Statistiques descriptives STA101

Analyse de données du terrorisme dans le monde en 2017

Cong Chen, Thomas Sadigh Rezvani

Introduction:

Ces dernières années, les attaques terroristes sont devenues un problème mondial.
Nous récupérons le jeu de données de l'Université du Maryland et la dernière année disponible : 2017 ainsi que certaines variables que nous avons séléctionnées pour tester nos hypothèses et répondre 2 problématiques larges: l'existence de relation entre les variables ou leurs modalités et l'existence de groupes parmi les attentats survenus en 2017 :

Ainsi :
Y-a-t-il une relation entre les variables géographiques et les types d'attaques survenues en 2017?
Y-a-t-il un lien entre la méthode d'attaque et les dommages causés?
Les attaques peuvent-elles êtres regroupées en différents profils?

Les données:

On récupère un fichier qui comporte les atatques survenues en 2017 (https://www.start.umd.edu/gtd/). Il y en a plus de 1000.
Sous le langage Python, on nettoye les données, séléctionne les variables d'intérêt et tire un échantillon plus petit que l'on utilisera en R.
Les données récupérées disposent en ligne (individus statistiques) : les attaques survenues et en colonnes (variables) : 16 variables : 11 variables catégorielles (l'une est le nom du pays dans lequel l'attaque a lieu, elle est finalement enlevée sous R car redondante avec la variable "regiontxt" qui est plus synthétique) et 6 variables quantitatives.
On a finalement un fichier de 340 individus et 17 variables. On considère dans la suite de l'étude que le nombre d'individus permet de supposer à priori la significativité des éventuelles relations que l'on trouverait parmi nos données (entre modalités par exemple...)
On présente les données :

Variable	Type de variable	Définition
country_txt	Qualitative	This field identifies the country or location where the incident occurred.
region_txt	Qualitative	This field identifies the region in which the incident occurred.
latitude	Quantitative	Numeric Variable
longitude	Quantitative	Numeric Variable
doubtterr	Qualitative	In certain cases there may be some uncertainty

		whether an incident meets all of the criteria for inclusion.
multiple	Qualitative	In those cases where several attacks are connected, but where the various actions do not constitute a single incident (either the time of occurrence of incidents or their locations are discontinuous – see Single Incident Determination section above), then “Yes” is selected to denote that the particular attack was part of a “multiple” incident.

success	Qualitative	Success of a terrorist strike is defined according to the tangible effects of the attack.
suicide	Qualitative	This variable is coded “Yes” in those cases where there is evidence that the perpetrator did not intend to escape from the attack alive.
attacktype1_txt	Qualitative	This field captures the general method of attack and often reflects the broad class of tactics used.
targtype1_txt	Qualitative	The target/victim type field captures the general type of target/victim.
natlty1_txt	Qualitative	This is the nationality of the target that was attacked, and is not necessarily the same as the country in which the incident occurred, although in most cases it is.
nperpcap	Quantitative	This field records the number of perpetrators taken into custody.
weaptype1_txt	Qualitative	Up to four weapon types are recorded for each incident.
nkill	Quantitative	This field stores the number of

		total confirmed fatalities for the incident.
nkillter	Quantitative	Limited to only perpetrator fatalities, this field follows the conventions of the “Total Number of Fatalities” field described above.
nwound	Quantitative	This field records the number of confirmed non-fatal injuries to both perpetrators and victims.

INT_LOG	Qualitative	This variable is based on a comparison between the nationality of the perpetrator group and the location of the attack. It indicates whether a perpetrator group crossed a border to carry out an attack.

Avant de commencer les analyses avec les différentes méthodes, il est important d’assurer que les données sont bien importées.

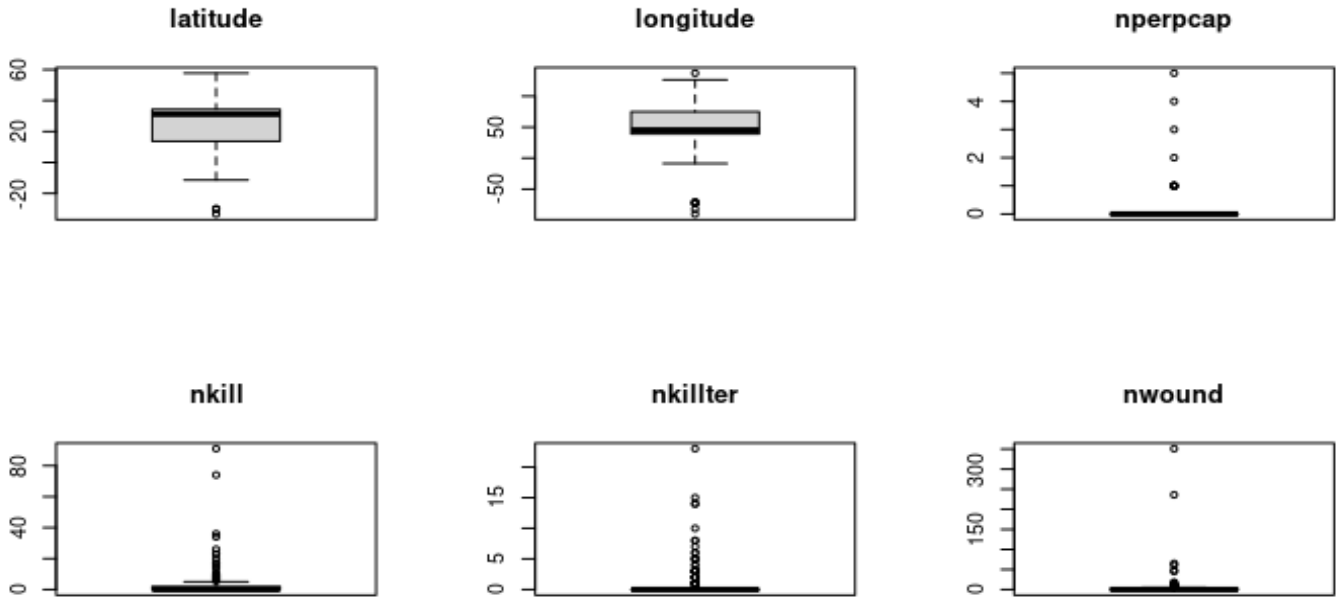
```
> head(df)
  country_txt               region_txt  latitude longitude doubtterr        multiple         success        suicide   attacktype1_txt
1        Iraq Middle East & North Africa 35.324825  43.76692      terr noseries_attack realized_attack     no_suicide Bombing/Explosion
2       Kenya         Sub-Saharan Africa -0.393767  40.32662      terr noseries_attack realized_attack     no_suicide Bombing/Explosion
3        Iraq Middle East & North Africa 36.158558  43.25510      terr noseries_attack realized_attack     no_suicide           Unknown
4      Sweden             Western Europe 61.305668  17.05815      terr noseries_attack realized_attack     no_suicide     Armed Assault
5       India                 South Asia 25.838593  84.57922      terr noseries_attack realized_attack     no_suicide     Assassination
6       India                 South Asia 33.966527  74.96422      terr   within_series       no_attack suicide_attack     Armed Assault
              targtype1_txt natlty1_txt nperpcap weaptype1_txt nkill nkillter nwound            INT_LOG
1                   Unknown        Iraq        0    Explosives     0        0      0     internat_unkown
2           Telecommunication       Kenya        0    Explosives     0        0      0 international_attack
3                  Military        Iraq        0       Unknown     3        0      0     domestic_attack
4 Private Citizens & Property      Sweden        0    Incendiary     0        0      0     internat_unkown
5        Government (General)       India        0      Firearms     1        0      0     internat_unkown
6                    Police       India        0    Explosives     8        3      3 international_attack
> summary(df)
      country_txt                     region_txt     latitude        longitude           doubtterr              multiple
 Iraq       :174   Middle East & North Africa:261   Min.   :-23.28   Min.   :-122.03   doubt_terr:132   noseries_attack:656
 Afghanistan: 85   South Asia                :235   1st Qu.: 15.07   1st Qu.:  41.15   terr      :576   within_series  : 52
 India      : 72   Sub-Saharan Africa        :114   Median : 31.96   Median :  44.62
 Pakistan   : 63   Southeast Asia            : 60   Mean   : 25.99   Mean   :  52.87
 Philippines: 44   Western Europe            : 16   3rd Qu.: 34.36   3rd Qu.:  70.65
 Nigeria    : 33   South America             :  8   Max.   : 61.31   Max.   : 137.12
 (Other)    :237   (Other)                   : 14
           success              suicide                          attacktype1_txt                    targtype1_txt      natlty1_txt
 no_attack      : 52   no_suicide    :656   Bombing/Explosion             :332    Private Citizens & Property:181   Iraq       :173
 realized_attack:656   suicide_attack: 52   Armed Assault                 :168    Military                   :146   Afghanistan: 83
                                            Assassination                 : 67    Police                     : 92   India      : 73
                                            Facility/Infrastructure Attack: 50    Unknown                    : 78   Pakistan   : 62
                                            Unknown                       : 45    Government (General)       : 66   Philippines: 43
                                            Hostage Taking (Kidnapping)   : 33    Business                   : 56   Nigeria    : 33
                                            (Other)                       : 13    (Other)                    : 89   (Other)    :241
    nperpcap                  weaptype1_txt     nkill           nkillter          nwound                       INT_LOG
 Min.   :-99.0000   Chemical          :  1   Min.   :  0.000   Min.   : 0.0000   Min.   :  0.000   domestic_attack     :328
 1st Qu.:  0.0000   Explosives        :352   1st Qu.:  0.000   1st Qu.: 0.0000   1st Qu.:  0.000   internat_unkown     :350
 Median :  0.0000   Firearms          :230   Median :  0.000   Median : 0.0000   Median :  0.000   international_attack: 30
 Mean   : -0.8771   Incendiary        : 45   Mean   :  2.831   Mean   : 0.8263   Mean   :  2.288
 3rd Qu.:  0.0000   Melee             : 22   3rd Qu.:  2.000   3rd Qu.: 0.0000   3rd Qu.:  2.000
 Max.   :  8.0000   Sabotage Equipment:  1   Max.   :266.000   Max.   :41.0000   Max.   :115.000
                    Unknown           : 57
```

L’apperçu global ne révèle pas de problème particulier.

1. Prise en main des données

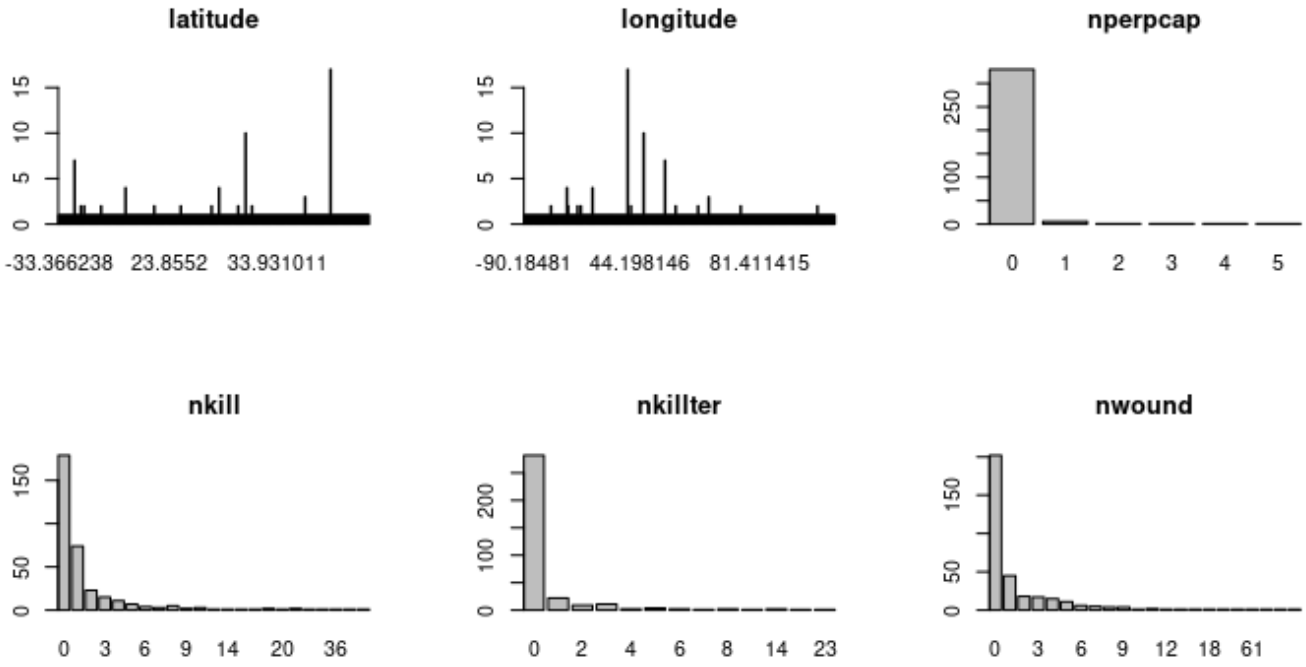
1.1 Analyses univariées
Les fonctions boxplot et barplot nous permettent d'avoir une visibilités sur chaque variable quantitative.

```
#Boxplots
par(mfrow = c(2, 3))
mapply(df[,quanti],
       FUN = function(xx,name){boxplot(xx, main = name,id.n = 4, ylab = "")},
       name = names(quanti))
```



On observe des spécificités, on complète avec l'analyse des "count" des variables en barplot

```
#barplots
par(mfrow = c(2, 3))
mapply(df[,quanti],
       FUN = function(xx,name){barplot(table(xx),main = name)},
       name = names(quanti))
```



Cela suggère certaines variables issues de distributions asymétriques.

On analyse dans un premier temps les variables quantitatives et qualitatives séparément avant de faire une analyse globale par AFDM.

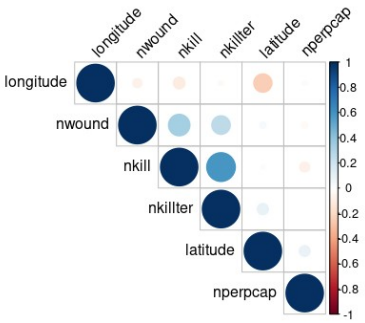
1.2 analyse bivariée

On étudie les variables quantitatives:

On applique un calcul du **coefficient de Spearman** qui mesure une corrélation sans hypothèse quant à la linéarité de la relation entre les variables. En effet, alors qu'on applique la fonction "barplot" il apparaît que plusieurs variables semblent issues de distributions très asymétrique et ayant une silhouette proche d'une décroissante exponentielle.
Les calculs du coefficients de Spearman sont finalement proches de ceux du coefficient de Pearson. Mais on utilise tout de même les coefficients de Spearman.

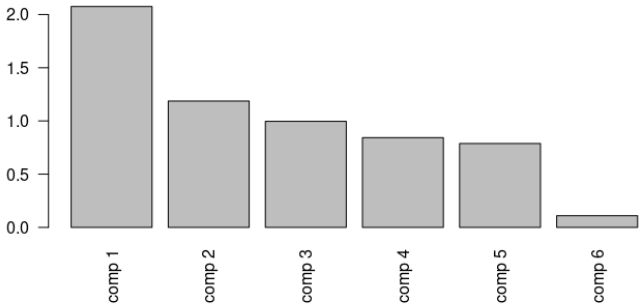
Il n'apparaît qu'une relation:
On voit une **relation entre le nombre de victimes civiles et le nombre de terroristes morts** lors de l'attaque. Elles sont corrélées à ~59%.
Cela n'est pas surprenant.

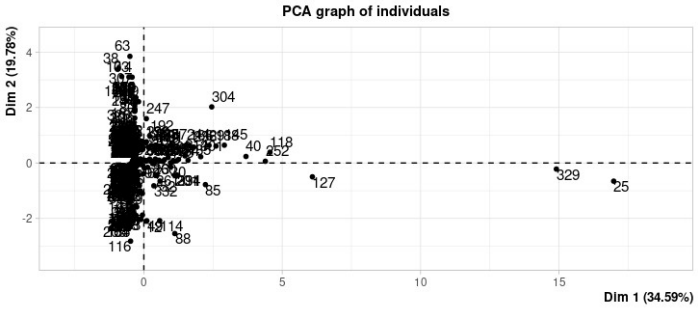
	latitude	longitude	nperpcap	nkill	nkillter
latitude	1.00000000	-0.24363343	0.088657332	-0.01136850	0.092488926
longitude	-0.24363343	1.00000000	-0.019057765	-0.10184212	-0.022150335
nperpcap	0.08865733	-0.01905777	1.000000000	-0.07312868	0.002821883
nkill	-0.01136850	-0.10184212	-0.073128679	1.00000000	0.589537709
nkillter	0.092	-0.01136850	15034	0.002821883	0.58953771	1.000000000
nwound	0.03130097	-0.06414360	-0.026983810	0.33517758	0.250938923

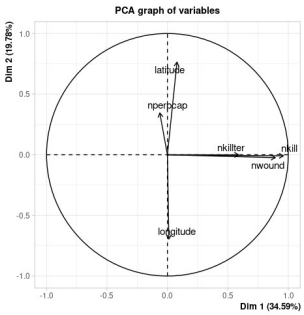


2. Partie ACP :

La structure du nuage de points des individus suggère une **absence de groupes biens définis**
On voit que la projection de 5 variables permet de **conserver une bonne information** car PC1 + PC2 ~**55% de variance** du nuage de points
L'allure de la variance capturées par les composantes montre aussi que les PC sont capables de capturer beaucoup de variance et donc décroître relativement rapidement.
Ainsi:





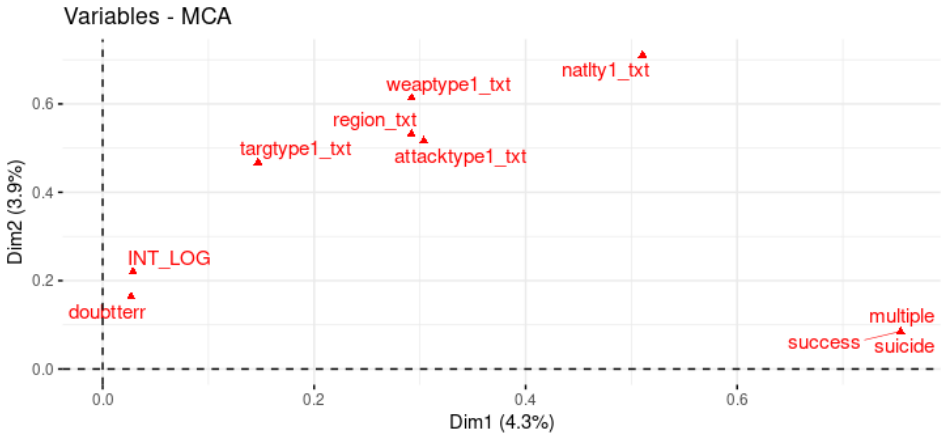


Le cercle des corrélations montre que : plus les individus sont projetés avec des valeurs positives sur PC1 : plus l'attaque est meutrière :
PC 2 disperse selon les coordonnées geographiques

Les variables "nkill", "nkillter" et "nwound" sont assez corrélées
On note une absence de groupes nets avec ces 6 variables quantitatives

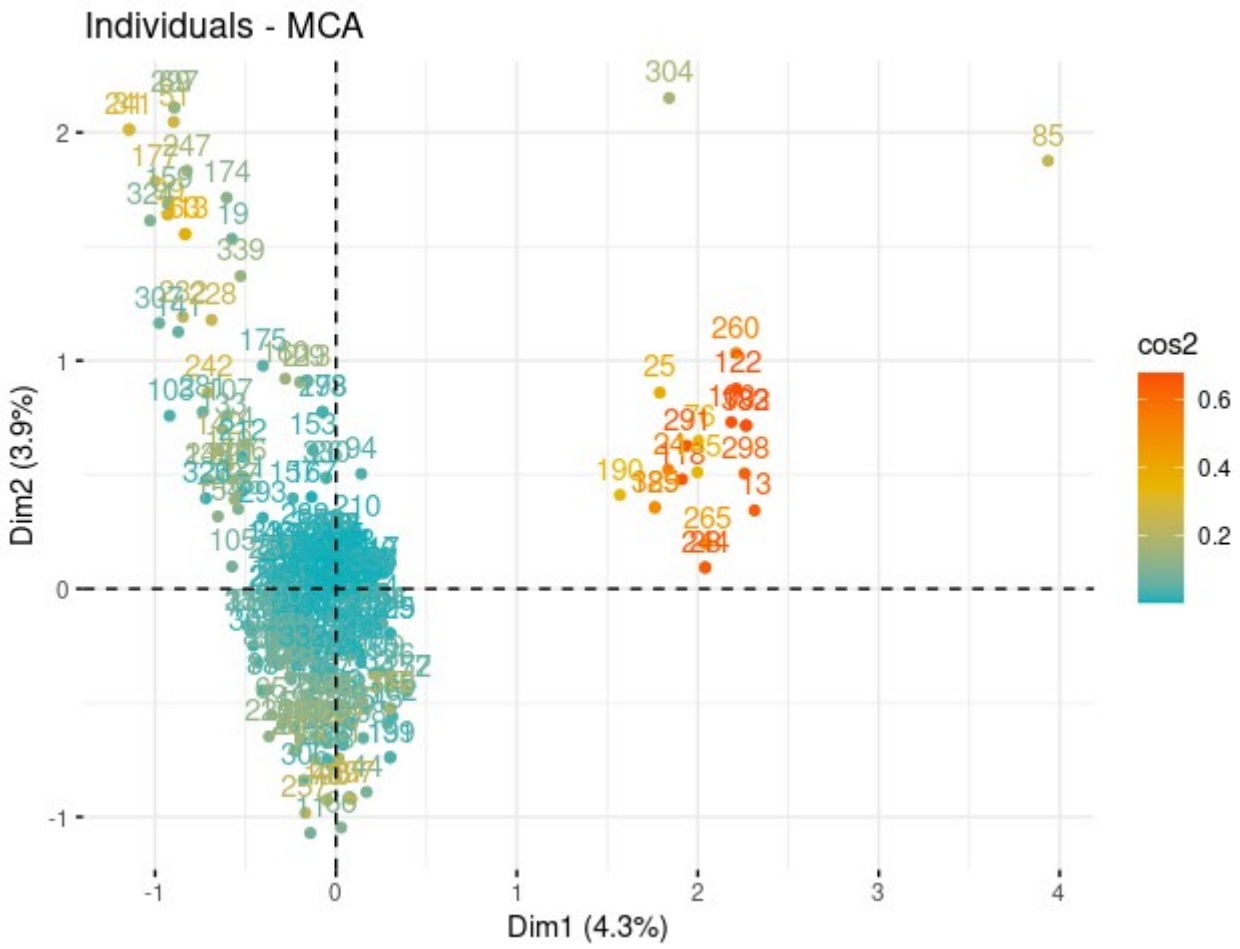
2. Partie ACM

On cherche à cette étape quelle information l'analyse des correspondances multiples peut apporter sur les données sur la base des seules variables qualitatives.
Les axes de projections ne capturent que ~8% de la variance des données.
On peut remarquer que la corrélations des variables avec les axes de projection apporte un élément de réponse à notre problématique : Y-a-t-il des liens entre les variables?

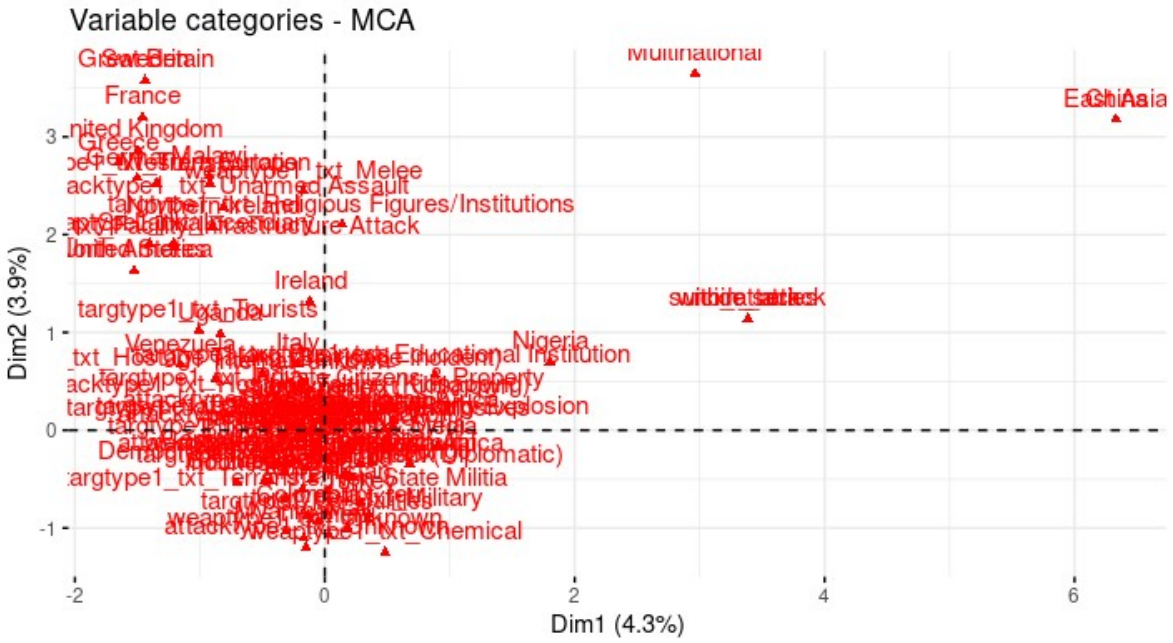


On voit que certaines variables sont proches e**n termes de leurs corrélations avec les axes** sur lesquels elles sont projetées:
en particulier : les variables catégorielles **region, attacktype et weapontype** sont proches entre elles?
Cela semble confirmer notre hypothèse de départ que la région géographique est une variable liée aux variables décrivant les attaques.
Peut-on dire que selon la région dans laquelle se produit l'attaque certains types d'attaques sont plus fréquents?
On se propose de répondre dans la partie AFDM

On se demande en outre si des groupes se dessinent parmi les attentats?
On projette les individus sur le premier plan factoriel



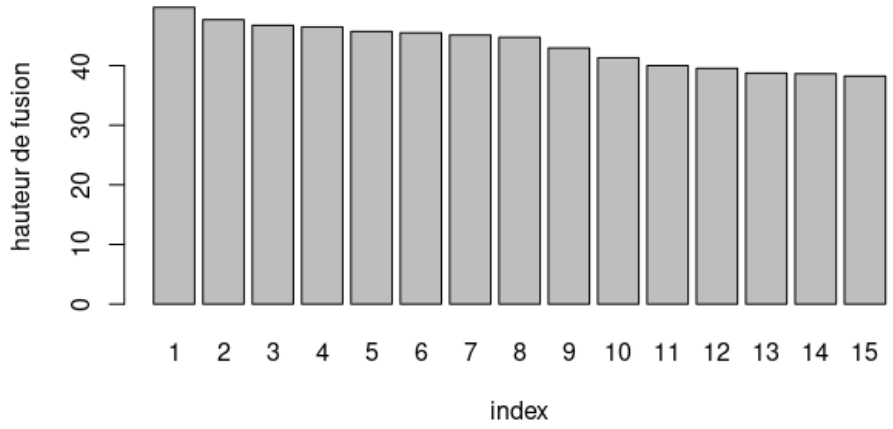
Les projections des variables sont illisibles car certaines ont trop de modalités. Mais on note la présence de modalités extèmes dans la mesure qu'elle sont projettées loin du reste du nuage de points.



3. Partie AFDM

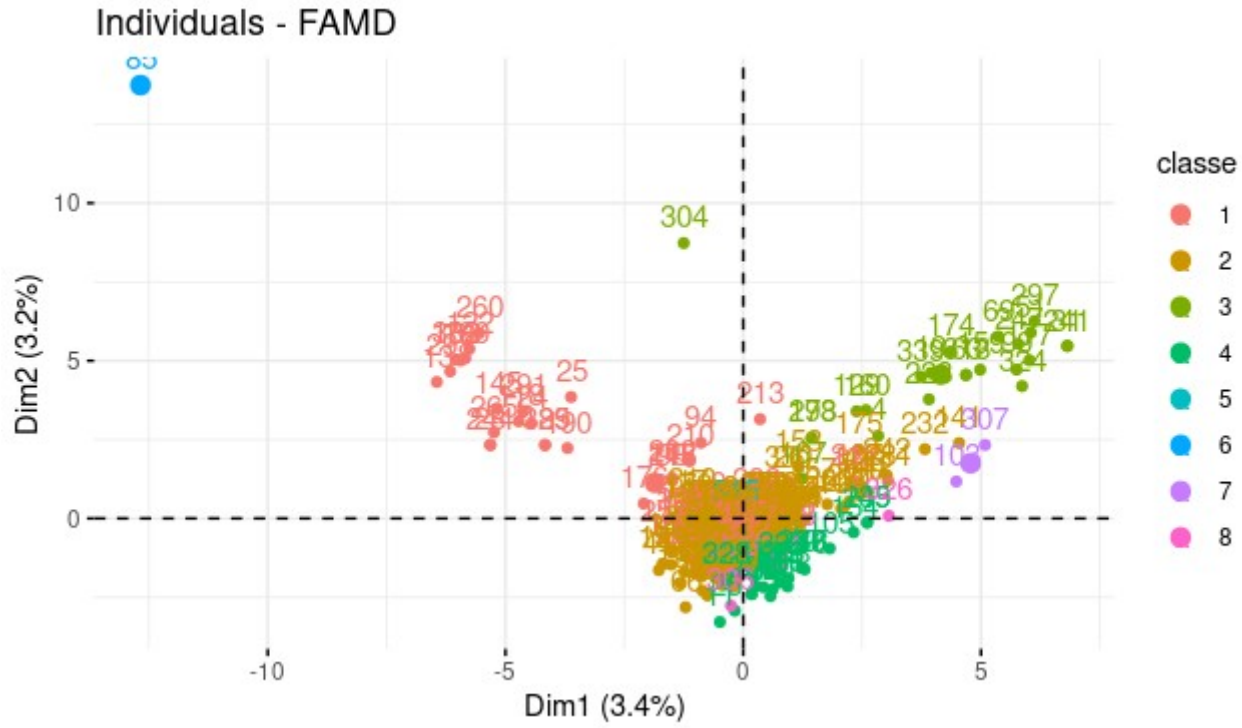
Les calculs de la CAH montrent que le premier plan factoriel capture ~ 3.6 % de la variance des données.
Les calculs de l'AFDM montrent que 55 premieres composantes accumulent ~90% de l'inertie des données donc on analyse uniquement ces composantes et exclue les autres les considérant comme du bruit.
Pour détecter un nombre de groupes parmi les individus
On utilise une mesure de la variance intergroupes grâce à la méthode de la CAH avec un critère d'aggégation dit de "Ward"

On observe qu'il y a ~ 8 groupes parmi nos individus.

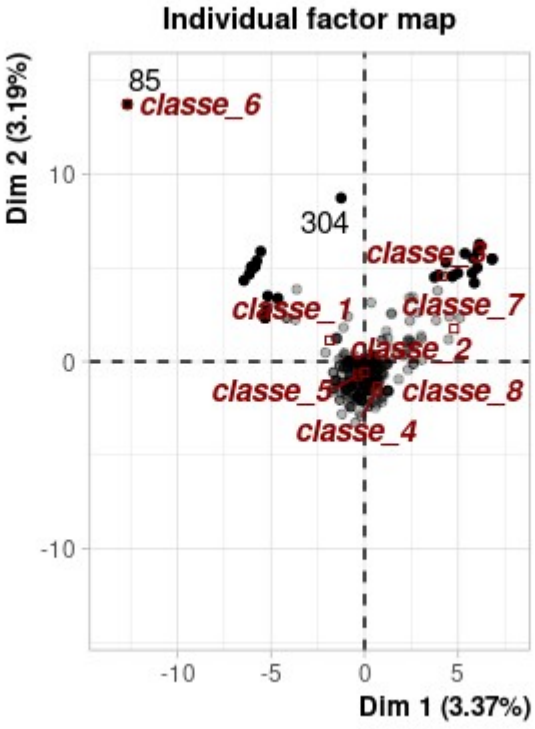


On réalise une CAH sur les 55 premières composantes mais Le clustering de la CAH etant trop lourd, on consolide par Kmeans une fois que la CAH a dégagé 8 groupes.

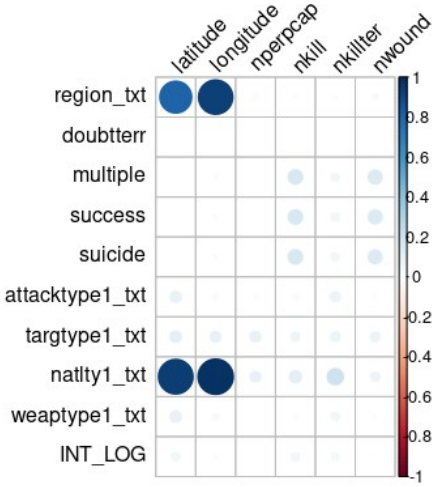
```
part.finale
  1   2   3   4   5   6   7   8
 62 204  23  40   3   1   2   5
```



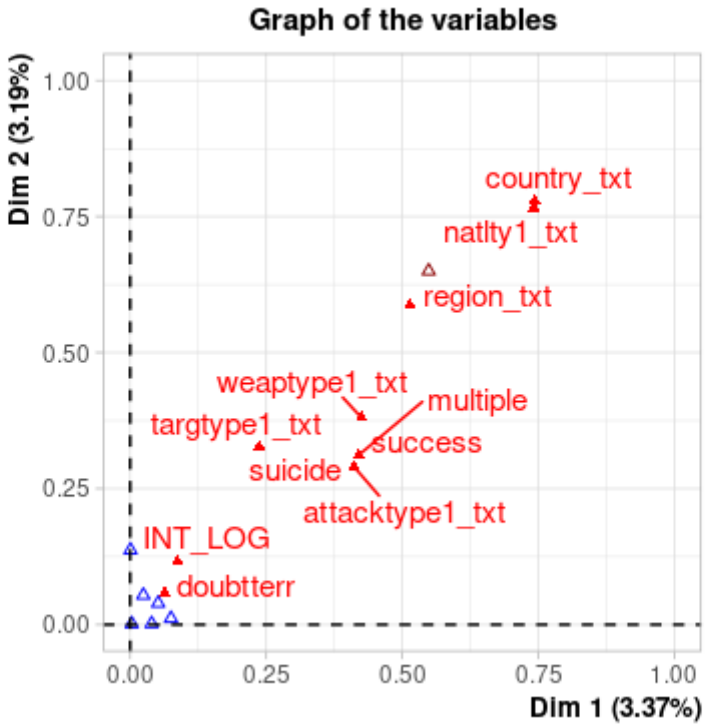
Il apparaît qu'il aurait peut être suffit de faire 6 groupes
Si on calcule les coordonnées des centroïdes, on peut tester la pertinence des groupes



Peut-on désormais conclure quant au lien entre modalités ou variables?



Cela n'est pas aisé mais en comparant la corrélation des différentes variables avec les 2 premiers axes factoriels :On observe que, comparés aux resultats de l'ACM, les variables "region_txt" et weaptype1, attacktype1 se sont éloignées ce qui porte un coup à notre hypothèse initiale. On ne peut conclure



Codes disponibles ici
https://github.com/ZukoTom/terrorism2017_analysis