RAPPORT DATA MINING

Accidents de la route

BAMBA William

CHUPIN Pierre-Henri

HOSSAIN Shajjad

JASSIGNEUX Marie

SOLOMON Maria

TREGER Pauline

30 JANVIER 2021

GROUPEX
Master IA

Table des matières

Introduction 2

Code 2

Présentation des données, statistiques et analyses préliminaires 2

- Présentation des données 2
- Statistiques et analyses préliminaires 2

Question(s) que l'on souhaite répondre 3

- Clustering 3
- Frequent Itemset 3

Pré-traitement des données 3

- Clustering 3
- Frequent itemset 4

Clustering 4

- Présentation de la méthode utilisée 4
- Première question 5
- Deuxième question 6
- Troisième question 7
- Quatrième question 8
- Cinquième question 9
- Sixième question 11
- Septième question 12

Frequent Itemset 14

- Présentation de la méthode utilisée 14
- Quels sont les types d'accidents qui se produise régulièrement ? 14

Discussion 15

Conclusion 15

Introduction

L'objectif de ce projet était de mettre en œuvre des approches complètes de fouille de données que nous avions étudiées en cours. Pour cela, nous avons exploité les données des accidents de la route arrivés en France de 2005 à 2016.

Notre but est donc d'utiliser des outils de fouille de données afin de faire apparaître des données pouvant répondre à des problématiques liées aux accidents de la route en France. Ces données pourraient être utilisées par différents acteurs que nous décrirons au fur et à mesure de ce rapport.

Code

En complément de ce rapport, vous trouverez le lien vers notre dépôt Forge ici.

Les indications pour utiliser le code sont fournies dans le README.

Pour la réalisation de ce projet, nous avons utilisé le logiciel KNINE afin d'appliquer les algorithmes de clustering. Nous avons aussi utilisé le langage python et ses librairies pour le pré-traitement des données ainsi que pour découvrir des motifs fréquents.

Présentation des données, statistiques et analyses préliminaires

Nous avons choisi notre jeu de données de notre choix sur le site KAGGLE, voici le lien vers ce jeu de données. Nous avons choisi de travailler sur les accidents en France de 2005 à 2016.

Présentation des données

Tout d'abord, il faut savoir que notre jeu de données est composé de cinq fichiers CSV :

- « *caracteristics.csv* » : sert à connaître le moment et la localisation de l'accident, ainsi que le type de collision (2 voitures, 3 voitures ou plus)
- « *holidays.csv* » : permet de mettre en correspondance les dates avec les jours de vacances et jours fériés en France.
- « *places.csv* » : permet de connaître le type de lieux où s'est produit l'accident (type de route, et surface de la route)
- « *users.csv* » : permet de connaître les informations des usagers touchés pas l'accident et leur état suite à l'accident (décédé, blessure grave, etc.). Il permet également de savoir quelles mesures de sécurités avaient été prises par l'usager (ceinture mise, casque mis, etc.).
- « *vehicles.csv* » : permet de connaître le type de véhicule impliqué dans l'accident (voiture, vélo, scooter, etc.).

Statistiques et analyses préliminaires

Nous avons fait quelques requête SQL pour faire ressortir certaines statistiques :

- Nous avons cherché à savoir quelle est l'année qui a eu le plus d'accident. Il s'avère que c'est l'année 2016 avec 11% des accidents totaux.

- Nous avons cherché à savoir quel type de véhicule a le plus d'accidents. Il s'avère que ce sont les voitures avec 85% des accidents totaux.

A l'aide de ses requêtes et de l'analyse du jeu de données, nous avons pu faire ressortir les questions les plus intéressantes qu'on pourrait se poser pour ce projet.

Question(s) que l'on souhaite répondre

Au cours de ce projet, on s'est posé différentes questions auxquelles nous souhaitons répondre. Nous avons décidé de répondre à différentes problématiques que nous nous sommes posées et qui nous semblaient intéressantes.

On a réparti ces questions en fonction de la méthode qu'on a utilisé pour répondre à nos problématiques.

Clustering

- Où se produisent la plupart des accidents graves de voiture pour chaque année ?
- Où se produisent la plupart des accidents de voiture grave en période de vacance ou de jours fériés ?
- Où se produisent la plupart des accidents de voiture grave la nuit sans éclairage ?
- Où se produisent la plupart des accidents de vélo ou de scooter ?
- Où se produisent la plupart des accidents graves ?
- Où se produisent la plupart des accidents graves en fonction de la condition atmosphérique ?
- Où se produisent la plupart des accidents graves en fonction de la surface de la route ?

Frequent Itemset

- Quels sont les types d'accidents qui se produise régulièrement ?

Maintenant nous allons voir plus en détail le travail qui a été effectué.

Pré-traitement des données

Dans cette première partie, nous allons montrer tous les pré-traitements des données qui nous ont été nécessaire d'effectuer afin de pouvoir utiliser les données. Nous avons effectué deux pré-traitements différents en fonction de l'algorithme utilisé.

Clustering

Le problème que nous avons rencontré pour le clustering, c'est que pour une large majorité des données nous n'avions pas la latitude et la longitude des accidents dans le fichier « *caracteristics.csv* ». Nous avions donc deux possibilités possibles :

- Soit on enlève ses données dont on avait des parties manquantes
- Soit on les complète

Nous avons fait le choix de les compléter. Pour cela nous avons ajouter les latitudes et les longitudes manquantes grâce aux adresses, le code de commune et le code du département. Ainsi nous avons pu compléter les informations manquantes et continuer de projet.

Pour faire cela, nous avons utilisé l'API du gouvernement *(api-adresse.data.gouv.fr)* en effectuant des requêtes sur 300 milles lignes. L'API du gouvernement nous renvoyait pour une adresse donnée une latitude, une longitude et un score. Ce score représente la confiance sur l'exactitude du résultat.

Une des difficultés que nous avons rencontré à ce stade, c'est que faire 300 milles requêtes, c'est long, environ 8h. Du coup pour aller plus vite nous avons utilisé des « Thread » tout en faisant attention que l'API ne refuse pas un nombre important de requête en simultanée. Nous avons fixé la limite à six requêtes en parallèle. De plus, nous avons sauvegardons les requêtes au fur et à mesure pour éviter de devoir tout recommencer en cas de problème.

Frequent itemset

Nous avons commencé par enlever les colonnes qui ne nous intéressait pas comme par exemple l'adresse, les coordonnées GPS, etc.

Ici, nous n'avons pas eu à compléter de donnée mais plutôt à filtrer nos données. En effet, nous avions des informations qui ne sont pas dans les descriptions des fichiers sur KAGGLE mais qui sont bien contenues dans les fichiers et nous ne voulions pas de ses informations supplémentaires. Nous les appellerons valeurs invalides pour la suite. Nous avons créé un fichier JSON pour décrire un intervalle de valeur autorisé par colonne. Ensuite, on a transformé les données en transaction exploitable pour extraire des motifs fréquents. A ce moment-là, on remplace les valeurs invalides par -1 pour pouvoir les enlever une fois que les transactions sont transformées en « binary dataset ». Ainsi, cela nous permet d'enlever uniquement les valeurs invalide plutôt que toute la ligne où la valeur invalide est présente.

Une fois le pré-traitement nécessaire fait, nous avons sauvegardé notre jeu de donnée complété et filtré. C'est sur ce jeu de donnée que nous utilisons ensuite.

Clustering

Présentation de la méthode utilisée

Nous avons dans un premier temps réalisé différents clustering sur les données. Pour cela nous avons utilisé l'algorithme DBSCAN. Cet algorithme de clustering basé sur la densité permet de créer des clusters de régions denses. Tout en ignorant les zones de faible densité, en les considérant comme du bruit.

Nous avons choisi d'utiliser cet algorithme à cause de ses nombreux avantages :

- Il fonctionne très bien pour les ensembles de données bruyant. Etant donné que nous travaillons sur les données des accidents de la route partout en France de 2005 à 2016, selon le type de clustering que nous allons faire, il y aura forcément du bruit dans les données.
- Il peut facilement identifier les valeurs aberrantes. De la même manière, vu la quantité de données que nous avons, même si nous avons réalisé un pré-traitement, il risque de subsister des valeurs aberrantes.

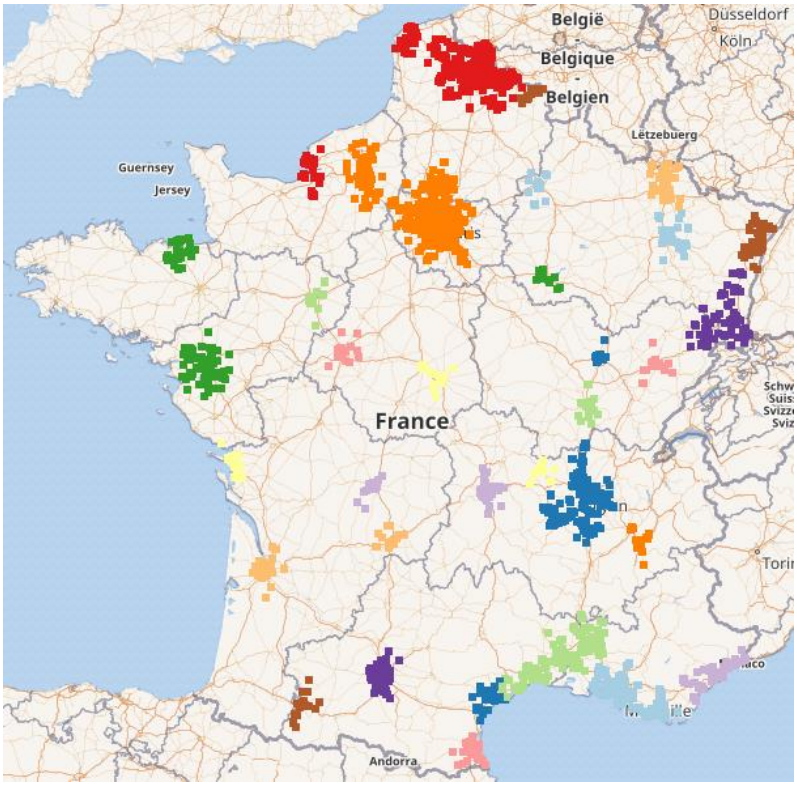
- Les clusters peuvent prendre n'importe quelle forme irrégulière, cet algorithme nous permet d'obtenir des clusters adaptés aux données que nous avons, nous pourrons ainsi obtenir des clusters plus significatifs.

Nous avons cherché à obtenir des clusters répondant aux différentes questions vu plus tôt. Pour chacune de ces problématiques que nous nous sommes posés nous avons eu le même procédé. Dans un premier temps, nous nous sommes demandé à qui s'adresserait ce genre de clusters. Puis nous avons réfléchis à comment nous pouvons afficher des clusters correspondants à cette problématique et au public que nous visons. Enfin, une fois que nous avons eu ces clusters, nous nous sommes demandé quelles informations supplémentaires pouvaient être ajoutées pour les rendre plus intéressantes et plus complètes.

Première question

La première problématique à laquelle nous avons essayé de répondre a été la suivante : « *Où se produisent la plupart des accidents graves de voiture pour chaque année ?* ». Nous avons donc ici essayé de trouver les clusters des accidents mortel ou/et avec blessure grave impliquant des voitures de 2005 à 2016. Nous avons réalisé une carte pour chaque année avec les clusters correspondant.

Nous avons obtenu les clusters suivants pour l'année 2009 :

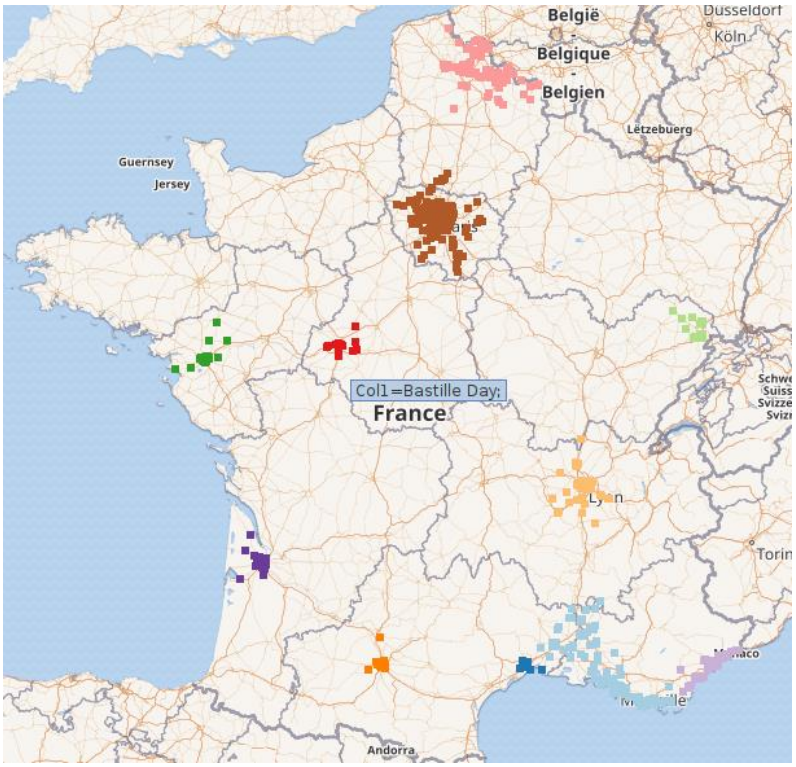


Nous pouvons remarquer qu'il existe de nombreux clusters partout en France. Pour cette année en particulier, les plus gros clusters se trouvent autour de Paris, Nord de la France et le bassin méditerranéen. Réalisé ce genre de carte pour chaque année permet de voir l'évolution des clusters. Cela permettrait aux organisations des routes d'adapter leur politique si des clusters subsistent au fil du temps. Cela leur permettrait également de voir si leur politique fonctionne dans le cas où les clusters diminuent.

Deuxième question

La deuxième problématique à laquelle nous avons essayé de répondre a été la suivante : « *Où se produisent la plupart des accidents de voiture grave en période de vacance ou de jours fériés ?* ». Nous avons donc ici essayé de trouver les clusters des accidents mortel ou/et avec blessure grave arrivés durant les vacances scolaires ou jours fériés.

Nous avons obtenu les clusters suivants :



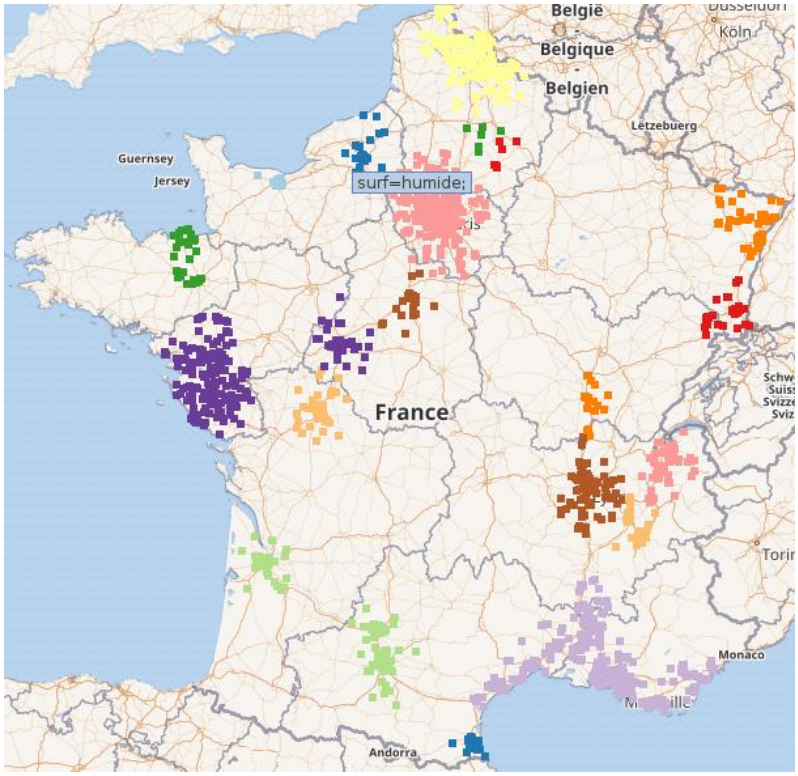
Nous pouvons tout d'abord constater que les lieux ayant le plus d'accidents lors de vacances ou de jours fériés se trouvent principalement dans les grandes villes de France. Cela est plutôt logique étant donné que les villes représentent de plus grosses densités de personnes. Lors de vacances et de jours fériés beaucoup de mouvements sont réalisés, et donc augmente les risques d'accidents graves. Nous pouvons également remarquer que les lieux avec les plus grandes densités d'accidents graves lors de vacances ou de jours fériés se trouvent autour de Paris et au niveau du bassin méditerranéen. Cela peut déjà donner une idée d'où se trouve les lieux avec le plus de mouvements lors de vacances ou de jours fériés.

Cela permettrait aux communes concernées d'améliorer leurs dispositifs de prévention d'accidents en adaptant leur communication ou/et définir les endroits où pourraient se positionner les gendarmes ou policiers. Chaque commune concernée pourrait regarder plus précisément dans leur ville où se situent les routes les plus dangereuses dans ces moments-là afin d'adapter au mieux leurs dispositifs.

Troisième question

La troisième problématique à laquelle nous avons voulu répondre est la suivante « *Où se produisent la plupart des accidents de voiture grave la nuit sans éclairage ?* ». Nous voulions pouvoir visualiser les routes de France les plus dangereuses la nuit.

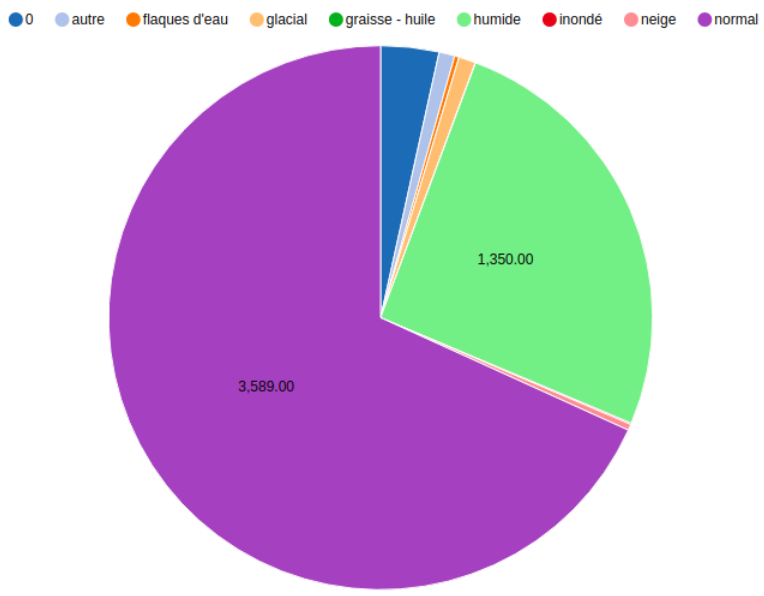
Nous avons obtenu les clusters suivants :



Nous pouvons ici apercevoir de nombreux clusters se trouvant partout en France. Cela démontre la dangerosité des routes de France la nuit lorsqu'il n'y a pas de lumière. Grâce à cette carte nous pouvons voir que les lieux ayant le plus d'accidents graves la nuit sans lumière sont autour de Paris, autour de Nantes et au niveau du bassin méditerranéen.

Cette carte seule aurait une utilité auprès des communes et des collectivités afin qu'elles puissent adapter l'éclairages de leurs routes en conséquence.

Nous avons tout de même essayé de chercher plus loin, afin de savoir si les communes et collectivités peuvent jouer sur d'autres facteurs afin de rendre les routes de France plus sûres. Pour cela, nous avons décidé d'afficher sous la forme d'un graphique circulaire la part des conditions de la surface de la route lors de ces accidents.



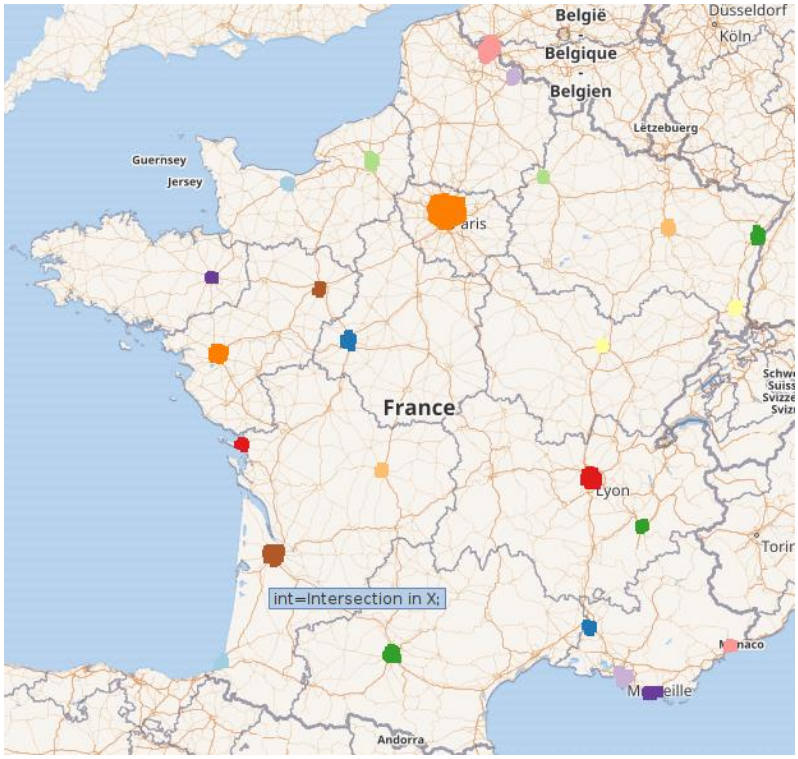
Nous pouvons voir que la plupart des accidents se déroulant la nuit lorsqu'il n'y a pas de lumière se déroulent dans des conditions de surface de route normales. Cependant, il faut noter que $\frac{1}{4}$ de ces accidents se déroulent lorsque la surface de la route est humide.

Cette information supplémentaire pourrait permettre aux communes et collectivités d'adapter leurs signalisations et leurs indications sur les panneaux afin d'avertir les conducteurs de faire particulièrement attention sur ces routes la nuit quand il pleut.

Quatrième question

La quatrième problématique à laquelle nous avons voulu répondre est la suivante : « *Où se produisent la plupart des accidents de vélo ou de scooter ?* ». L'objectif de visualiser ces clusters était de pouvoir visualiser les lieux où le plus d'accidents de vélo et scooter se produisaient. Nous permettant de pouvoir repérer sur quels types d'intersections ce type d'accidents se produisent.

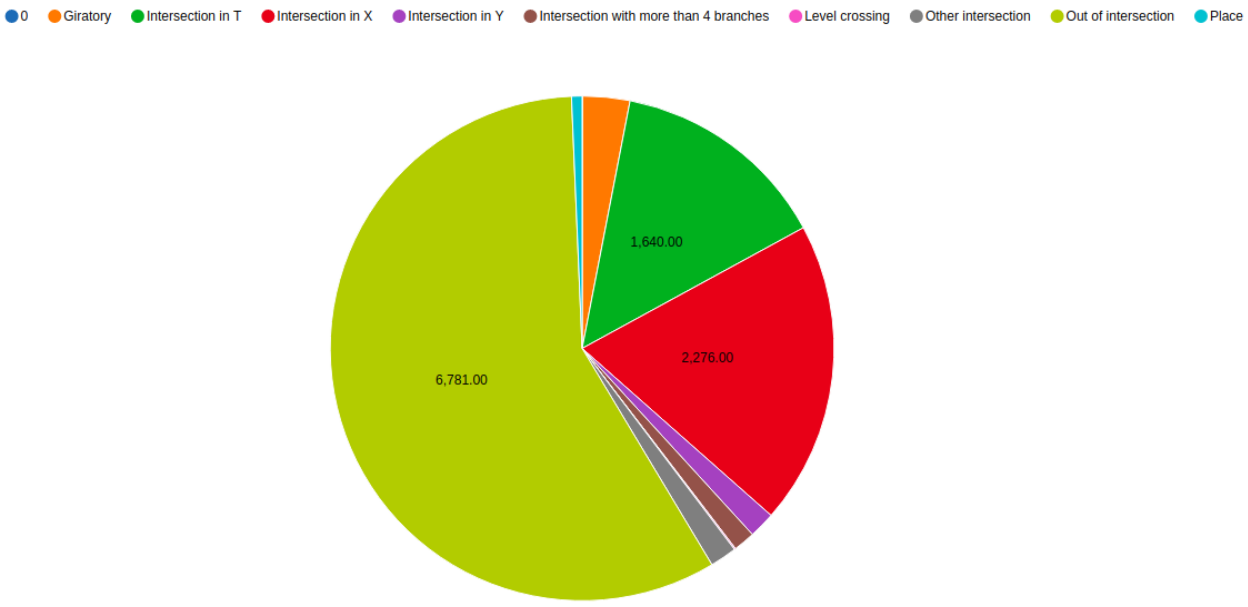
Nous avons obtenu les clusters suivants :



Nous pouvons voir que les clusters sont plus petits et plus denses. Le plus gros cluster se trouve à Paris. Nous voyons que de nombreuses routes et intersections en France sont sources d'accidents de vélo et de scooter.

Ces données permettraient à des communes ou collectivités de pouvoir adapter leurs dispositifs sur les routes et de sécuriser les intersections les plus problématiques pour les vélos et scooter.

Nous avons donc voulu connaître la part d'accident liée à chaque type d'intersection.

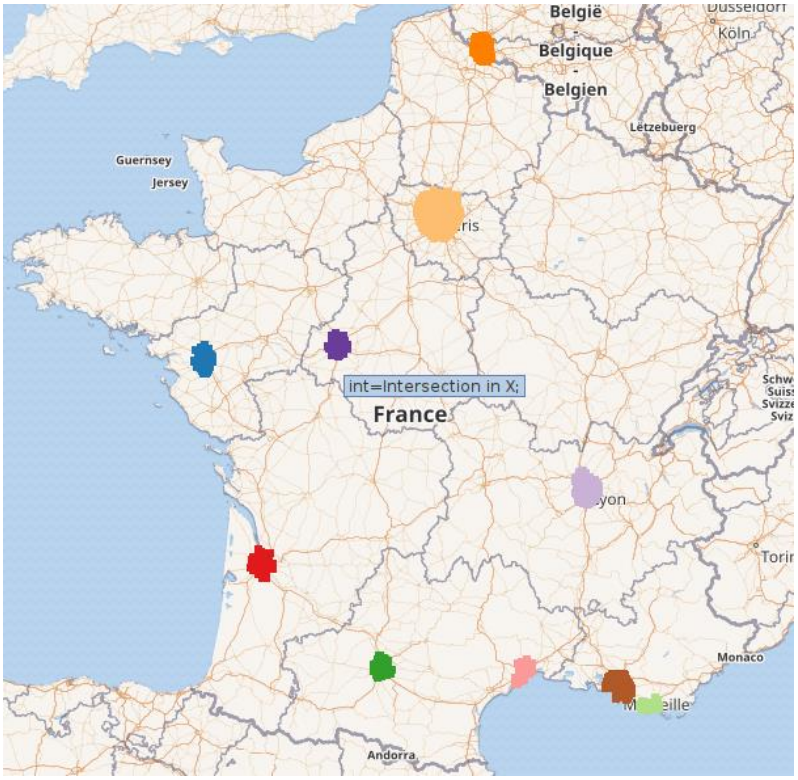


Cette visualisation permettrait de sécuriser les intersections au niveau national en connaissant les types d'intersections les plus dangereuses pour les vélos et les scooters. Par exemple en évitant qu'une piste cyclable emprunte ce genre d'intersection. Nous voyons donc ici que la plupart des accidents de vélo et de scooter se déroulent en dehors des intersections. Cependant, il y a une grande part d'accidents se déroulant dans des intersections en X et dans des intersections en T. Cela permettrait aux collectivités de connaître les intersections à éviter pour les pistes cyclables.

Cinquième question

La cinquième problématique à laquelle nous avons voulu répondre est la suivante : « *Où se produisent la plupart des accidents graves ?* ». Nous avons voulu créer une visualisation qui permettrait de montrer où les accidents les plus graves se produisaient. Nous avons peaufiné cette problématique pour mettre en évidence les clusters selon le nombre de véhicule impacté dans l'accident. Dans notre projet vous pourrez donc voir aussi les accident entre deux véhicules, ou trois véhicules et plus.

Nous avons obtenu les clusters suivants pour tout type de collision :

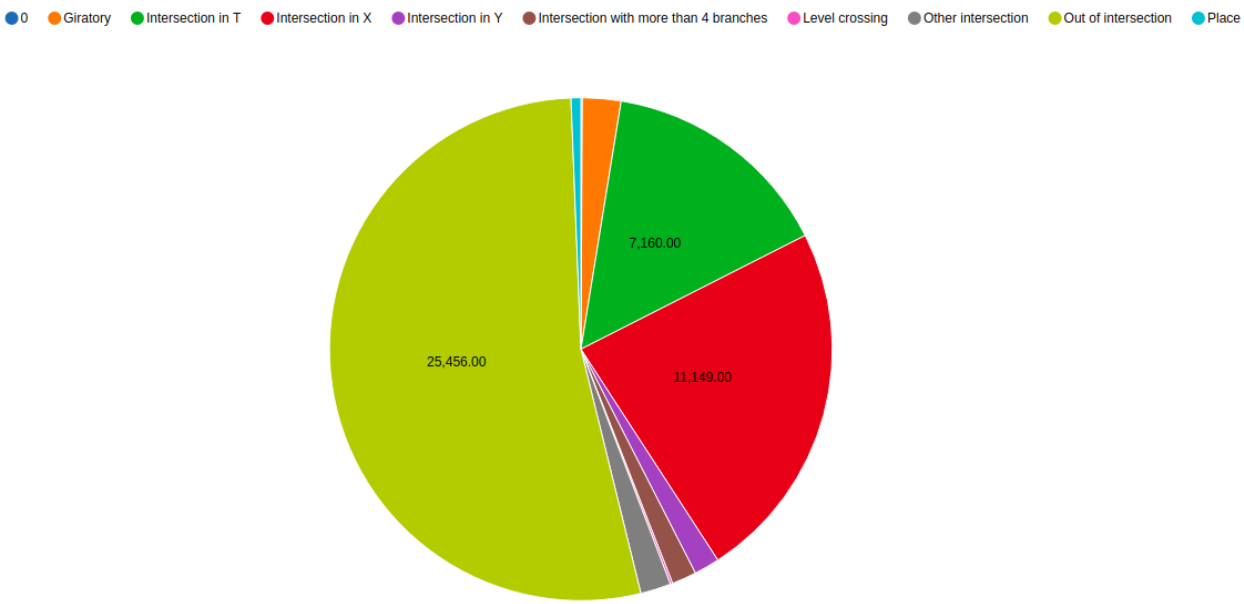


A l'aide de cette carte nous pouvons remarquer que les accidents les plus graves se localisent dans les grandes villes. Nous pouvons voir que Paris est le plus gros cluster où se situent les accidents les plus graves.

Cette carte permettrait aux collectivités et communes de pouvoir connaître les routes les plus dangereuses dans leur commune afin d'adapter les rénovations des routes prévues.

En complément de cette carte, nous avons voulu montrer sur quels types d'intersections se déroulent ces accidents les plus graves.

Grâce à cette visualisation, nous pouvons remarquer que la plupart des accidents graves se déroulent hors des intersections. Nous pouvons voir que les intersections en X et en T sont les plus dangereuses et causent le plus d'accidents.

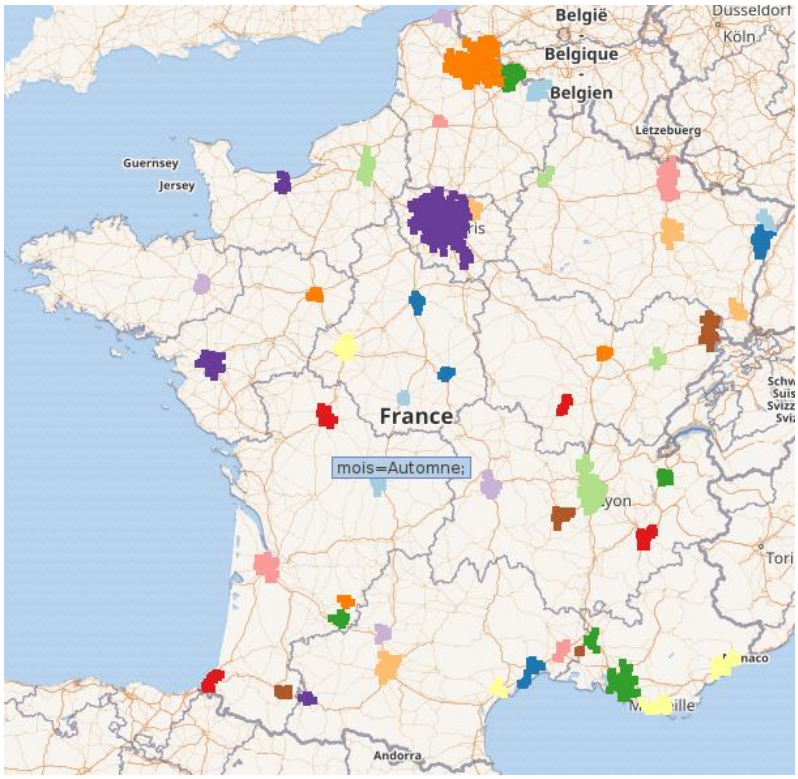


Ces informations pourraient permettre aux collectivités ou communes de modifier leurs intersections afin de les rendre plus sécurisées.

Sixième question

La sixième problématique à laquelle nous avons voulu répondre est la suivante : « *Où se produisent la plupart des accidents graves en fonction de la condition atmosphérique ?* ». Nous cherchons à connaître où se situent les routes les plus dangereuses selon les conditions atmosphériques. Nous avons effectué cet algorithme plusieurs fois dans cette problématique. Nous avons un résultat général et des résultats pour chaque condition atmosphérique.

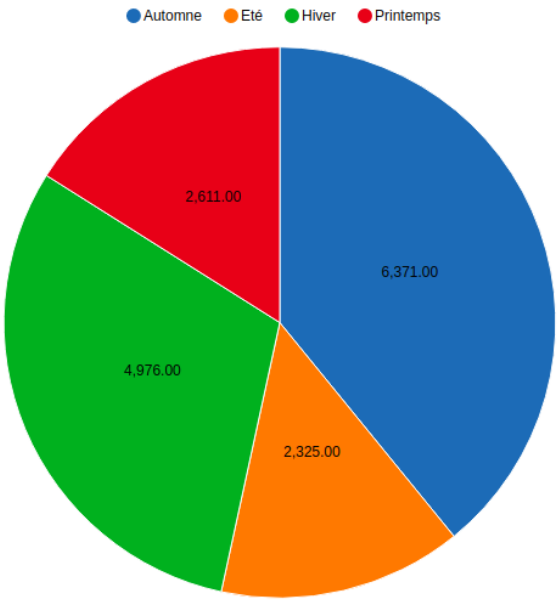
Nous avons obtenu les clusters suivants pour le résultat général pour toutes les conditions atmosphériques autre que « Normal » :



Cette carte permet de visualiser les clusters des accidents les plus dangereux. Nous pouvons visualiser qu'il existe beaucoup de clusters très petits correspondants à des lieux précis de certaines routes très dangereuses. En passant notre souris sur les points, la saison de l'année à laquelle l'accident s'est produit apparaît.

Nous avons voulu connaître à quels saisons les accidents se produisent le plus.

Ce graphique montre que les accidents se produisent le plus en automne et en hiver. Cela peut s'expliquer par les conditions météos de ces deux saisons. En automne il pleut beaucoup, et en hiver il peut neiger voir geler ce qui peut induire beaucoup d'accidents.

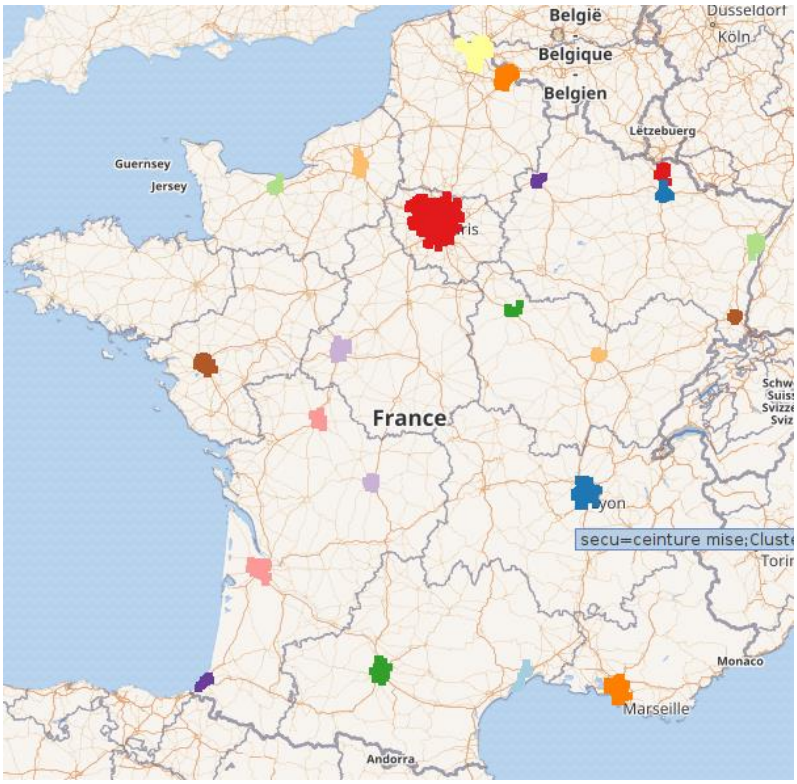


Toutes ces données peuvent permettre aux organisations gérant les routes de s'adapter et faire en sorte d'avertir les conducteurs lors de certaines conditions atmosphériques ou de refaire les routes pour qu'elles soient plus adaptées.

Septième question

La dernière problématique à laquelle nous avons voulu répondre est la suivante : : « *Où se produisent la plupart des accidents graves en fonction de la surface de la route ?* ». Nous voulions déterminer où se situaient les routes les plus dangereuses dans le cas de surfaces de la route pas normales. Nous avons effectué cet algorithme plusieurs fois dans cette problématique. Nous avons un résultat général et des résultats pour chaque type de surface.

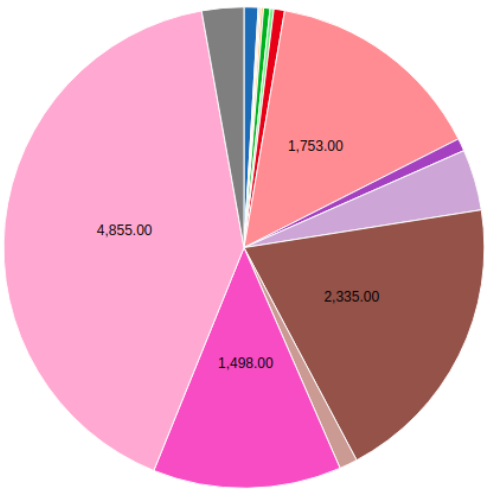
Nous avons obtenu les clusters suivants pour une surface humide :



Cette carte permet de visualiser les lieux où se produisent les accidents graves lorsque les conditions de surface de la route sont difficiles. Le plus gros cluster se situe à Paris. Les villes pourraient utiliser cela pour redéfinir leur politique d'entretien et de rénovation de leurs routes.

A cela nous avons voulu regarder quelles mesures de sécurités étaient prises par les conducteurs lors de leurs accidents quand la surface est humide.

Ce graphique permet de voir que la plupart des accidents se réalisent alors que les conducteurs ont pris leurs précautions (ceinture mise, casque mis, autre mise en place indéterminable).



Nous pouvons remarquer que $\frac{1}{6}$ des mesures de sécurité prise ne permet pas de savoir si la ceinture a été mise ou non. Ce graphique pourrait permettre aux organisations d'adapter les préventions à réaliser sur les routes concernées.

Ces différents clusters ont permis de mettre en évidence les améliorations possibles que les grandes villes, organisations, communes, et collectivités peuvent mettre en place pour limiter les accidents de la route en France.

Frequent Itemset

Présentation de la méthode utilisée

Dans cette partie nous avons extrait des motifs fréquents sur l'ensemble du jeu de donnée pré-traité. Pour cela, nous avons testé plusieurs algorithmes notamment Apriori, fpmax, et fpgrowth. Nous avons choisi fpmax car c'est celui qui nous donnent les plus grands motifs fréquents.

Une des difficultés rencontrées, c'est qu'une fois la fusion des trois fichiers users.csv, caracteristic.csv et vehicles.csv, notre base de données dépassait 3 millions de lignes. De plus, les algorithmes comme Apriori sont très lents avec autant de lignes. Pour résoudre ce problème, nous nous sommes inspirés de l'article vu en TP. Nous avons choisi des échantillons d'une taille de 100milles éléments aléatoirement en fonction de la taille de leur ensemble des parties. Cela nous permet de faire ressortir des motifs intéressants très rapidement.

Quels sont les types d'accidents qui se produise régulièrement ?

Afin de découvrir divers motifs fréquents intéressants, il est possible, en changeant légèrement les paramètres, d'obtenir de nouveaux résultats.

Voici un exemple des patterns que nous pouvons obtenir à l'aide de notre code :

```
([('atm', 1, 'Normal'), ('agg', 2, 'En agglomération')], [0.53939])

([('agg', 2, 'En agglomération'), ('actp', 0, 'non spécifié ou non applicable')], [0.54887])

([('atm', 1, 'Normal'), ('sexe', 1, 'Homme')], [0.538])

([('place', 1, 'Place occupée'), ('actp', 0, 'non spécifié ou non applicable'), ('catu', 1, 'Pilote'), ('sexe', 1, 'H
omme')], [0.55031])

([('int', 1, 'Hors intersection'), ('lum', 1, 'Plein Jour')], [0.50117])

([('place', 1, 'Place occupée'), ('actp', 0, 'non spécifié ou non applicable'), ('lum', 1, 'Plein Jour'), ('catu', 1,
'Pilote')], [0.53981])

([('atm', 1, 'Normal'), ('actp', 0, 'non spécifié ou non applicable'), ('lum', 1, 'Plein Jour')], [0.52283])

([('catv', 7, 'VL uniquement'), ('int', 1, 'Hors intersection')], [0.52187])

([('place', 1, 'Place occupée'), ('actp', 0, 'non spécifié ou non applicable'), ('int', 1, 'Hors intersection'), ('ca
tu', 1, 'Pilote')], [0.53031])

([('atm', 1, 'Normal'), ('actp', 0, 'non spécifié ou non applicable'), ('int', 1, 'Hors intersection')], [0.50137])

([('catv', 7, 'VL uniquement'), ('actp', 0, 'non spécifié ou non applicable'), ('catu', 1, 'Pilote'), ('place', 1, 'P
lace occupée')], [0.53015])

([('atm', 1, 'Normal'), ('catv', 7, 'VL uniquement'), ('actp', 0, 'non spécifié ou non applicable')], [0.50637])

([('atm', 1, 'Normal'), ('place', 1, 'Place occupée'), ('actp', 0, 'non spécifié ou non applicable'), ('catu', 1, 'Pi
lote')], [0.61277])

([('atm', 1, 'Normal'), ('int', 1, 'Hors intersection')], [0.5638])

([('catv', 7, 'VL uniquement'), ('lum', 1, 'Plein Jour')], [0.50043])
```

Nous pouvons voir ici un ensemble de motifs fréquents associés à leur support.

Nous allons prendre l'exemple de trois motifs fréquents que nous obtenons :

- Nous pouvons voir que 52% des accidents se déroulent hors des intersections et impliquent des voitures.
- Nous constatons également 53% des accidents impliquent des hommes lors de conditions météorologiques normales.
- Enfin 61 % des accidents se déroulent lors de conditions atmosphériques normales et impactent le conducteur.

En modifiant un peu les paramètres de notre code, nous pouvons obtenir des résultats de motifs fréquents différents comme par exemple ce résultat :

```
([('grav', 1, 'Indemne'), ('actp', 0, 'non spécifié ou non applicable')], [0.41472])

([('trajet', 5, 'Promenade - loisirs')], [0.42622])

([('agg', 2, 'En agglomération'), ('int', 1, 'Hors intersection')], [0.40003])

([('agg', 2, 'En agglomération'), ('sexe', 1, 'Homme')], [0.4245])

([('catv', 7, 'VL uniquement'), ('agg', 2, 'En agglomération')], [0.45489])

([('agg', 2, 'En agglomération'), ('lum', 1, 'Plein Jour')], [0.46438])

([('actp', 0, 'non spécifié ou non applicable'), ('catu', 1, 'Pilote'), ('atm', 1, 'Normal'), ('place', 1, 'Place occ
upée'), ('agg', 2, 'En agglomération')], [0.40017])

([('actp', 0, 'non spécifié ou non applicable'), ('lum', 1, 'Plein Jour'), ('sexe', 1, 'Homme')], [0.41886])

([('catv', 7, 'VL uniquement'), ('actp', 0, 'non spécifié ou non applicable'), ('sexe', 1, 'Homme')], [0.41847])

([('actp', 0, 'non spécifié ou non applicable'), ('int', 1, 'Hors intersection'), ('sexe', 1, 'Homme')], [0.43107])

([('actp', 0, 'non spécifié ou non applicable'), ('catu', 1, 'Pilote'), ('atm', 1, 'Normal'), ('sexe', 1, 'Homme'),
('place', 1, 'Place occupée')], [0.44633])

([('atm', 1, 'Normal'), ('int', 1, 'Hors intersection'), ('lum', 1, 'Plein Jour')], [0.41529])

([('actp', 0, 'non spécifié ou non applicable'), ('int', 1, 'Hors intersection'), ('lum', 1, 'Plein Jour')], [0.4406
1])
```

Cette méthode permet donc d'obtenir des motifs fréquents pouvant être utilisés par un expert pour déduire les causes les plus courantes des accidents de la route.

Discussion

Notre projet a nécessité un travail conséquent et nous sommes arrivés à des résultats satisfaisants. Néanmoins, nous y voyons plusieurs axes d'amélioration.

Pour la partie de recherche des motifs fréquents, il serait intéressant de créer des règles d'associations. Cela pourrait être intéressant pour trouver des corrélations entre les différents paramètres déclencheurs d'accidents. De plus, nous pourrions réfléchir à des méthodes de visualisation de motif. Cela permettrait d'avoir des résultats plus exploitables et plus compréhensibles pour tous.

Conclusion

Ce projet nous aura permis d'utiliser des méthodes apprises durant l'UE Data Mining. Nous avons mis en évidence différents types de clusters d'accidents s'étant déroulés en France de 2005 à 2016. Nous avons également cherché des motifs fréquents pour ces accidents.

Ces informations pourraient permettre à des organisations, des communes ou des collectivités de modifier leurs politiques d'entretien des routes. Ils pourraient également adapter voir améliorer leurs systèmes de prévention des accidents de la route.