# 

# Etude de l’évolution du taux de suicide pour 100 000 habitants segmenté par pays, par sexe et par classe d’âge

Thaïs Perdriel et Fernanda Tchouacheu

ADD2 Avril 2019

### 

### Exercice 2 :

En prenant un exemple de jeu de données d’au moins 1000 observations, proposez une étude de votre choix incorporant des notions de cours et des notions extérieures si vous le souhaitez. Les travaux originaux, non plagiés, seront valorisés.

# Sommaire

Sommaire 2

Introduction 3

I) Analyses descriptives traditionnelles 3

II) Clustering non supervisé 10

Réduction du nombre de dimensions du data set 10

Analyses exploratoires 14

Variables retenues : Taux de suicide, année, PIB par année et PIB par habitant 14

Variables retenues : Taux de suicide, Sexe, Catégorie d’âge, Année 16

Variables retenues : Taux de suicide, Sexe, Catégorie d’âge, Année, PIB par année 18

Variables retenues : Taux de suicide, Sexe, Catégorie d’âge, Année, PIB par année, PIB par habitant (Suppression de la variable Sexe) 20

K-means 22

III) Régression linéaire 23

Modèle initiale 23

Amélioration du modèle 24

Niveau 1 : Modification de la variable prédite 24

Niveau 2 : Implémentation des pénalisations Ridge et Lasso 25

Bibliographie 26

Citations 26

Documents consultés 26

# Introduction

Vous nous présentons ci-après un jeu de donnée nommé Suicide Rates Overview 1985 to 2016(1) issu du site internet « kaggle.com » et mis en ligne en décembre 2018 par « Rusty ». Elle présente l’évolution du taux de suicide associé à des indicateurs socio-économiques. Cette base de donnée a été réalisée à partir des données compilées de 4 bases de données principales :

* United Nations Development Program. (2018). Human development index (HDI)(2)
* World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016 (3)
* [Szamil]. (2017). Suicide in the Twenty-First Century [dataset](4)
* World Health Organization. (2018). Suicide prevention(5)

# I) Analyses descriptives traditionnelles

Nous allons étudier une base de donnée composée de 27 820 lignes et de 7 variables. Les données sont disponibles de 1985 a 2016 pour 101 pays différents. L’âge est présenté par catégorie : 5-15 ans, 15-24 ans, 25-24 ans, 35-54 ans, 55-74 ans et 75 ans et plus. Il n’y a pas de valeurs manquantes. A noter : il n’y a pas de données disponibles pour les pays Chine et Inde.

**Tableau 1.** Variables de la base de donnée

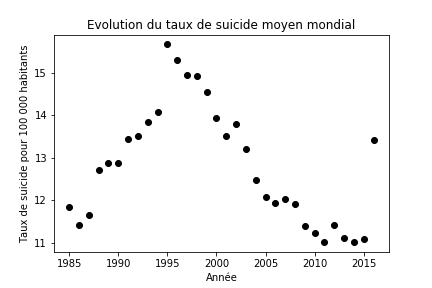
|  |  |
| --- | --- |
| Variable | Nom de la variable  Au sein de la base de donnée |
| Année | year |
| Sexe | sex |
| Classe d’âge | age |
| Taux de suicide | scd\_r |
| Produit intérieur brut (PIB) par année | gdp\_yr |
| Produit intérieur brut (PIB) par habitant | gdp\_cp |
| Pays | Country |

**Tableau 2.** Cinq premières lignes de la base de données



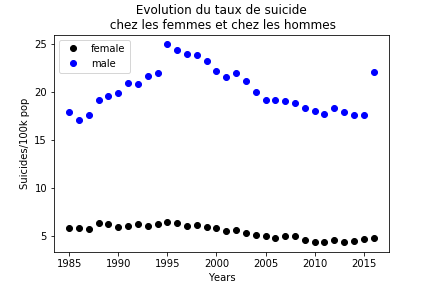
**Tableau 3.** Description des variables quantitatives





**Figure 1.** Evolution du taux de suicide moyen mondial entre 1985 et 2016.

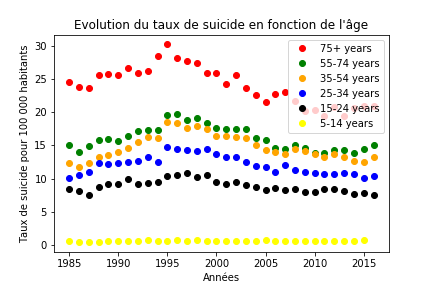
**Analyse :**

* Le taux de suicide mondial par 100k habitants a connu un pic en 1995 à 15.7 par 100 000 (100k) habitants avant de redescendre avec un minimum en 2014 avec 11 suicides par 100k habitants.
* 

**Figure 2.** Evolution du taux de suicide moyen mondial par sexe

**Analyse :**

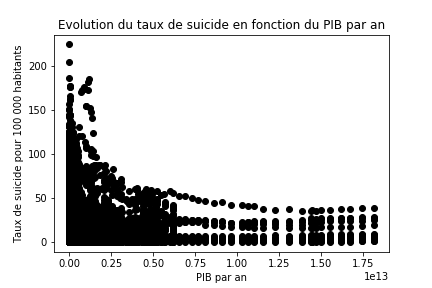
* Les hommes rendent compte de la grande majorité du taux de suicide. Les femmes ont vu une diminution régulière de leur taux de suicide, atteignant un minimum en 2011 avec un ratio à 4.3 pour 100 000 habitants.

****

**Figure 3.** Evolution du taux de suicide moyen mondial par âge

**Analyse :**

* Les jeunes de 5 à 14 ans présente le taux de suicide le plus faible alors que les personnes âgées de plus de 75 ans présentent le taux de suicide le plus élevé.



**Figure 4.** Evolution du taux de suicide en fonction du PIB par an.

**Analyse :**

* Seuls les pays ayant un PIB par an inférieur ou égal à 0.\*10^13 équivalent dollar peuvent présenter un taux de suicide supérieur à 50 pour 100k habitants.
* La plus part des pays des observations semble avoir à la fois un PIB par an bas (inférieur à 0.6\*10^13 par an) et un taux de suicide bas inférieur à 50 pour 100k habitants.

## 

**Figure 5.** Evolution du taux de suicide en fonction du PIB par habitant.

**Analyse :**

* L’impact du PIB par habitant sur le taux de suicide semble être similaire à celui du PIB par an.

# II) Clustering non supervisé

## Réduction du nombre de dimensions du data set

Afin de réduire le nombre de dimensions de notre data set nous allons implémenter une analyse en composante principale (ACP). A noter que notre data set possédant des variables qualitatives il aurait été intéressant de réaliser une analyse factorielle des correspondances (AFC). Cependant le seul module AFC disponible n’étant pas vectorisé son implémentation n’est pas possible du fait d’un délai de calcul trop important.

Techniquement parlant nous allons :

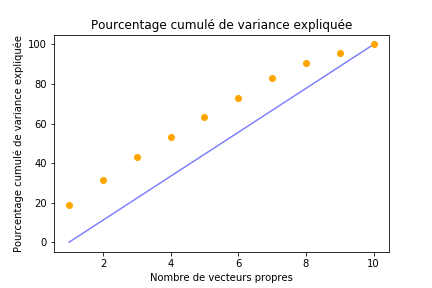
* transformer nos variables quantitatives en variables qualitatives grâce au module DictVectorizer.
* Supprimer la variable «country » qui rajouterait un trop grand nombre de colonne et nous ferait perdre de la lisibilité lors de notre ACP.

**Tableau 4. Résultats de notre première ACP**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dimension** | **Valeur propre** | **Pourcentage de variance expliquée** | **Pourcentage cumulé de variance expliquée** |
| Dim1 | 251.4 | 19.0 | 19.0 |
| Dim2 | 204.6 | 13.0 | 31.0 |
| Dim3 | 196.6 | 12.0 | 43.0 |
| Dim4 | 182.7 | 10.0 | 53.0 |
| Dim5 | 182.7 | 10.0 | 63.0 |
| Dim6 | 182.7 | 10.0 | 73.0 |
| Dim7 | 182.7 | 10.0 | 83.0 |
| Dim8 | 159.0 | 8.0 | 91.0 |
| Dim9 | 128.2 | 5.0 | 96.0 |
| Dim10 | 121.6 | 4.0 | 100.0 |

**Analyse :**

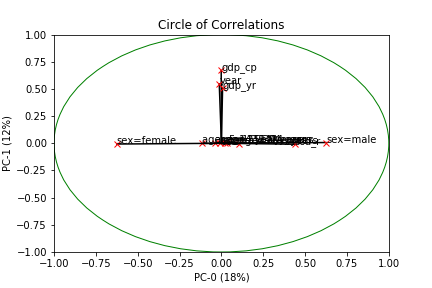
* On observe que l’ACP dans ce cas là n’est pas très efficace
* Les 3 premiers vecteurs sur 10 apportent une variance expliquée entre 12 et 19%.
* Ainsi les différents dimensions de notre data-set apportent chacune une information spécifique avec peu de recoupement.



**Figure 6. Représentation graphique** du pourcentage cumulé de variance expliquée en fonction du nombre de vecteurs propres considéré.

**Analyse :**

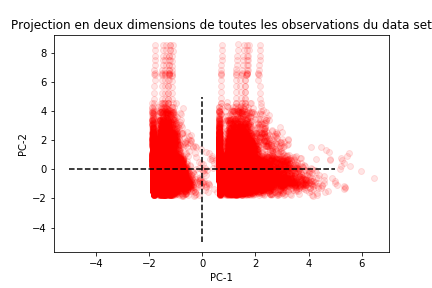
* Cette représentation graphique confirme la faible efficacité de l’ACP
* La courbe de variance expliquée cumulée est pratiquement parallèle à la droite x=y.



**Figure 7. Projection en deux dimensions** des différentes dimensions de notre data set avec en abscisse l’axe du premier vecteur propre et en ordonnée l’axe du deuxième vecteur propre.

**Analyse :**

* Du fait de la superposition des différentes dimensions cette représentation graphique n’est pas très lisible.
* Le PIB par habitant, l’année et le PIB par année semblent apporter le même type d’information.
* Les vecteurs femme et homme divisent le data set en deux comme on pouvait s’y attendre.

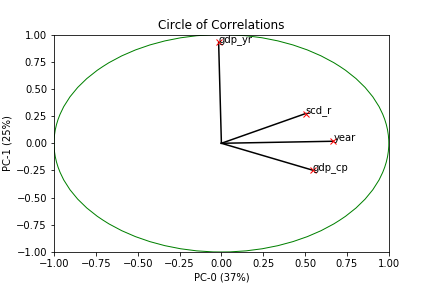


**Figure 8.** Projection en deux dimensions de toutes les observations du data set avec en abscisse PC-1 soit l’axe du premier vecteur propre et en ordonnée PC-2 soit l’axe du deuxième vecteur propre.

## Analyses exploratoires

A titre exploratoire nous vous proposons ci-après d’explorer grâce à une ACP notre data-set réduit à quelques variables. Pour chaque représentation nous vous indiquerons les variables retenues puis nous vous proposerons une analyse.

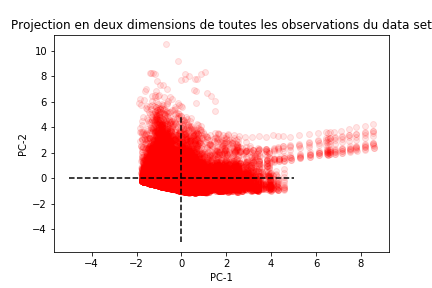
### Variables retenues : Taux de suicide, année, PIB par année et PIB par habitant



**Figure 9.** Projection en deux dimensions des différentes dimensions de notre data set avec en abscisse l’axe du premier vecteur propre et en ordonnée l’axe du deuxième vecteur propre.

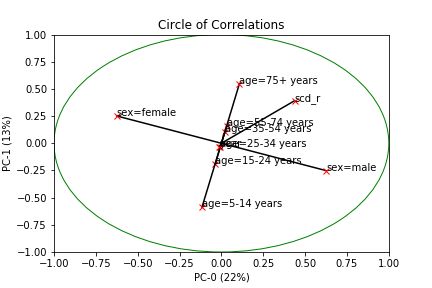
**Analyse :**

* Faible efficacité de l’ACP : les deux premiers vecteurs propres permettent d’expliquer 52% de la variance sur une base de données à 4 dimensions initialement.
* Le PIB par année semble avoir peu d’impact sur le taux de suicide.
* Plus on se rapproche de 2016 et plus le PIB par habitant augmente, plus le taux de suicide semble augmenter. Ceci est étonnant et contraste avec les analyses exploratoires réalisées précédemment.

****

**Figure 10.** Projection en deux dimensions de toutes les observations de notre data set avec en abscisse l’axe du premier vecteur propre et en ordonnée l’axe du deuxième vecteur propre.

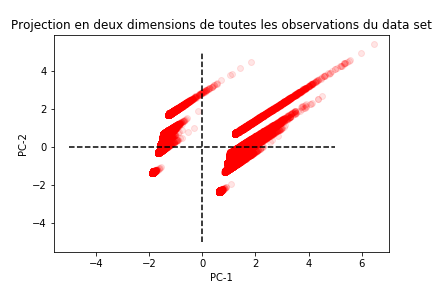
### Variables retenues : Taux de suicide, Sexe, Catégorie d’âge, Année

****

**Figure 11.** Projection en deux dimensions des différentes dimensions de notre data set avec en abscisse l’axe du premier vecteur propre et en ordonnée l’axe du deuxième vecteur propre.

**Analyse :**

* Bonne efficacité de l’ACP : les deux premiers vecteurs propres expliquent 35% de la variance sur une base de données à 10 dimensions
* Ceci est cohérent avec le fait que nous avons artificiellement augmenté le nombre de dimension en distribuant la colonne âge en 6 colonnes afin de rendre son information quantitative.
* Cette projection est en cohérence avec nos observations précédentes : les hommes et les personnes âgées présentent un taux de suicide plus important que les femmes et les jeunes.

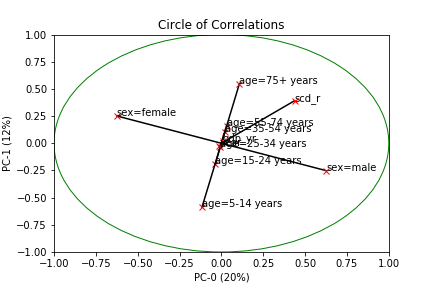
****

**Figure 12.** Projection en deux dimensions de toutes les observations du data set avec en abscisse l’axe du premier vecteur propre et en ordonnée l’axe du deuxième vecteur propre.

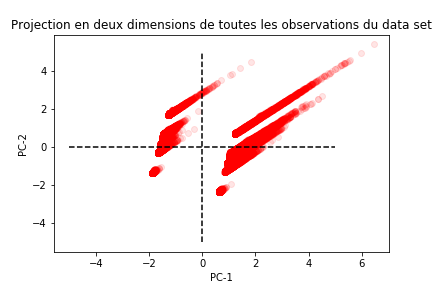
**Analyse :**

* En cohérence avec le cercle de corrélation présenté ci-avant les observations sont segmentés en deux groupes semblant correspondre à gauche aux femmes et à droite aux homme. Chaque groupe est ensuite à nouveau segmenté en 5 avec en bas le groupe semblant correspondre aux âges les plus jeunes et en haut ceux semblant correspondre aux âges les plus avancés.

### Variables retenues : Taux de suicide, Sexe, Catégorie d’âge, Année, PIB par année

****

**Figure 13.** Projection en deux dimensions des différentes dimensions du data set avec en abscisse l’axe du premier vecteur propre et en ordonnée l’axe du deuxième vecteur propre.

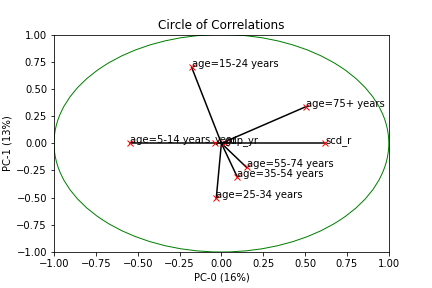


**Figure 14.** Projection en deux dimensions de toutes les observations de notre data set avec en abscisse l’axe du premier vecteur propre et en ordonnée l’axe du deuxième vecteur propre.

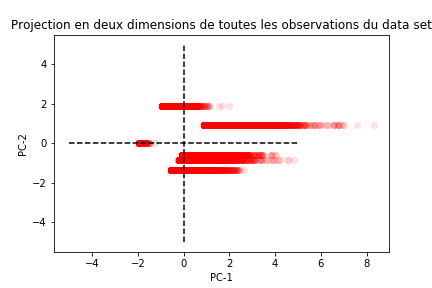
**Analyse :**

* En cohérence avec la Figure 4 la représentation est peu modifiée, le vecteur PIB par année n’apporte que peu d’information supplémentaire par rapport à l’âge et au sexe. De la même manière les deux premiers vecteurs propres expliquent toujours 32% de la variance malgré le rajout d’une dimension à la base de donnée initiale.

### Variables retenues : Taux de suicide, Sexe, Catégorie d’âge, Année, PIB par année, PIB par habitant (Suppression de la variable Sexe)

****

**Figure 15.** Projection en deux dimensions des différentes dimensions de notre data set avec en abscisse l’axe du premier vecteur propre et en ordonnée l’axe du deuxième vecteur propre.



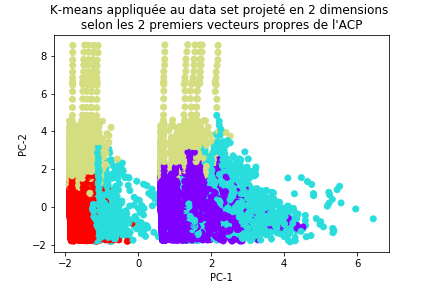
**Figure 16.** Projection en deux dimensions de toutes les observations de notre data set avec en abscisse l’axe du premier vecteur propre et en ordonnée l’axe du deuxième vecteur propre.

**Analyse**

* On observe des variables séparées en 6 groupes distincts semblant correspondre à chaque groupe d’âge d’après le cercle de corrélation. La représentation en groupes distincts est cohérente avec le fait que la variable âge est codée de manière binaire dans la base de donnée initiale et que chaque observation ne peut prendre 1 que pour une seule catégorie d’âge.

## K-means

Pour cloturer ce chapitre exploratoire réalisons un K-means sur l’intégralité des données (c’est à dire les 6 variables) projetées en deux dimensions les deux premiers vecteurs propres de l’ACP (comme réalisé pour la **Figure 8**).



**Figure 17.** K-means appliquée au data set contenant toutes les variables, projeté en deux dimensions selon les deux premiers vecteurs propres obtenue par application d’une ACP.

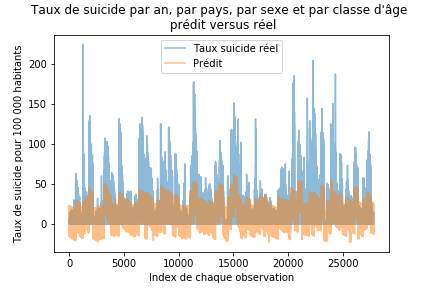
**Analyse :**

* La couleur verte semble représenter les observations des pays à hauts revenus.
* La couleur rouge semble représenter les femmes jeunes des pays à bas revenus.
* La couleur violette semble représenter les hommes jeunes des pays à bas revenus
* La couleur bleue semble représenter les personnes âgées.

# III) Régression linéaire

## Modèle initiale

Nous allons maintenant implémenter une technique de prédiction basique : la régression linéaire.

****

**Figure 18 :** Taux de suicide prédit versus réel de chacune des 27 820 observations.

**Analyse :**

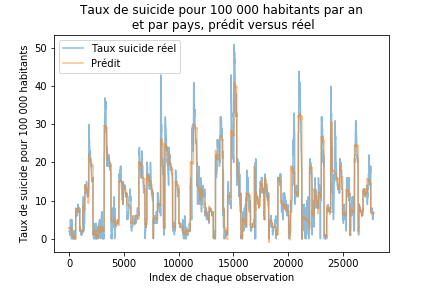
* Notre modèle est peu efficient avec un coefficient de détermination R^2 à 0.51.
* Il existe un biais avec un taux de suicide prédit pratiquement systématiquement sous-estimé avec des résultats régulièrement aberrants (taux de suicide prédit inférieur à 0).
* Cependant on observe un certain respect des relations d’ordre avec : si taux\_suicide\_a > taux\_de suicide\_b alors taux\_suicide\_prédit\_a > taux\_suicide\_prédit\_b.

## Amélioration du modèle

### Niveau 1 : Modification de la variable prédite

Une des failles de notre modèle est le manque de variables explicatives. En effet le taux de suicide est présenté par pays, par an, par sex et par classe d’âge. Les deux seules variables indépendantes disponibles sont le PIB par an dont on a vu le faible impacte sur le taux de suicide dont on a vu le faible impact sur le taux de suicide par rapport aux variables sexe et classe d’âge, et le PIB par habitant.

Nous allons donc essayer de prédire le taux de suicide moyen par pays et par année.



**Figure 19.** Taux de suicide pour 100 000 habitants par an et par pays, prédit versus réel, de chacune des 27 820 observations.

**Analyse :**

* Comme attendu notre modèle est beaucoup plus performant avec un coefficient de corrélation R^2 à 0.87, c’est à dire que notre courbe de régression est capable de déterminer 87% de la distribution des points.
* Il reste cependant peu efficient pour la prédiction des valeurs extrêmes supérieures.

### Niveau 2 : Implémentation des pénalisations Ridge et Lasso

Après implémentation d’une technique de cross-validation nous observons que l’erreur absolue à la moyenne de notre régression linéaire est de 2.12.

Evaluons l’évolution de l’erreur absolue à la moyenne après implémentation des pénalisations Ridge et Lasso pour des valeurs de coefficient alpha évoluant entre 10^-6 et 10^4.

**Tableau 5.** Erreur absolue à la moyenne de la régression linéaire avec pénalisation de type Lasso et Ridge en fonction du coefficient alpha.



**Analyse :**

* Pour un coefficient alpha inférieur ou égale à 5 pour la pénalisation ridge et inférieur ou égale à 0.001 pour la pénalisation lasso, l’erreur absolue à la moyenne est inférieure est inférieure que celle d’une régression linéaire standard. Les erreurs moyennes à la moyenne les plus faibles sont obtenues pour les coefficients alpha les plus bas avec des efficacités équivalentes pour les pénalisations type lasso et ridge.

# Bibliographie

## Citations

1. Rusty. Suicide Rates Overview 1985 to 2016 [Internet]. Kaggle. 2018 [cité 29 avr 2019]. Disponible sur: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

2. United Nations Developtment Programme. Human Development Index (HDI) [Internet]. Human Development Reports. [cité 29 avr 2019]. Disponible sur: http://hdr.undp.org/en/indicators/137506

3. The World Bank. World Development Indicators [Internet]. Data Bank from the World Bank. [cité 29 avr 2019]. Disponible sur: https://databank.worldbank.org/data/source/world-development-indicators#

4. WHO Suicide Statistics [Internet]. Kaggle. 2018 [cité 29 avr 2019]. Disponible sur: https://www.kaggle.com/szamil/who-suicide-statistics

5. WHO. Latest Data on Suicide [Internet]. World Health Organization. 2018 [cité 29 avr 2019]. Disponible sur: https://www.who.int/mental\_health/suicide-prevention/en/

## Documents consultés

6. VanderPlas J. Python Data Science Handbook. O’Reilly. United States of America: Dawn Schanafelt; 2016. 530 p.

7. Scikit-Learn [Internet]. Disponible sur: https://scikit-learn.org/stable/

8. Stack Overflow [Internet]. Disponible sur: https://stackoverflow.com/

9. Liam M. World Health Organisation: Global Suicide Trends & Analysis [Internet]. Kaggle. 2019 [cité 29 avr 2019]. Disponible sur: https://www.kaggle.com/lmorgan95/r-suicide-rates-in-depth-stats-insights