Projet_ModIA

Introduction

Brève introduction sur le sujet, la problématique principale , les questions que l'on se pose et comment on va essayer de trouver une solution (quels données on aura besoin).

Jeu de données

1) Récupération des données et analyse introductoire

```
# on ouvre le jeu de données
data = read.table("LifeExpectationData-Etudiants.csv")
```

```
# on affiche les 6 premières lignes
head(data)
```

```
##                      Country Year     Status Life.expectancy Adult.Mortality
## Afghanistan-2014 Afghanistan 2014 Developing            59.9             271
## Afghanistan-2013 Afghanistan 2013 Developing            59.9             268
## Afghanistan-2012 Afghanistan 2012 Developing            59.5             272
## Afghanistan-2011 Afghanistan 2011 Developing            59.2             275
## Afghanistan-2010 Afghanistan 2010 Developing            58.8             279
## Afghanistan-2009 Afghanistan 2009 Developing            58.6             281
##                  infant.deaths Alcohol percentage.expenditure Hepatitis.B
## Afghanistan-2014            64    0.01              73.523582          62
## Afghanistan-2013            66    0.01              73.219243          64
## Afghanistan-2012            69    0.01              78.184215          67
## Afghanistan-2011            71    0.01               7.097109          68
## Afghanistan-2010            74    0.01              79.679367          66
## Afghanistan-2009            77    0.01              56.762217          63
##                  Measles  BMI under.five.deaths Polio Total.expenditure
## Afghanistan-2014     492 18.6                86    58              8.18
## Afghanistan-2013     430 18.1                89    62              8.13
## Afghanistan-2012    2787 17.6                93    67              8.52
## Afghanistan-2011    3013 17.2                97    68              7.87
## Afghanistan-2010    1989 16.7               102    66              9.20
## Afghanistan-2009    2861 16.2               106    63              9.42
##                  Diphtheria HIV.AIDS       GDP Population thinness..1.19.years
## Afghanistan-2014         62      0.1 612.69651     327582                 17.5
## Afghanistan-2013         64      0.1 631.74498   31731688                 17.7
## Afghanistan-2012         67      0.1 669.95900    3696958                 17.9
## Afghanistan-2011         68      0.1  63.53723    2978599                 18.2
```

```
## Afghanistan-2010         66      0.1 553.32894    2883167                 18.4
## Afghanistan-2009         63      0.1 445.89330     284331                 18.6
##                   thinness.5.9.years Schooling
## Afghanistan-2014               17.5      10.0
## Afghanistan-2013               17.7       9.9
## Afghanistan-2012               18.0       9.8
## Afghanistan-2011               18.2       9.5
## Afghanistan-2010               18.4       9.2
## Afghanistan-2009               18.7       8.9
```

(petit commentaire sur le jeu de données)

2) Exploration du jeu de données

(expliquer en bref l'objectif de cette partie (regarder s'il y a des données manquantes, reduire nbr de variables, identifier les outliers, enlever les valeurs aberrantes, etc.)) ### Summary des données

```
summary(data)
```

```
##    Country               Year         Status          Life.expectancy
##  Length:1647        Min.   :2000   Length:1647        Min.   :44.00
##  Class :character   1st Qu.:2005   Class :character   1st Qu.:64.35
##  Mode  :character   Median :2008   Mode  :character   Median :71.70
##                     Mean   :2008                      Mean   :69.30
##                     3rd Qu.:2011                      3rd Qu.:75.00
##                     Max.   :2014                      Max.   :89.00
##  Adult.Mortality infant.deaths        Alcohol       percentage.expenditure
##  Min.   :  1.0   Min.   :   0.00   Min.   : 0.010   Min.   :    0.0
##  1st Qu.: 77.0   1st Qu.:   1.00   1st Qu.: 0.815   1st Qu.:   37.3
##  Median :148.0   Median :   3.00   Median : 3.790   Median :  145.1
##  Mean   :168.2   Mean   :  32.55   Mean   : 4.536   Mean   :  699.6
##  3rd Qu.:227.0   3rd Qu.:  22.00   3rd Qu.: 7.345   3rd Qu.:  510.0
##  Max.   :723.0   Max.   :1600.00   Max.   :17.870   Max.   :18961.3
##   Hepatitis.B       Measles              BMI        under.five.deaths
##  Min.   : 2.00   Min.   :     0.0   Min.   : 2.00   Min.   :   0.00
##  1st Qu.:74.00   1st Qu.:     0.0   1st Qu.:19.50   1st Qu.:   1.00
##  Median :89.00   Median :    15.0   Median :43.70   Median :   4.00
##  Mean   :79.21   Mean   :  2226.5   Mean   :38.13   Mean   :  44.22
##  3rd Qu.:96.00   3rd Qu.:   372.5   3rd Qu.:55.80   3rd Qu.:  29.00
##  Max.   :99.00   Max.   :131441.0   Max.   :77.10   Max.   :2100.00
##      Polio      Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.0   Min.   : 0.740    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:81.0   1st Qu.: 4.405    1st Qu.:82.00   1st Qu.: 0.100
##  Median :93.0   Median : 5.840    Median :92.00   Median : 0.100
##  Mean   :83.6   Mean   : 5.955    Mean   :84.16   Mean   : 1.986
##  3rd Qu.:97.0   3rd Qu.: 7.465    3rd Qu.:97.00   3rd Qu.: 0.700
##  Max.   :99.0   Max.   :14.390    Max.   :99.00   Max.   :50.600
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.100
##  1st Qu.:   461.94   1st Qu.:1.929e+05   1st Qu.: 1.600
##  Median :  1592.57   Median :1.420e+06   Median : 3.000
##  Mean   :  5570.04   Mean   :1.465e+07   Mean   : 4.845
```

```
##  3rd Qu.:  4727.00   3rd Qu.:7.626e+06   3rd Qu.: 7.050
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.200
##  thinness.5.9.years   Schooling
##  Min.   : 0.100     Min.   : 4.20
##  1st Qu.: 1.700     1st Qu.:10.35
##  Median : 3.200     Median :12.30
##  Mean   : 4.902     Mean   :12.12
##  3rd Qu.: 7.100     3rd Qu.:14.00
##  Max.   :28.200     Max.   :20.70
```

Commentaires : 1. Il peut y avoir des valeurs manquantes 2. On a pas les données de tous les pays pour chaque anné 3. status : au lieu de chaîne de caractères, on peut assigner des booléens 4. on peut regrouper la couverture vaccinale de Hepatitis.B, Polio et HIV.AIDS en une seule variable (si l'écart n'est pas trop grand) 5. percentage.expenditure : plusieurs pourcentages dépassent 100 (il faut peut être l'enlevé)

```
colSums(is.na(data))
```

Valeurs manquantes

```
##                Country                   Year                 Status
##                      0                      0                      0
##        Life.expectancy        Adult.Mortality          infant.deaths
##                      0                      0                      0
##                Alcohol percentage.expenditure            Hepatitis.B
##                      0                      0                      0
##                Measles                    BMI      under.five.deaths
##                      0                      0                      0
##                  Polio      Total.expenditure             Diphtheria
##                      0                      0                      0
##               HIV.AIDS                    GDP             Population
##                      0                      0                      0
##   thinness..1.19.years     thinness.5.9.years              Schooling
##                      0                      0                      0
```

Il n'y a donc pas de valeurs manquantes dans le jeu de données.

Variable Year On compte le nombre d'occurences pour chaque pays. On devrait avoir 133.

```
table(data$Year)
```

```
##
## 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
##   61   66   81   95  103  110  114  120  123  126  128  130  129  130  131
```

On remarque qu'il y a 2 pays qui n'ont pas de données pour 2014. (on doit les identifier et s'îl leur manque d'autres années, on les vire (on peut aussi simplement les virer directement vu que notre étude se base principalement en 2014 jsp))

```
# Identification des pays :
```

Variable Status (on peut remplacer developed par 1 et developping par 0)

Variables Hepatitis.B, Polio et HIV.AIDS (On explique notre idée que les pays auront en général des couvertures vaccinales similaires pour ces 3 vaccins et donc de regrouper les 3 variables en une seule en prenant la moyenne) (On peut également controler qu'il n'y a pas de grandes variations dans les couvertures au cours des années)

```
country_2014_polio = data[which(data$Year == 2014),]$Polio
country_2014_dipht = data[which(data$Year == 2014),]$Diphtheria
country_2014_hB = data[which(data$Year == 2014),]$Hepatitis.B

summary(country_2014_polio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.0    78.0    92.0    83.5    97.0    99.0
```

```
summary(country_2014_dipht)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   80.00   92.00   83.89   97.00   99.00
```

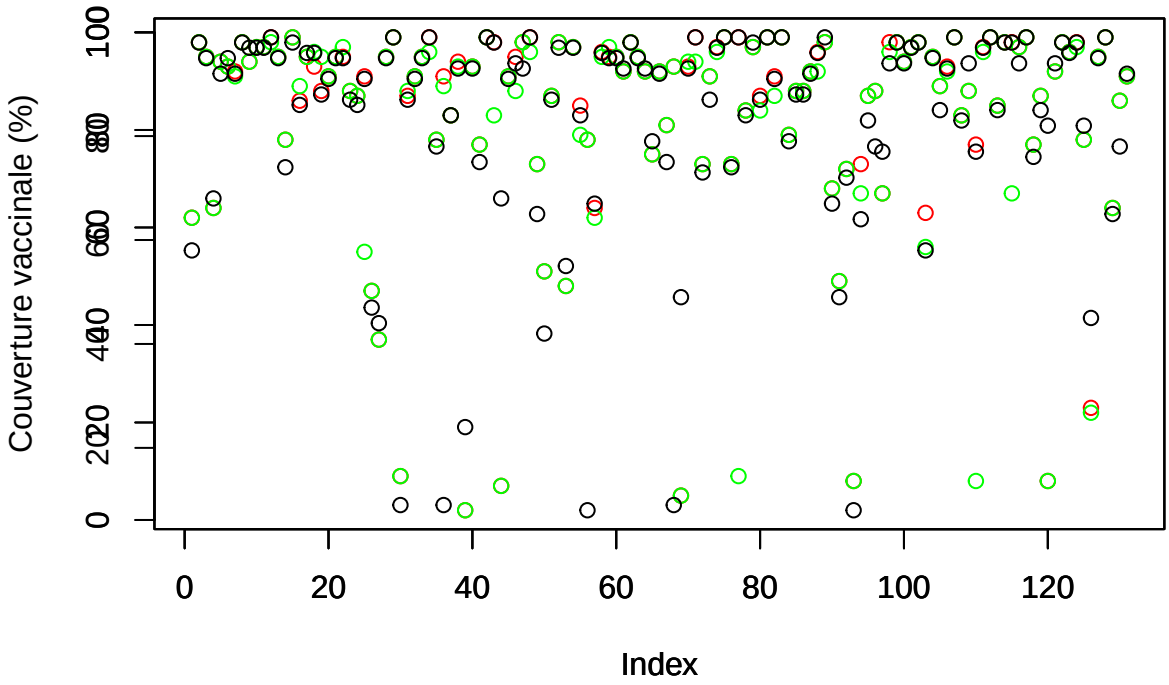
```
summary(country_2014_hB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   78.00   91.00   81.71   96.00   99.00
```

Les valeurs sont très similaires.

(faudrait essayer de faire avec ggplot car c'est tres approximatif ce graphique, les échelles sont pas les mêmes)

```
plot(country_2014_dipht, col = "red", ylab = "")
par(new = TRUE)
plot(country_2014_hB, col = "green", ylab = "")
par(new = TRUE)
plot(country_2014_polio, col = "black", ylab = "Couverture vaccinale (%)")
```



(on peut dire que de façon générale, on peut prendre la moyenne)

```
# création de la nouvelle variable couverture vaccinale qui est la moyenne des 3 autres pour chaque pay
```