Nous avons tout d’abord supprimer à la main les colonnes qui nous semblait inutile.

Ensuite à l’aide de ce script python (voir fichier calcul\_cel\_vide.py) nous avons lister les colonnes qui ont plus de 100 000 cellule vide, sachant que notre fichier initial possède 181 690lignes. Nous avons eu ce resultat (prendre des screen de l’excel génére par le script). Nous avons fait cette étape pour avoir une première vision des colonnes potentiellement supprimable.

Nous avons par la suite de ce script, créer un nouveau qui lui vas maintenant supprimer les colonnes mentionner dans le fichier excel dont nous en avons parler précedemment.

**Nouvelle version : (prendre en compte ceci)**

Grâce au script, nous supprimons les colonnes qui ont plus de 100 000 lignes vides.

Ensuite, nous allons traiter les colonnes une à une (du moins celle qui represente des incohérence dans leur données) :

* Colonnes imonth : nous supprimons les lignes qui ont pour mois « 0 » car ce n’est pas logique, il n’existe pas de mois 0 (20 lignes impacter)
* Colonnes iday : pareil que pour imonth (892 lignes impacter)
* Colonnes atacktype1 : on supprime les lignes vides car non pertinent (30116 lignes impacter)
* Colonnes attacktype1\_txt : on supprime les lignes vides car non pertinent (30116 lignes impacter)
* Colonnes targetype1 : on supprime les lignes vides car non pertinent (30116 lignes impacter)
* Colonnes targetype1\_txt : on supprime les lignes vides car non pertinent (30116 lignes impacter)
* Colonnes INT\_LOG, INT\_IDEO, INT\_MISC, INT\_ANY : on les supprime car elle nous n’apporte pas d’informations
* Colonnes crit1, crit2, crit3  : on les supprime car elle nous n’apporte pas d’informations
* Colonnes nperps : on remplace les lignes a valeur négative et les lignes vides par -9 (pour dire que l’on ne connait pas le nombre de terroriste qui est impliqué dans l’attaque)
* Colonne claimed : on remplace les valeur vides par -9 (pour dire que l’on n’a pas d’info si l’attaque est revendiqué ou non)
* Colonne weatype : on remplace les lignes vides par 13
* Colonne weatype\_txt : on remplace les lignes vides par « Unknown »

Fichier de base : 135 colonnes et 181 690 lignes

Fichier nettoyés : 47 colonnnes et 150 728 lignes