Sorbonne université

Rapport

Analyse et prédiction de criminalité aux USA

Sidhoum Imad

Merrouche Aymen

Juin 2020

Contents

Chapter 1

INTRODUCTION

Bien que les chiffres montrent que les taux de criminalité diminuent régulièrement depuis quelques années, les sondages d'opinion ont révélé que la criminalité aux états unis est un sujet qui préoccupe fortement la population. L'inquiétude grandissante suscitée par le crime a poussé le gouvernement américain a multiplié les financements pour des programmes destinés à palier à ce problème. Dans ce projet, nous essayons de comprendre la relation entre la criminalité aux états-unis et différents facteurs sociaux et économiques. Nous essayerons aussi d'utiliser l'analyse de médias sociaux pour comprendre et prédire la criminalité. Dans cette optique-là, nous nous sommes posé les questions suivantes :

- Comment la criminalité aux états-unis évolue-t-elle à travers le temps ?
- Quels sont les facteurs influant sur la criminalité aux états-unis ?
  - Quel est l'impact du niveau d'éducation des habitants sur la criminalité ?
  - Quel est l'impact du taux de pauvreté sur la criminalité ?
  - Les états les plus diversifiés sur le plan ethnique sont-ils moins stables ?
- Peut-on prédire le taux de criminalité dans un état à partir de ces facteurs ?
- Les indicateurs sur le crime mesurés dans un état durant les années prétendantes sont-ils proportionnels au nombre de tweets parlant de crimes émis à partir de ce même état ?
- Peut-on dire d'un état s'il est sure ou non en se basant sur le nombre de tweets parlant de criminalité émis à partir de ce dernier ?
- Comment l'analyse spatio-temporelle des médias sociaux peut-elle aider à identifier les tendances de criminalité ainsi que les événements qui l'influencent ?
- Peut-on prédire quelle catégorie de crime est la plus probable pour une date et une ville donnée des états unis (non traitée)?

Chapter 2

Données :

Pour construire notre datawarehouse nous nous sommes basés sur plusieurs sources hétérogènes. Nous avons récolté plusieurs indicateurs de criminalité sur plusieurs années (2001 à 2016) : nombre de prisonniers, nombre de crimes par types, etc. Nous avons aussi récolté sur plusieurs années (2009 à 2016) des données sur des facteurs sociaux et économiques : pauvreté, éducation et diversité ethnique. Pour l'analyse de médias sociaux, nous avons récolté des tweets : la méthode est indiquée ci-dessous. Nous avons utilisé aussi d'autres sources de données complémentaires.

Récolte de tweets : L'API officielle de Twitter ayant l'inconvénient d'être limitée dans le temps, nous ne pouvions pas récupérer de tweets datant de plus d'une semaine. Pour contourner ce problème, nous avons utilisé l'API GetOldTweets3 qui nous permet de collecter les anciens tweets avec des critères de localisation et sur le contenu des tweets. Alors, nous avons collectés des tweets parlants de crime pour chaque état comme suit :

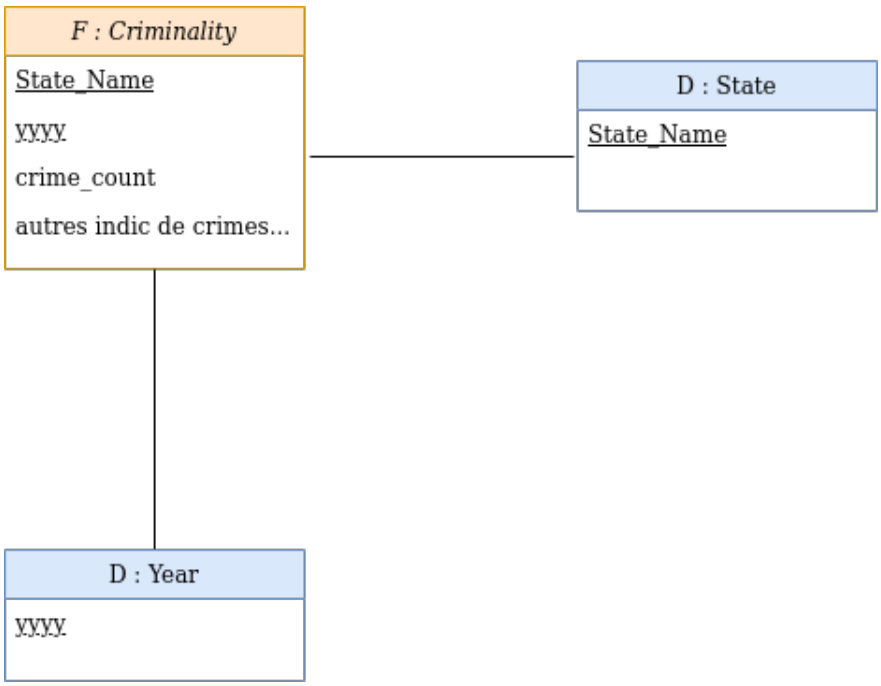
- Récupérer des tweets sur une durée définie.
- Nettoyer les tweets.
- On utilise ensuite un dictionnaire qui contient le vocabulaire relié au crime pour filtrer et détecter les tweets parlants de crimes.
- Organiser les tweets par état en utilisant les informations de géo-localisation.
- On compte ensuite le nombre de tweets parlants de crimes pour chaque état.

Chapter 3

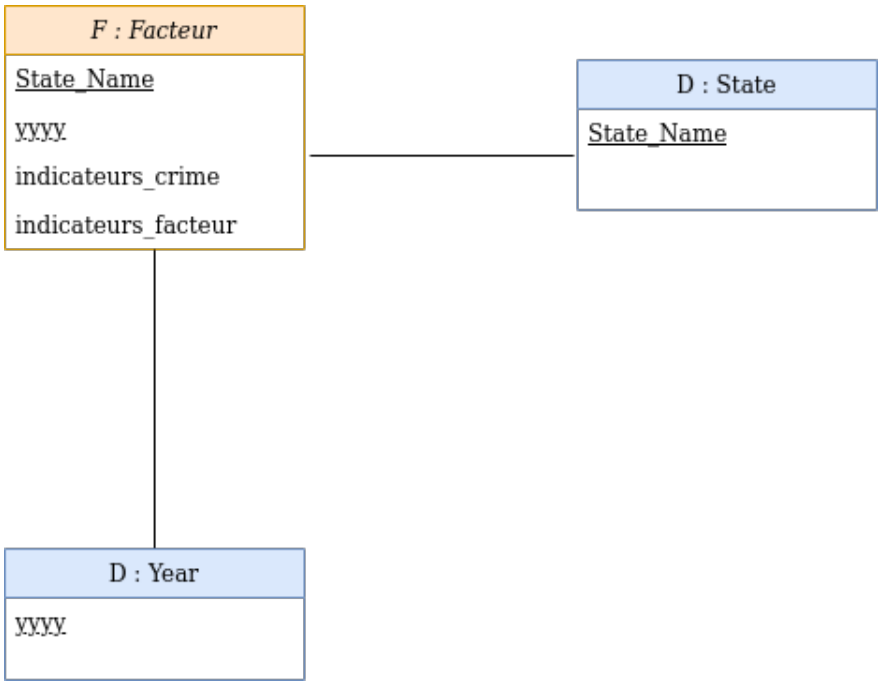
Modélisation :

Après l'extraction et la collecte des données, nous avons défini les faits et les dimensions suivantes :

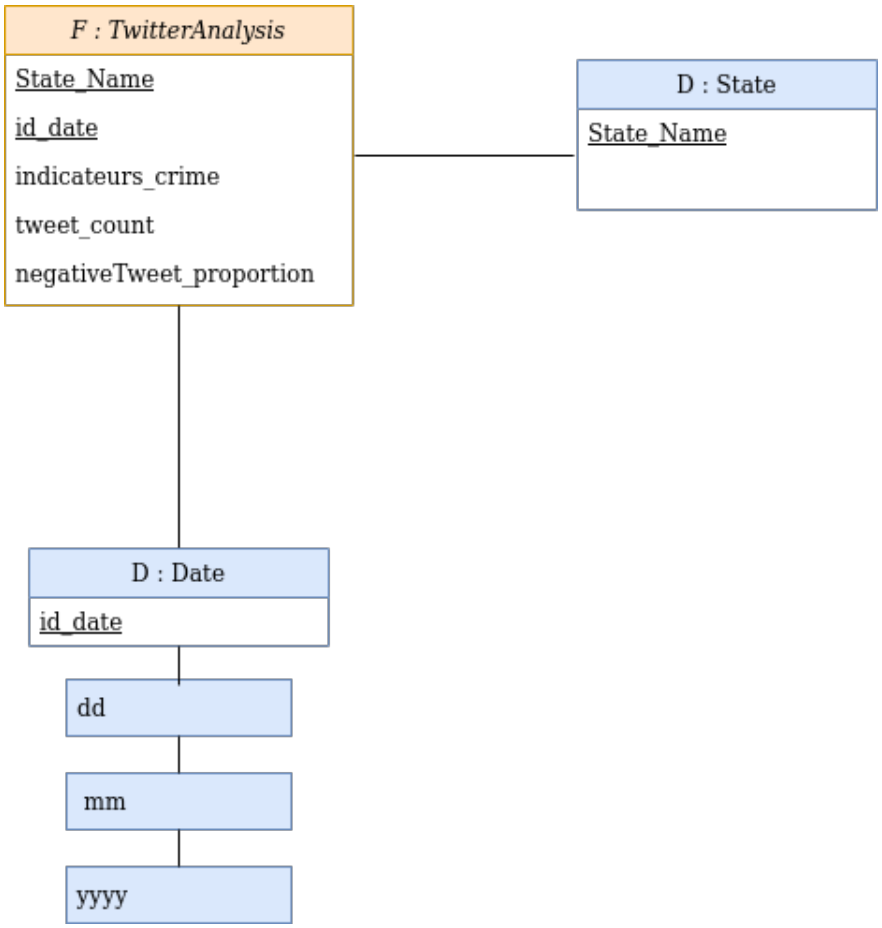
- On veut analyser la criminalité par année et par état. Le schéma en étoile ci-dessous correspond au fait Criminality. On analyse donc différents indicateurs de criminalité (nombre de crimes, nombre de crimes violents...) qui sont donc le sujet de l'analyse et par conséquent les mesures du fait, en fonction de l'état et de l'année à laquelle ils correspondent ce qui nous donne les dimensions. L'état est représenté par son nom et la date par l'année elle-même.



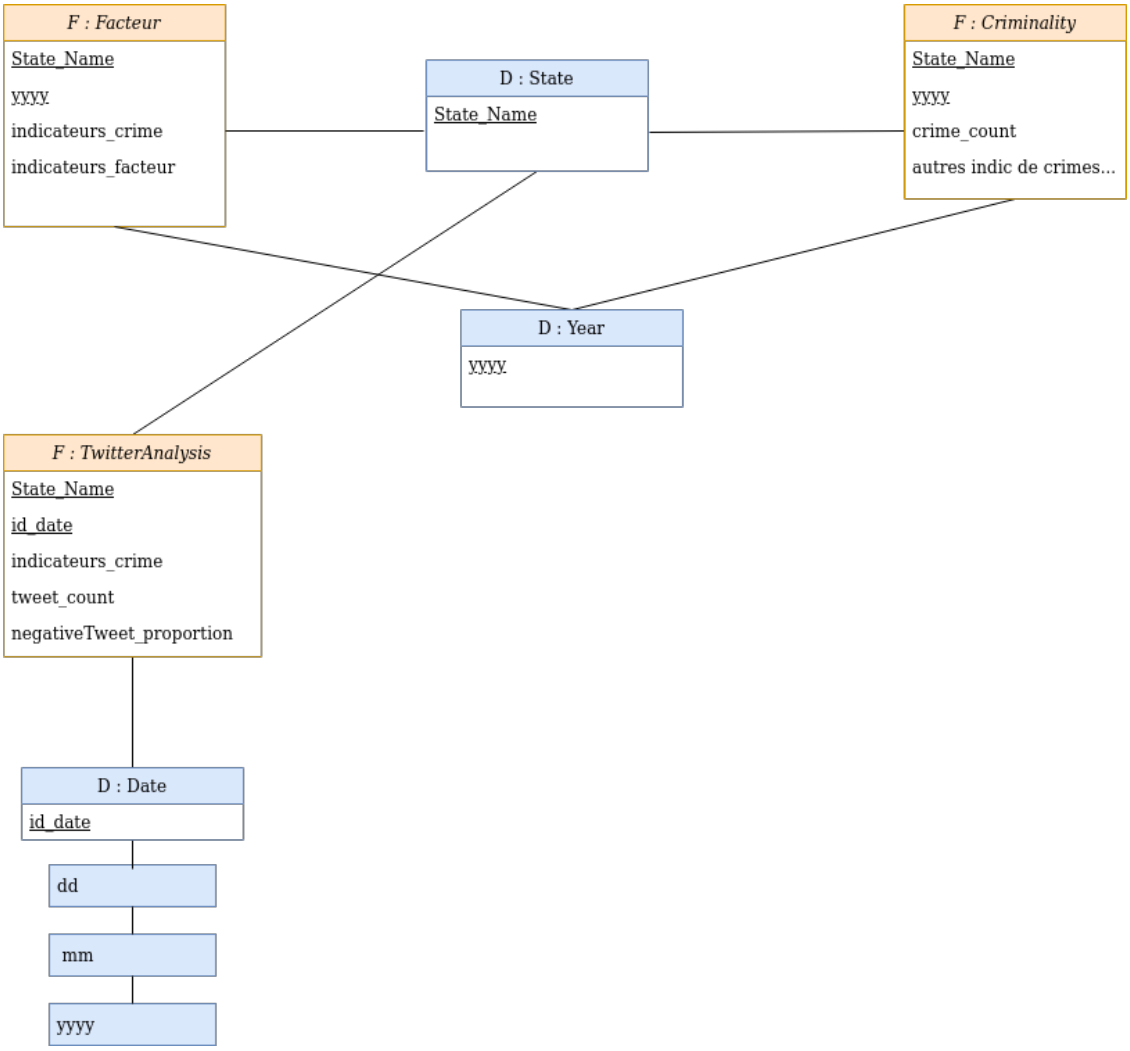
- On veut analyser la criminalité avec un facteur quelconque (la pauvreté, le taux de chômage, l'éducation ...) par année et par état. Le schéma en étoile ci-dessous correspond au fait Facteur. On analyse donc un (des) indicateur de criminalité avec l'indicateur du facteur considéré (taux de pauvreté par exemple) qui sont donc le sujet de l'analyse et par conséquent les mesures du fait, en fonction de l'état et de l'année à laquelle ils correspondent ce qui nous donne les dimensions. L'état est représenté par son nom et la date par l'année elle-même.



- On veut analyser la criminalité avec des statistiques qu'on a calculé sur les tweets par date et par état. Le schéma en étoile ci-dessous correspond au fait TwitterAnalysis. On analyse donc un (des) indicateur de criminalité avec les statistiques calculées sur les tweets (tweet count et negativeTweet proportion) qui sont donc le sujet de l'analyse et par conséquent les mesures du fait, en fonction de l'état et de la date à laquelle ils correspondent ce qui nous donne les dimensions. L'état est représenté par son nom et la date par un identifiant unique (la date complète).



- On obtient donc le schéma en constellation suivant :



On obtient le schéma relationnel normalisé suivant :

- $Facteur(\underline{State\_name, yyyy}, indicateurs\_crime, indicateurs\_facteur)$
- $Criminality(\underline{State\_name, yyyy}, crime\_count, autre\_indicateurs\_crime)$
- $TwitterAnalysis(\underline{State\_name, id\_date}, indeicateurs\_crime, tweet\_count, negativeTweet\_proportion)$
- $State(\underline{State\_name})$
- $Year(\underline{yyyy})$
- $DateJour(\underline{id\_date}, dd, \#mm)$
- $DateMois(\underline{mm}, \#yyyy)$
- $DateAnnee(\underline{yyyy})$

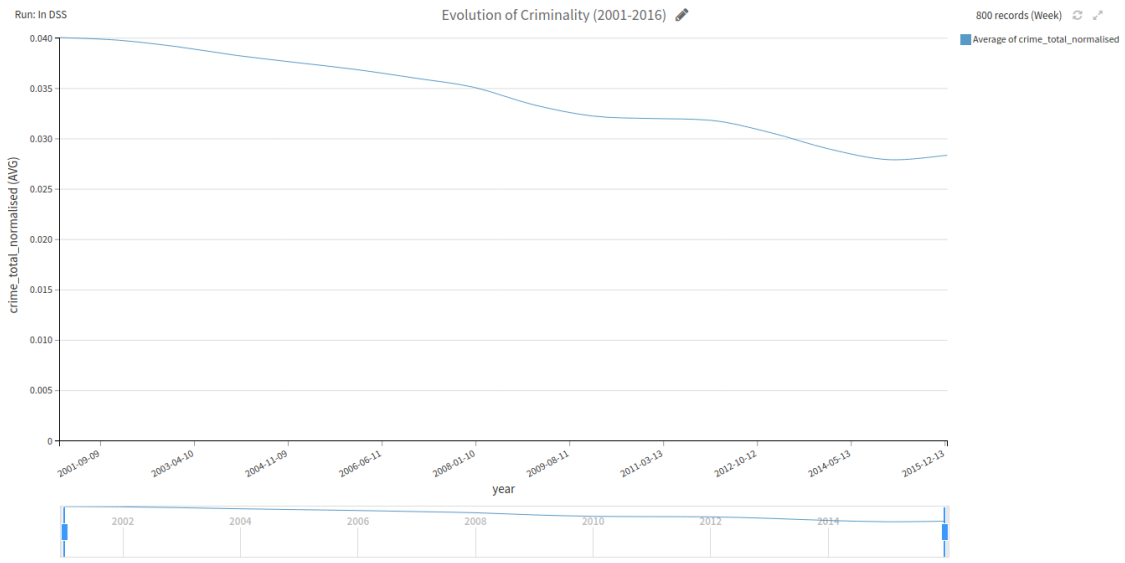
Chapter 4

Résultats et Visualisations :

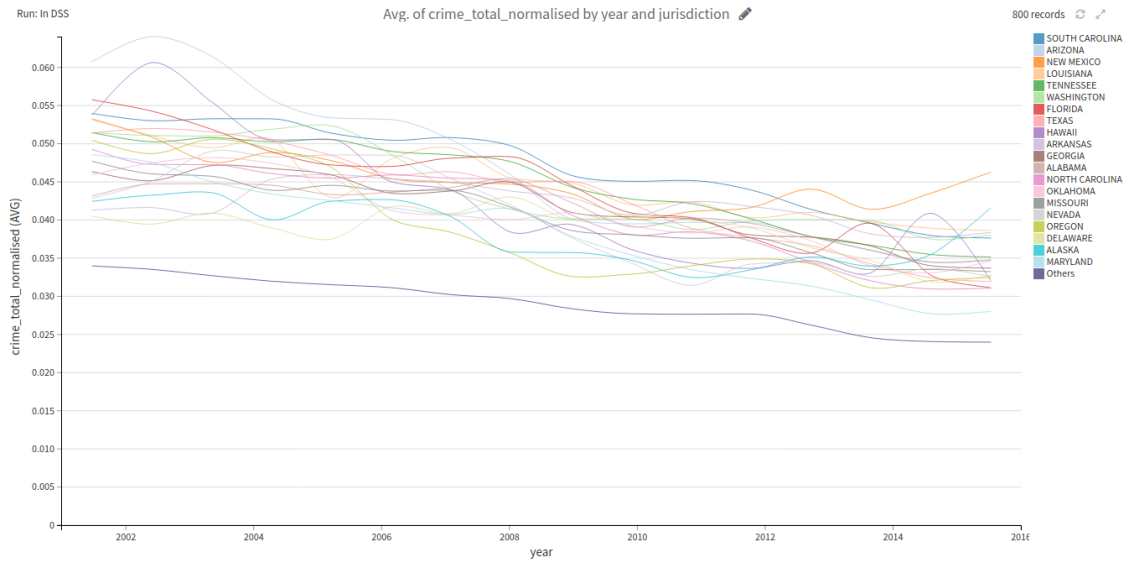
4.1 Analyse et prédiction de la criminalité aux USA :

4.1.1 Comment la criminalité aux états unis évolue-t-elle à travers le temps ?

Évolution du nombre de crimes aux états unis de 2001 à 2016



Évolution du nombre de crimes aux états unis par état de 2001 à 2016

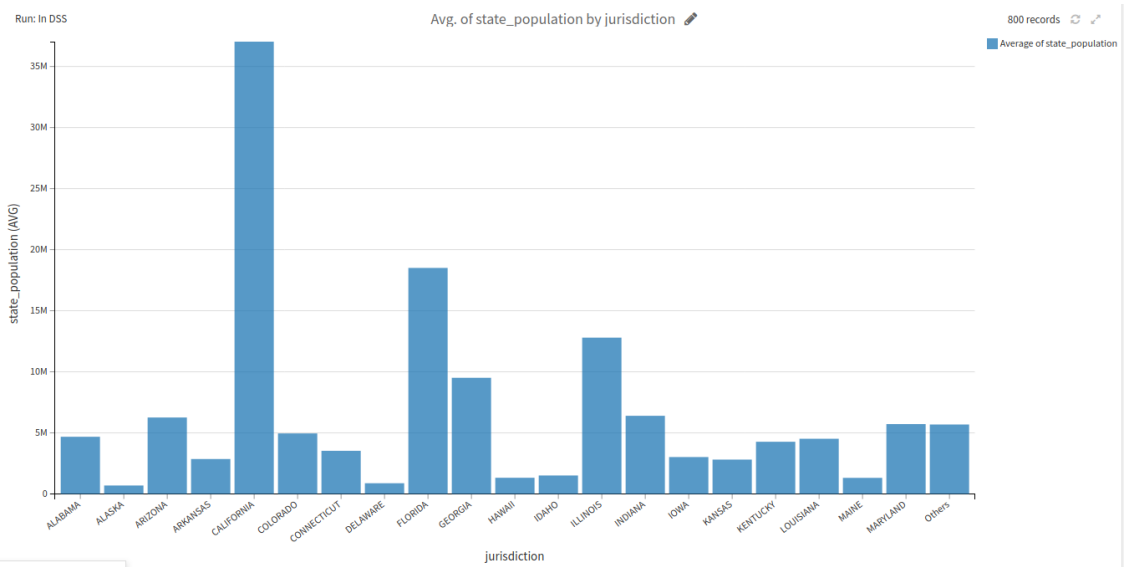


Les figures représentent l'évolution du nombre de crimes aux états unis de 2001 à 2016, sur tout le territoire sur la première figure et par état sur la deuxième. Il est clair que la criminalité baisse régulièrement depuis plusieurs années aux états unis, et ce, dans la plupart des états. Il existe cependant quelques exceptions par exemple l'état du nouveau mexique qui enregistre une hausse du crime depuis 2013.

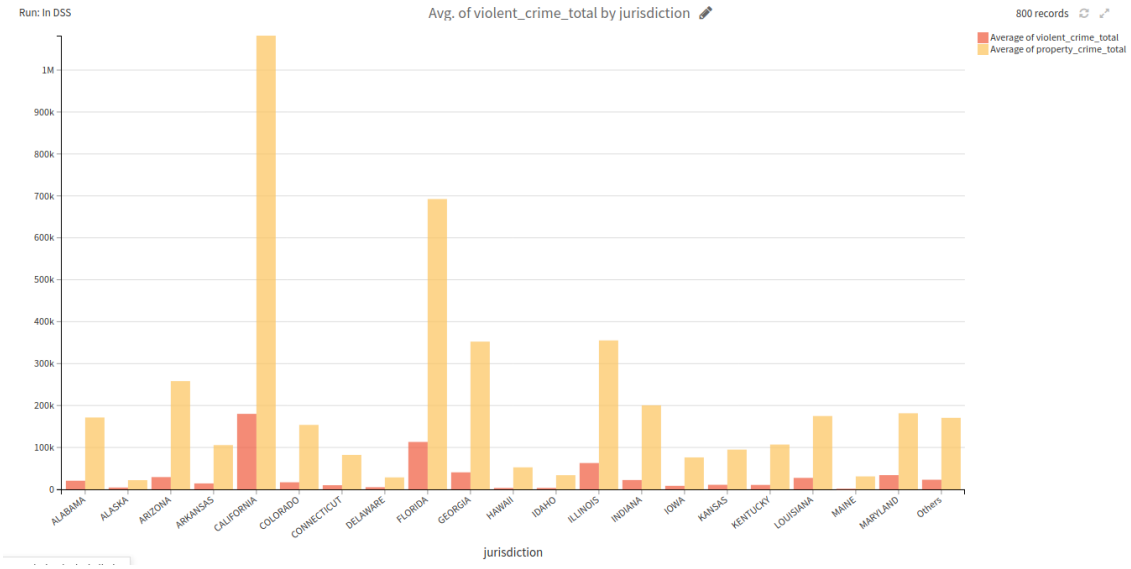
4.1.2 Quels sont les facteurs influant sur la criminalité aux états unis ?

Dans cette partie, nous essayons d'identifier quelques facteurs socio-économiques qui influent sur la criminalité. Nous analyserons la relation de la criminalité avec trois facteurs : le niveau d'éducation, la pauvreté et la diversité ethnique, mais avant cela, on va s'intéresser à la relation entre la criminalité et la population :

Population par état



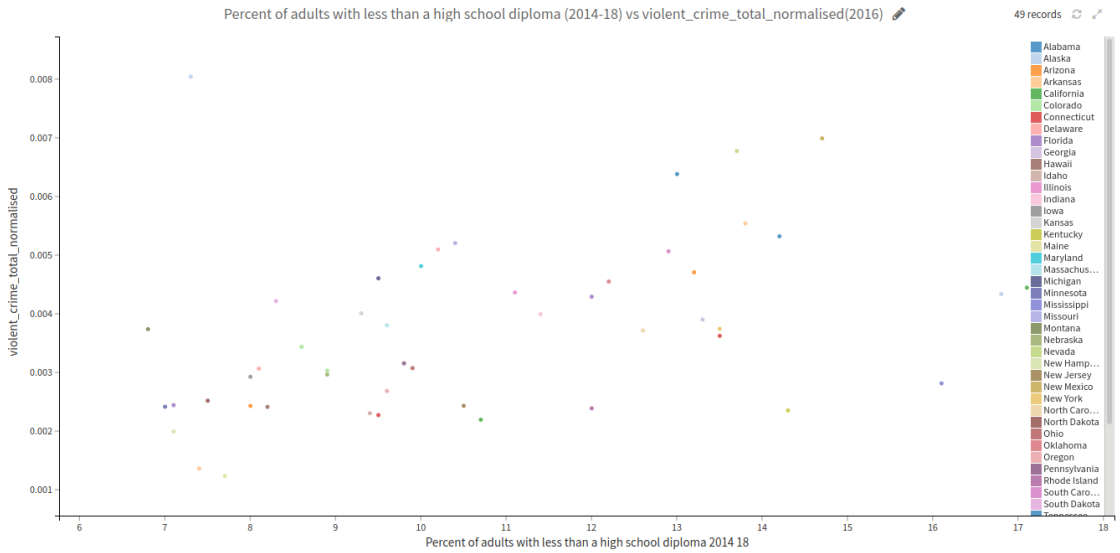
Nombre de crimes violents/de propriété par état



La première figure représente la population de chaque état. La deuxième figure représente le nombre de crimes de propriété (vol, casse ...) ainsi que le nombre de crimes violents (meurtre, viol ...) par état. Il est naturel que plus la population est importante plus le nombre de crimes est important, mais pas toujours. Si on prend par exemple les deux états "Iowa" et "Kansas", bien que l'Iowa a une population plus importante que celle du Kansas, le nombre de crimes violents ainsi que le nombre de crimes de propriété y sont moins important. Ceci prouve qu'il y a réellement des états plus dangereux que d'autres ce qui met en exergue l'existence de facteurs qui influent sur la criminalité.
Pour éliminer tout biais confondant pouvant être induit par la population, nous avons normalisé tous les indicateurs lors des analyses suivantes.

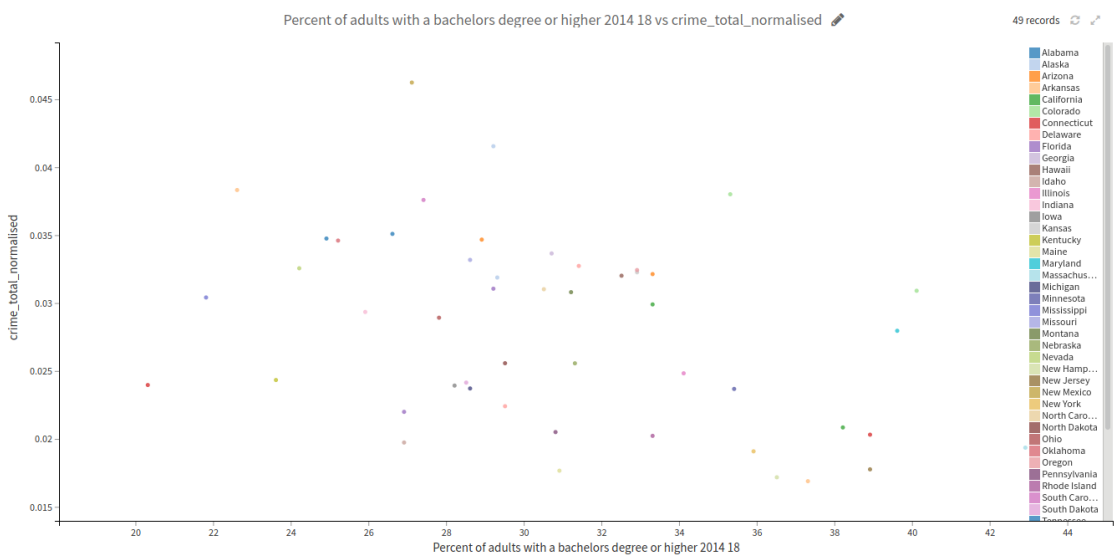
Quel est l'impact du niveau d'éducation des habitants sur la criminalité ?

Criminalité en fonction du pourcentage de population sans baccalauréat



La figure représente le nombre de crimes violents commis en 2016 en fonction du pourcentage de la population sans diplôme du baccalauréat (d'un niveau lycée ou inférieur) durant la même période. On remarque que dans la plupart des cas, plus la proportion de population sans baccalauréat augmente plus le nombre de crimes violents croit.

Criminalité en fonction du pourcentage de population avec licence ou plus



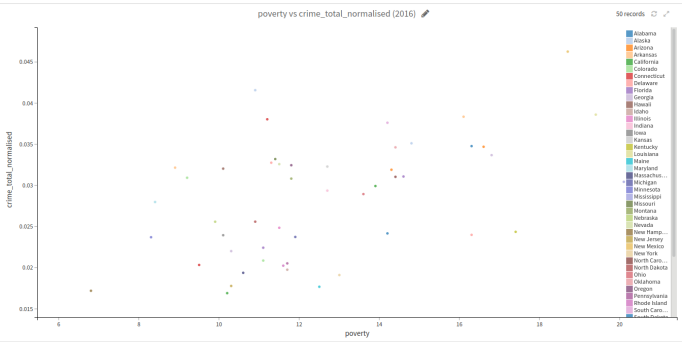
La figure représente le nombre de crimes violents commis en 2016 en fonction du pourcentage de la population avec un niveau licence ou supérieur durant la même période. On remarque que dans la plupart

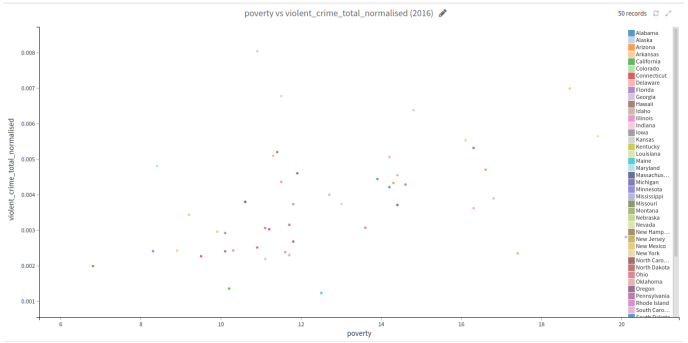
des cas et à l'inverse de la figure précédente plus la proportion de population avec un niveau licence ou supérieur augmente plus le nombre de crimes violents décroît.

Le niveau d'éducation de la population à un impact évident sur la criminalité.

Quel est l'impact du taux de pauvreté sur la criminalité ?

Nombre de crimes/crimes violents en fonction du taux de pauvreté

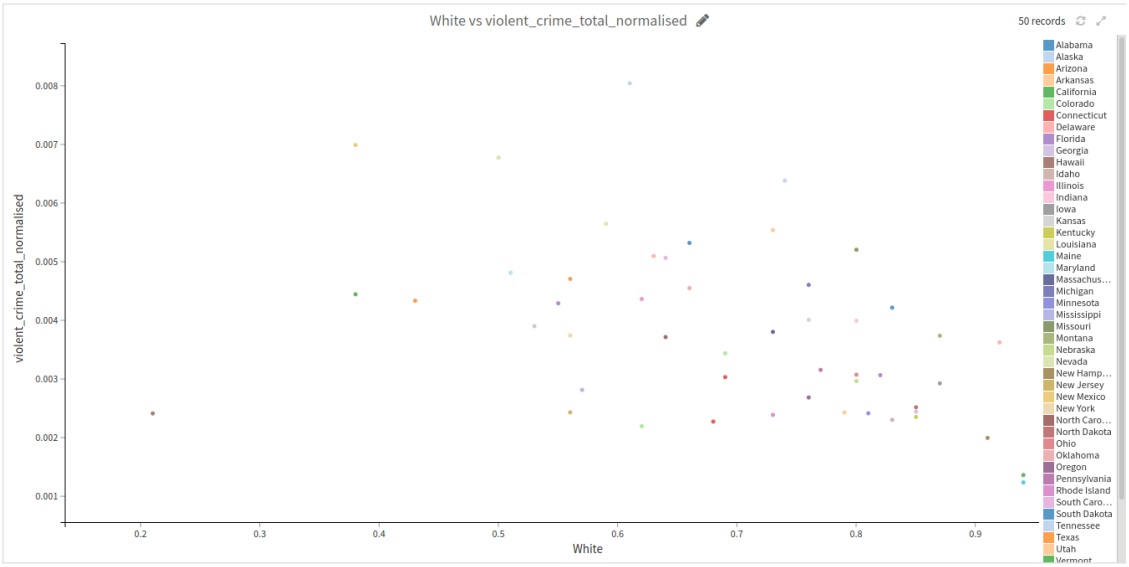




La figure de droite représente le nombre de crimes commis en 2016 en fonction des taux de pauvreté de la même année, on remarque que plus le taux de pauvreté augmente plus le nombre de crimes augmente. Cette corrélation est d'autant plus significative sur la figure de gauche qui représente le nombre de crimes **violents** commis en fonction du taux de pauvreté. **On peut donc conclure que la pauvreté a un impact significatif sur la criminalité**.

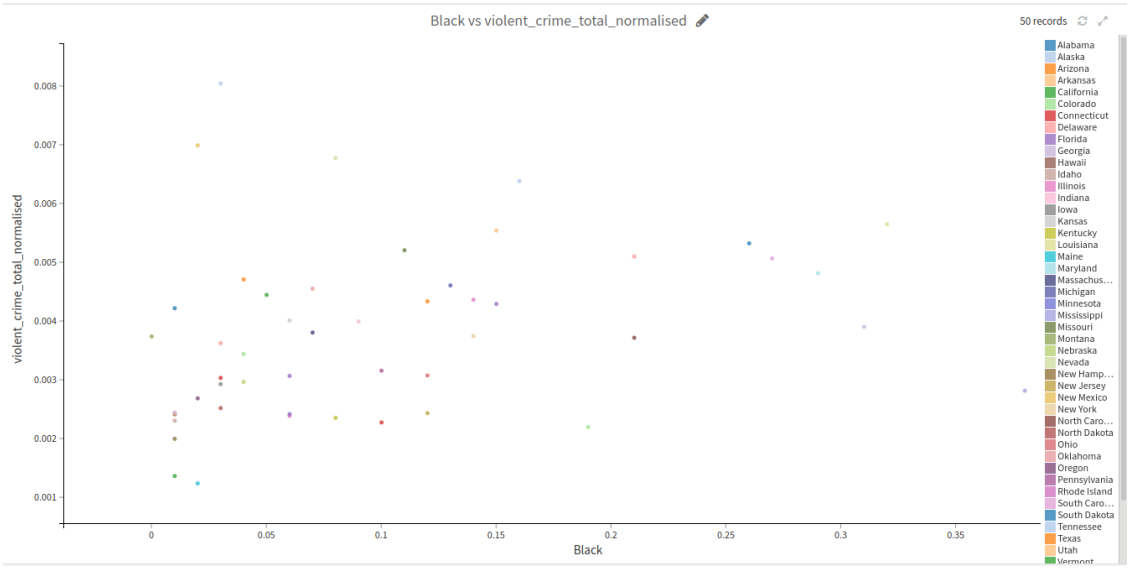
Les états les plus diversifiés sur le plan ethnique sont-ils moins stables ?

Évolution du nombre de crimes en fonction du pourcentage de blancs



La figure représente le taux de criminalité de 2016 en fonction du pourcentage de population blanche dans chaque état durant la même année. Un pourcentage de blancs élevé indique que cet état ne présente pas de diversité ethnique. On remarque que plus la part de population blanche augmente, plus le taux de criminalité diminue.

Évolution du nombre de crimes en fonction du pourcentage de noirs



La figure représente le taux de criminalité de 2016 en fonction du pourcentage de population afro-américaine dans chaque état durant la même année. On remarque que plus la part de population afro-américaine augmente, plus le taux de criminalité croit.

D'après les analyses précédentes, il est clair que les états les plus diversifiés sur le plan ethnique sont ceux qui enregistrent les taux de criminalité les plus élevés. **Cependant nous ne tirons évidemment aucune relation de cause à effet.**. La corrélation n'implique pas la causalité, il existe plusieurs facteurs confondants qui introduisent un biais de confusion (confounding bias) et qui pourraient apporter une explication à une telle corrélation. Les milieux diversifiés sont souvent issus de l'immigration donc ce sont des milieux pauvres, ou bien, il s'agit de milieux où la communauté afro-américaine est importante, cette dernière est historiquement défavorisée. (un biais de confusion est typiquement introduit quand on a deux phénomènes qui ont une même cause, ces deux phénomènes sont corrélés sans pour autant qu'il y a relation de cause à effet.)

4.1.3 Peut-on prédire le taux de criminalité dans un état à partir de ces facteurs ?

Après avoir déterminé quelques facteurs socio-économiques influant sur la criminalité nous essayons de prédire le nombre de crimes commis par année pour un état. Nous avons récolté des données sur différents facteurs de 2009 à 2016. On entraîne un modèle de régression (Gradient boosting : XGBOOST, mesure R2 pour l'ajustement) sur les données de 2009 à 2015, et on teste sur les données de 2016.

Résultats de la prédiction du nombre de crimes par état en 2016

jurisdiction	crime_total	prediction	error	error_decile	abs_error_decile	relative_error
string	double	double	double	bigint	bigint	double
US State	Decimal	Decimal	Decimal	Integer	Integer	Decimal
ALABAMA	169137.0	168055.55	-1081.453125	1	0	-0.0063939476578158535
ALASKA	30842.0	25921.977	-4920.0234375	0	0	-0.15952348866805005
ARIZONA	239859.0	245031.77	5172.765625	1	0	0.021565860046944246
ARKANSAS	114655.0	111628.625	-3026.375	0	0	-0.026395490820286947
CALIFORNIA	1176866.0	1190704.1	13838.125	2	1	0.0117584542335321 1
COLORADO	171176.0	160328.44	-10847.5625	0	1	-0.06337081424966116
CONNECTICUT	73044.0	73149.04	105.0390625	1	0	0.0014380245126225288
DELAWARE	31229.0	32891.97	1662.96875	1	0	0.05325078452720228
FLORIDA	642512.0	663961.94	21449.9375	3	2	0.033384493207908955
GEORGIA	347573.0	348142.6	569.59375	1	0	0.001638774444505183
HAWAII	45805.0	46849.1	1044.1015625	1	0	0.02279448886584434
IDAHO	33233.0	32585.16	-647.83984375	1	0	-0.019493871866819126
ILLINOIS	319310.0	312380.34	-6929.65625	0	0	-0.02170197065547587
INDIANA	194976.0	197742.3	2766.296875	1	0	0.014187884021623173
IOWA	75058.0	73304.7	-1753.296875	0	0	-0.02335922719763383
KANSAS	93958.0	91190.33	-2767.671875	0	0	-0.02945647922476

evs	mae	mse	mape	rmse	rmsle	r2	pearson
double	double	double	double	double	double	double	double
Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal	Decimal
0.9957812496890076	7570.411172672194	2.128888217645892E8	0.07354238448249208	14590.710118585359	0.11280938820300507	0.995410864682233 9	0.9982108017458897

La table représente les résultats de la prédiction de la variable cible "crime total", le résultat de la régression est dans la colonne "prediction" , la colonne error donne la différence entre la valeur prédite et la valeur réelle.

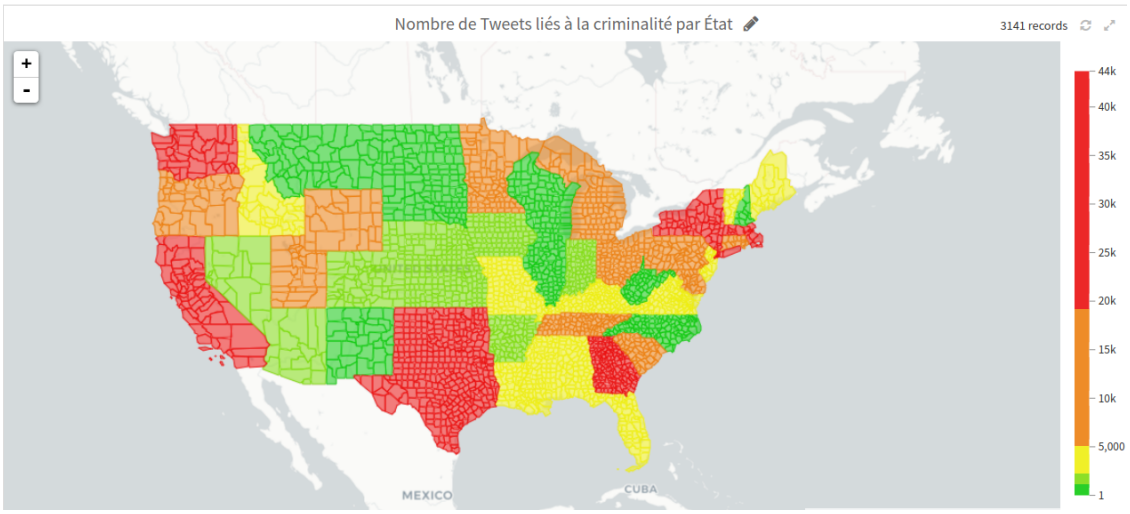
Un autre modèle intéressant serait de considérer les données sur la criminalité comme une time series, puisque nous avons récolté des données de 2001 à 2016 (plusieurs time series, pour chaque état une time series). Et utiliser un modèle de forecasting pour prédire les valeurs futures sur la base des valeurs précédemment observées. Ceci peut être utile par exemple pour un gouvernement pour attribuer un budget à chaque état pour la sécurité.

4.2 Utilisation de l'analyse de médias sociaux :

Nous voulons maintenant utiliser les médias sociaux comme twitter pour étudier la criminalité aux états unis. Nous avons récolté et compté les tweets parlants de crimes pour chaque état des USA comme précisé dans le chapitre données. Le nombre de tweets parlant de criminalité émis à partir d'un état devrait normalement être proportionnel au nombre de crimes commis dans ce même état. Cependant et vu les moyens très limités avec lesquels nous avons collecté les tweets les résultats peuvent ne pas être concluants (de plus, nous avons collecté une petite quantité de tweets et sur une petite durée.)

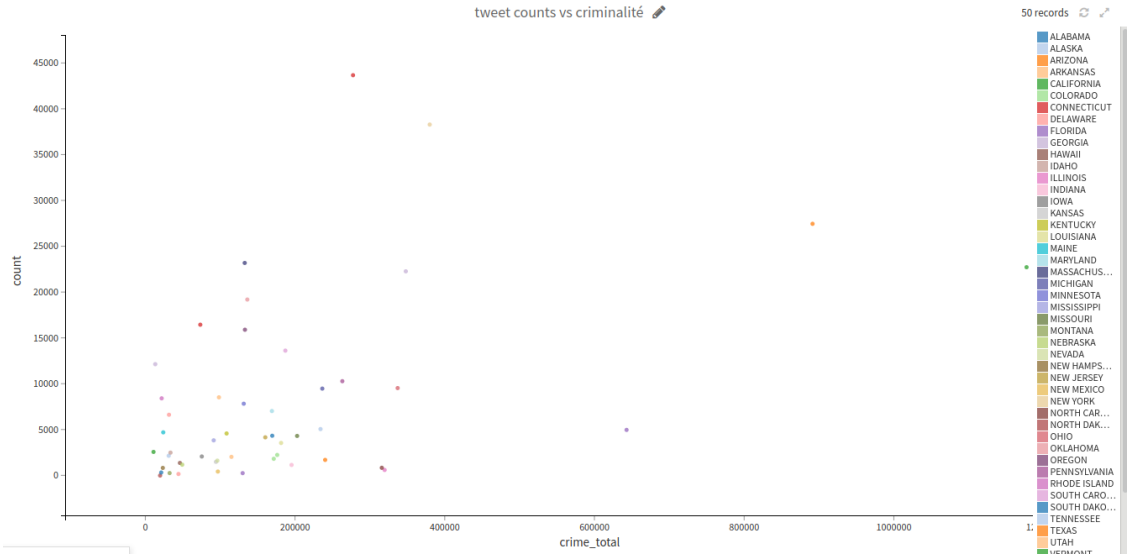
4.2.1 Les indicateurs sur le crime mesurés dans un état durant les années prétendantes sont-ils proportionnels au nombre de tweets parlant de crimes émis à partir de ce même état ?

Nombre de Tweets liés à la criminalité par État



Cette figure représente le nombre de tweets liés à la criminalité par état américain. Les zones rouges sont les zones où il y a eu le plus de tweets liés à la criminalité. Nous constatons que certains états ont un nombre très élevé de tweets liés à la criminalité (Washington, Californie, Texas, New York), en nous fiant aux tweets récoltés et en supposant que le nombre de crimes commis est proportionnel au nombre de tweets liés à la criminalité, nous classons ces états comme ceux où il y a le plus de crimes.

La corrélation entre les tweets collectes parlant de criminalité et le nombre réel de crimes commis par état

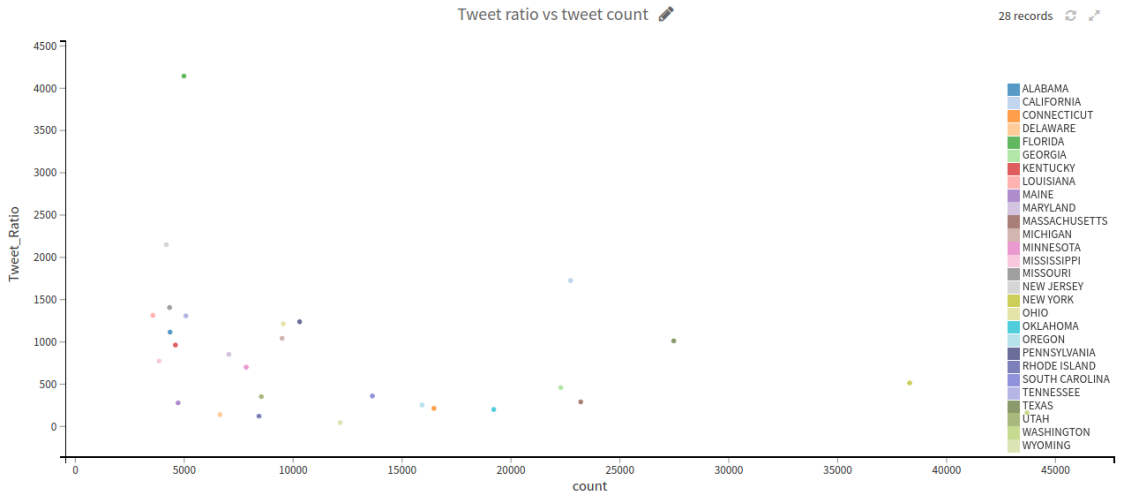


La figure ci-dessus montre la corrélation entre le nombre de tweets liés à la criminalité collectés et le nombre réel de crimes commis par état. Les moyens de collecte des tweets étant limités, les résultats peuvent ne pas être cohérents.

4.2.2 Peut-on dire d'un état s'il est sure ou non en se basant sur le nombre de tweets parlants de criminalité émis à partir de ce dernier ?

Cependant si le nombre de tweets parlant de crime est élevé dans une ville cela ne veut pas dire qu'il s'agit d'une ville dangereuse, il faut prendre d'autres critères en considération et principalement la population puisqu'une population plus élevée implique un nombre de tweets plus élevé. Pour palier à ce problème, on analyse aussi la proportion des tweets sur la population totale de l'état.

La proportion des tweets sur la population totale de chaque etat

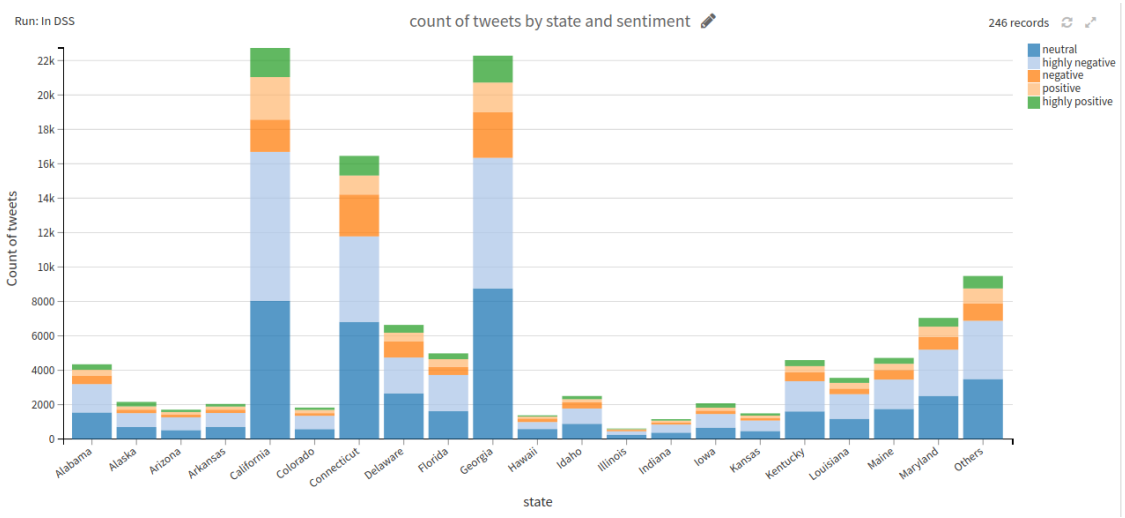


Chaque point de la figure ci-dessus représente un état, les états les plus sûrs sont ceux qui se trouvent en haut à gauche , ils ont donc une population importante et un petit nombre de tweet sur la criminalité. Les États les plus dangereux sont ceux qui se trouvent en bas à droite, avec un nombre très élevé de tweets sur la criminalité, et une petite population.

4.2.3 Comment l'analyse spatio-temporelle des médias sociaux peut-elle aider à identifier les tendances de criminalité ainsi que les événements qui l'influencent ?

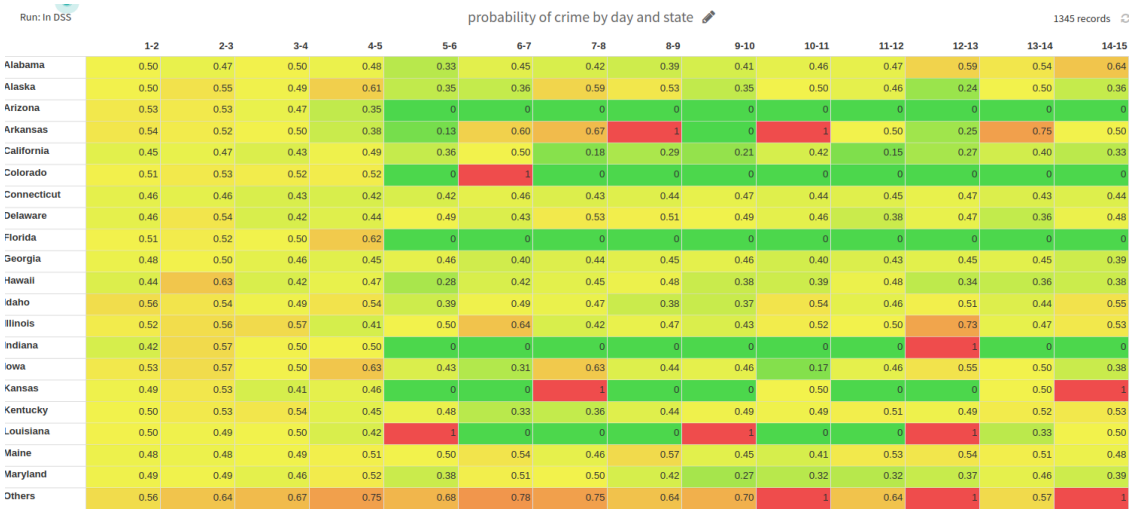
On réalise une analyse de sentiments sur les tweets parlants de criminalité récoltés. L'analyse de sentiments renvoie 5 valeurs (très négatif, négatif, neutre, positif, très positif). Les tweets négatifs et très négatifs indiquent que le crime dont parle le tweet revêt une importance particulière, ça peut aussi filtrer les tweets qui utilisent un vocabulaire relié au crime, mais qui ne traduisent pas un crime. Nous allons exploiter cette information pour "monitorer" la criminalité dans un état des USA. On doit calculer pour chaque ville la proportion des tweets parlant de crimes violents/importants. (probabilité qu'un tweet traduit un sentiment négatif ou très négatif)

Comptage des tweets par état et par sentiment



La figure montre le nombre de tweets parlant de la criminalité par état et par sentiment empilé. Les tweets parlant du crime de manière très positive, positive ou neutre sont très rares, ce qui est logique puisque le sujet est rarement positif (la mort d'un meurtrier peut être considérée comme positive, quelqu'un qui parle en utilisant le champ lexical du crime sans parler de crime ...). D'autre part, on constate que le nombre de tweets parlant du crime de manière négative ou très négative est assez important.

La proportion de tweets négatifs par jour et par État.



	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-13	13-14	14-15
Alabama	0.50	0.47	0.50	0.48	0.33	0.45	0.42	0.39	0.41	0.46	0.47	0.59	0.54	0.64
Alaska	0.50	0.55	0.49	0.61	0.35	0.36	0.59	0.53	0.35	0.50	0.46	0.24	0.50	0.36
Arizona	0.53	0.53	0.47	0.35	0	0	0	0	0	0	0	0	0	0
Arkansas	0.54	0.52	0.50	0.38	0.13	0.60	0.67	1	0	1	0.50	0.25	0.75	0.50
California	0.45	0.47	0.43	0.49	0.36	0.50	0.18	0.29	0.21	0.42	0.15	0.27	0.40	0.33
Colorado	0.51	0.53	0.52	0.52	0	1	0	0	0	0	0	0	0	0
Connecticut	0.46	0.46	0.43	0.42	0.42	0.46	0.43	0.44	0.47	0.44	0.45	0.47	0.43	0.44
Delaware	0.46	0.54	0.42	0.44	0.49	0.43	0.53	0.51	0.49	0.46	0.38	0.47	0.36	0.48
Florida	0.51	0.52	0.50	0.62	0	0	0	0	0	0	0	0	0	0
Georgia	0.48	0.50	0.46	0.45	0.46	0.40	0.44	0.45	0.46	0.40	0.43	0.45	0.45	0.39
Hawaii	0.44	0.63	0.42	0.47	0.28	0.42	0.45	0.48	0.38	0.39	0.48	0.34	0.36	0.38
Idaho	0.56	0.54	0.49	0.54	0.39	0.49	0.47	0.38	0.37	0.54	0.46	0.51	0.44	0.55
Illinois	0.52	0.56	0.57	0.41	0.50	0.64	0.42	0.47	0.43	0.52	0.50	0.73	0.47	0.53
Indiana	0.42	0.57	0.50	0.50	0	0	0	0	0	0	0	1	0	0
Iowa	0.53	0.57	0.50	0.63	0.43	0.31	0.63	0.44	0.46	0.17	0.46	0.55	0.50	0.38
Kansas	0.49	0.53	0.41	0.46	0	0	1	0	0	0.50	0	0	0.50	1
Kentucky	0.50	0.53	0.54	0.45	0.48	0.33	0.36	0.44	0.49	0.49	0.51	0.49	0.52	0.53
Louisiana	0.50	0.49	0.50	0.42	1	0	0	0	1	0	0	1	0.33	0.50
Maine	0.48	0.48	0.49	0.51	0.50	0.54	0.46	0.57	0.45	0.41	0.53	0.54	0.51	0.48
Maryland	0.49	0.49	0.46	0.52	0.38	0.51	0.50	0.42	0.27	0.32	0.32	0.37	0.46	0.39
Others	0.56	0.64	0.67	0.75	0.68	0.78	0.75	0.64	0.70	1	0.64	1	0.57	1

En raison du manque de tweets nous avons additionner (pour les états où on a le plus de tweets) tous les tweets émis pour un même jour du mois (les tweets émis le 1er du mois, les tweets émis le 2ème du mois...) La figure ci-dessus montre les États des États-Unis en lignes, en colonnes les jours du mois. Quant aux cases, elles contiennent la proportion de tweets traduisant un sentiment négatif ou très négatif (crimes importants). À partir des données que nous avons récoltées, on remarque que les débuts du mois enregistrent une activité criminelle importantes.

Ça aurait été intéressant d'avoir des datasets de tweets datés, c'est-à-dire pour chaque état, pour chaque jour les tweets parlants de crimes pour ce jour-là. De cette façon, on pourra suivre l'évolution de la proportion des tweets parlant de crimes violents (ou importants) pour identifier pour chaque état les jours où les crimes les plus graves ont été commis ou bien les jours où les gens sont le plus affecté par un ou plusieurs crimes en particulier. Ceci permettrait d'identifier par exemple les jours de la semaine (les mois ou les saisons) où les crimes sont plus intenses. (identifier des tendances)

Pour un gouvernement : Avoir déjà une base de données pour chaque ville qui comprend pour chaque jour le nombre de tweets parlants de crimes ainsi que la proportion de tweets parlant de crimes violents ou importants (négatif à très négatif) Ensuite, on sonde chaque jour cette ville pour comparer les résultats du jour et les statistiques habituelles. Ce genre de monitoring permettra d'identifier au plus vite un pic de crime dans une ville. Cette méthode permet de traiter en particulier les cas où la criminalité est affectée par un événement émergeant (manifestations, élections, ...)

4.2.4 Peut-on prédire quelle catégorie de crime est la plus probable pour une date et une ville donnée des états unis ?(non traitée)

Afin de prévoir quelle catégorie de crime est la plus probable pour une date et une ville données, nous suggérons de scraper les réseaux sociaux (pages Facebook, Tweets, Instagram...), les sites d'information et d'utiliser la OpenData du FBI et des organisations américaines pour collecter les données nécessaires (statistiques des crimes passés avec leurs types, les facteurs de chaque type de crime et les textes parlants de chaque type de crime). Grâce à ces données et aux modèles de traitement du langage naturel, nous pourrons construire une base de données spatio-temporelle qui nous permettra d'entraîner nos modèles à prédire les types de crimes les plus probables pour un jour donné dans une ville donnée.

Chapter 5

Conclusion

Dans le cadre de ce projet, nous nous sommes attelés à la problématique de l'étude de criminalité aux états-unis. Dans un premier temps, nous avons analysé la criminalité aux états-unis en détectant des facteurs socio-économiques qui influent sur la criminalité et essayé de prédire cette dernière en nous basant sur ces mêmes facteurs. Par la suite, nous avons exploité les medias sociaux (twitter) d'une part pour juger si un état est dangereux sur la base des tweets émis à partir de ce dernier et d'autre part pour identifier les tendances de criminalité ainsi que les événements qui l'influencent en utilisant la détection de sentiment et en effectuant une analyse spatio-temporelle.