# 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.

`* XXX TODO: include total income/total value in dicusssion. XXX`

# 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.
     XXX TODO: discuss the random initialization of the algorithm; it may yield really different clusters depending on its initialization.
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 multidisciplinary 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.

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

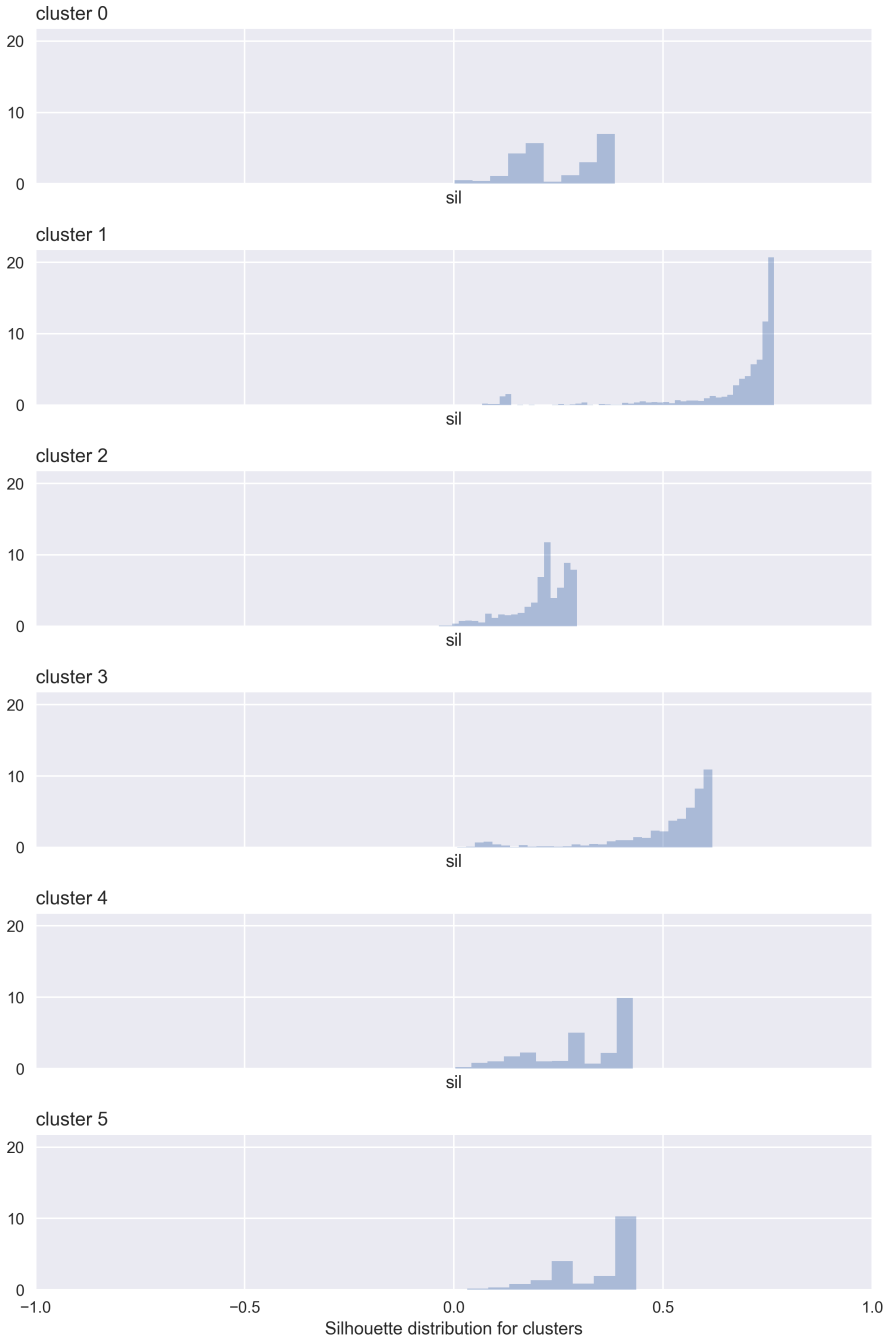
cluster 0
20
10
0
sil
cluster 1
20
10
0
sil
cluster 2
20
10
0
sil
cluster 3
20
10
0
sil
cluster 4
20
10
0
sil
cluster 5
20
10
0
−1.0
−0.5
0.0
0.5
1.0
Silhouette distribution for clusters

Figure 1:

### Cluster 0

*Distinctive features:*

1. Has the most concentration of other marital status (that is, it is neither married nor single).

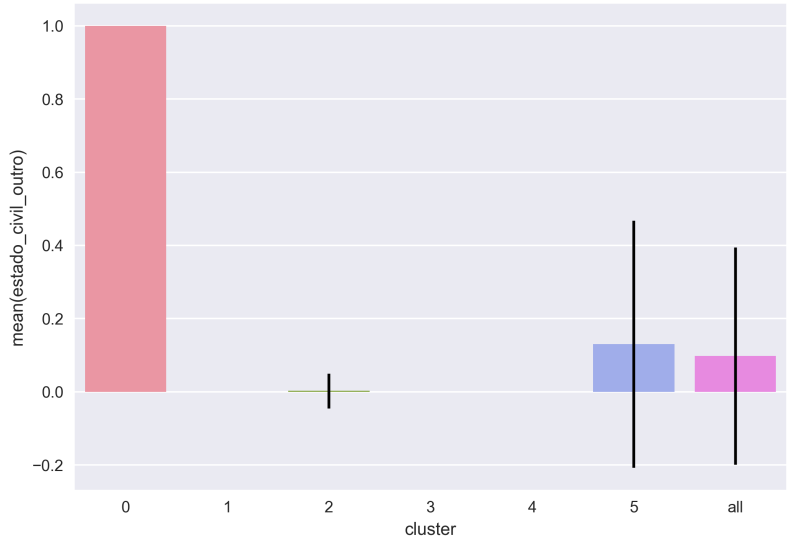


Figure 2:

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).

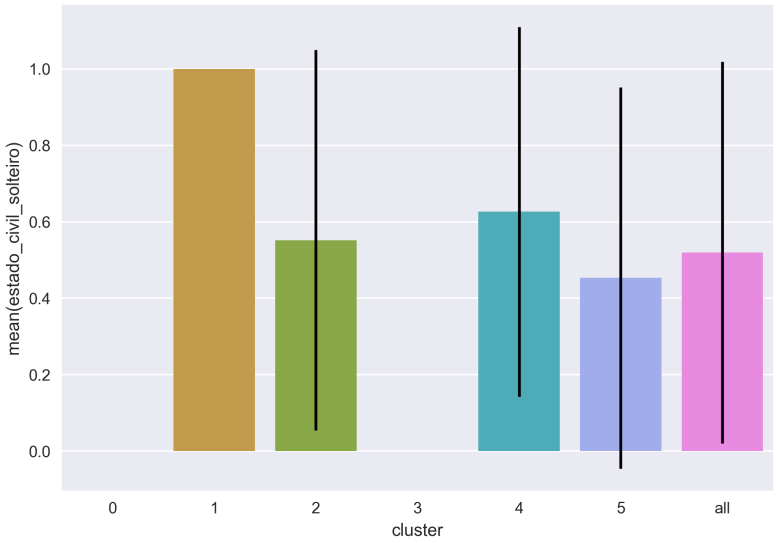


Figure 3:

**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 much 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').

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:*

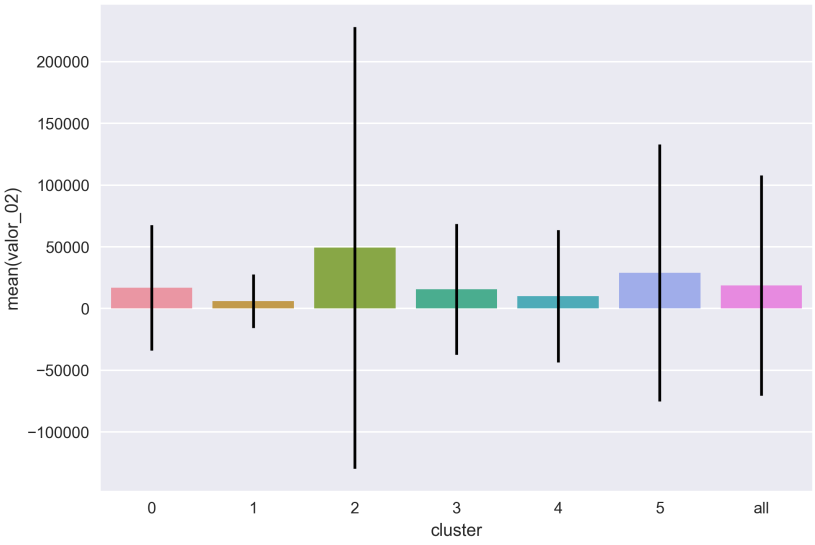


Figure 4:

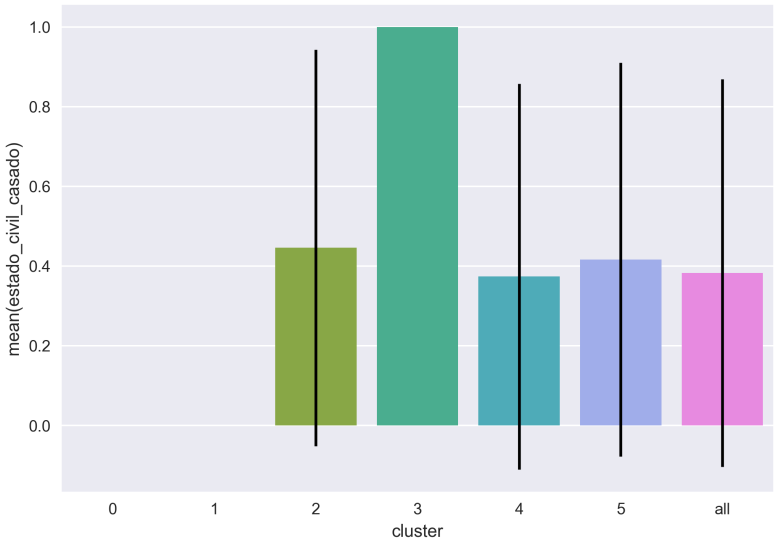


Figure 5:

1. Comprised solely of female subjects:

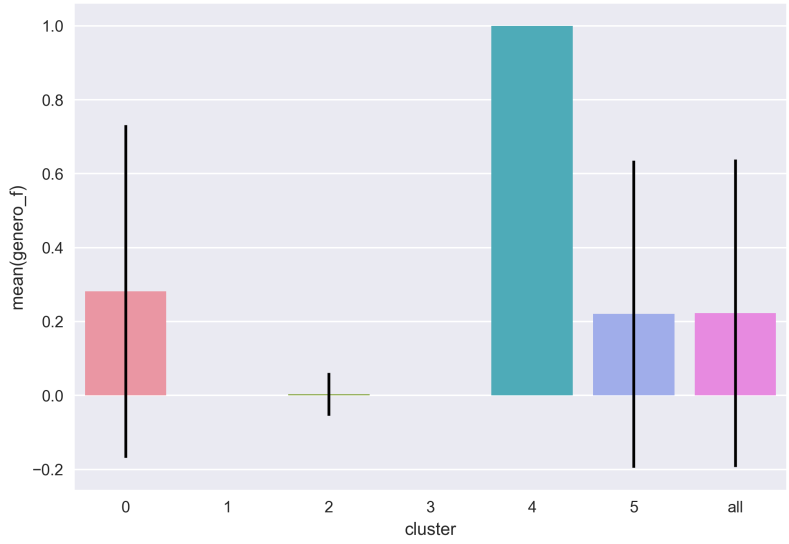


Figure 6:

1. XXX TODO

**Cluster 5**

*Distinctive features:*

1. Comprised solely of profile C:

(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.
2. XXX TODO

# Summary

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

- A satisfactory value of silhouette (XXX TODO comment further XXX)

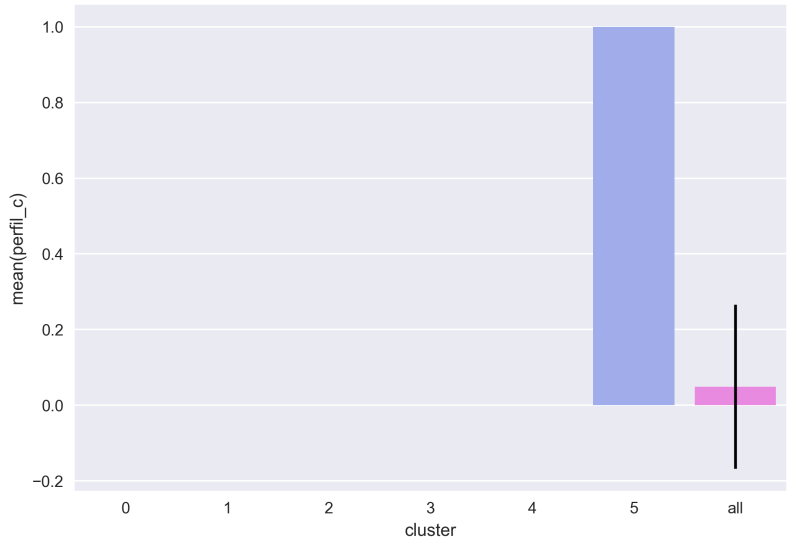


Figure 7:

- Some very sharp separations, some of which are coupled and yield easily interpretable results:
  - a (coupled) XXX TODO
  - b (coupled)
  - c (single)
  - d (single)
- A reasonable amount of clusters, facilitating the communication and interpretation of the results (one of the strenghts of the algorithm)

A quick summary of each cluster's characteristics are:

- Cluster 0: (XXX TODO one liner XXX)
- Cluster 1: (XXX TODO one liner XXX)
- Cluster 2: (XXX TODO one liner XXX)
- Cluster 3: (XXX TODO one liner XXX)
- Cluster 4: (XXX TODO one liner XXX)
- Cluster 5: (XXX TODO one liner XXX)

# 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

- Comment on the data set ; suggest improvements. XXX TODO

# 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

XXX TODO

### Clusters

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

XXX TODO

## Code output (stdout)

XXX TODO

# Bibliography

XXX TODO Improve XXX

## K-Means algorithm

1. https://en.wikipedia.org/wiki/K-means_clustering

## 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. XXX TODO