

RAPPORT : SOUTENANCE DE MATHÉMATIQUES POUR BIG DATA

CORENTIN GARNIER
STEVEN BOUCHE
LINA BELKARFA

Sujet : GUN VIOLENCE DATA
Date de rendu : 25/01/2021

I. Descriptif du sujet et de ses finalités

Enoncé :

Il s'agira de mener une série de statistiques sur l'archive du même intitulé qui se trouve ici.

Cet archive (au format CSV) contient des entrées relatives aux faits divers causés par des armes à feu qui ont eu lieu aux Etats Unis entre le 1er janvier 2013 et le 31 mars 2018. Il s'agira de montrer l'évolution d'un certain nombre de variables pendant la période concernée, d'en calculer la moyenne et l'écart type (quand cela est possible). L'ensemble V des variables qui nous intéressent contient :

- nombre de blessées ;*
- nombre de morts ;*
- nombre de malfaiteurs ;*
- âge des malfaiteurs ;*
- état dans lequel le fait divers a eu lieu.*

Questions :

1. Est-ce qu'il y a **des corrélations (linéaires)** entre les variables de V ?
2. Prenez les données de la période entre le **1er janvier 2013 et le 31 décembre 2017** pour calculer **les quantités statistiques (moyenne, écart type, etc)** et considérez les données **entre le 1 janvier 2018 et le 31 mars 2018** comme un **échantillon**. Est-ce que l'on peut dire que **les valeurs moyennes des variables dans V ont significativement changé** par rapport au passé ?
3. Si l'on prend en compte aussi **le mois de l'année dans lequel le fait divers a été commis**, est-ce qu'il y a une **forte corrélation entre le nombre de faits divers et le mois de l'année** ? Quelles conclusions en tirez-vous ?

Finalités du sujet :

- *Traitement des données afin qu'elles soient exploitables*
- *Création des colonnes manquantes (si nécessaire)*
- *Etude des données en répondant aux problématiques posées*
- *Conclusion à tirer de cette étude et informations supplémentaires*

Les fichiers :

- *PRE est le dossier contenant les fichiers de scripts de pré-traitement des données suivant :*
 - *1_PréTraitementDeLaBase.R*
 - *2_ModificationTypeVariables.R*
 - *R est le dossier contenant les fichiers de scripts d'études statistiques (répondant au question) :*
 - *1_Corrélations_Linéaires.R*
 - *2 Changements Significatifs.R*
 - *3_Crimes_et_mois.R*
 - *FIG contient les scripts R pour générer les graphiques contenu dans le rapport et des fichiers image:*
 - *VisualisationDesDonnées.R*
 - *Des fichiers png des graphiques générés*
 - *REP le rapport final au format html*
-

II. Descriptif des données utilisées

Les données ont été chargées depuis Kaggle au format csv. Voici une vue d'une partie des données :

incident_id	date	state	city_or_county	address	n_killed	n_injured	incident_url
461105	2013-01-01	Pennsylvania	McKeesport	1506 Versailles Avenue and Coursin Street	0	4	http://www.gunviolencea
460726	2013-01-01	California	Hawthorne	13500 block of Cerise Avenue	1	3	http://www.gunviolencea
478855	2013-01-01	Ohio	Lorain	1776 East 28th Street	1	3	http://www.gunviolencea
478925	2013-01-05	Colorado	Aurora	16000 block of East Ithaca Place	4	0	http://www.gunviolencea
478959	2013-01-07	North Carolina	Greensboro	307 Mourning Dove Terrace	2	2	http://www.gunviolencea
478948	2013-01-07	Oklahoma	Tulsa	6000 block of South Owasso	4	0	http://www.gunviolencea
479363	2013-01-19	New Mexico	Albuquerque	2806 Long Lane	5	0	http://www.gunviolencea
479374	2013-01-21	Louisiana	New Orleans	LaSalle Street and Martin Luther King Jr. Boulevard	0	5	http://www.gunviolencea
479389	2013-01-21	California	Brentwood	1100 block of Breton Drive	0	4	http://www.gunviolencea
492151	2013-01-23	Maryland	Baltimore	1500 block of W. Fayette St.	1	6	http://www.gunviolencea
491674	2013-01-23	Tennessee	Chattanooga	1501 Dodds Ave	1	3	http://www.gunviolencea
479413	2013-01-25	Missouri	Saint Louis	W Florissant Ave and Riverview Blvd	1	3	http://www.gunviolencea
479561	2013-01-26	Louisiana	Charenton	1000 block of Flat Town Road	2	3	http://www.gunviolencea
479554	2013-01-26	District of Columbia	Washington	2403 Benning Road Northeast	0	5	http://www.gunviolencea
479460	2013-01-26	Ohio	Springfield	601 West Main Street	1	3	http://www.gunviolencea

En faisant la commande “summary()”, on peut également observer le type de nos données ainsi que quelques valeurs statistiques.

> summary(DATA)											
incident_id	date	state	city_or_county	address	n_killed	n_injured	incident_url	source_url	incident_url_fields_missing	congressional_district	
Min. : 92114	Length:239677	Length:239677	Length:239677	Length:239677	Min. : 0.0000	Min. : 0.000	Length:239677	Length:239677	Length:239677	Min. : 0.000	
1st Qu.: 308545	Class :character	Class :character	Class :character	Class :character	1st Qu.: 0.0000	1st Qu.: 0.000	Class :character	Class :character	Class :character	1st Qu.: 2.000	
Median : 543587	Mode :character	Mode :character	Mode :character	Mode :character	Median : 0.0000	Median : 0.000	Mode :character	Mode :character	Mode :character	Median : 5.000	
Mean : 559334					Mean : 0.2523	Mean : 0.494				Mean : 8.001	
3rd Qu.: 817228					3rd Qu.: 0.0000	3rd Qu.: 1.000				3rd Qu.:10.000	
Max. :1083472					Max. :50.0000	Max. :53.000				Max. :53.000	
										NA's :11944	
gun_stolen	gun_type	incident_characteristics	latitude	location_description	longitude	n_guns_involved	notes	participant_age	participant_age_group	participant_gender	
Length:239677	Length:239677	Length:239677	Min. :19.11	Length:239677	Min. :-171.43	Min. : 1.00	Length:239677	Length:239677	Length:239677	Length:239677	Length:239677
Class :character	Class :character	Class :character	1st Qu.:33.90	Class :character	1st Qu.: -94.16	1st Qu.: 1.00	Class :character	Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Median :38.57	Mode :character	Median : -86.25	Median : 1.00	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character
			Mean :37.55		Mean : -89.34	Mean : 1.37					
			3rd Qu.:41.44		3rd Qu.: -80.05	3rd Qu.: 1.00					
			Max. :71.34		Max. : 97.43	Max. :400.00					
			NA's :7923		NA's :7923	NA's :99451					
participant_name	participant_relationship	participant_status	participant_type	sources	state_house_district	state_senate_district					
Length:239677	Length:239677	Length:239677	Length:239677	Length:239677	Min. : 1.00	Min. : 1.00					
Class :character	Class :character	Class :character	Class :character	Class :character	1st Qu.: 21.00	1st Qu.: 9.00					
Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Median :47.00	Median :19.00					
					Mean :55.45	Mean :20.48					
					3rd Qu.:84.00	3rd Qu.:30.00					
					Max. :901.00	Max. :94.00					
					NA's :38772	NA's :32335					

Nous n’avons pas besoin de toutes les colonnes, en tout il y en a 29 pour 239 677 lignes :

```
> ncol(DATA)
[1] 29
> nrow(DATA)
[1] 239677
```

Il nous faut :

- nombre de blessées ; “n_injured”
- nombre de morts ; “n_killed”
- nombre de malfaiteurs ; à créer à partir de “participant_type”
- âge des malfaiteurs ; à créer à partir de “participant_type” et “participant_age” ou “participant_age_group”
- état dans lequel le fait divers a eu lieu ; “state”

Voici une vue de la base après avoir gardé les colonnes nécessaires :

incident_id	state	n_killed	n_injured	participant_age	participant_age_group	participant_type
461105	Pennsylvania	0	4	0:20	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Subject-Suspect
460726	California	1	3	0:20	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Subject-Suspect
478855	Ohio	1	3	0:25 1:31 2:33 3:34 4:33	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0:Subject-Suspect 1:Subject-Suspect 2:Victim 3:Victim 4:Subject-Suspect
478925	Colorado	4	0	0:29 1:33 2:56 3:33	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+	0:Victim 1:Victim 2:Victim 3:Subject-Suspect
478959	North Carolina	2	2	0:18 1:46 2:14 3:47	0:Adult 18+ 1:Adult 18+ 2:Teen 12-17 3:Adult 18+	0:Victim 1:Victim 2:Victim 3:Subject-Suspect
478948	Oklahoma	4	0	0:23 1:23 2:33 3:55	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Subject-Suspect 5:Subject-Suspect
479363	New Mexico	5	0	0:51 1:40 2:9 3:5 4:2 5:15	0:Adult 18+ 1:Adult 18+ 2:Child 0-11 3:Child 0-11 4:Child 12-17 5:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Victim 5:Subject-Suspect
479374	Louisiana	0	5	NA	NA	0:Victim 1:Victim 2:Victim 3:Victim 4:Victim 5:Subject-Suspect
479389	California	0	4	NA	0:Teen 12-17 1:Teen 12-17 2:Teen 12-17 4:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Subject-Suspect
492151	Maryland	1	6	0:15	0:Teen 12-17 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Victim 5:Victim 6:Subject-Suspect
491674	Tennessee	1	3	0:19	0:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Subject-Suspect
479413	Missouri	1	3	0:28	0:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Subject-Suspect
479561	Louisiana	2	3	3:78 4:48	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Subject-Suspect
479554	District of Columbia	0	5	NA	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Victim 5:Subject-Suspect
479460	Ohio	1	3	0:34 1:28 2:23 3:29 4:29	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Subject-Suspect
479573	Tennessee	0	5	5:24	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 4:Adult 18+ 5:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Victim 5:Subject-Suspect
479580	California	1	3	0:20 4:25 5:18 6:19	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Subject-Suspect 5:Subject-Suspect
479592	Illinois	0	4	0:18 1:41 2:28 3:28	0:Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0:Victim 1:Victim 2:Victim 3:Victim 4:Subject-Suspect 5:Subject-Suspect

Nous avons choisi dans un premier temps de créer la colonne contenant **le nombre de malfaiteurs**, car elle est la plus rapide à créer. Nous avons utilisé la fonction `str_count()` de `tidyverse`, qui compte le nombre de fois qu'apparaît une string choisie.

Dans un second temps, nous avons décidé que nous utiliserons la colonne **participant_age_group**, car elle contient plus d'information que `participant_age`. En effet, la colonne de l'âge contient **39% de valeur manquante**. Elle en contient en réalité **bien plus**, car dans le détail, chaque cellule de cette colonne contient l'âge de chaque participant séparé par des " || ". Or, dans beaucoup de cas, il y a **l'âge d'un participant ou plus**, ce qui signifie qu'en réalité une cellule remplie n'est pas toujours complète. Dans 61% des cellules il y a une information, mais pas nécessairement celle du ou des malfaiteurs.

incident_id	date	state	# n_killed	# n_injured	participant_age	participant_age_g...	participant_type
	Date of crime	State of crime	Number of people killed	Number of people injured	Age of participant(s) at the time of crime	Age group of participant(s) at the time of crime	Type of participant
		Illinois 7% California 7% Other (205815) 86%			[null] 39% 0::24 2% Other (143565) 60%	0::Adult 18+ 39% 0::Adult 18+ 1:Ad... 21% Other (95733) 40%	0::Victim 24% 0::Victim 1:Subjec... 21% Other (130534) 54%
461185	2013-01-01	Pennsylvania	0	4	0::20	0::Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0::Victim 1::Victim 2::Victim 3::Victim 4::Subject-Suspect
468726	2013-01-01	California	1	3	0::20	0::Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0::Victim 1::Victim 2::Victim 3::Victim 4::Subject-Suspect
478855	2013-01-01	Ohio	1	3	0::25 1::31 2::33 3::34 4::33	0::Adult 18+ 1:Adult 18+ 2:Adult 18+ 3:Adult 18+ 4:Adult 18+	0::Subject-Suspect 1::Subject-Suspect 2::Victim 3::Victim 4::Victim

Notre choix a donc été d'utiliser la colonne de **groupe d'âge**, qui contient 3 catégories d'âge - Adult, Teen, Child :

```
0::Adult
18+||1::Adult
18+||2::Child
0-11||3::Child
0-11||4::Child
0-11||5::Teen 12-17
```

```
0::Adult 18+||
1::Adult 18+||
2::Child 0-11||
3::Child 0-11||
4::Child 0-11||
5::Teen 12-17
```

Pour faire ce tri, nous avons utilisé des fonctions de `tidyverse`. Nous avons appliqué dans un premier temps **la fonction `str_split()`** sur nos colonnes catégorie d'âge et sur le type. Cette fonction permet de séparer les données par le séparateur choisie : dans notre cas " || ". Or, R ne reconnaît pas ces caractères comme tels, mais comme un opérateur logique (peu importe comment nous l'écrivons, il ne le détecte pas).

Nous avons donc choisi de **modifier ce séparateur** en le remplaçant par "et". Cela n'est pas une méthode très formelle, mais il y a beaucoup de symboles dans ce tableau, il est donc difficile d'en choisir un sans risquer de tout détruire et devoir recommencer.

Après quoi, il a été très simple d'utiliser `str_split`, et d'obtenir à partir des colonnes, de **nouvelles matrices dont chaque ligne est un crime, et chaque colonne est à la position de l'identifiant du participant -1** (car il commence à 0 et dans R les colonnes commencent à 1). Dans une matrice il y a **le type du participant** dans chaque cellule, et dans l'autre **il y a sa catégorie d'âge**.

Par la suite, il a suffi alors de **croiser les deux matrices** en gardant les indices de lignes/colonnes de ceux étant suspect et étant Adult (ensuite suspect et Teen, puis enfin suspect et Child).

La base finale ressemble alors à cela :

incident_id	state	n_killed	n_injured	ColonneNbMalfaiteur	Adult	Teen	Child
461105	Pennsylvania	0	4	1	1	0	0
460726	California	1	3	1	0	0	0
478855	Ohio	1	3	2	2	0	0
478925	Colorado	4	0	1	1	0	0
478959	North Carolina	2	2	1	1	0	0
478948	Oklahoma	4	0	2	2	0	0
479363	New Mexico	5	0	1	0	1	0
479374	Louisiana	0	5	1	0	0	0
479389	California	0	4	1	0	0	0
492151	Maryland	1	6	0	0	0	0
491674	Tennessee	1	3	1	0	0	0
479413	Missouri	1	3	1	0	0	0
479561	Louisiana	2	3	1	1	0	0
479554	District of Columbia	0	5	1	1	0	0
479460	Ohio	1	3	1	1	0	0
479573	Tennessee	0	5	1	0	0	0
479580	California	1	3	3	3	0	0
479592	Illinois	0	4	2	2	0	0
479603	Louisiana	0	4	3	3	0	0

- **incident_id** : contient l'id du crime (qui nous sera utile pour récupérer les dates par exemple)
- **state** : l'état dans lequel le crime a été commis
- **n_killed** : le nombre de personnes tués lors du crime
- **n_injured** : le nombre de blessés lors du crime
- **ColonneNbMalfaiteur** : le nombre de malfaiteur lors du crime
- **Adult** : le nombre de malfaiteurs adultes (plus de 18 ans)
- **Teen** : le nombre de malfaiteurs adolescents (entre 12 et 17 ans)
- **Child** : le nombre de malfaiteurs enfants (entre 0 et 11 ans)

Le code de ces traitements est contenu dans le fichier "1_PréTraitementDeLaBase.R".

III. Méthodologie : réponses aux questions et résultats

1. Est-ce qu'il y a des corrélations (linéaires) entre les variables de V ?

Tout d'abord, nous avons encodé nos data, car les états ne l'étaient pas, pour cela, il a suffit d'utiliser la fonction `one_hot()` afin d'obtenir une colonne par état, avec 0 si ce n'est pas dans cet état que le crime a été commis, et 1 dans le cas contraire (traitement dans le fichiers `2_ModificationTypeVariables.R`) :

```
#Encodage : cols est la colonne qu'on encode
#
EncodingData<-one_hot(dt=DATA, cols = "state", sparsifyNAs = FALSE, naCols = FALSE, dropCols = TRUE, dropUnusedLevels = FALSE)
view(EncodingData)
```

Nous allons maintenant utiliser la méthode "`cor()`" pour répondre à cette question : **`cor(data, method = c("pearson", "kendall", "spearman"))`**. Cette fonction nous permet d'obtenir la **matrice des coefficients de corrélation** entre les différentes **paires possibles de variables**.

Nous choisirons d'utiliser le paramètre **Pearson**. Le coefficient de corrélation de pearson mesure **une corrélation linéaire entre deux variables**, c'est-à-dire, la dépendance linéaire entre deux variables.

Ce coefficient de corrélation est compris entre -1 (pour une **forte corrélation négative**) et 1 (pour une **forte corrélation positive**). En effet, plus il est proche de 1, plus la corrélation est importante, et plus il est proche de 0, plus la corrélation est faible.

Cependant, il nous faut également **évaluer la significativité de la corrélation**. Si la **p-value est inférieure à 5% ($p < 0.05$)**, la corrélation est dite **significative**.

Ainsi, une faible corrélation peut cependant s'avérer certaine, et une forte corrélation s'avérer incertaine.

La formule du coefficient de corrélation de Pearson est la suivante :

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

x et y sont les observations, et m_x , m_y sont les moyennes de X et Y.

La p-value, c'est à dire le niveau de significativité de chaque corrélation, peut être déterminé à partir de la table de Student :

d.d./a	0.9	0.5	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0.158	1	2	3.078	6.314	12.706	31.821	64	637
2	0.142	0.816	1.386	1.886	2.92	4.303	6.965	10	31.598
3	0.137	0.765	1.25	1.638	2.353	3.182	4.541	5.841	12.929
4	0.134	0.741	1.19	1.533	2.132	2.776	3.747	4.604	8.61
5	0.132	0.727	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.718	1.134	1.44	1.943	2.447	3.143	3.707	5.959
7	0.13	0.711	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.13	0.706	1.108	1.397	1.86	2.306	2.896	3.355	5.041
9	0.129	0.703	1.1	1.383	1.833	2.263	2.821	3.25	4.781
10	0.129	0.7	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.129	0.697	1.088	1.363	1.796	2.201	2.718	3.106	4.437

C'est cette méthode qu'utilise la fonction **rcorr()** que nous utiliserons également pour calculer les p-values de chaque paire de variables.

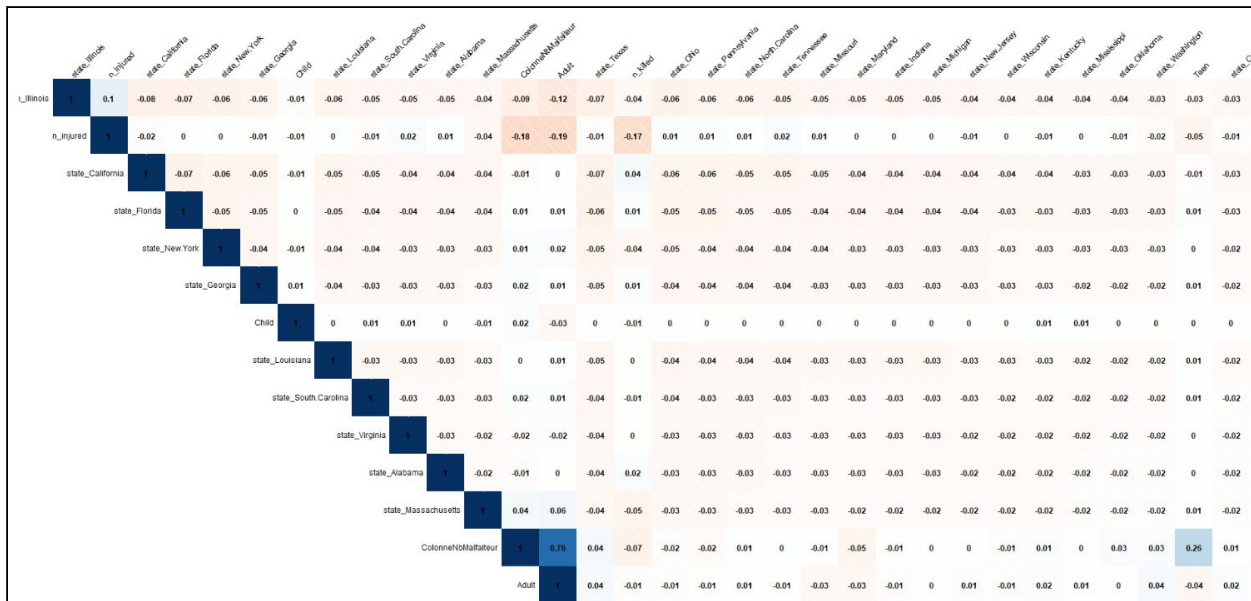
Nous prenons soin de retirer la colonnes des ID des crimes, et appliquer la fonction **cor()** pour obtenir la matrice de corrélation que voici :

	state_Delaware	state_District.of.Columbia	state_Florida	state_Georgia	state_Hawaii	state_Idaho	state_Illinois	state_Indiana	state_I
state_South.Dakota	-0.0037680821	-0.004613496	-0.0109971890	-0.008575581	-0.001523798	-0.0023158865	-0.012641834	-0.0070951542	-0.004
state_Tennessee	-0.0157040772	-0.019227483	-0.0456325213	-0.035740089	-0.006350667	-0.0096518226	-0.052686840	-0.0295701753	-0.017
state_Texas	-0.0209362054	-0.025633504	-0.0611025447	-0.047647616	-0.008466519	-0.0128675208	-0.070240517	-0.0394220721	-0.023
state_Utah	-0.0056237654	-0.006885527	-0.0164130209	-0.012798834	-0.002274229	-0.0034564008	-0.018867611	-0.0105893346	-0.006
state_Vermont	-0.0037175907	-0.004551679	-0.0108496292	-0.008460670	-0.001503379	-0.0022846541	-0.012472436	-0.0070000808	-0.004
state_Virginia	-0.0136102575	-0.016908761	-0.0403053876	-0.031430043	-0.005584814	-0.0084878693	-0.046333115	-0.0260041853	-0.015
state_Washington	-0.0103268522	-0.012643610	-0.0301390311	-0.023502343	-0.004176139	-0.0063469469	-0.034646366	-0.0194450667	-0.011
state_West.Virginia	-0.0070836779	-0.008672989	-0.0206737914	-0.016121372	-0.002864612	-0.0043536720	-0.023765586	-0.0133382939	-0.008
state_Wisconsin	-0.0116892614	-0.014311893	-0.0341152375	-0.026602979	-0.004727092	-0.0071842921	-0.039217220	-0.0220104311	-0.013
state_Wyoming	-0.0026372701	-0.003228975	-0.0076969017	-0.006002025	-0.001066502	-0.0016208824	-0.008847984	-0.0049658785	-0.002
n_killed	-0.0225711149	-0.016512653	0.0125459398	0.010821521	-0.002379491	-0.0002422073	-0.039095167	0.0038977059	-0.025
n_injured	-0.0008204107	0.007416106	0.0024621349	-0.009500743	-0.010247294	-0.0211255004	0.104227634	-0.0020644425	-0.018
ColonneNbMalfaiteur	-0.0090803722	-0.010492400	0.0063149360	0.018850100	0.006371730	0.0089125585	-0.085916285	-0.0096107782	0.0116
Adult	-0.0024804880	-0.013500740	0.0054024783	0.005072316	0.012581890	0.0164926337	-0.116629537	-0.0110982475	0.0023
Teen	0.0026827718	-0.008769808	0.0149964716	0.006757180	-0.003919537	0.0006250931	-0.025090076	0.0006428571	0.0126
Child	-0.0032961287	-0.003610875	0.0003909584	0.009259334	-0.001749973	0.0073425714	-0.010231530	0.0007599717	-0.002

En effet, nous choisissons d'afficher une petite partie car notre matrice de corrélation comporte 57 colonnes fois 57 lignes.

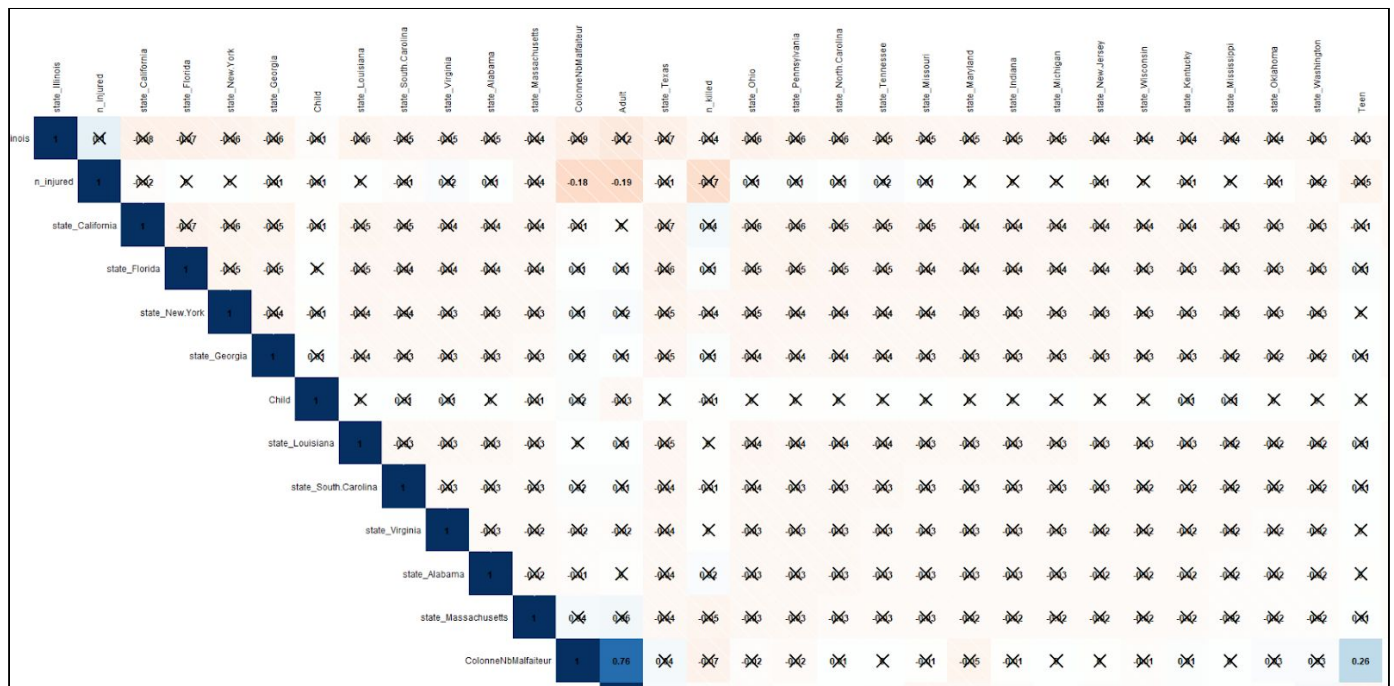
```
> ncol(EncodingData)
[1] 58
```

Nous choisissons alors de la représenter de manière plus propre, en créant un fichier png à partir de R, au dimension que l'on souhaite (fichier Corrélation_Linéaire.PNG), voici un zoom sur les variables corrélées :



A première vue, il semble y avoir quelques corrélations linéaires plus ou moins importantes. Cependant, nous devons observer les corrélations significatives (de la même manière, nous créons un second fichier “*Corrélation_Significativite_0,01.PNG*”) :

```
png(file="Corrélation_Significativite.png", width=4000, height=4000)
corrplot(Matcor, type="upper", order="hclust", tl.col="black",method="shade",outline = FALSE,addCoef.col = "black",p.mat = Pvalue$P, sig.level = 0.01)
dev.off()
```



Nous pouvons désormais voir les corrélations suivantes(significatives au niveau $p < 0,01$) :

```
# -Corrélation positive entre les variables Adult et ColonneNbMalfaiteur -> 0,76
# -Corrélation positive entre les variables Teen et ColonneNbMalfaiteur -> 0,26
# -Corrélation négative entre les variables Adult et n_injured -> -0.19
# -Corrélation négative entre les variables ColonneNbMalfaiteur et n_injured -> -0.18
```

Affichons de plus prêt les coefficients de corrélations et leur p_value associée:

```

n_injured      n_injured ColonneNbMalfaiteur      Adult      Teen
n_injured      1.0000000      -0.1821953 -0.18783374 -0.05487770
ColonneNbMalfaiteur -0.1821953      1.0000000      0.76399812 0.25991095
Adult      -0.1878337      0.7639981      1.00000000 -0.04013131
Teen      -0.0548777      0.2599110      -0.04013131 1.00000000
> Pvalue$P[c(53,54,55,56),c(53,54,55,56)]
n_injured      n_injured ColonneNbMalfaiteur      Adult      Teen
n_injured      NA      0.002561041 0.002054628 0.300916957
ColonneNbMalfaiteur 0.002561041      NA      0.000000000 0.006853091
Adult      0.002054628      0.000000000      NA      0.572306756
Teen      0.300916957      0.006853091 0.572306756      NA

```

Nous avons bien nos p-values en dessous de 0.01 pour les corrélations qui nous intéressent.

Cela signifie qu'il y a une **forte corrélation positive et significative au niveau $p < 0.01$, entre le nombre de malfaiteurs durant un crime, et le nombre de malfaiteurs Adult**. Plus il y a de malfaiteurs, plus il y a de chance qu'il y ait des adultes (coefficient de corrélation = 0,76).

On peut également effectuer le test de corrélation avec cor.test qui nous donne également un intervalle de confiance à 95%.

```
> cor.test(DATA$Adult, DATA$ColonneNbMalfaiteur)

Pearson's product-moment correlation

data: DATA$Adult and DATA$ColonneNbMalfaiteur
t = 548.8, df = 214812, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7622319 0.7657529
sample estimates:
      cor
0.7639981
```

Il y a également une corrélation **entre le nombre de malfaiteurs, et le nombre de malfaiteurs adolescents**, mais celle-ci est moins forte que la précédente. **Elle est positive et significative au niveau de $p < 0.01$** . Plus il y a de malfaiteurs, plus il y a de chance qu'il y ait des adolescents, mais l'influence de cette variable est moins importante que celle Adult.

Cela nous informe que de manière plus globale **il y a une corrélation entre l'âge et le nombre de malfaiteurs** : Plus il y a de malfaiteur, plus la moyenne d'âge sera élevée (corrélation positive,et les adultes ayant une grande influence, les ados une petite, et l'enfant pas de corrélation significative).

Il y a également de plus petites corrélations : **corrélation négative et significative au niveau $p < 0.01$ entre le nombre de malfaiteurs et le nombre de personnes blessées**. Cela peut paraître paradoxale, car on imagine que plus il y a d'agresseurs, plus il y a de blessés. Or, il est possible que cette légère corrélation montre que :

- Plus il y a d'agresseur plus il y a de mort (donc moins dans la catégorie injured et plus dans n_killed)
- Ou plus il y a d'agresseur, moins ils sont obligés d'agresser pour arriver à leur fin

On ne peut pas émettre d'affirmation à ce sujet, cependant il **n'apparaît pas de corrélation significative entre le nombre de mort et le nombre d'agresseur** :

```
> Pvalue$P[c(54,52),c(54,52)]
               ColonneNbMalfaiteur  n_killed
ColonneNbMalfaiteur                NA 0.4342099
n_killed                        0.4342099      NA
> MatCor[c(54,52),c(54,52)]
               ColonneNbMalfaiteur  n_killed
ColonneNbMalfaiteur                1.0000000 -0.06737866
n_killed                        -0.06737866  1.00000000
```

On peut donc supposer que la seconde raison peut être la bonne: plus il y a d'agresseur, moins il y a de blessé, probablement pour des raisons psychologiques/sociales (les victimes ont peur du grand nombre d'agresseur alors elles se laissent faire par exemple, ou les agresseurs étant plus en confiance ont moins tendance à agresser).

De la même manière, **une petite corrélation négative mais significative entre le nombre de malfaiteurs adultes et le nombre de personnes blessées**. Cela semble logique, étant donné les conclusions que nous venons de tirer : Plus il y a de malfaiteur Adulte, moins il y a de blessés. On peut conclure la même chose mais plus précisément :

Soit plus il y a un grand nombre d'agresseurs adultes, plus les victimes ont peur et ne sont pas difficiles et donc il y a moins de blessés. Soit plus il y a d'agresseurs adultes, moins ils agressent (par maturité par exemple, contrairement à des ados qui pourraient paniquer). Cet argument semble intéressant puisque nous avons vu plus haut que les crimes étaient principalement commis par des adultes. Il faut cependant toujours garder à l'esprit que cela peut être une variable externe qui influence également ce résultat (âge légal pour détenir une arme à feu par exemple).

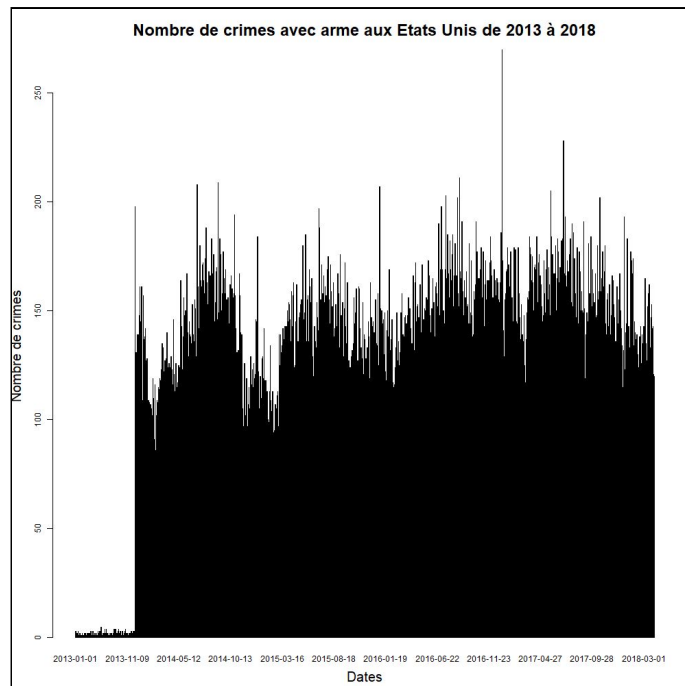
2. Prenez les données de la période entre le 1er janvier 2013 et le 31 décembre 2017 pour calculer les quantités statistiques (moyenne, écart type, etc) et considérez les données entre le 1 janvier 2018 et le 31 mars 2018 comme un échantillon. Est-ce que l'on peut dire que les valeurs moyennes des variables dans V ont significativement changé par rapport au passé ?

Dans un premier temps, nous récupérons grâce au ID des crimes, les dates de chacun d'entre eux (et les lat et longitude pour créer une map plus tard).

Ainsi, nous avons les données tel que :

date	latitude	longitude	incident_id	state_Alabama	state_Alaska	state_Arizona	state_Arkansas	state_California
2013-01-01	40.3467	-79.8559	461105	0	0	0	0	0
2013-01-01	33.9090	-118.3330	460726	0	0	0	0	1
2013-01-01	41.4455	-82.1377	478855	0	0	0	0	0
2013-01-05	39.6518	-104.8020	478925	0	0	0	0	0
2013-01-07	36.1140	-79.9569	478959	0	0	0	0	0
2013-01-07	36.2405	-95.9768	478948	0	0	0	0	0
2013-01-19	34.9791	-106.7160	479363	0	0	0	0	0
2013-01-21	29.9435	-90.0836	479374	0	0	0	0	0
2013-01-21	37.9656	-121.7180	479389	0	0	0	0	1
2013-01-23	39.2899	-76.6412	492151	0	0	0	0	0

Après avoir formalisé les dates, nous voudrions voir une représentation dans le temps. Si nous affichons le nombre de crime dans le temps nous obtenons une représentation un peu brouillon :

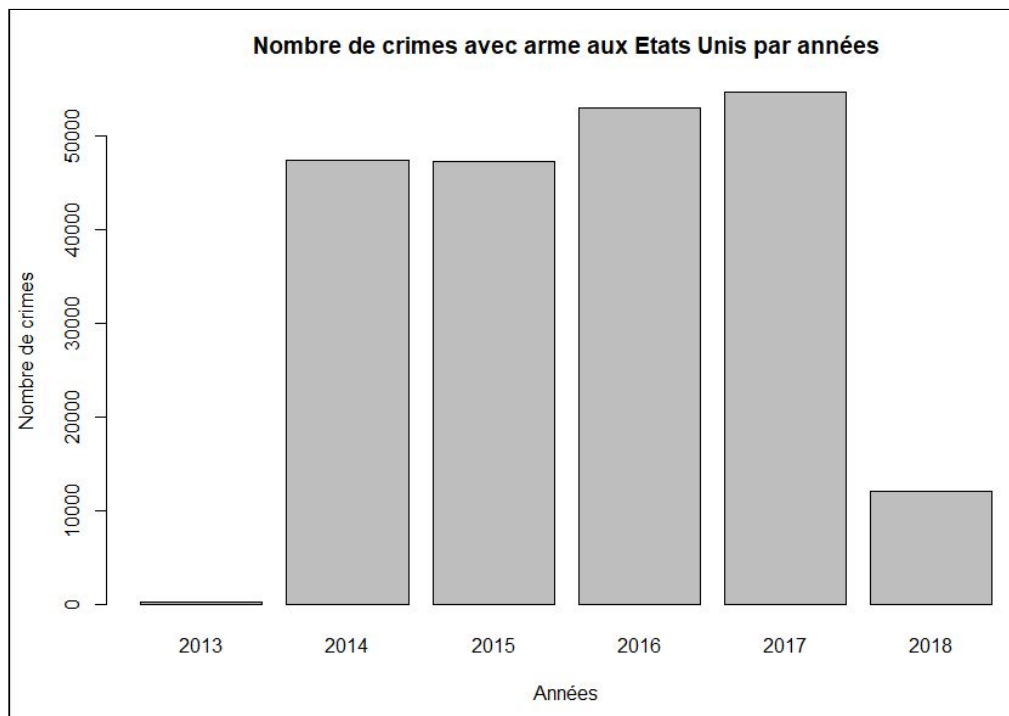


Cela nous permet de comprendre qu'il nous faut faire un traitement afin d'avoir des colonnes de mois et d'années en plus :

Voici la base de données avec désormais les années et mois également :

Annees	Mois	date	latitude	longitude	incident_id	state_Alabama	state_Alaska	stat
2013	janvier	2013-01-01	40.3467	-79.8559	461105	0	0	0
2013	janvier	2013-01-01	33.9090	-118.3330	460726	0	0	0
2013	janvier	2013-01-01	41.4455	-82.1377	478855	0	0	0
2013	janvier	2013-01-05	39.6518	-104.8020	478925	0	0	0
2013	janvier	2013-01-07	36.1140	-79.9569	478959	0	0	0
2013	janvier	2013-01-07	36.2405	-95.9768	478948	0	0	0
2013	janvier	2013-01-19	34.9791	-106.7160	479363	0	0	0
2013	janvier	2013-01-21	29.9435	-90.0636	479374	0	0	0
2013	janvier	2013-01-21	37.9656	-121.7180	479389	0	0	0
2013	janvier	2013-01-23	39.2899	-76.6412	492151	0	0	0
2013	janvier	2013-01-23	35.0221	-85.2697	491674	0	0	0
2013	janvier	2013-01-25	38.7067	-90.2494	479413	0	0	0
2013	janvier	2013-01-26	29.8816	-91.5251	479561	0	0	0
2013	janvier	2013-01-26	38.8836	-75.8547	478551	0	0	0

Nous pouvons désormais constater que nos données sont plus lisibles, car nous pouvons afficher nos crimes par années :



En 2013, il y a moins d'observations que les autres années, car il y avait beaucoup de valeurs manquantes (que nous avons retiré pour une étude des données statistiques plus justes). En 2018, il s'agit de Janvier à Mars, donc $\frac{1}{4}$ de l'année, ce qui paraît logique à vue d'œil.

Nous avons alors coupé en deux nos données : une partie statistiques (2013 à 2017) et une partie échantillon (les 3 premiers mois de l'année 2018).

De plus, nous avons mis nos dates sous forme de **facteur**, afin d'avoir une logique temporelle. Par exemple, si l'on fait un summary de nos données de 2013 à 2017, on obtient un rangement correct pour les dates :

```
> #Quantités statistiques
> summary(Statistiques2013_2017)
```

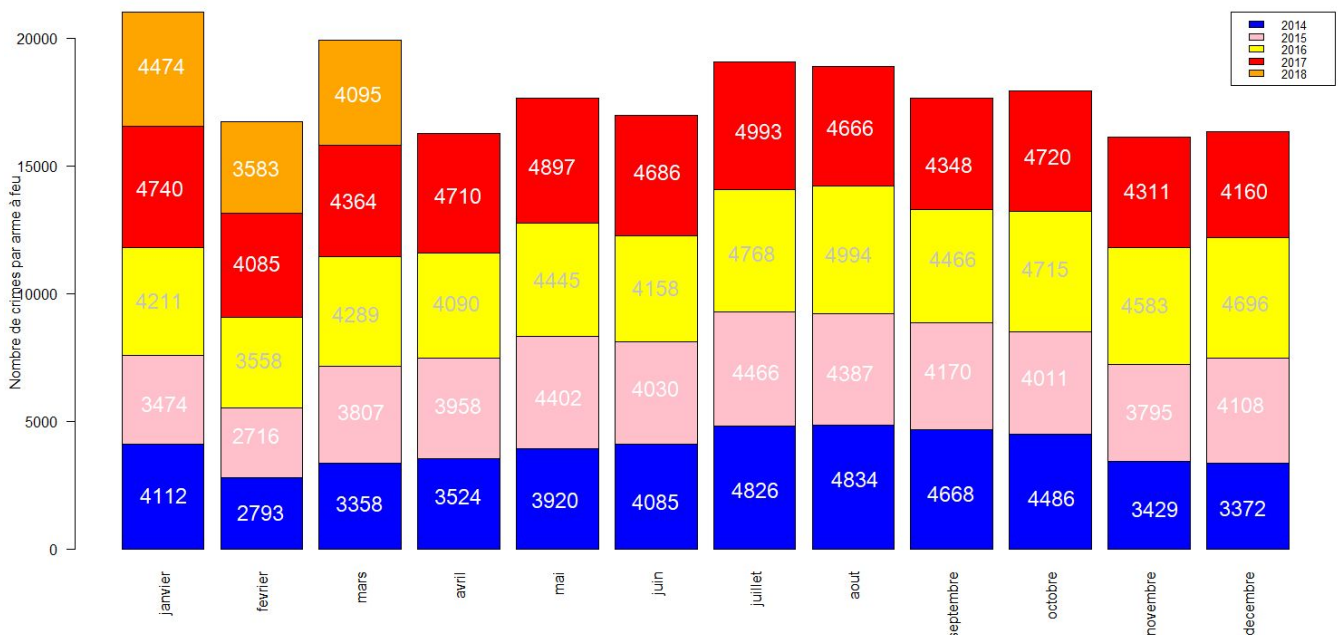
Années	Mois	date	latitude	longitude	incident_id	state_Alabama
Min. :2013	Length:202662	Min. :2013-01-01	Min. :19.11	Min. : -171.43	Length:202662	Min. :0.0000
1st Qu.:2015	Class :character	1st Qu.:2015-01-27	1st Qu.:33.92	1st Qu.: -93.78	Class :character	1st Qu.:0.0000
Median :2016	Mode :character	Median :2016-02-18	Median :38.51	Median : -86.25	Mode :character	Median :0.0000
Mean :2016		Mean :2016-01-27	Mean :37.51	Mean : -89.28		Mean :0.0238
3rd Qu.:2017		3rd Qu.:2017-01-27	3rd Qu.:41.33	3rd Qu.: -80.02		3rd Qu.:0.0000
Max. :2017		Max. :2017-12-31	Max. :71.34	Max. : -67.27		Max. :1.0000

state_Alaska	state_Arizona	state_Arkansas	state_California	state_Colorado	state_Connecticut	state_Delaware
Min. :0.000000	Min. :0.000000	Min. :0.000000	Min. :0.000000	Min. :0.000000	Min. :0.000000	Min. :0.000000
1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000
Median :0.000000	Median :0.000000	Median :0.000000	Median :0.000000	Median :0.000000	Median :0.000000	Median :0.000000
Mean :0.005077	Mean :0.009597	Mean :0.01189	Mean :0.06995	Mean :0.01227	Mean :0.01265	Mean :0.007278
3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000
Max. :1.000000	Max. :1.000000	Max. :1.000000	Max. :1.000000	Max. :1.000000	Max. :1.000000	Max. :1.000000

state_Texas	state_Utah	state_Vermont	state_Virginia	state_Washington	state_West.Virginia	state_Wisconsin
Min. :0.000000	Min. :0.000000	Min. :0.000000	Min. :0.000000	Min. :0.000000	Min. :0.000000	Min. :0.000000
1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000	1st Qu.:0.000000
Median :0.000000	Median :0.000000	Median :0.000000	Median :0.000000	Median :0.000000	Median :0.000000	Median :0.000000
Mean :0.05639	Mean :0.004303	Mean :0.001895	Mean :0.02543	Mean :0.01424	Mean :0.006809	Mean :0.01807
3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000	3rd Qu.:0.000000
Max. :1.000000	Max. :1.000000	Max. :1.000000	Max. :1.000000	Max. :1.000000	Max. :1.000000	Max. :1.000000

state_Wyoming	n_killed	n_injured	Colonnenbmalfaiteur	Adult	Teen	child
Min. :0.0000000	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.000000	Min. :0.000000
1st Qu.:0.0000000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.000000	1st Qu.:0.000000
Median :0.0000000	Median :0.0000	Median :0.0000	Median :1.0000	Median :1.0000	Median :0.000000	Median :0.000000
Mean :0.0009474	Mean :0.2809	Mean :0.5538	Mean :0.9258	Mean :0.6816	Mean :0.05766	Mean :0.00261
3rd Qu.:0.0000000	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:0.000000	3rd Qu.:0.000000
Max. :1.0000000	Max. :50.0000	Max. :53.0000	Max. :63.0000	Max. :63.0000	Max. :10.00000	Max. :3.00000

Nous choisissons également de représenter nos données par mois et années, afin de voir ce que nous pouvons observer comme répartition (nous retirons donc l'année 2013) :



Pour voir si des valeurs ont changé de manière significative, nous devons effectuer un test de student sur les deux ensembles de valeurs que l'on souhaite comparer. Si la

p-value est inférieure à 0.05, nous pouvons affirmer de la significativité du test et donc du changement significatif des variables.

Nous allons donc comparer dans un premier temps les moyennes, puis dans un second temps les écarts-types, enfin, nous avons choisi d'appliquer également ce test au valeur tout court (en créant une fonction qui fait des test de student sur chaque colonne en comparant avant et après).

Calcul des moyennes de chaque colonne (des data statistiques 2013 à 2017) :

```
> Moyennes<-colMeans(as.matrix(Statistiques2013_2017[, -c(1,2,3,4,5,6)]))
> Moyennes
```

state_Alabama	state_Alaska	state_Arizona	state_Arkansas	state_California
0.0237982454	0.0050774195	0.0095972605	0.0118917212	0.0699539134
state_Colorado	state_Connecticut	state_Delaware	state_District.of.Columbia	state_Florida
0.0122716642	0.0126466728	0.0072781281	0.0109492653	0.0587332603
state_Georgia	state_Hawaii	state_Idaho	state_Illinois	state_Indiana
0.0369778251	0.0011891721	0.0027188126	0.0770938805	0.0254512439
state_Iowa	state_Kansas	state_Kentucky	state_Louisiana	state_Maine
0.0093850845	0.0085462494	0.0176747491	0.0362722168	0.0030938212
state_Maryland	state_Massachusetts	state_Michigan	state_Minnesota	state_Mississippi
0.0254413753	0.0220218887	0.0246617521	0.0084080883	0.0151878497
state_Missouri	state_Montana	state_Nebraska	state_Nevada	state_New.Hampshire
0.0283778236	0.0024770307	0.0064837019	0.0082699273	0.0037352834
state_New.Jersey	state_New.Mexico	state_New.York	state_North.Carolina	state_North.Dakota
0.0235071202	0.0066465346	0.0432986944	0.0367212403	0.0022549861
state_Ohio	state_Oklahoma	state_Oregon	state_Pennsylvania	state_Rhode.Island
0.0444533262	0.0146746800	0.0086400016	0.0390453070	0.0035872537
state_South.Carolina	state_South.Dakota	state_Tennessee	state_Texas	state_Utah
0.0288707306	0.0019293207	0.0326553572	0.0563894563	0.0043027307
state_Vermont	state_Virginia	state_Washington	state_West.Virginia	state_Wisconsin
0.0018947805	0.0254315066	0.0142404595	0.0068093673	0.0180744293
state_Wyoming	n_killed	n_injured	ColonneNbMalfaiteur	Adult
0.0009473902	0.2809357452	0.5537841332	0.9258124365	0.6815683256
Teen	child			
0.0576575776	0.0026102575			

Calcul des écart-types pour les quantités statistiques (2013 à 2017) :

```
> EcartType
```

state_Alabama	state_Alaska	state_Arizona	state_Arkansas	state_California
0.15242048	0.07107506	0.09749462	0.10839911	0.25506996
state_Colorado	state_Connecticut	state_Delaware	state_District.of.Columbia	state_Florida
0.11009600	0.11174433	0.08500113	0.10406456	0.23512536
state_Georgia	state_Hawaii	state_Idaho	state_Illinois	state_Indiana
0.18870782	0.03446395	0.05207143	0.26674101	0.15749159
state_Iowa	state_Kansas	state_Kentucky	state_Louisiana	state_Maine
0.09642121	0.09205027	0.13176660	0.18696715	0.05553616
state_Maryland	state_Massachusetts	state_Michigan	state_Minnesota	state_Mississippi
0.15746185	0.14675500	0.15509245	0.09130955	0.12229985
state_Missouri	state_Montana	state_Nebraska	state_Nevada	state_New.Hampshire
0.16593652	0.04970822	0.08026017	0.09056255	0.06100286
state_New.Jersey	state_New.Mexico	state_New.York	state_North.Carolina	state_North.Dakota
0.15150792	0.08125510	0.20352917	0.18807702	0.04743324
state_Ohio	state_Oklahoma	state_Oregon	state_Pennsylvania	state_Rhode.Island
0.20610055	0.12024727	0.09254941	0.19370327	0.05978631
state_South.Carolina	state_South.Dakota	state_Tennessee	state_Texas	state_Utah
0.16744357	0.04388175	0.17773334	0.23067282	0.06545409
state_Vermont	state_Virginia	state_Washington	state_West.Virginia	state_Wisconsin
0.04348792	0.15743210	0.11848096	0.08223766	0.13322099
state_Wyoming	n_killed	n_injured	ColonneNbMalfaiteur	Adult
0.03076520	0.54283427	0.75331658	0.98558799	0.87545103
Teen	child			
0.31023508	0.05245462			

Calcul des moyennes et écart-type pour 2018 (échantillon à étudier) :

```

> Moyennes2018<-colMeans(as.matrix(Echantillon2018[,~c(1,2,3,4,5,6)]))
> Moyennes2018
state_Alabama      state_Alaska      state_Arizona      state_Arkansas      state_California
0.0272383147      0.004442130      0.0118499013      0.0139071758      0.087576037
state_Colorado      state_Connecticut    state_Delaware      state_District.of.Columbia 0.0102121907
0.0157998683      0.0114384463      0.0076530612      0.0063765635
state_Georgia      state_Hawaii      state_Idaho      state_Indiana
0.0318466096      0.0013989467      0.0036208032      0.0246872943
state_Iowa      state_Kansas      state_Kentucky      state_Louisiana
0.0077353522      0.0106155365      0.0162936142      0.0348913759
state_Maryland      state_Massachusetts    state_Michigan      state_Minnesota
0.0295424621      0.0239466754      0.0277320606      0.0158821593
state_Missouri      state_Montana      state_Nevada      state_New.Hampshire
0.0285549704      0.0029624753      0.0066655695      0.0035385122
state_New.Jersey      state_New.Mexico      state_New.York      state_North.Carolina
0.0167050691      0.0064186965      0.0258393680      0.0348090849
state_Ohio      state_Oklahoma      state_Oregon      state_Pennsylvania
0.0427913101      0.0152238315      0.0093811718      0.0381007242
state_South.Carolina      state_South.Dakota      state_Tennessee      state_Texas
0.0272383147      0.0018926926      0.0290487163      0.0540651745
state_Vermont      state_Virginia      state_Washington      state_West.Virginia
0.0015635286      0.0227946017      0.0152238315      0.0062541145
state_Wyoming      n_killed      n_injured      colonnbnMalfaitteur
0.0009052008      0.2907340355      0.5078176432      0.9575378539
Teen      Child      Adult
0.0723337722      0.0038676761      0.6931369322
> EcartType2018<-apply(as.matrix(Echantillon2018[,~c(1,2,3,4,5,6)]),2,sd)
> EcartType2018
state_Alabama      state_Alaska      state_Arizona      state_Arkansas      state_California
0.16278381      0.06631564      0.10821481      0.1171061      0.28266207
state_Colorado      state_Connecticut    state_Delaware      state_District.of.Columbia 0.10010077
0.12470570      0.10634161      0.08714997      0.0063765635
state_Georgia      state_Hawaii      state_Idaho      state_Indiana
0.17559881      0.03737786      0.06006655      0.15517672
state_Iowa      state_Kansas      state_Kentucky      state_Louisiana
0.08761363      0.10248762      0.12660747      0.18351223
state_Maryland      state_Massachusetts    state_Michigan      state_Minnesota
0.16932827      0.15288936      0.16421088      0.12502481
state_Missouri      state_Montana      state_Nevada      state_New.Hampshire
0.16655890      0.05435018      0.08137373      0.09255592
state_New.Jersey      state_New.Mexico      state_New.York      state_North.Carolina
0.12816927      0.07986252      0.15866243      0.18330351
state_Ohio      state_Oklahoma      state_Oregon      state_Pennsylvania
0.20239463      0.12244713      0.09640503      0.19144732
state_South.Carolina      state_South.Dakota      state_Tennessee      state_Texas
0.16278381      0.04346568      0.16795002      0.22615557
state_Vermont      state_Virginia      state_Washington      state_West.Virginia
0.03951218      0.14925428      0.12244713      0.07883852
state_Wyoming      n_killed      n_injured      colonnbnMalfaitteur
0.03007417      0.55685231      0.69712941      0.96792820
Teen      Child      Adult
0.33497207      0.06338478      0.88200714

```

Explication de la méthode :

Nous cherchons à savoir si **les valeurs moyennes des variables** sont différentes par rapport au passé.

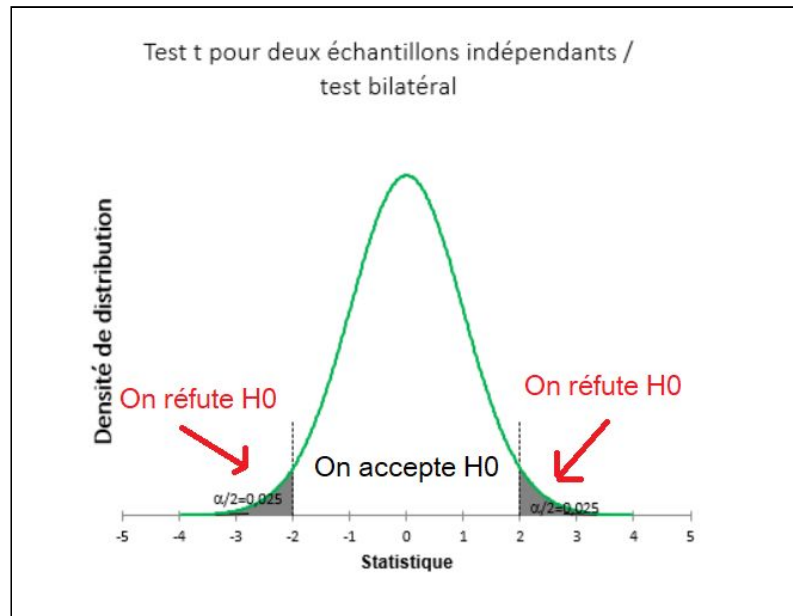
Nos quantités statistiques sont représentées par “**Mean**”, c'est-à-dire les moyennes des variables de 2013 à 2017.

Notre échantillon est représentés par “**Échantillon**”, c'est-à-dire les moyennes des variables des trois premiers mois de 2018.

Un **test de Student bilatéral** est associé à une d'hypothèse alternative selon laquelle le **signe de la différence potentielle est inconnu**. Avant de mettre en place l'expérimentation et de lancer le test, nous ne savons pas avec certitude si **Mean** serait supérieur à **Échantillon** ou le contraire dans la situation où une différence entre les deux serait mise en relief par le test. Ceci nous conduit à opter pour un **test de Student bilatéral**.

Nous souhaitons tester l'hypothèse suivante :

- ***H0 : Échantillon = Mean (les valeurs moyennes des variables sont très semblables aux années précédentes)***
- ***H1 : Échantillon ≠ Mean (les valeurs moyennes des variables sont très différentes des années précédentes)***



Dans la zone grisée : l'hypothèse H_0 est rejetée

Dans la zone blanche : l'hypothèse H_0 est acceptée

L'intervalle de confiance bilatéral de μ au niveau de confiance $1 - \alpha$ est donné par :

$$\left[\bar{X} - t_{\alpha/2}^{n-1} \frac{S}{\sqrt{n}}; \bar{X} + t_{\alpha/2}^{n-1} \frac{S}{\sqrt{n}} \right],$$

Dans notre cas, l'intervalle de confiance au niveau 95% est donc :

$$[\text{Mean} - 1,96 * \text{Ecart-type}/\sqrt{n} ; \text{Mean} + 1,96 * \text{Ecart-type}/\sqrt{n}]$$

Si **Échantillon** (la moyenne du nouvel échantillon) appartient à l'intervalle de confiance, on **accepte H_0** , l'hypothèse est vraie. C'est-à-dire, **que les valeurs moyennes des variables ne sont pas différentes des années précédentes.**

Nous effectuons nos tests :

Test sur les Moyennes

```
Paired t-test

data: x and y
t = -0.3366, df = 56, p-value = 0.7377
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.002812166  0.002003063
sample estimates:
mean of the differences
      -0.0004045515
```

Dans le résultat ci-dessus :

- **t** est la **statistique de Student** ($t = -0.3366$),
- **df** est le **degré de liberté** ($df = 56$), **p-value** est le degré de significativité du test ($p\text{-value} = 0.7377$).
- **L'intervalle de confiance de la différence des moyennes à 95%** est également montrée [-0,002812 ; 0.002003] (son calcul
- et enfin, on a la **valeur moyenne de la différence des deux séries de moyennes** (-0.0004045515).

La p-value du test est de 0.7377. Ce qui est **largement supérieur à 0.05**. On conclut que l'hypothèse H_0 n'est pas rejetée, nous l'acceptons et pouvons alors dire que **les valeurs moyennes des variables de 2013 à 2017 ne sont pas significativement différentes de celles de 2018**.

Nous avons décidé d'apporter un petit plus à cette étude en effectuant aussi une observation **des différences entre les écarts-types également**, ainsi que sur **les valeurs directement, par curiosité :**

Test sur les écart types moyens de chaque variable :

```
> z<-EcartType
> # Ecart type de 2018
> w<-EcartType2018
> #res le test statistiques de student
> res<-t.test(z, w, paired=TRUE) #paired=TRUE car les colonnes désigne les mêmes variables
> res

Paired t-test

data: z and w
t = -0.09341, df = 56, p-value = 0.9259
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.003993164  0.003637357
sample estimates:
mean of the differences
      -0.0001779036
```

La p-value du test est de 0.9259. Ce qui est **largement supérieur à 0.05**. On conclut que **les écarts-types moyens des variables de 2013 à 2017 ne sont pas significativement différents de ceux de 2018**.

Test sur les Valeurs

Nous avons choisi d'observer également **la différence colonne par colonne entre les valeurs** (on compare la colonne state Alabama de 2013 à 2017 avec celle de state Alabama de 2018 par exemple).

Nous créons alors une fonction permettant de réaliser des tests de Student sur les colonnes de deux tables (les colonnes retirées sont les indices, la latitude etc...).

Nous devons y indiquer que les échantillons que nous allons comparer ne comportent pas des valeurs appariées, grâce au **test t de Student indépendant (ou non apparié)**.

Dans ce cas de figure, il s'agit de **comparer deux moyennes observées**, lorsque les deux groupes d'échantillons (A et B) à comparer n'ont aucun lien.

Par exemple, dans le cas précédent, nous comparions **la moyenne de ColonneNbMalfaiteurs de 2013-2017 à la moyenne ColonneNbMalfaiteurs de 2018**, or dans ce cas, nous allons comparer **chacune de ses colonnes** et non pas leur moyenne.

La différence de résultat que nous obtiendrons relève alors de **l'écart-type** qui est calculé **pour chaque colonne par rapport à leur moyenne respective**. Il est donc possible que nous obtenions des résultats différents que précédemment :

```
##### TEST par colonnes #####
x<-Statistiques2013_2017[,~c(1,2,3,4,5,6)]
y<-Echantillon2018[,~c(1,2,3,4,5,6)]

TestStudent<-function(x,y){
  MatricePvalue<-c()
  pvalue=0
  for (i in 1:ncol(x)){
    Test<-t.test(x[,i], y[,i], paired=FALSE) #paired=FALSE car ici nous avons des échantillons de taille différentes
    pvalue=Test$p.value
    MatricePvalue<-append(MatricePvalue,as.numeric(pvalue))
  }
  variable<-colnames(x)
  MatricePvalue<-cbind(variable,MatricePvalue)
}
TestStudent(x,y)
as.numeric(MatricePvalue[,2])
```

Cette fonction stocke les p-values de chaque variables dans une nouvelle matrice, et nous obtenons alors :

```
> MatricePvalue
  variable MatricePvalue
[1,] "state_Alabama"      "0.0231825967245189"
[2,] "state_Alaska"      "0.309630782691736"
[3,] "state_Arizona"     "0.025053199200921"
[4,] "state_Arkansas"    "0.0643044800506454"
[5,] "state_California"  "2.11506905785508e-11"
[6,] "state_Colorado"    "0.00230510221216756"
[7,] "state_Connecticut" "0.225164553583335"
[8,] "state_Delaware"    "0.644605265136437"
[9,] "state_District.of.Columbia" "0.37720174156904"
[10,] "state_Florida"     "0.106226703058351"
[11,] "state_Georgia"    "0.00184206783063327"
[12,] "state_Hawaii"     "0.546197133360544"
[13,] "state_Idaho"      "0.105410584395408"
[14,] "state_Illinois"   "6.02156669132628e-07"
[15,] "state_Indiana"    "0.598423347436295"
[16,] "state_Iowa"       "0.0450674200018533"
[17,] "state_Kansas"     "0.0297390910471805"
[18,] "state_Kentucky"   "0.243919229436418"
[19,] "state_Kentucky"   "0.243919229436418"
[20,] "state_Kentucky"   "0.243919229436418"
```

Or, ce qui nous intéresse, c'est les changements significatifs, on peut alors récupérer les valeurs inférieures à 0,05 :

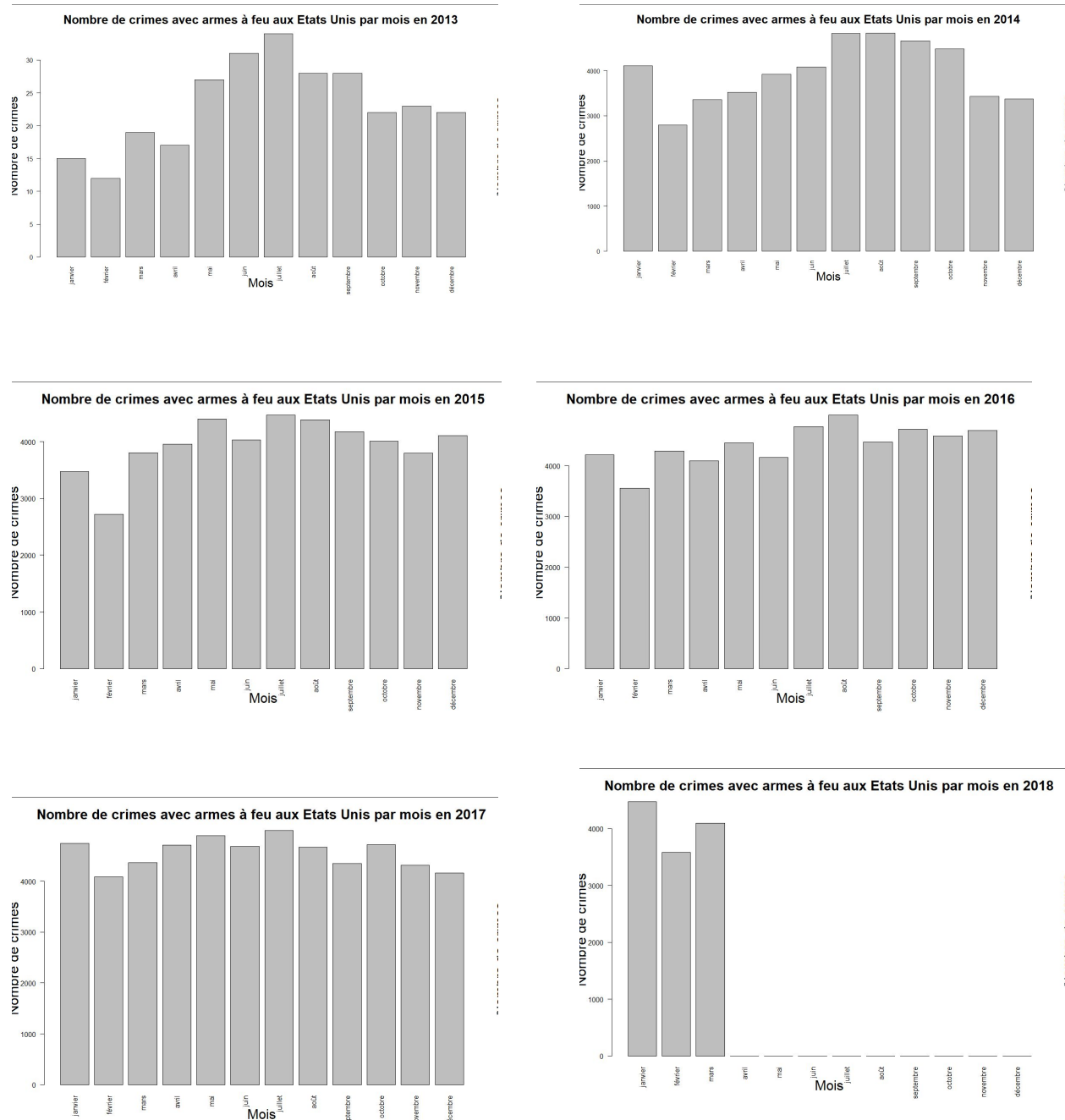
```
> as.numeric(MatricePvalue[,2])
[1] 2.318260e-02 3.096308e-01 2.505320e-02 6.430448e-02 2.115069e-11 2.305102e-03 2.251646e-01 6.446053e-01 3.772017e-01 1.062267e-01
[11] 1.842068e-03 5.461971e-01 1.054106e-01 6.021567e-07 5.984233e-01 4.506742e-02 2.973909e-02 2.439192e-01 4.209458e-01 9.240884e-01
[21] 9.244583e-03 1.767211e-01 4.465065e-02 9.070600e-11 4.740982e-06 8.889602e-01 3.366673e-01 8.107308e-01 6.677310e-01 7.231610e-01
[31] 1.949468e-08 7.602474e-01 7.694029e-31 2.647499e-01 9.399735e-01 3.797759e-01 6.307485e-01 4.093774e-01 5.975547e-01 8.380632e-01
[41] 2.837484e-01 9.281453e-01 2.194523e-02 2.717128e-01 5.480855e-01 3.722320e-01 5.935914e-02 3.890031e-01 4.519246e-01 2.513039e-02
[51] 8.807594e-01 5.922334e-02 2.211846e-12 4.566374e-04 1.600504e-01 2.499700e-06 3.210851e-02
> which(as.numeric(MatricePvalue[,2])<0.05)
[1] 1 3 5 6 11 14 16 17 21 23 24 25 31 33 43 50 53 54 56 57
```

```
> MatricePvalue[Indice,]
  variable MatricePvalue
[1,] "state_Alabama"      "0.0231825967245189"
[2,] "state_Arizona"     "0.025053199200921"
[3,] "state_California"  "2.11506905785508e-11"
[4,] "state_Colorado"    "0.00230510221216756"
[5,] "state_Georgia"    "0.00184206783063327"
[6,] "state_Illinois"   "6.02156669132628e-07"
[7,] "state_Iowa"       "0.0450674200018533"
[8,] "state_Kansas"     "0.0297390910471805"
[9,] "state_Maryland"    "0.00924458346138175"
[10,] "state_Michigan"   "0.0446506455827089"
[11,] "state_Minnesota"  "9.07060022246771e-11"
[12,] "state_Mississippi" "4.74098192292744e-06"
[13,] "state_New.Jersey" "1.94946752976689e-08"
[14,] "state_New.York"   "7.69402946928034e-31"
[15,] "state_Tennessee" "0.0219452279603087"
[16,] "state_Wisconsin"  "0.0251303907346181"
[17,] "n_injured"       "2.21184581569038e-12"
[18,] "ColonnesNbMalfaiteur" "0.000456637434173685"
[19,] "Teen"            "2.49969978891074e-06"
[20,] "child"           "0.0321085096842281"
```

Il semblerait que **20 variables** aient changé de manière significative. Cela signifie que malgré que de manière globale les variables n'aient pas changé de manière significative (leurs moyennes et écart-types restent les mêmes), mais de manière ciblée, il semblerait qu'en 2018, leur répartition ne soit plus vraiment semblable à la période de 2013 à 2017.

3. Si l'on prend en compte aussi le mois de l'année dans lequel le fait divers a été commis, est-ce qu'il y a une forte corrélation entre le nombre de faits divers et le mois de l'année ? Quelles conclusions en tirez-vous ?

Tout d'abord, représentons par mois, pour chaque année, le nombre de crimes :



Par simple observation, nous pouvons noter que **certains mois de l'année comportent plus de faits-divers que d'autres**, notamment en milieu d'année (même si le mois de janvier semble faire exception).

Visuellement, on peut donc voir qu'il y a une corrélation entre la période de l'année et le nombre de faits-divers. Certains mois ont plus de crimes que d'autres.

Cependant nous devons également nous demander, **est-ce qu'une augmentation de crime sur l'année (donc générale), implique une augmentation des crimes en milieu d'année ou en janvier ? Autrement dit, une augmentation du nombre de faits-divers sur une année est-elle dû aux mois de milieu d'année ou à janvier nécessairement?**


Pour répondre à cette question, nous devons calculer **le nombre de crime par années**, afin **d'observer le nombre de crime par mois**, et **le nombre de crime cette année-là**.

Car nous pouvons en effet avoir le mois de janvier qui comporte toujours plus de crime que les autres (à cause du nouvel an et du manque de la baisse du pouvoir d'achat "après-fête" par exemple), mais avoir une année où il y a eut des attentats un certains mois (autre que janvier), et donc le mois de janvier ne serait pas lié à l'augmentation du nombre de crime cette année-là.

Ainsi, nous pourrions mettre en avant les corrélations linéaires de ces valeurs s'il y en a réellement, voici les données de nombre de faits-divers auxquelles nous avons ajouté les sommes :

```
> DataByMounth<-cbind(DataByMounth,somme)
> DataByMounth
```

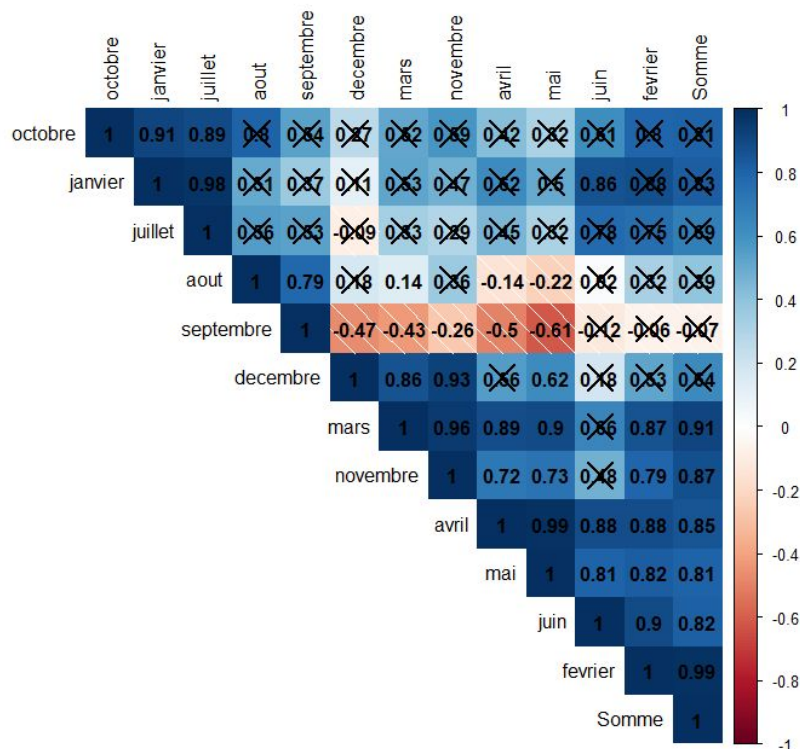
	Annees	janvier	fevrier	mars	avril	mai	juin	juillet	août	septembre	octobre	novembre	decembre	Somme
1	2014	4112	2793	3358	3524	3920	4085	4826	4834	4668	4486	3429	3372	47407
2	2015	3474	2716	3807	3958	4402	4030	4466	4387	4170	4011	3795	4108	47324
3	2016	4211	3558	4289	4090	4445	4158	4768	4994	4466	4715	4583	4696	52973
4	2017	4740	4085	4364	4710	4897	4686	4993	4666	4348	4720	4311	4160	54680
5	2018	4474	3583	4095	0	0	0	0	0	0	0	0	0	12152



Nous retirons l'année 2018 pour l'étude des corrélations, **car elle ne présente que 3 mois**. L'algorithme pourrait prendre les zéro comme des valeurs à prendre en compte.

De plus, nous avons choisi de ne pas prédire les données après mars 2018 car cela **pourrait influencer nos résultats** et mettre en évidence de fausses corrélations.

Observons dès à présent les corrélations entre la Somme et les mois :



Cette matrice de corrélation met en évidence les corrélations suivantes :

- **Forte Corrélation positive**, entre le nombre de crimes et le mois de février (corrélacion significative car non-barré), le coefficient de corrélation est de 0.99)
- **Forte Corrélation positive**, entre le nombre de crimes et le mois de mars (corrélacion significative car non-barré), le coefficient de corrélation est de 0.91)
- **Forte Corrélation positive**, entre le nombre de crimes et le mois d'avril (corrélacion significative car non-barré), le coefficient de corrélation est de 0.85)
- **Forte Corrélation positive**, entre le nombre de crimes et le mois de mai (corrélacion significative car non-barré), le coefficient de corrélation est de 0.81)
- **Forte Corrélation positive**, entre le nombre de crimes et le mois de juin (corrélacion significative car non-barré), le coefficient de corrélation est de 0.82)
- **Forte Corrélation positive**, entre le nombre de crimes et le mois de novembre (corrélacion significative car non-barré), le coefficient de corrélation est de 0.87)

D'une manière globale, lorsque les crimes ont augmenté chaque année (entre 2014 et 2017), le nombre de crimes en février, mars, avril, mai, juin et novembre a augmenté également. Nous avons vu visuellement que certains mois semblaient comporter plus de faits divers que d'autres (janvier, et alentour de milieu d'année). Or, il semblerait qu'une augmentation globale (c'est-à-dire sur une année) soit **plutôt due à une augmentation dans les autres mois** (cités précédemment). Cela peut s'expliquer par des événements spécifiques, une crise économique, des faits sociaux etc...

IV. Conclusion

Après avoir étudié les données des crimes commis par arme à feu aux Etats-Unis de 2013 à 2018, nous avons **observé des corrélations positives entre l'âge des malfaiteurs et le nombre de malfaiteurs**. Cette corrélation signifie **que plus il y a de malfaiteurs, plus la moyenne d'âge des malfaiteurs est élevée** (dans la catégorie Adulte 18 et plus). Cependant, nous avons également noté que même s'il y a une majorité d'adultes commettant des crimes, il y a une partie des crimes commis par **les "teen", c'est-à-dire les adolescents entre 12 et 17 ans**.

De plus, **une légère corrélation négative apparaît entre le nombre de malfaiteurs et le nombre de personnes blessées**. Elle est faible mais significative d'après les tests. Cela peut être dû à **une variable extérieure**, mais nous avons émis l'hypothèse des **comportements sociaux** : plus il y a de malfaiteurs, moins ils ont besoin de blesser des personnes. Car nous pouvons supposer que les personnes soient plus coopératives face à un nombre de malfaiteurs élevés, et que les malfaiteurs se sentant plus en confiance aient donc moins besoin de blesser des personnes.

Par ailleurs, nous avons observé que d'une manière globale, **les moyennes et écart-types des variables n'ont pas significativement changé en 2018** (de janvier à mars), lorsque l'on compare à la période 2013-2017. Nous avons réalisé **des tests bilatéraux de Student appariés** pour constater cela.

Or, dans un second temps, nous avons réalisé des **tests bilatéraux de Student non appariés**. Cela nous a conduit à une vingtaine de variables dont nous savons qu'elles ont changé significativement lorsqu'on observe leur répartition une à une.

L'écart-type moyen calculé lors du test appariés est peut-être le même que dans le passé, mais **les valeurs qui composent cet écart-type** sont soit devenues **plus extrêmes**, soit devenues **moins extrêmes**. La répartition des écart-types de ces 20 variables a changé.

Enfin, nous avons constaté que lorsque nous considérons les mois de l'année, il y avait clairement **un maximum atteint chaque milieu d'année** (en juillet principalement et parfois août), **et chaque mois de janvier** (fin-début d'année).

Cependant, nous avons constaté que ces valeurs **sont plutôt stables** lorsque l'on compare aux augmentations **de l'année totale**. En effectuant des **tests de corrélations** entre **le nombre de crimes dans l'année et le nombre de crimes dans le mois**, nous avons conclu qu'une **augmentation des crimes dans l'année** a plutôt tendance à impliquer **une augmentation des crimes les autres mois que ceux des extrêmes de saison**. Cela peut signifier qu'une augmentation des crimes, effective une année, impliquerait de "nouveaux crimes" que l'on pourrait qualifier d'inhabituels, car n'étant pas stable d'une année à une autre. Les variations ont tendance à se trouver dans les périodes mi-saisons, et probablement dû à des faits économiques ou sociaux.

V. Présentation des membres du groupes

GARNIER CORENTIN s'est occupé de la partie concernant les corrélations entre crime et mois de l'année, et de l'élaboration des conclusions générales.

BOUCHE STEVEN s'est occupé de la partie concernant les corrélations entre les variables de V , et de l'élaboration de la carte en WEB, qui reconduit notre projet sur une future étude.

BELKARFA LINA s'est occupée de la partie concernant les différences significatives dans le temps, et du pré-traitement à apporter aux données.

Le travail de rédaction a été commun, tout comme la réalisation et l'entraînement pour la soutenance orale.