09/12/2017

Elaboré par :

* **Fatma NAJJAR**
* **Marwen HAMROUNI**
* **Rym BEN KHAMSA**
* **Souha AMRI**

Rapport du projet d’analyse de données

Analyse d’un jeu de données portant sur le terrorisme international

**1. Introduction :**

Depuis les attentats du 11 septembre 2001 à New York et Washington qui ont fait 3052 morts, la question du terrorisme occupe une place essentielle dans la vie internationale. Les actions que l'on attribue à M. Oussama Ben Laden et à son groupe Al-Qaida depuis le 11 septembre, révèlent la naissance d'un terrorisme transnational qui ne soit plus directement lié à un conflit géographique précis.

Il suffit d'examiner la liste de principaux attentats survenus depuis cette date pour s'en rendre compte : les attentats de Bali en Indonésie en octobre 2002, de Casablanca au Maroc en mai 2003, d'Istanbul en Turquie en novembre 2003, de Madrid en Espagne en mars 2004, d'Arabie Saoudite en mai 2004, et tout récemment les attentats perpétrés en Europe ces deux dernières années avec l’apparition de Daech.

En effet, Le terrorisme implique le recours à la violence pour terroriser une population ou un gouvernement à des fins politiques, religieuses ou idéologiques.

La menace terroriste est devenue une préoccupation mondiale avec une escalade importante ces dernières années dans attaques terroristes fréquentes dans plusieurs parties du monde.

S'inquiétant peu des vies humaines, les terroristes continuent de frapper en toute impunité, laissant une traînée de mort et de destruction partout où ils choisissent d'infliger leurs coups.

Et donc cette monté grandissante de l'insécurité due au terrorisme et à ses actions nécessite de pouvoir analyser les faits passés dans l'optique de pouvoir contrer le terrorisme et prévoir son évolution future.

**2. Le jeu de données :**

**1. La source des données :**

Le jeu de données qu’on a choisi pour ce projet est extrait d’une base de données qui s’appelle “Global Terrorism Database” (GTD) qui est une base open source contenant des informations sur les attentats terroristes dans le monde de 1970 à 2016 (avec des mises à jour annuelles prévues pour l'avenir).

La GTD comprend des données systématiques sur les incidents terroristes nationaux et internationaux qui ont eu lieu au cours de cette période et comprend maintenant plus de 170 000 cas.

La base de données est maintenue par des chercheurs du Consortium national pour l'étude du terrorisme et des réponses au terrorisme, dont le siège est à l'Université du Maryland.

Nous avons choisi de focaliser notre analyse sur les 16 dernières années puisqu’elles ont connu une montée alarmante de ce fléau mondial.

**2. Les éléments étudiés :**

Pour analyser ce jeu de données, nous avons choisi de centrer notre analyse sur 3 parties :

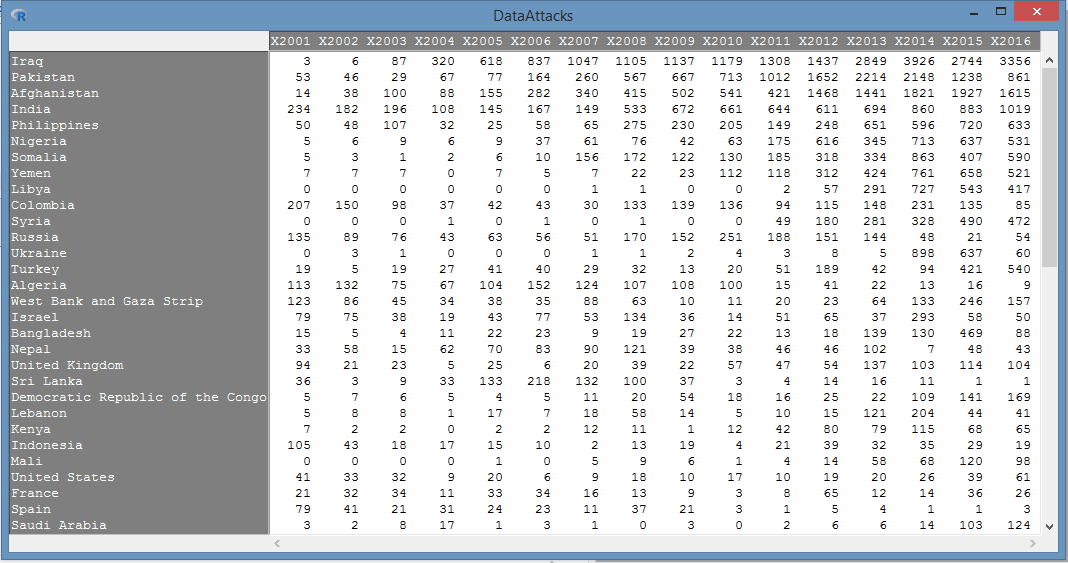
* Analyse de nombre des attaques dans chaque pays depuis 2001
* Analyse de nombre des victimes des attentats terroristes dans chaque pays depuis 2001
* Analyse des types d’attaques choisis par les groupes terroristes durant les 16 dernières années

Afin de bien visualiser et comprendre cette analyse nous avons réduit le nombre de pays vers 69 états des différentes régions dans le monde.

**3. Analyse de nombre des attaques :**

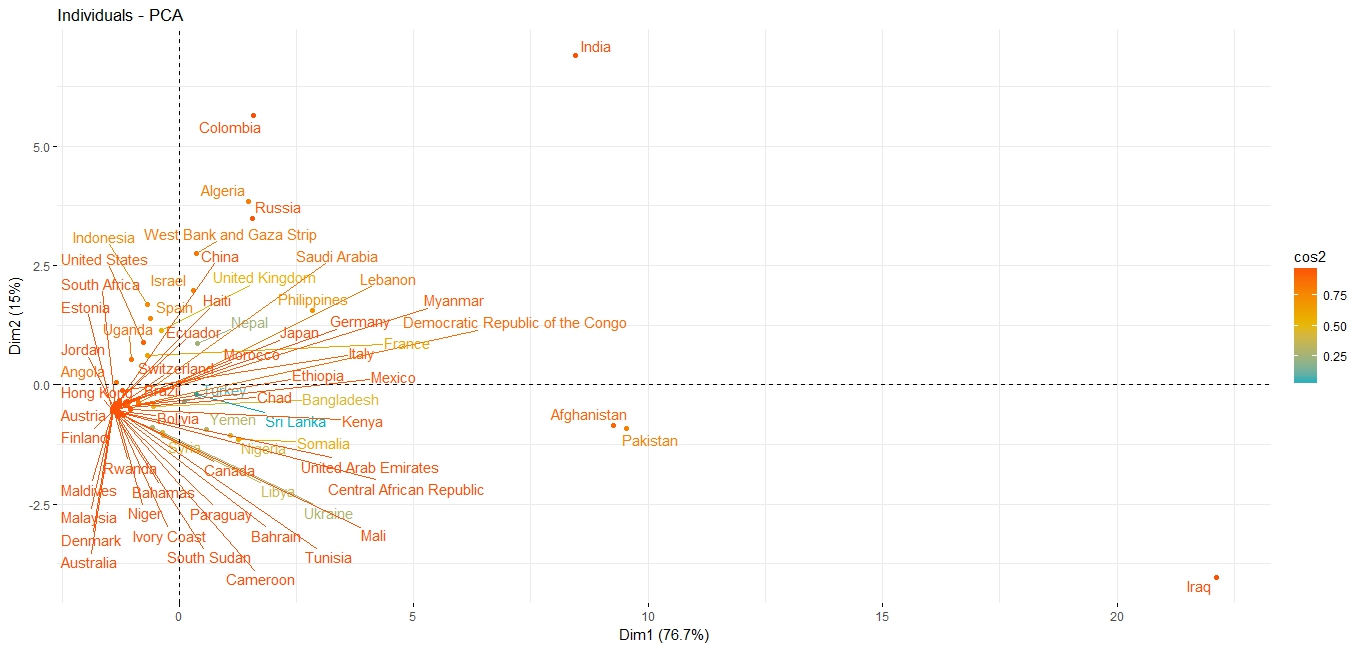
Pour mieux étudier les données en relation avec cette analyse, nous avons opté pour une analyse de composantes principales ACP qui saura nous fournir une image simplifiée du nuage de points résumant au mieux les données.

Le tableau de données qu’on va utiliser est visualisé comme suit :



Pour mieux visualiser les graphes demandés pour l’ACP, nous avons eu recours aux méthodes de la bibliothèque **factoextra** qui nous offre des méthodes de visualisation accompagné d’indications sur les individus telle que la qualité de leurs projections sur le graphe.

Le graphe des individus suivant est obtenu grâce à la commande suivante : fviz\_pca\_ind(PCA(DataAttaques), col.ind = "cos2",gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE)



Le graphe des individus montre une forte concentration d’un grand nombre de pays des différentes régions du monde sur une partie du graphe et ceci explique la fréquence modeste des attentats dans ces états entre 2001 et 2016.

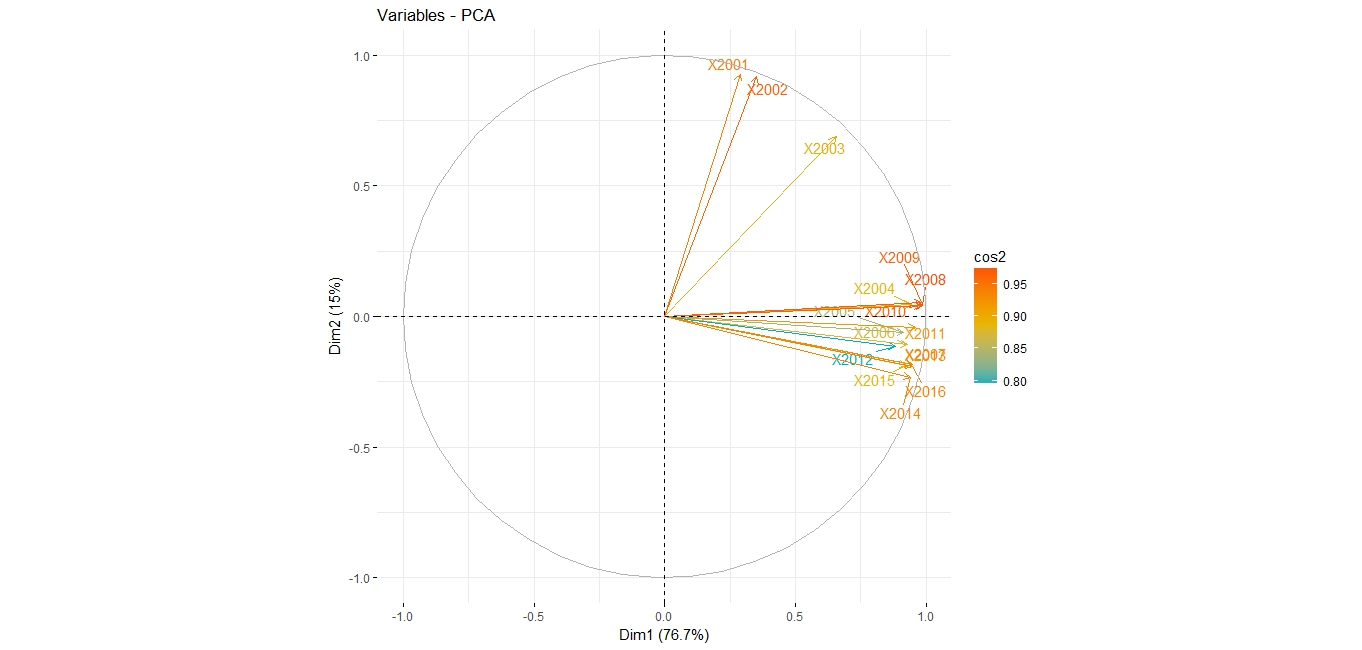
Particulièrement, l’Irak se trouve à l’extrémité de l’axe 1 vu le grand nombre des attaques qui ont touché le pays après le déclenchement de la guerre menée par les Etats-Unis en 2003.

On remarque aussi la présence de l’Afghanistan et le Pakistan sur une même zone dans le graphe des individus expliquée par l’escalade des attaques terroristes sur le territoire Afghan qui impacte directement son voisin Pakistan qui connait une montée sans précédent des attentats terroristes.

Quant à l’axe 2, on remarque la présence de l’Inde et la Colombie dans l’extrémité de ce dernier.

Le cercle de corrélations est obtenu grâce à la commande suivante : fviz\_pca\_var(PCA(DataAttaques), col.var = "cos2",gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)

Le graphe obtenu est le suivant :

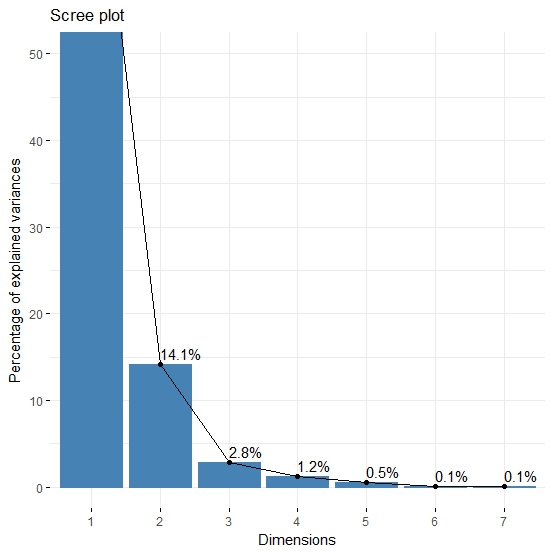


On observe une forte corrélation entre les années 2001 et 2002 et ceci est expliqué par le début de la propagation du terrorisme durant ces deux années qui ont suivi les attentats de 11 Septembre.

Les variables les moins présentées sur le cercle sont 2012, 2005 et 2003.

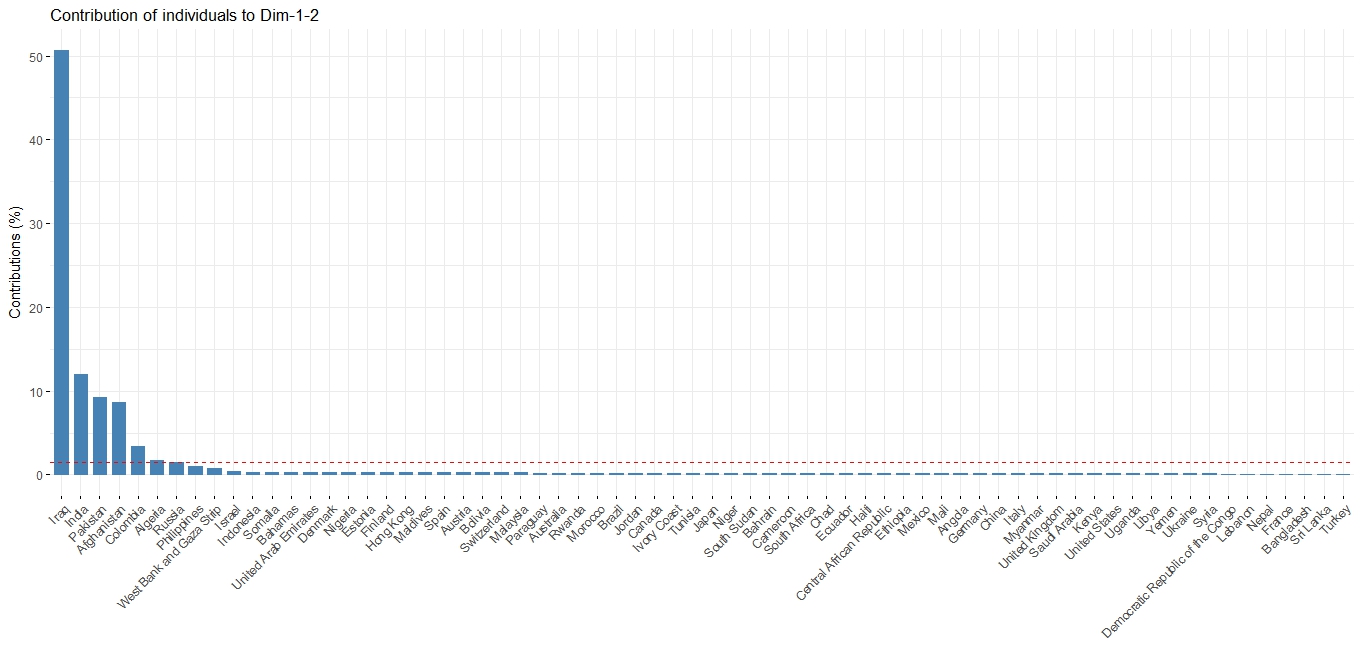
Les autres variables sont bien présentées sur le cercle avec une forte corrélation des années 2015 et 2016 qui montrent la forte occurrence des attentats terroristes durant ces deux dernières années.

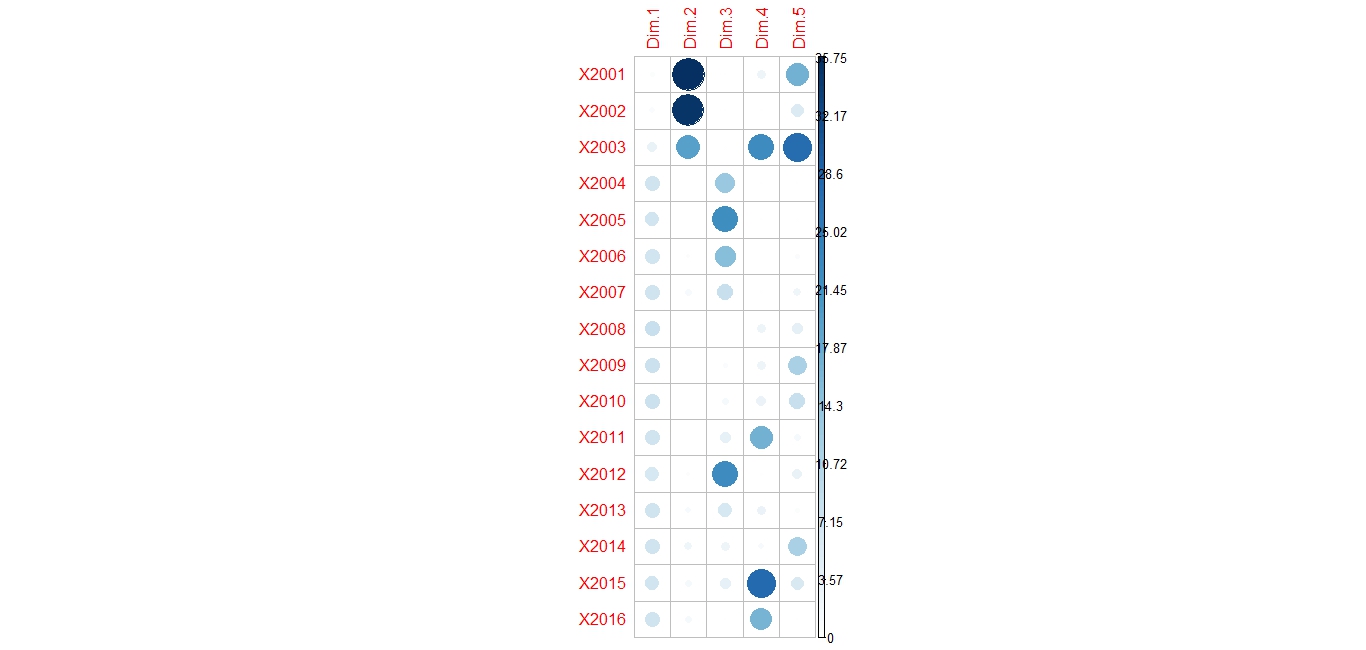
La figure suivante met en exergue Le pourcentage de variance expliquée par chaque composante principale:



On voit que les deux premiers axes principaux renferment le plus d’informations sur les données en comparaison avec le reste des axes et ce résultat est prévisible puisqu’ils sont associés aux plus grandes valeurs propres.

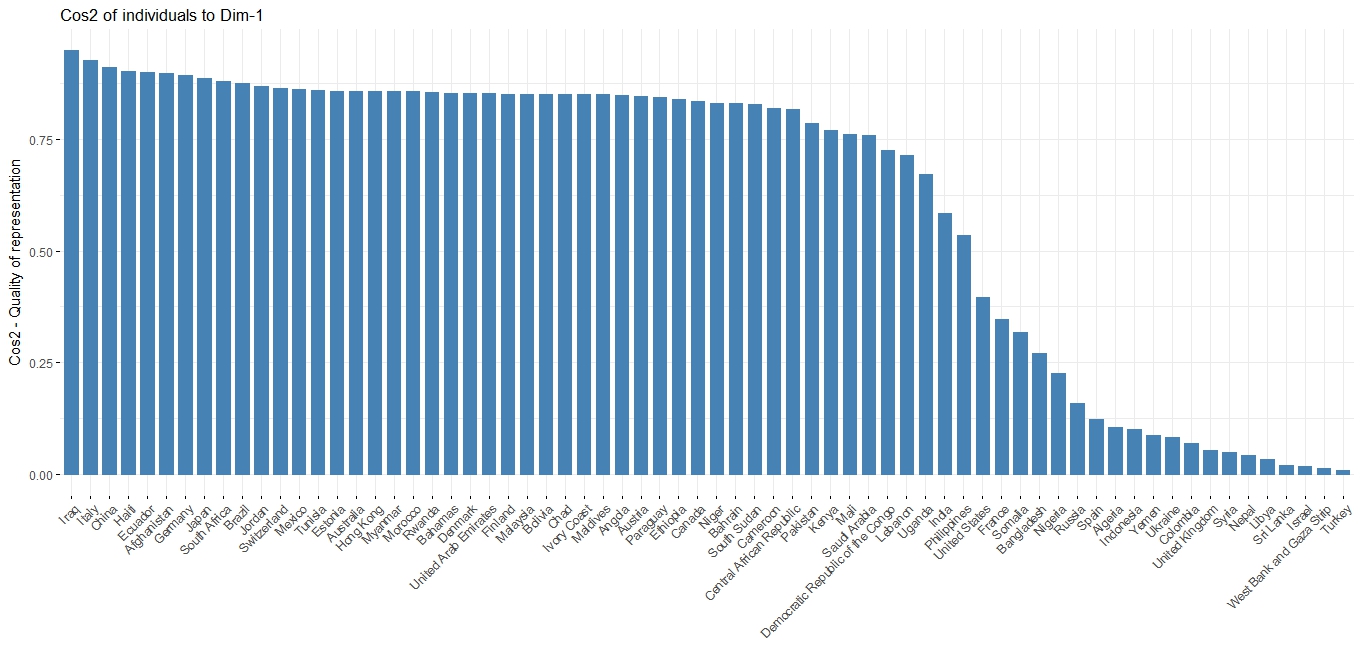
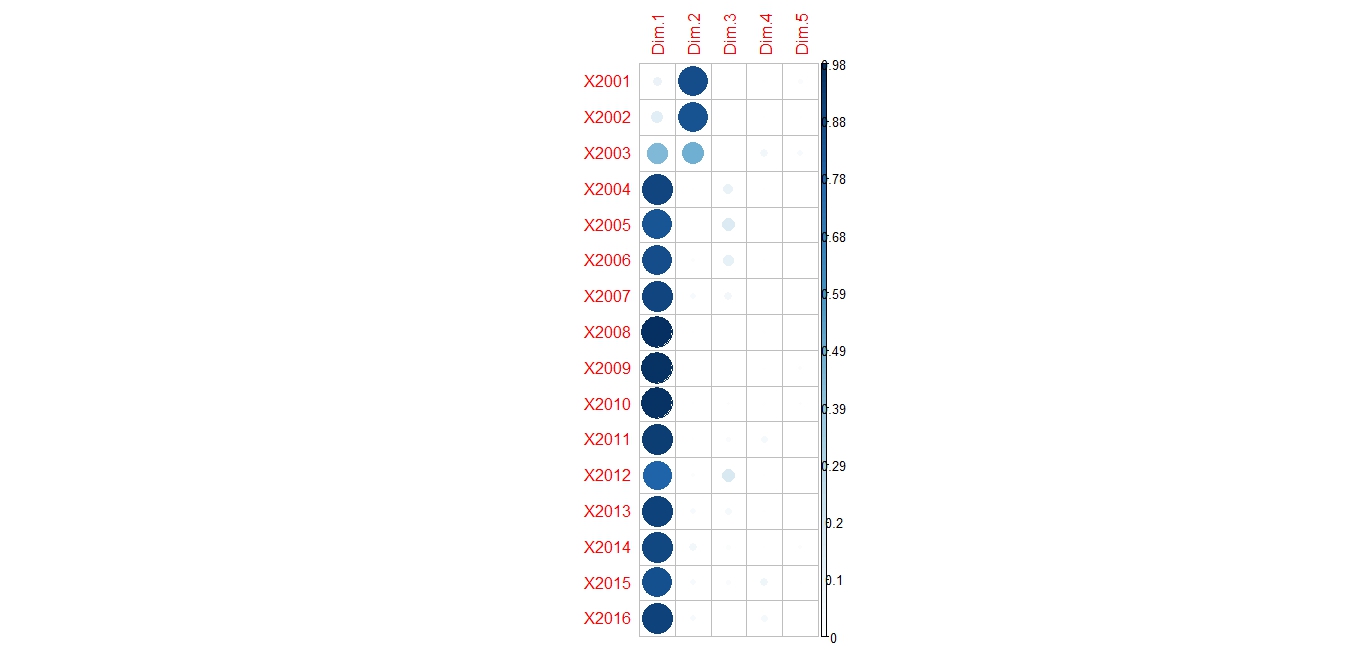
Quant à la contribution des individus et les variables dans la construction des deux axes, les deux graphes suivant illustrent cette fonctionnalité.





On remarque une contribution extrême de l’Irak dans la construction des axes ainsi qu’une contribution importante des années 2001 et 2002 à cette construction.

Passons maintenant à la qualité de projection des individus et des variables de notre étude :



On voit bien qu’un grand nombre de pays est bien présenté sur le graphe.

Quant aux variables, on trouve que les années 2008 et 2009 sont très bien présentés sur le graphe.

* **Interprétation des résultats :**

On remarque qu’en parcourant l’axe 1, on trouve que l’Afghanistan, le Pakistan et l’Irak sont les pays les plus touchés par les attaques terroristes plus particulièrement le terrorisme ‘djihadiste’ depuis 2001 avec une valuation importante pour l’Irak.

Quant à l’axe 2, l’Inde et la Colombie constituent les individus qui ont contribué le plus dans la construction de cet axe et ceci est expliqué par la différence de la fréquence des attaques dans ces deux pays comparée aux autres pays.

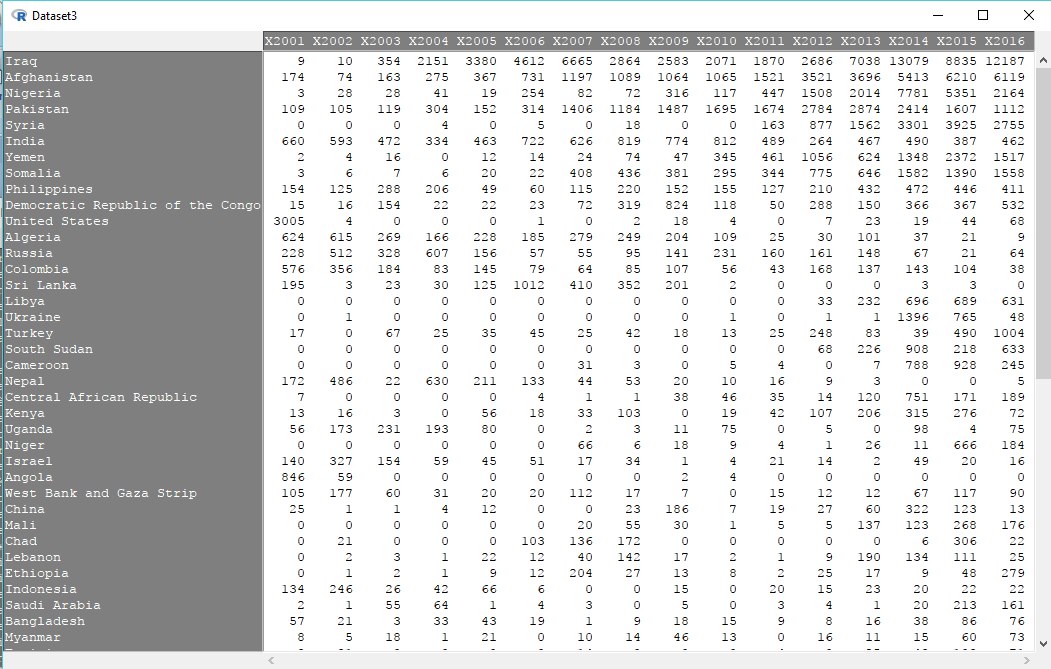
Les pays corrélés négativement par rapport aux deux axes sont des pays touchés par les attentats moins fréquemment que le reste des pays.

Quant aux variables, on constate une forte corrélation positive entre les années 2001 et 2002 qui ont anticipé le début de la montée du terrorisme.

Mais à partir de l’année 2013, on trouve une forte corrélation entre les années qui ont suivi cette dernière, celle de l’apparition du groupe idéologique Daesh, et qui ont connu la propagation des attaques dans les différentes zones du monde

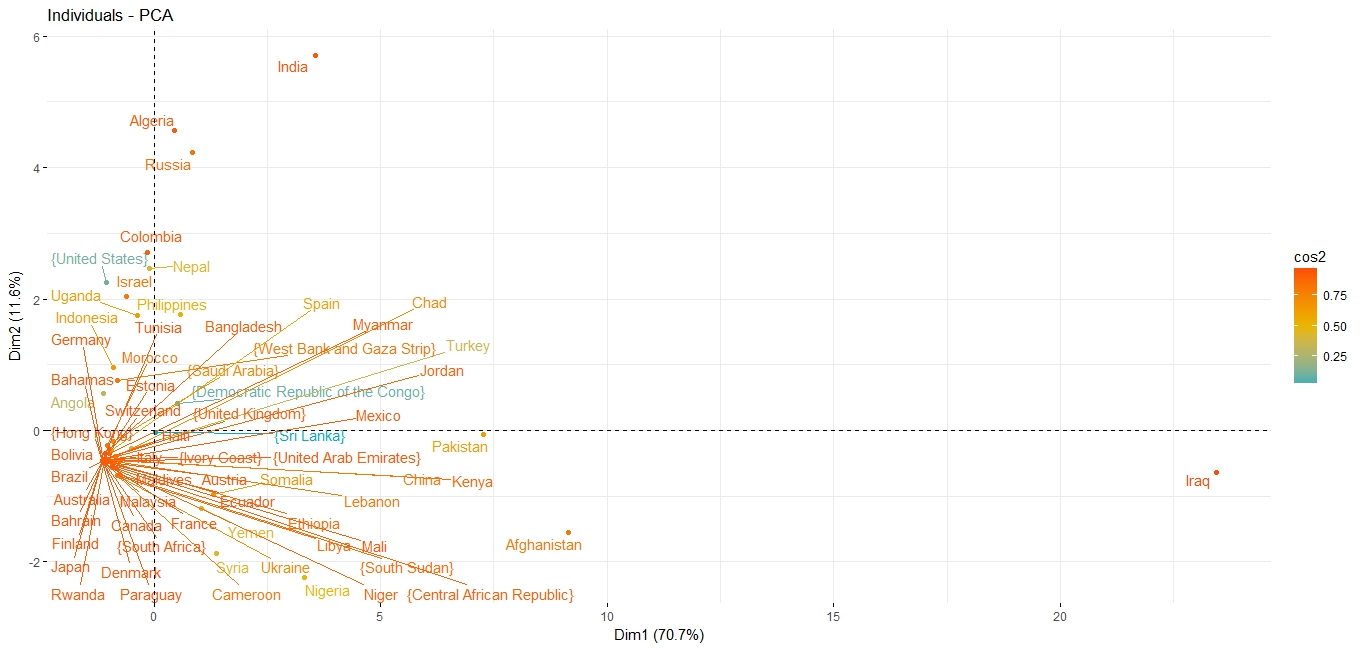
**4. Analyse de nombre des victimes des attaques terroristes :**

Pour mieux étudier les données, nous avons opté pour une analyse ACP de nombre des victimes des attaques terroristes, les données sont comme suit :

****

La fonction fviz\_pca\_ind() est utilisée pour produire le graphique des individus. Comme les variables, il est également possible de colorer les individus en fonction de leurs valeurs de cos2, on a utilisé la commande suivante

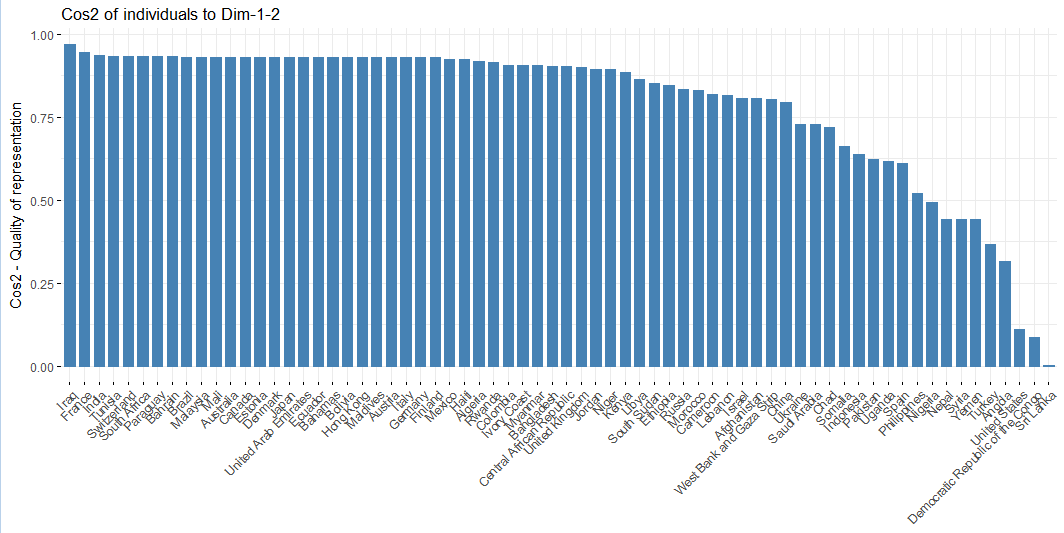
fviz\_pca\_ind (res.pca, col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE )



Puisque le nombre des victimes est en relation avec le nombre des attaques terroristes, on remarque un graphe des individus très similaire au graphe de nombre d’attaques terroristes ; une forte concentration d’un grand nombre de pays des différentes régions du monde sur une partie du graphe, l’Irak se trouve à l’extrémité de l’axe 1, aussi la présence de l’Afghanistan et le Pakistan sur une même zone dans le graphe des individus et on a la présence de l’Inde et la Colombie dans l’extrémité de l’axe 2, tous les positions des pays dans ce graphe sont très similaire au graphe de nombre des attaques.

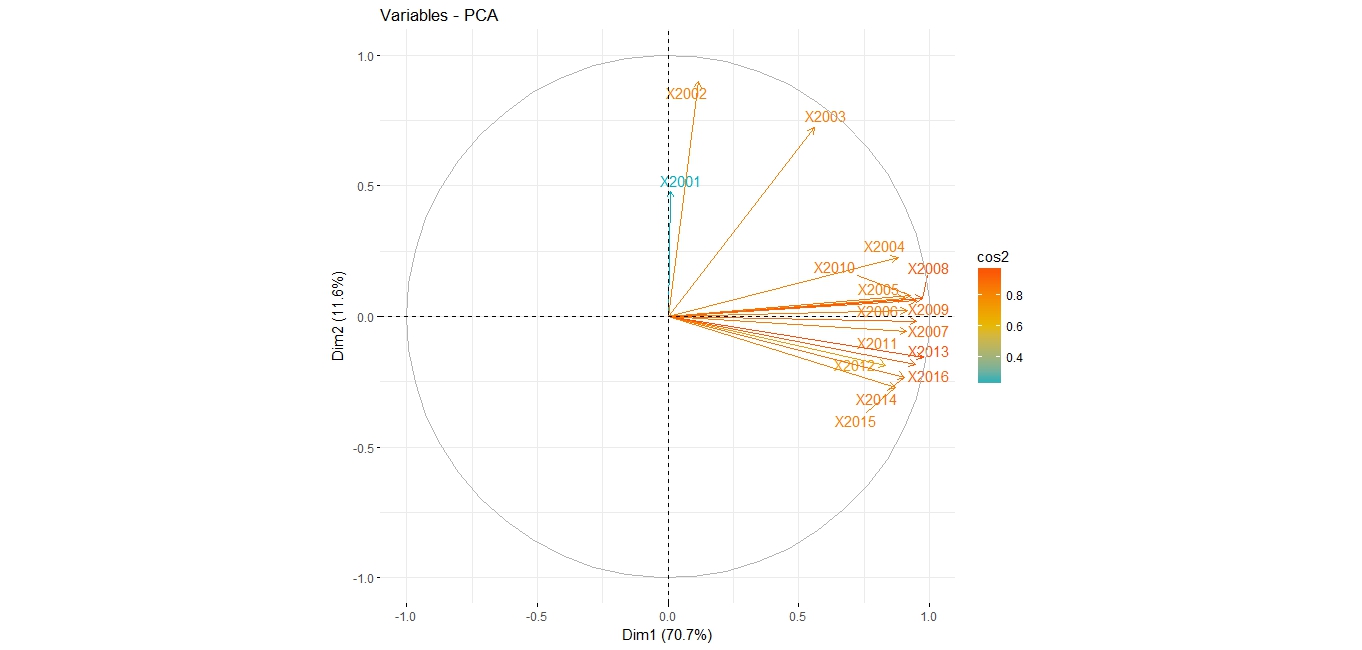
La corrélation entre une variable et une composante principale (PC) est utilisée comme coordonnées de la variable sur la composante principale. La représentation des variables diffère de celle des observations: les observations sont représentées par leurs projections, mais les variables sont représentées par leurs corrélations

On remarque que tous les individus sont bien présentés et le graphe suivant montre leurs cos2 :



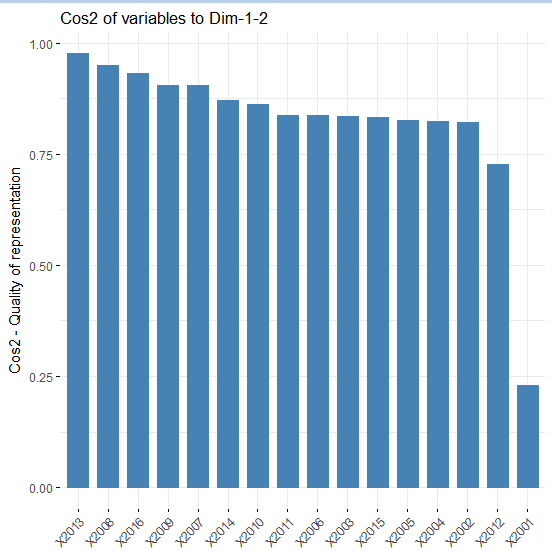
On a visualisé le graphe de Nuage des variables en utilisant la commande suivante :

fviz\_pca\_var(res.pca, col.var = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)

****

Les valeurs de cos2 sont utilisées pour estimer la qualité de la représentation donc plus une variable est proche du cercle de corrélation, meilleure est sa représentation sur la carte de l’ACP

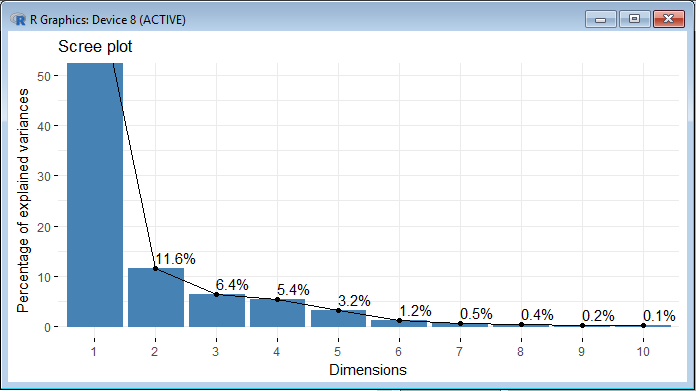
D’après le nuage des variables tous les variables sont bien présentés sauf pour 2001. Le graphe suivant représente la qualité de représentation des variables



On remarque que le nuage des variables ressemble à celui de nombre d’attaques terroristes car le nombre des victimes est lié au nombre des attaques.

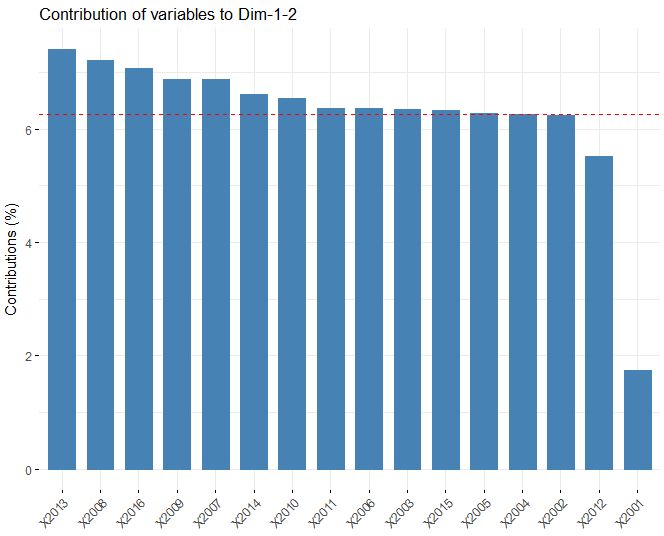
Pour déterminer le nombre de composantes principales est de regarder le graphique des valeurs propres. Le nombre d’axes est déterminé par le point, au-delà duquel les valeurs propres restantes sont toutes relativement petites et de tailles comparables

Le graphique des valeurs propres peut être généré à l’aide de la fonction *fviz\_eig* () ou *fviz\_screeplot* ()

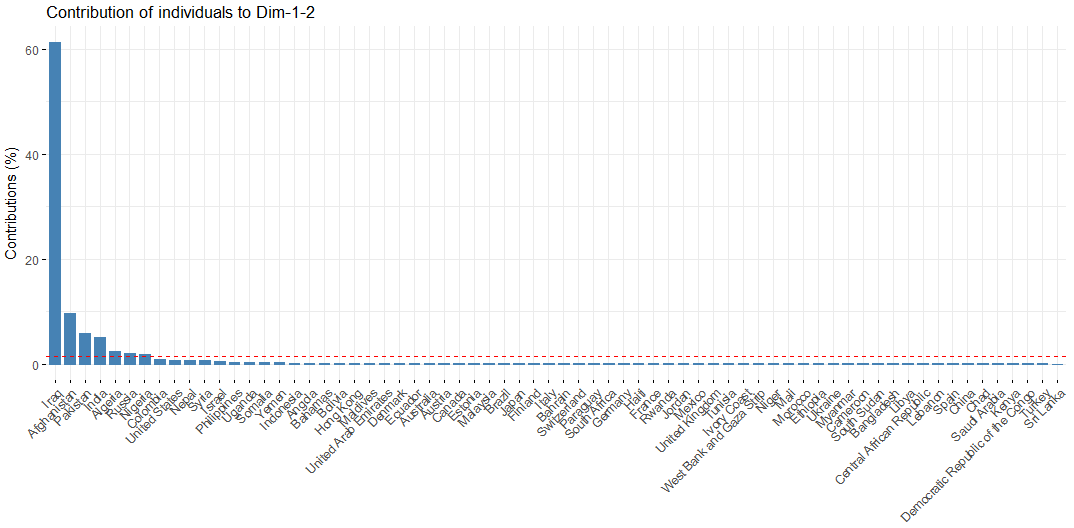


Dans notre analyse, les deux premières composantes principales expliquent 72,3% de la variation. C’est un pourcentage acceptable.

Quant à la contribution des individus et les variables dans la construction des deux axes, les deux graphes suivant illustrent cette fonctionnalité.



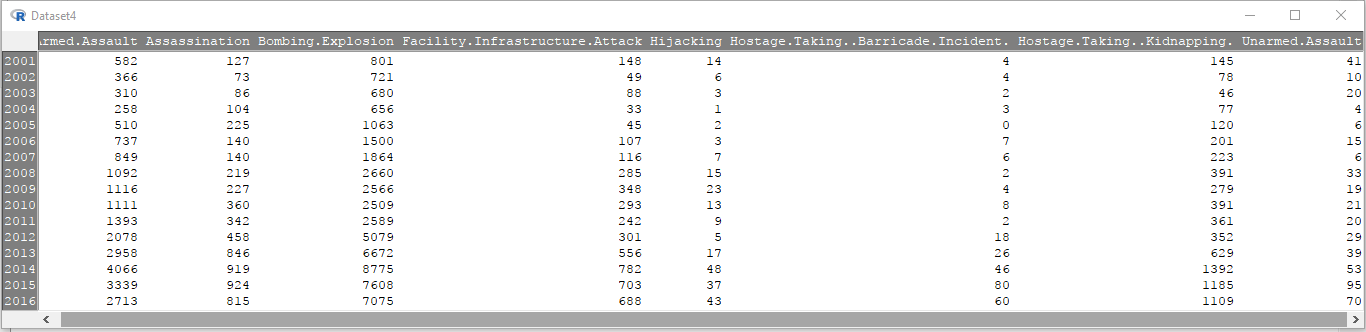
On remarque que 2002 2003 et 2001 contribues da

****

**5. Analyse des types des attaques terroristes :**

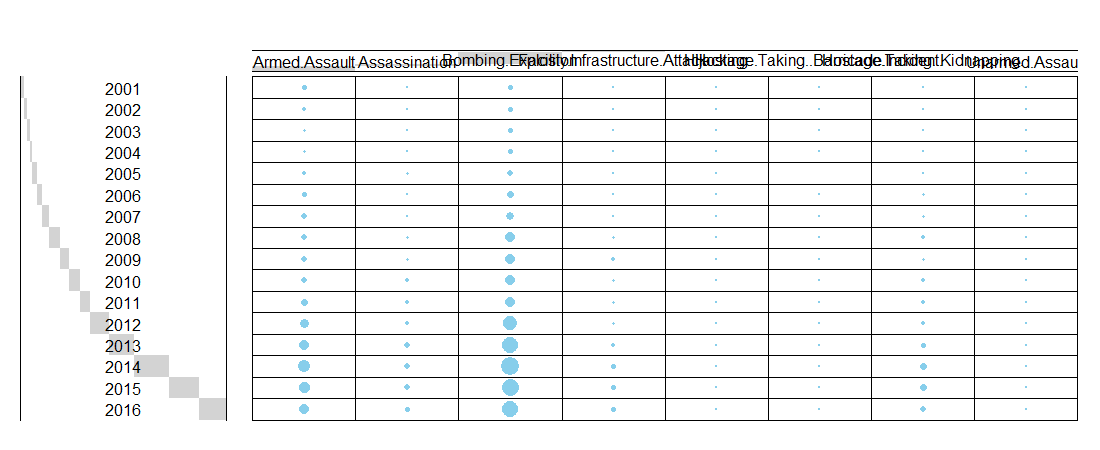
L’objectif d’une AFC est de déterminer les correspondances possibles entre les deux variables qualitatives du tableau de données, à savoir les possibles correspondances entre les années et les types des attaques terroristes

Il s’agit d’un tableau de contingence X\*Y, où la variable X est définie par des années (modalités=années) et la variable Y représente les types des attaques. Ces variables sont qualitatives.

****

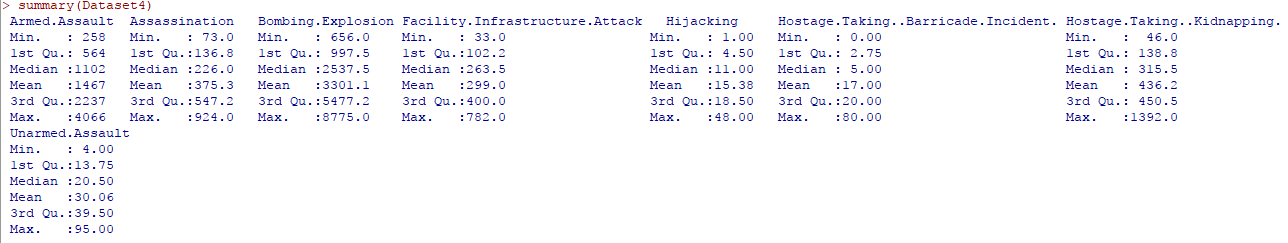
Le tableau de contingence ci-dessus n’est pas très gros. Par conséquent, il est facile d’inspecter et d’interpréter visuellement les profils des lignes et des colonnes:

Le tableau de contingence peut être visualisé en utilisant les fonctions balloonplot()

****

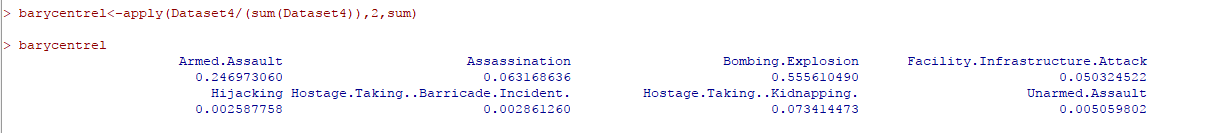
On commence en général notre étude par illustrer les résultats de statistiques descriptives :

Summary(Dataset) donne

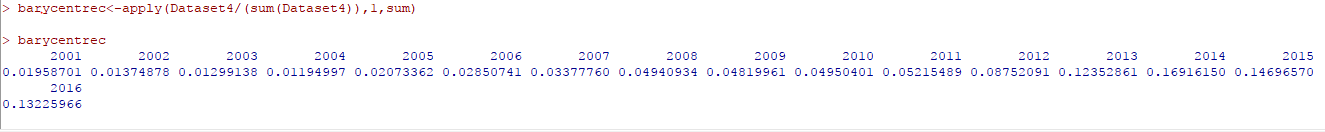
****

On a l’effective totale de C:\Users\Dell\Desktop\projetanal\AFC\sumAFC.PNG

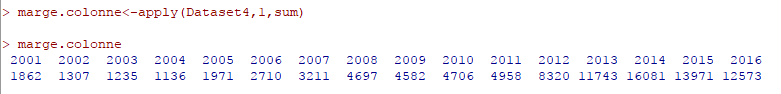
Le barycentre est calculé comme suit : pour les types des attaques

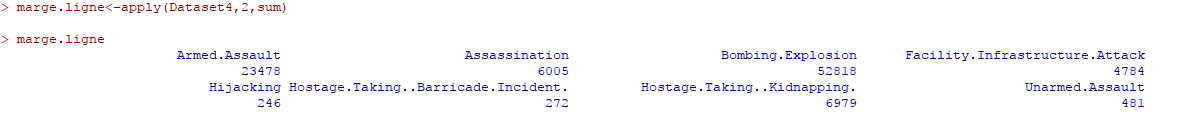
****

Pour les années

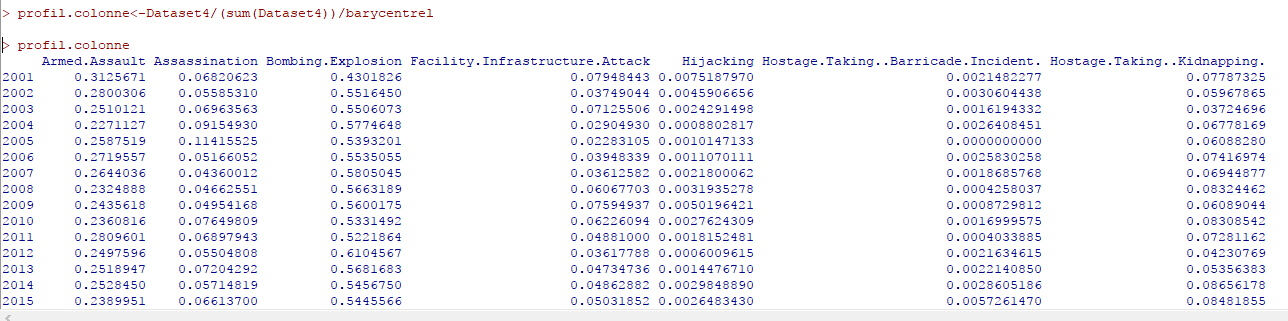
****

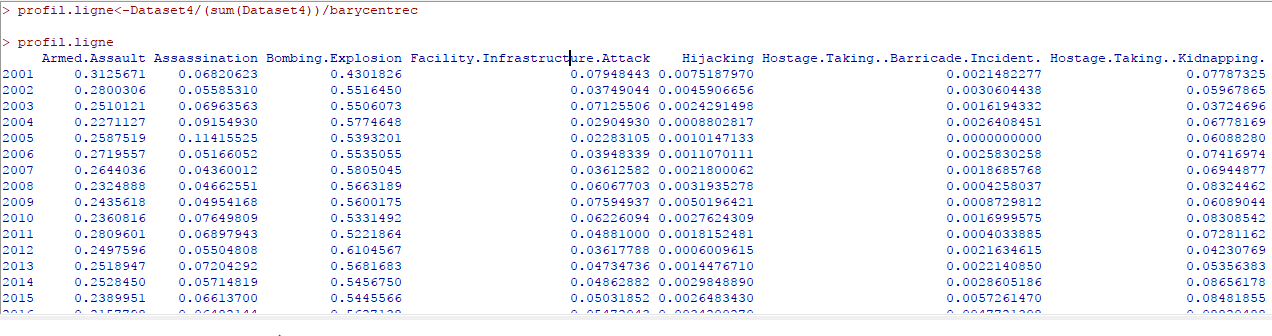
Les marges colonnes et les marges lignes sont les suivants

****

****

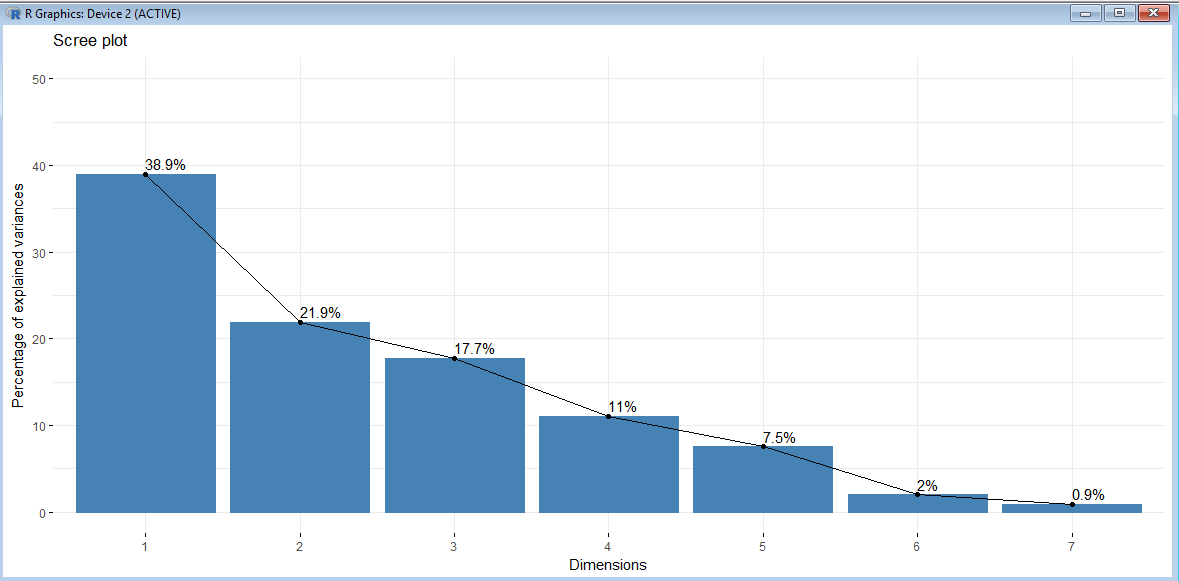
Les profiles colonnes et lignes sont :

****

****

A partir de ces constructions, on peut envisager de n’est pas éliminer car on a commencé par lancer une première analyse sur l’ensemble de toutes les données afin d’être sûr de ne pas laisser passer quelques petites informations qui pourraient être pertinentes dans la synthèse des travaux.

Le graphique des valeurs propres généré est :

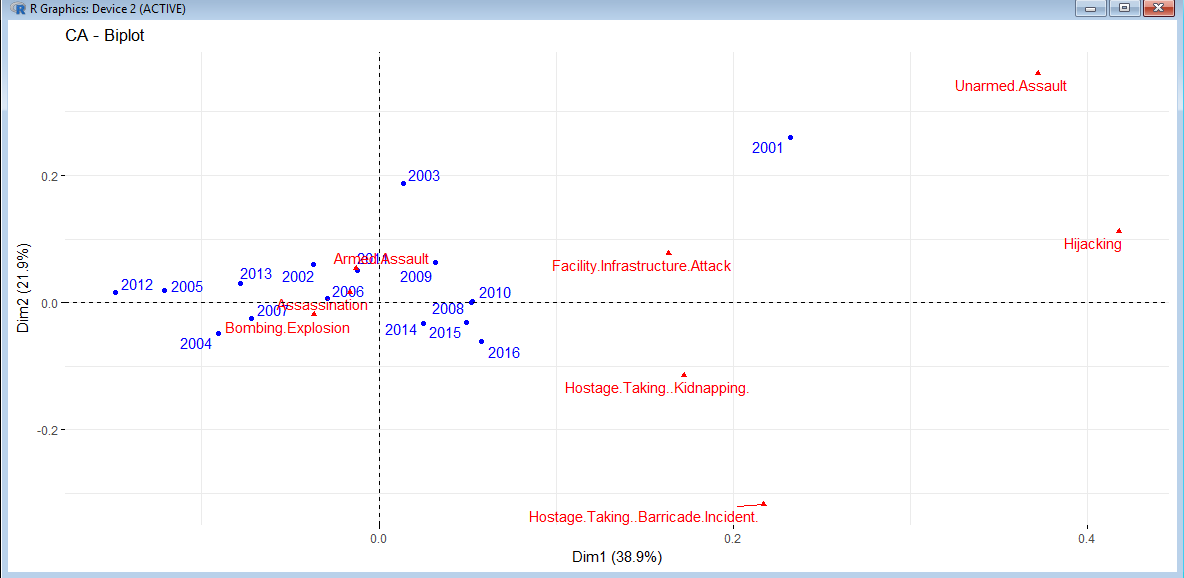
****

Dans notre analyse, les deux premières composantes principales expliquent 60.8% de la variation. C’est un pourcentage acceptable.

On fait appel à la fonction res.ca <-CA(Dataset, graph=FALSE). On obtient la statistique Chi2, la p valeur :

Chi2=1370.168 (p-value = 3.154003e-219 )

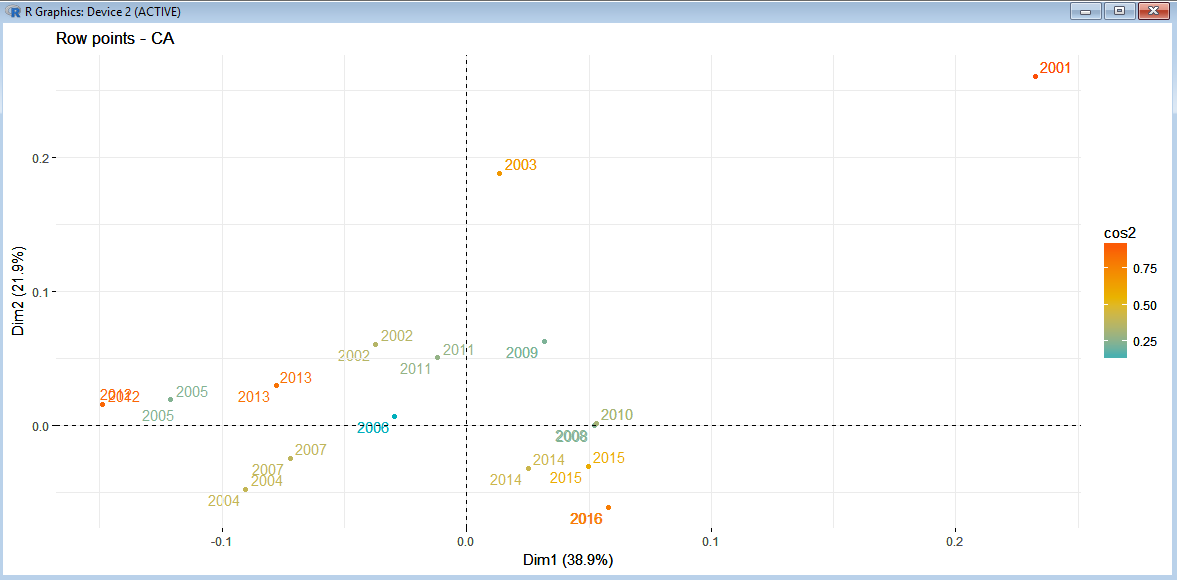
On a Chi2 grande et p value petite alors on a dépendances entre les variables

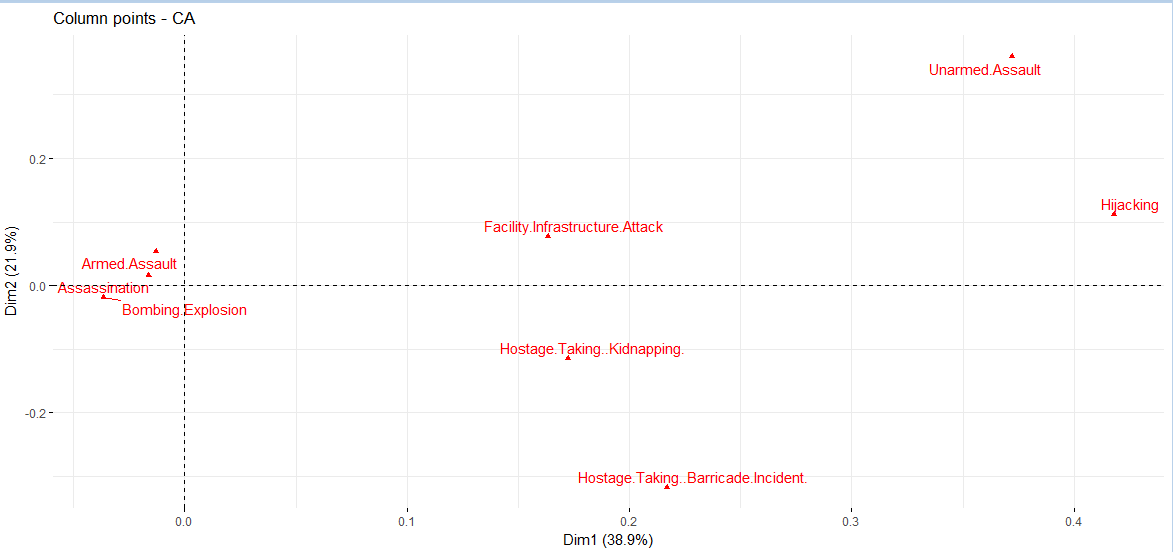


Le groupement entre lignes qui se caractérise par des points très proches sur le graphique s’explique par une certaine similarité entre ces profils lignes et vice versa. Le graphique représentant le nuage des lignes permet de déduire un certain groupement entre les années 2001 est clairement à l’écart, il en est de même pour 2003. Par contre, une similarité est nettement visible entre les années 2016,2015 et 2014. En général, des points positionnés dans des quadrants opposés s’interprètent par des corrélations négatives, et donc sont des points qui se comportent de façon opposée. On pourra donc combiner cela à la visualisation déjà tirée précédemment, en considérant ;Les années dernières 2014 2015 et 2016 : plus d’attaques terroristes

-2001 à l’écart car le début des attaques terroristes

La distance entre les points et l’origine mesure la qualité de représentation sur le graphique. Autant les points s’éloignent de l’origine, mieux ils sont représentés. Ici aussi, en s’appuyant sur ce qui a été déjà mentionné, on note la qualité de l’année 2001 et aussi celle de 2003

****



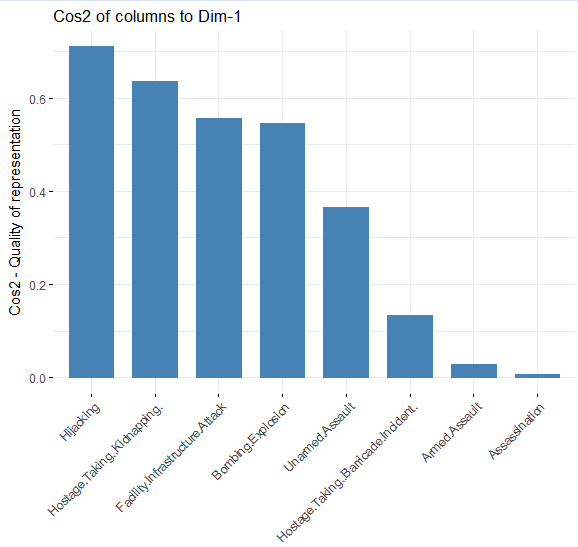
Le graphique représentant le nuage des colonnes permet 3 groupements entre les types d’attaques.

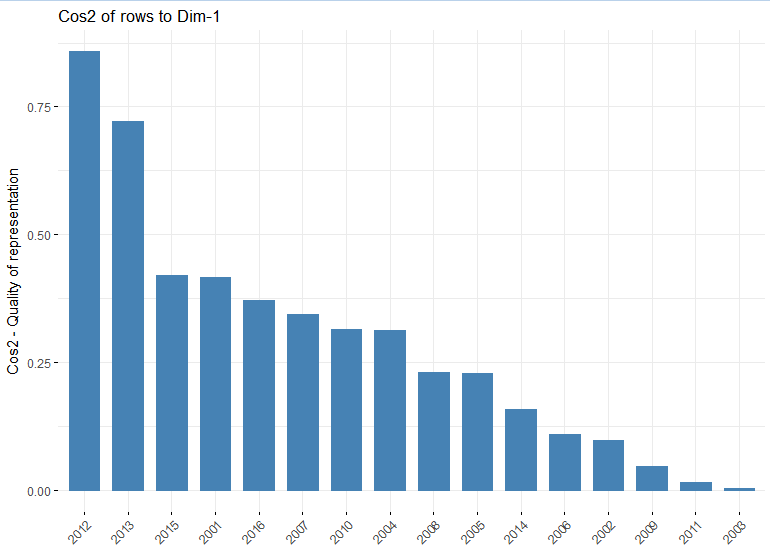
Le rapprochement du premier groupement des points (Armed Assault, Assassination et Bombing) illustre une similarité en augmentant d’une manière significative d’une année à une autre, ce qui peut être traduit par l’augmentation de l’accessibilité aux armes à feu suite aux guerres.

Pour les points « Hostage Taking Kidnapping », « Hostage Taking Barricade Incident » et « Facility infrastructure Attack » situés au milieu, les nombres d’attaques pour ces types ont augmenté graduellement au fil du temps.

Pour le « Hijacking » et « unarmed assault » sont situés à l’écart, ce qui signifie qu’ils ont resté stables et non variés.

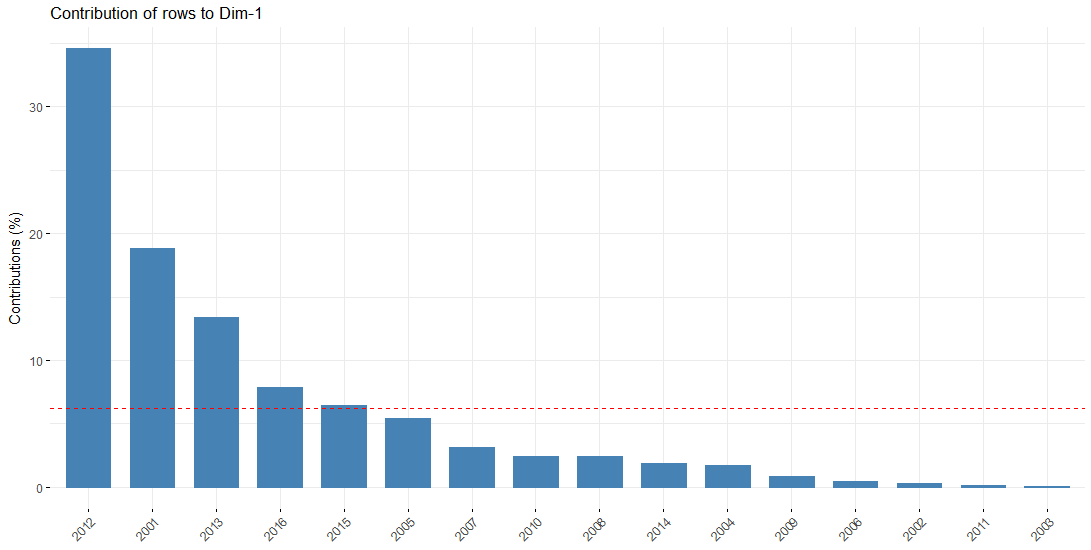
Notons que l’inertie du nuage a été le premier critère sélectionné afin de limiter l’étude de ce tableau de données aux axes de dimension 1 et 2, en déterminant les inerties expliquées par chaque axe. Afin de mieux visualiser la qualité de représentation des points aussi bien ceux du profil lignes que ceux du profil colonnes, nous pouvons visualiser en diagramme à batons une sélection des points les mieux représentés sur les axes principaux. Cette sélection permet de mieux se focaliser sur les points les mieux représentés et donc ceux qui dégageront le plus d’informations et sur lesquels se basera une interprétation synthétique. Pour cela, on exécute par exemple la commande fviz\_cos2(res.ca, choice = "col", axes = 1)

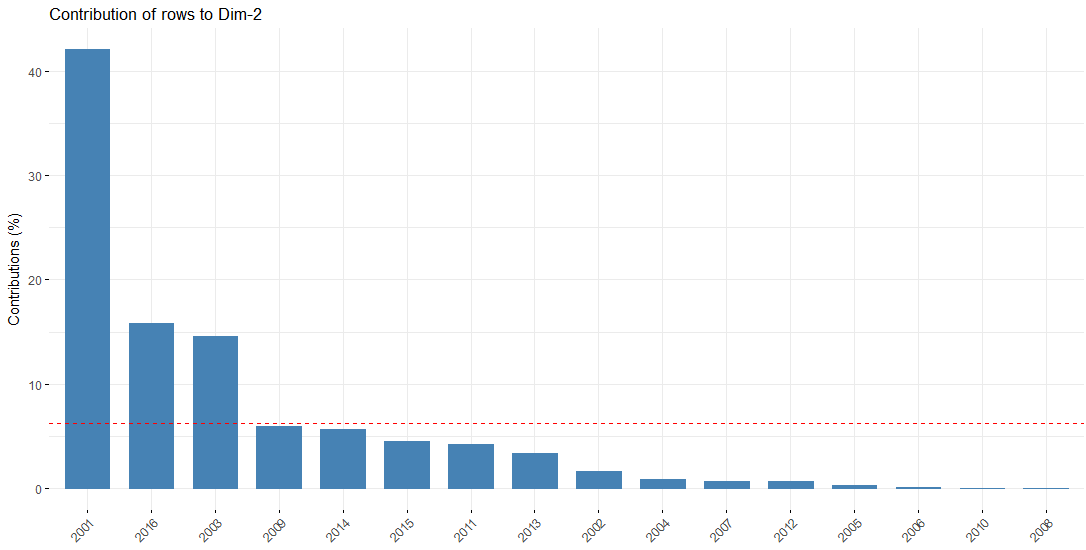
****

****

D’autre part, la contribution des lignes (resp. des colonnes) à la construction des axes est un atout d’interprétation. En effet, les lignes qui contribuent le plus à la construction du plan principal sont les plus importantes et doivent être tenues dans l’interprétation de la variabilité des données. Une sélection pourrait d’ailleurs être faite dans ce sens, en négligeant les lignes (resp. les colonnes) dont la contribution aux axes principaux est assez faible. Ces informations pertinentes pourraient être dégagées directement par l’exécution de la commande summary(Dataset). Cependant, l’abondance des données pourraient mettre l’analyste en difficulté.

On obtient les graphiques ci-dessous :





**3. Conclusion Générale :**

* Le terrorisme est devenu un risque imminent qui ne laisse aucune région du monde épargnée
* Le pays le plus impacté par les attentats terroristes est l’Irak avec un nombre pharaonique d’attaques terroristes entre 2001 et 2016 résultant vers des pertes humaines très importantes.
* Parmi les régions du monde, on constate que la région du moyen orient est la plus touchée par les attaques terroristes surtout avec la succession des attentats revendiqués par Le groupe de l’Etat Islamique.
* Les modes opératoires les plus utilisés ces dernières années par les groupes terroristes sont l’assassinat, les explosions et les attaques armées.
* De nombreux pays ont connu une diminution du taux des attentats contrairement au reste des pays du monde.
* Généralement, on associe le terrorisme aux groupes à titre djihadiste ou Islamiste mais il s’est avéré que le terrorisme implique aussi les crimes de haines et les conflits entre les groupes ethniques ainsi que les combats réguliers dans quelques pays contre les forces armés plus particulièrement dans les états connus pour le trafic de drogues.