256
Number of individuals per cluster for 2 clusters ---------------------
(2392,)
(2580,)
Silhouette average for 3 clusters 0.310
Number of individuals per cluster for 3 clusters ---------------------
(1887,)
(1979,)
(1106,)
Silhouette average for 4 clusters 0.349
Number of individuals per cluster for 4 clusters ---------------------
(1103,)
(1492,)
(1405,)
(972,)
Silhouette average for 5 clusters 0.374
Number of individuals per cluster for 5 clusters ---------------------
(1431,)
(972,)
(1072,)
(133,)
(1364,)
Silhouette average for 6 clusters 0.433
Number of individuals per cluster for 6 clusters ---------------------
(451,)
(1407,)
(884,)
(1062,)
(923,)
(245,)
Silhouette average for 7 clusters 0.426
Number of individuals per cluster for 7 clusters ---------------------
(23,)
(1547,)
(1017,)
(952,)
(133,)
```

```
(244,)
(1056,)
Silhouette average for 8 clusters 0.468
Silhouette average for 9 clusters 0.485

----------------------------------------------------------------------
Cluster standard deviation of variables   ----------------------------
----------------------------------------------------------------------
               id  geo_referencia   valor_01    valor_02   valor_03  \
cluster
0        1432.085         318.310  18855.499   50776.256  17566.300
1        1444.033         301.413   1273.152   21754.206   3742.253
2        1429.445         300.858   3528.072  178904.471  83183.573
3        1440.205         298.530   2208.455   53021.604   9307.479
4        1431.331         294.305   1391.051   53714.303   5459.086
5        1417.395         291.092   3051.785  104054.907  44757.829
all      1435.437         300.712   6098.823   89180.835  37645.943

          valor_04  data_nascimento_int     age  estado_civil_solteiro  \
cluster
0        27998.787            3.714e+17  11.769                  0.000
1         8408.652            1.972e+17   6.248                  0.000
2        33028.168            3.354e+17  10.628                  0.498
3        26499.920            3.286e+17  10.414                  0.000
4        21677.157            3.467e+17  10.985                  0.484
5        16546.183            3.947e+17  12.508                  0.499
all      23246.037            3.406e+17  10.792                  0.500

         estado_civil_casado  estado_civil_outro  genero_m  genero_f  \
cluster
0                      0.000               0.000     0.450     0.450
1                      0.000               0.000     0.000     0.000
2                      0.497               0.048     0.058     0.058
3                      0.000               0.000     0.000     0.000
4                      0.484               0.000     0.000     0.000
5                      0.494               0.338     0.415     0.415
all                    0.486               0.297     0.416     0.416

         perfil_a  perfil_b  perfil_c  perfil_d
cluster
0           0.433     0.422     0.000     0.132
1           0.204     0.000     0.000     0.204
2           0.034     0.034     0.000     0.000
3           0.186     0.000     0.000     0.186
4           0.363     0.334     0.000     0.166
5           0.000     0.000     0.000     0.000
```

```
all         0.458     0.416     0.216     0.161

----------------------------------------------------------------------
Cluster mean of variables   ------------------------------------------
----------------------------------------------------------------------
               id  geo_referencia  valor_01   valor_02   valor_03  valor_04  \
cluster
0        2452.268         360.517  3176.098  16914.336   3676.908  6541.581
1        2524.927         336.119  1241.121   5949.802    472.001  2373.197
2        2457.329         339.374  2737.195  49328.855  15918.938  8769.796
3        2469.815         349.920  2329.157  15720.792   1585.815  5992.588
4        2497.011         320.521  1448.414  10067.777    631.819  3835.935
5        2466.812         292.376  2645.085  28873.244  10001.609  4562.969
all      2486.500         336.808  2022.698  18638.059   4246.265  5041.123

         data_nascimento_int     age  estado_civil_solteiro  \
cluster
0                  1.856e+17  41.120                  0.000
1                  5.707e+17  28.918                  1.000
2                  3.262e+17  36.665                  0.552
3                  2.436e+17  39.281                  0.000
4                  4.143e+17  33.871                  0.626
5                  2.291e+17  39.741                  0.453
all                3.766e+17  35.068                  0.520

         estado_civil_casado  estado_civil_outro  genero_m  genero_f  \
cluster
0                      0.000               1.000     0.718     0.282
1                      0.000               0.000     1.000     0.000
2                      0.446               0.002     0.997     0.003
3                      1.000               0.000     1.000     0.000
4                      0.374               0.000     0.000     1.000
5                      0.416               0.131     0.780     0.220
all                    0.383               0.098     0.777     0.223

         perfil_a  perfil_b  perfil_c  perfil_d
cluster
0           0.752     0.231     0.000     0.018
1           0.957     0.000     0.000     0.043
2           0.001     0.999     0.000     0.000
3           0.964     0.000     0.000     0.036
4           0.844     0.128     0.000     0.028
5           0.000     0.000     1.000     0.000
all         0.702     0.222     0.049     0.027

----------------------------------------------------------------------
```

```
Normalized cluster standard deviation of variables   -----------------
-----------------------------------------------------------------------
            id  geo_referencia  valor_01  valor_02  valor_03  valor_04  \
cluster
0        0.992           1.000     1.000     0.284     0.211     0.848
1        1.000           0.947     0.068     0.122     0.045     0.255
2        0.990           0.945     0.187     1.000     1.000     1.000
3        0.997           0.938     0.117     0.296     0.112     0.802
4        0.991           0.925     0.074     0.300     0.066     0.656
5        0.982           0.914     0.162     0.582     0.538     0.501
all      0.994           0.945     0.323     0.498     0.453     0.704

         data_nascimento_int    age  estado_civil_solteiro  \
cluster
0                      0.941  0.941                  0.000
1                      0.499  0.499                  0.000
2                      0.850  0.850                  0.996
3                      0.833  0.833                  0.000
4                      0.878  0.878                  0.969
5                      1.000  1.000                  0.998
all                    0.863  0.863                  1.000

         estado_civil_casado  estado_civil_outro  genero_m  genero_f  \
cluster
0                      0.000               0.000     1.000     1.000
1                      0.000               0.000     0.000     0.000
2                      1.000               0.141     0.129     0.129
3                      0.000               0.000     0.000     0.000
4                      0.973               0.000     0.000     0.000
5                      0.993               1.000     0.922     0.922
all                    0.977               0.879     0.924     0.924

         perfil_a  perfil_b  perfil_c  perfil_d
cluster
0           0.945     1.000       0.0     0.649
1           0.445     0.000       0.0     1.000
2           0.074     0.080       0.0     0.000
3           0.406     0.000       0.0     0.912
4           0.794     0.792       0.0     0.813
5           0.000     0.000       0.0     0.000
all         1.000     0.986       1.0     0.792

-----------------------------------------------------------------------
Normalized cluster mean   --------------------------------------------
-----------------------------------------------------------------------
            id  geo_referencia  valor_01  valor_02  valor_03  valor_04  \
```

```
cluster
0        0.971            1.000     1.000     0.343     0.231     0.746
1        1.000            0.932     0.391     0.121     0.030     0.271
2        0.973            0.941     0.862     1.000     1.000     1.000
3        0.978            0.971     0.733     0.319     0.100     0.683
4        0.989            0.889     0.456     0.204     0.040     0.437
5        0.977            0.811     0.833     0.585     0.628     0.520

         data_nascimento_int    age  estado_civil_solteiro  \
cluster
0                      0.325  1.000                  0.000
1                      1.000  0.703                  1.000
2                      0.572  0.892                  0.552
3                      0.427  0.955                  0.000
4                      0.726  0.824                  0.626
5                      0.401  0.966                  0.453

         estado_civil_casado  estado_civil_outro  genero_m  genero_f  \
cluster
0                      0.000               1.000     0.718     0.282
1                      0.000               0.000     1.000     0.000
2                      0.446               0.002     0.997     0.003
3                      1.000               0.000     1.000     0.000
4                      0.374               0.000     0.000     1.000
5                      0.416               0.131     0.780     0.220

         perfil_a  perfil_b  perfil_c  perfil_d
cluster
0           0.780     0.231       0.0     0.409
1           0.992     0.000       0.0     1.000
2           0.001     1.000       0.0     0.000
3           1.000     0.000       0.0     0.825
4           0.875     0.128       0.0     0.650
5           0.000     0.000       1.0     0.000

----------------------------------------------------------------------
Cluster sum of variables   -------------------------------------------
----------------------------------------------------------------------
                id  geo_referencia   valor_01   valor_02   valor_03  \
cluster
0        1.106e+06        162593.0  1.432e+06  7.628e+06  1.658e+06
1        3.553e+06        472919.0  1.746e+06  8.371e+06  6.641e+05
2        2.172e+06        300007.0  2.420e+06  4.361e+07  1.407e+07
3        2.623e+06        371615.0  2.474e+06  1.670e+07  1.684e+06
4        2.305e+06        295841.0  1.337e+06  9.293e+06  5.832e+05
5        6.044e+05         71632.0  6.480e+05  7.074e+06  2.450e+06
```

```
         valor_04  data_nascimento_int        age  estado_civil_solteiro  \
cluster
0       2.950e+06            8.370e+19  18544.975                    0.0
1       3.339e+06            8.029e+20  40687.039                 1407.0
2       7.752e+06            2.883e+20  32411.595                  488.0
3       6.364e+06            2.587e+20  41715.986                    0.0
4       3.541e+06            3.824e+20  31263.225                  578.0
5       1.118e+06            5.613e+19   9736.619                  111.0

         estado_civil_casado  estado_civil_outro  genero_m  genero_f  \
cluster
0                        0.0               451.0     324.0     127.0
1                        0.0                 0.0    1407.0       0.0
2                      394.0                 2.0     881.0       3.0
3                     1062.0                 0.0    1062.0       0.0
4                      345.0                 0.0       0.0     923.0
5                      102.0                32.0     191.0      54.0

         perfil_a  perfil_b  perfil_c  perfil_d
cluster
0           339.0     104.0       0.0       8.0
1          1346.0       0.0       0.0      61.0
2             1.0     883.0       0.0       0.0
3          1024.0       0.0       0.0      38.0
4           779.0     118.0       0.0      26.0
5             0.0       0.0     245.0       0.0
```

Bibliography

K-Means algorithm

1. https://en.wikipedia.org/wiki/K-means_clustering
2. http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans

Silhouette analysis

1. http://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient
2. http://www.sciencedirect.com/science/article/pii/0377042787901257

Preprocessing

1. http://scikit-learn.org/stable/modules/preprocessing.html
2. https://stats.stackexchange.com/questions/21222/are-mean-normalization-and-feature-scaling-needed-for-k-means-clustering