Projet A3 - Big Data

Antonin SOQUET, Maxence LAURENT, Martin LOBEL

16 juin 2023

ISEN
école
d'ingénieurs

Table des matières

1 Introduction

Ce projet est découpé en 3 matières : Big Data / IA / Web et à pour principal objectif de concevoir et développer une application d'étude des accidents de la route.

Dans le cadre de notre étude à l'ISEN Nantes il nous permettra d'approfondir les compétences acquises dans les modules Big Data, Intelligence Artificielle, Développement Web et Base de Données à travers une application complète de traitements et de visualisation de données concernant les accidents corporels de la circulation routière en France.

Pour cela, nous avons du realiser des modification et des analyse des données de la base de donnée : Tout d'abord, nous avons vérifié la qualité des données en identifiant les éventuelles valeurs manquantes ou aberrantes. Nous avons ensuite effectué des transformations nécessaires pour rendre les données cohérentes et exploitables. Par exemple, nous avons converti les variables de type caractère en facteurs ou les variables de type date en format adéquat.

Ensuite, nous avons procédé à une exploration des données en calculant des statisques sur certaines valeurs de la table, comme par exemple le nombre d'accidents en fonction de la description de la surface. Cela nous a permis d'obtenir une vision générale des caractéristiques des accidents de la route en 2009.

Par la suite, nous avons réalisé des analyses plus approfondies en utilisant des techniques statistiques avancées. Par exemple, nous avons effectué des tests pour évaluer les différences significatives entre certains groupes de variables. De plus, nous avons réalisé des modèles de régression pour étudier les facteurs qui peuvent influencer la gravité des accidents.

2 Organisation

Nous estimons notre organisation durant ce projet de Big Data plutot bonne, puisque nous n'avons pas eu de soucis particuliers que ce soit dans la communication au sein du groupe ou dans la répartition des taches par exemple.

Comme évoqué précedemment, nous avons réussi à bien répartir les tâches entre les membres de notre équipe, ce qui nous a permis d'avancer efficacement dans notre travail. De plus, nous avons mis un point d'honneur à communiquer clairement sur les besoins spécifiques de chacun.

2.1 Outils

Pour mener ce projet a bien nous avons utilisé plusieurs outils. Comme environnement de developpement nous avons utilisé VS Code ou RStudio (dépend selon les membres du groupe). Pour un travail collaboratif et partagé nous avons utilisé GitHub qui est un service web d'hébergement et de gestion de développement de logiciels, utilisant le logiciel de gestion de versions Git.

FIGURE 1 – GitHub

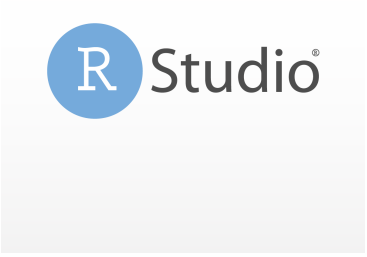


FIGURE 2 – RStudio

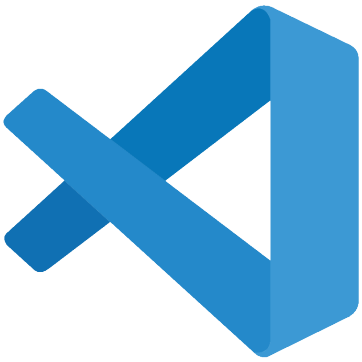


FIGURE 3 – Visual Studio Code

2.2 Répartition du travail

Nous avons décidé de créer un diagramme de Gantt sur GitHub pour des raisons organisationnelles. En utilisant GitHub, nous avons pu créer un référentiel centralisé pour notre projet, ce qui facilite la collaboration et la gestion des tâches. Chaque tache peut etre commenté et ainsi associée a un ou plusieurs commit, ce qui permet de suivre facilement les modifications et confère une traçabilité. En utilisant cette approche, nous pouvons efficacement coordonner nos travaux et travailler de manière organisée et construite.

Si vous souhaitez consulter le diagramme de Gantt il est visualisable via ce lien

On peut également voir la répartition des taches attribuées au debut du projet ainsi :

Antonin	Maxence	Martin
Répartition des tâches	Création du git	Répartition des tâches
Préparation des données P1	Répartition des tâches	Préparation des données P1
Visualisation des données P1	Visualisation des données P2	Préparation des données P2
Analyse des données P1	Rédaction du Rapport	Analyse des données P1
Rédaction du Rapport	Préparation soutenance	Rédaction du Rapport
Préparation soutenance	Réalisation de la diapo	Préparation soutenance
	Analyse des données P2	

P1 – Partie 1
P2 – Partie 2

3 Préparation des données

3.1 Description de la base de donnée

La base de données utilisée est le fichier stat_acc_V3.csv, qui est une base de données contenant toutes les informations sur les accidents de la route survenus en 2009. Cette base de données provient du site du gouvernement "Bases de données annuelles des accidents corporels de la circulation routière", qui regroupe toutes les informations sur les accidents de la période de 2005 à 2021.

La base de données est composée de 21 champs contenant des informations sur les 73 643 accidents survenus en 2009. Ces champs fournissent des informations sur le type de véhicule, le type de route et les conditions météorologiques lors de l'accident. Les valeurs dans ces champs peuvent être de différents types, tels que des caractères, des nombres, des dates ou encore des entiers.

Les parametre étudiés sont les suivants :

Num_Acc : Identifiant de l'accident.

num_veh : Identifiant du véhicule impliqué dans l'accident.

id_usa : Identifiant de l'usager impliqué dans l'accident.

date : Date de l'accident.

ville : Nom de la ville où l'accident a eu lieu.

id_code_insee : Identifiant INSEE de la ville où l'accident a eu lieu.

latitude : Latitude géographique de l'accident.

longitude : Longitude géographique de l'accident.

descr_cat_veh : Description de la catégorie du véhicule.

descr_agglo : Description du type d'agglomération où l'accident a eu lieu.

descr_athmo : Description des conditions atmosphériques au moment de l'accident.

descr_lum : Description de la luminosité au moment de l'accident.

descr_etat_surf : Description de l'état de surface de la route au moment de l'accident.

description_intersection : Description du type d'intersection où l'accident a eu lieu.

an_nais : Année de naissance de l'usager impliqué dans l'accident.

age : Âge de l'usager impliqué dans l'accident.

place : Place occupée par l'usager dans le véhicule.

descr_dispo_secu : Description de la disposition de sécurité utilisée par l'usager.

descr_grav : Description de la gravité de l'accident.

descr_motif_traj : Description du motif ou de la raison du trajet.

descr_type_col : Description du type de collision.

Pour analyser les données et effectuer divers traitements statistiques, nous avons utilisé le langage de programmation R. Nous avons importé la base de données sous le nom 'data' et avons procédé à différentes opérations de nettoyage et de préparation des données.

3.2 Partie 1 : Traitements

La partie du traitement des données avait pour objectif de nettoyer le jeu de données afin de le rendre utilisable pour les traitements ultérieurs. Dans cette étape, nous avons effectué plusieurs modifications sur les variables afin de les adapter à nos besoins afin de le rentre utilisable et ainsi pouvoir traiter les donnees par la suite.

Tout d'abord, nous avons dû convertir certaines variables de type caractère en format entier. Les variables 'an_nais', 'age' et 'place' étaient initialement en format caractère dans la table de données, ce qui nous empêchait de les manipuler correctement. Nous avons donc utilisé la fonction de conversion pour les transformer en format entier, ce qui nous a permis de les traiter plus facilement par la suite. De plus, nous avons également modifié le format de la variable de date en utilisant le format POSIXct, ce qui nous a permis de conserver non seulement la date de l'accident, mais aussi l'heure exacte, le format date normal ne permet de ne garder que la date.

Une autre partie importante du nettoyage des données consistait à gérer les valeurs manquantes. Nous avons remarqué que certaines valeurs étaient non assignées, indiquées par 'NA' dans la table. Pour remédier à cela, nous avons utilisé la fonction 'mean' pour calculer la valeur moyenne de chaque colonne, en excluant les valeurs manquantes. Nous avons ensuite assigné cette valeur moyenne à chaque valeur 'NA' dans les colonnes 'an_nais', 'age' et 'place'. Cette modification a permis de remplir les valeurs manquantes dans le jeu de données, le rendant ainsi utilisable pour nos analyses.

Une autre modification que nous avons apportée à la table était celle de la variable 'age', qui représentait l'âge des personnes impliquées dans les accidents. Cependant, les valeurs d'âge ne se limitaient pas à l'année de l'accident en 2009, mais allaient jusqu'en 2023. Afin de traiter correctement ces données, nous avons ajusté les valeurs d'âge pour qu'elles correspondent à l'année de l'accident.

Enfin, nous avons également identifié des valeurs aberrantes dans la table, notamment les coordonnées (latitude et longitude) des arrondissements des villes de Paris, Marseille et Lyon. Pour résoudre ce problème, nous avons choisi de remplacer ces coordonnées respectives par la localisation géographique générale de Paris, Marseille et Lyon, tout en conservant le numéro de l'arrondissement.

ville	latitude
PARIS 18	2009
PARIS 18	2009
PARIS 12	2009
PARIS 10	2009
PARIS 12	2009
PARIS 13	2009
PARIS 12	2009

Nous avons donc choisi de modifier toutes les coordonnes respectives des arrondissements de Paris, Marseille et Lyon avec la localisation de Paris, Merseille et Lyon. Cela a permis de changer la coordonner afin d'exploiter la donnee tout en gardant le numero d'arrondissement.

3.3 Partie 2 : Construction jeu de données séries chronologiques

Dans cette partie de traitement des données nous avons construit des plots à partir de séries chronologiques pour pourvoir visualiser l'évolution du nombre d'accidents par mois et par semaine. Nous avons analysé les données que nous avons collecté pour determiner le niveau d'agrégation le plus approprié pour réaliser des prévisions de qualité à l'aide d'une régression linéaire.

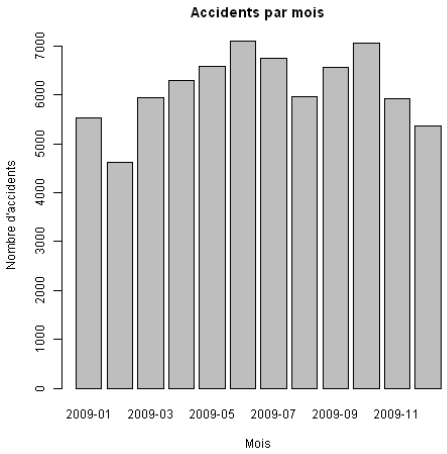


FIGURE 4 – Modèle pour les mois

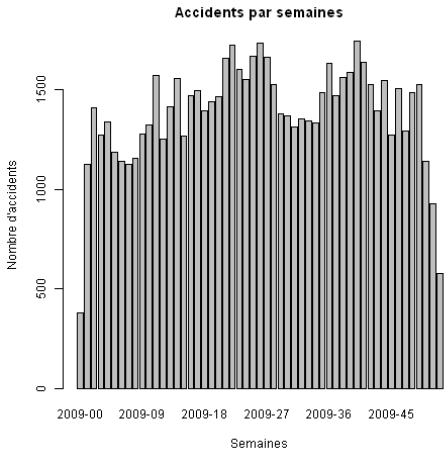


FIGURE 5 – Modèle pour les semaines

On remarque en comparant ces deux graphiques que les tendences sont les mêmes. Ceci est plutot logique puisque 4 semaines constituent 1 mois. On peut donc interpreter chaque mois comme étant une moyenne des 4 semaines qui le constitue. Le modèle par semaine est donc plus précis mais contiendra 4 fois plus de variables. On le voit bien d'ailleurs avec les deux images ci dessus.

Par la suite, nous avons créé un jeu de données comprenant le nombre d'accidents par gravité pour 100 000 habitants par région. Nous l'utiliserons dans une Analyse en Composantes Principales (ACP) discutée dans la section 3 de l'analyse de données. Plusieurs étapes sont necessaires pour créer ce jeu de donnée. Tout d'abord il nous a fallu trouver un autre fichier CSV comportant les regions de chaques villes. Ensuite nous avons pu "merge" ces deux CSV de manière a avoir les informations de régions sur notre CSV des accidents. Cela nous servira aussi pour la partie visualisation.

Ainsi, on obtient un jeu de donnée dans le format suivant :

(row)	descr_grav	region_name	number
1	0	auvergne-rhône-alpes	3475
2	0	bourgogne-franche-comté	1130
3	0	bretagne	1222
4	0	centre-val de loire	919
5	0	corse	120
6	0	grand est	3091
7	0	guadeloupe	110
8	0	guyane	16
9	0	hauts-de-france	2603
10	0	la réunion	283
11	0	martinique	86
12	0	normandie	1690
13	0	nouvelle-aquitaine	2773
14	0	occitanie	2532
15	0	pays de la loire	939
16	0	provence-alpes-côte d'azur	1646
17	0	île-de-france	8369
18	1	auvergne-rhône-alpes	2874
19	1	bourgogne-franche-comté	807
20	1	bretagne	1058

FIGURE 6 – 20 premières lignes du jeu de données

Ces modifications apportées à la table ont permis de nettoyer et de préparer les données pour les analyses ultérieures.

4 Visualisation des données

4.1 Visualisation 1

Pour réaliser les visualisations des données, nous avons exploré les interactions entre les variables et choisi les types de graphiques les plus appropriés. Grâce à la préparation préalable des données, nous avons pu utiliser l'ensemble des valeurs disponibles. Voici une description des trois premières visualisations que nous avons réalisées en utilisant des diagrammes en camembert (pie charts)

Les diagrammes en camembert sont particulièrement adaptés pour représenter des proportions et permettent une visualisation claire et intuitive des différentes catégories. Cependant, les diagrammes en camembert peuvent présenter certaines limitations, notamment lorsqu'il y a un grand nombre de catégories ou lorsque les proportions sont proches les unes des autres. Dans ces cas, d'autres types de graphiques, tels que les diagrammes à barres peuvent être plus appropriés pour une meilleure lisibilité des données.

Toutefois, nous avons également utilisé des histogrammes ou des diagrammes en barres dans d'autres cas de visualisations. Ces types de graphiques offrent plusieurs avantages, notamment la capacité à comparer plusieurs valeurs sur une même échelle et à mettre en évidence des différences significatives. Ils se révèlent être plus versatiles que les diagrammes en camembert, notamment lorsqu'il s'agit de créer des tranches, comme nous l'avons fait avec les heures des accidents, ou pour une visualisation plus chronologique en fonction des semaines ou des mois. Les histogrammes et les diagrammes en barres sont des outils adaptés pour comparer des valeurs avec la même échelle et rendre compte de variations significatives. Ils sont particulièrement utiles pour l'analyse de tranches de données ou de tendances temporelles. Ainsi, ils offrent une représentation visuelle précise et permettent une meilleure compréhension des données.

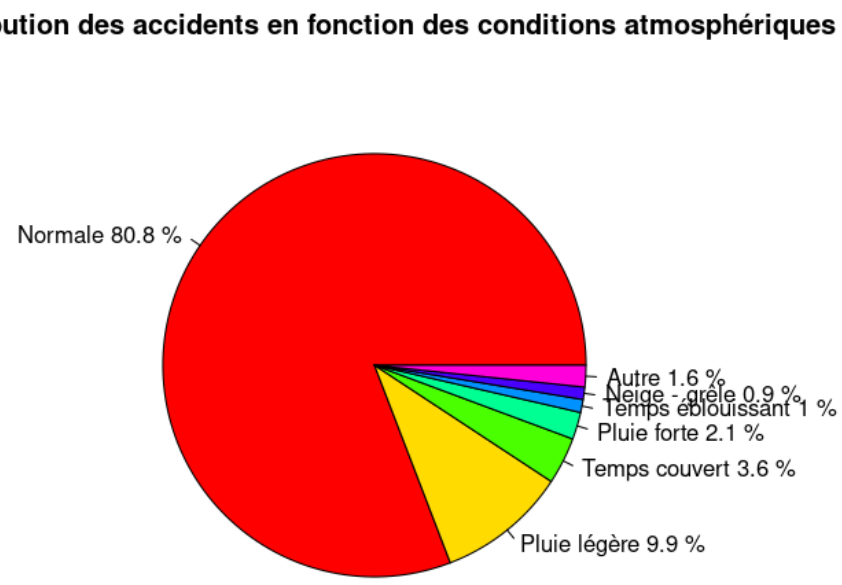


Diagramme en camembert des conditions météorologiques : Ce graphique montre la répartition des conditions météorologiques lors des accidents. Chaque part du camembert représente la proportion d'accidents survenus dans une condition météorologique spécifique, offrant une vue d'ensemble des conditions météorologiques les plus fréquentes lors des accidents. En examinant le nombre d'accidents en fonction des conditions atmosphériques, nous constatons que la majorité des accidents se produisent dans des conditions atmosphériques dites "normales" ou avec une pluie légère. À première vue, il ne semble pas évident de conclure que les conditions atmosphériques sont des facteurs propices aux accidents, contrairement à ce que l'on pourrait penser.

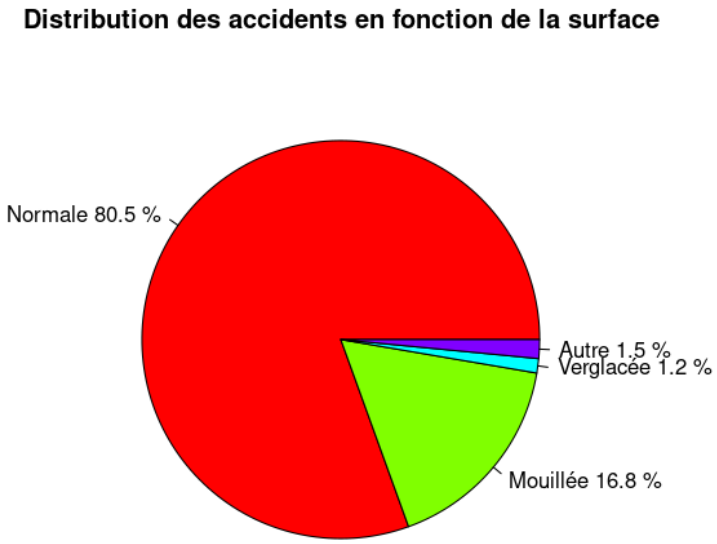
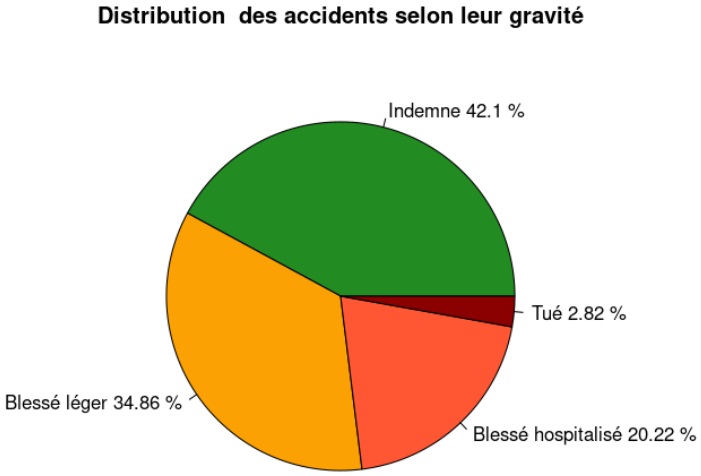
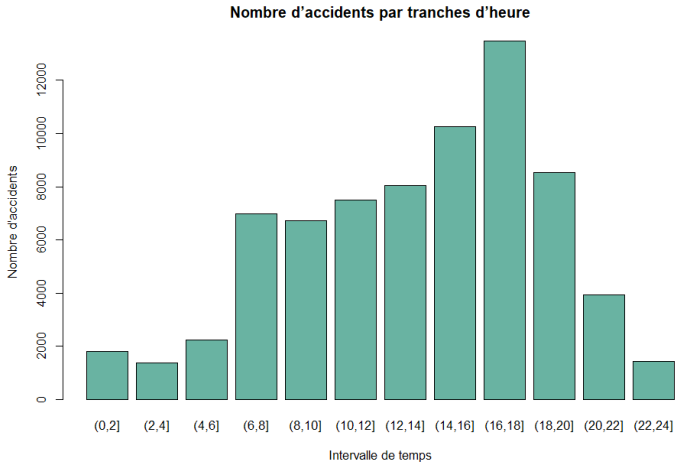


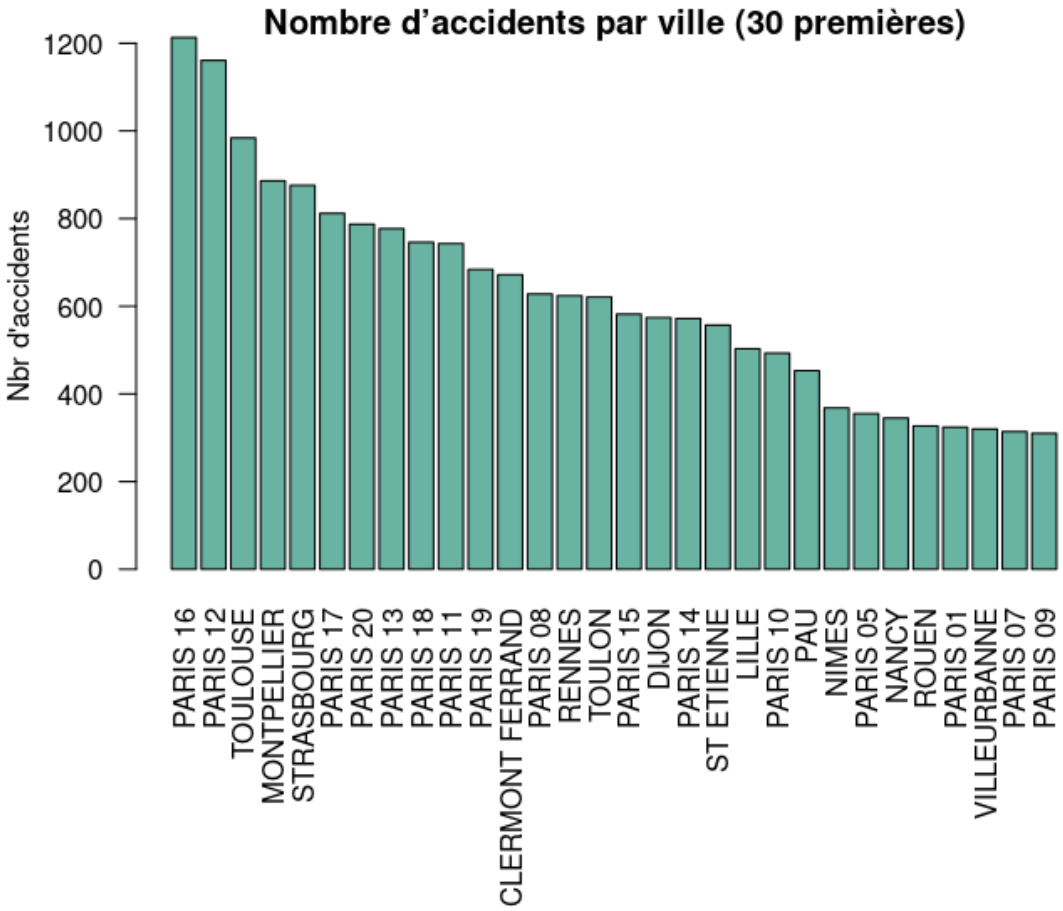
Diagramme en camembert des types de routes : Ce graphique présente la répartition des surfaces de routes où les accidents se sont produits. Chaque part du camembert représente la proportion d'accidents sur une surface de route donné. De manière similaire au cas précédent, nous avons remarqué que dans la majorité des cas (plus de 80%), la description de la surface de l'accident est considérée comme "normale". Les accidents liés à des routes verglacées, par exemple, ne représentent qu'une très faible proportion des accidents.



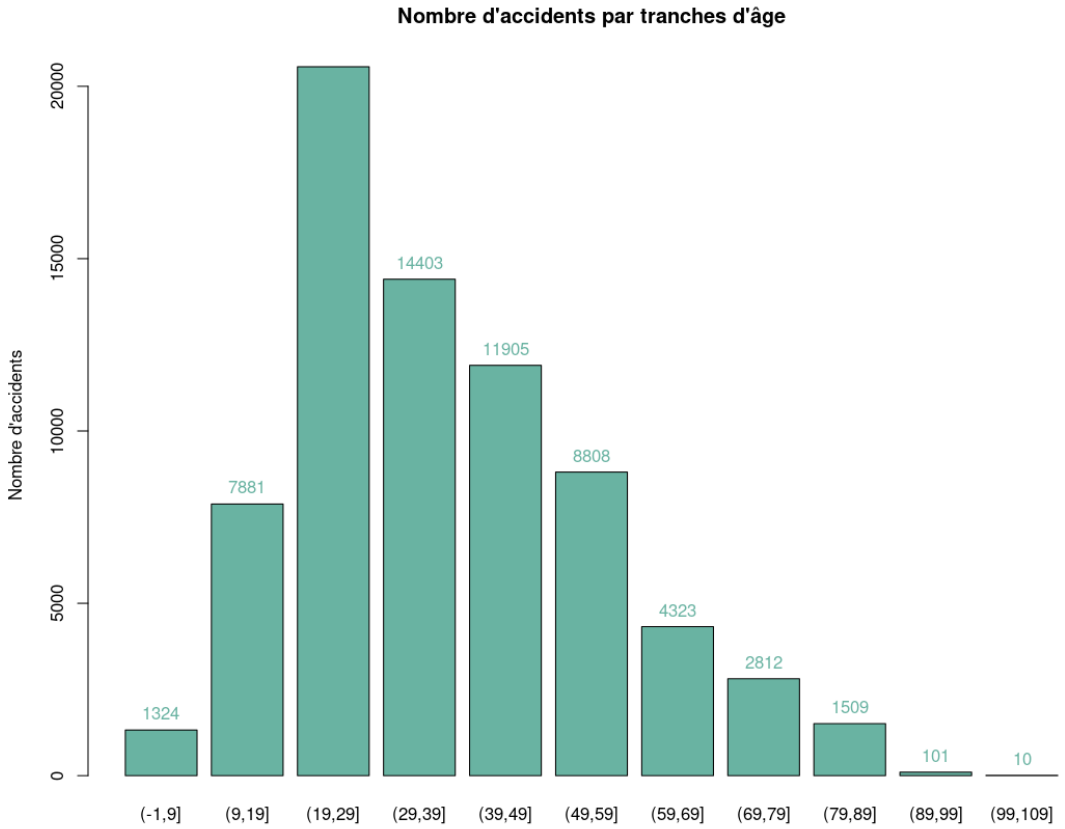
Le diagramme des accidents en fonction de la gravité permet de classifier les accidents en fonction de l'état de la personne victime de l'accident. On observe que dans près de 75% des cas, la personne est indemne ou seulement légèrement blessée. Cependant, il est important de noter que même si le pourcentage de décès sur les routes en 2009 est de 2,82%, cela représente, sur le nombre total d'accidents, un nombre considérable de décès chaque année en France.



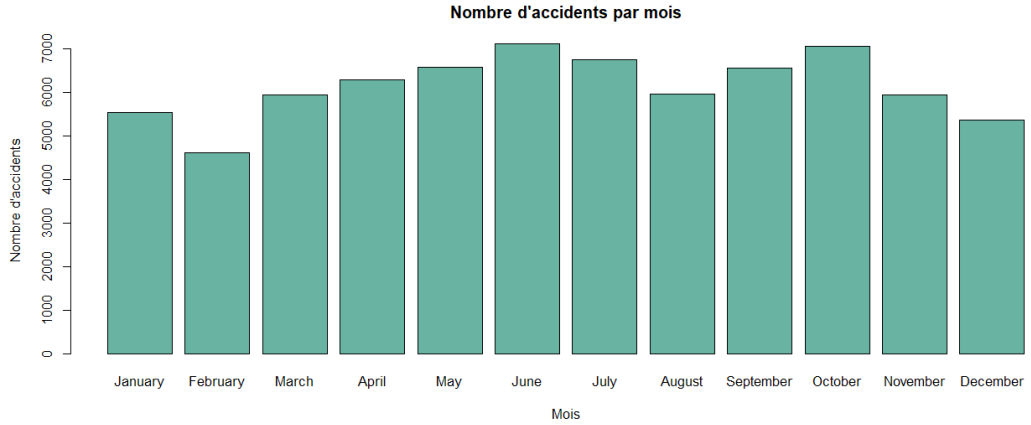
Le diagramme des accidents par tranches d'heure permet de visualiser les moments de la journée où les accidents sont les plus fréquents. Dans ce diagramme, qui est divisé en 12 tranches de 2 heures, nous pouvons rapidement observer que les accidents surviennent principalement en journée, avec un pic entre 16 et 18 heures, correspondant probablement à l'heure de sortie du travail et des écoles, entraînant un trafic plus dense.



L'histogramme du nombre d'accidents par ville permet de représenter l'ordre des villes ou arrondissements qui enregistrent le plus grand nombre d'accidents. Pour rendre le diagramme clair et significatif, nous avons sélectionné les 30 villes avec le plus d'accidents et les avons triées par ordre décroissant. Sans surprise, les grandes métropoles françaises se retrouvent en tête de liste, avec le plus grand nombre d'accidents. Cette observation met en évidence le fait que les zones urbaines densément peuplées et les centres urbains sont souvent associés à un volume plus élevé d'accidents.



Le diagramme des accidents par tranches d'âge permet d'estimer l'âge des victimes d'un accident. Dans cet histogramme, les valeurs d'âge des victimes de la base de données ont été divisées en 11 catégories, représentant chaque décennie. Une observation importante est que les personnes âgées de 20 à 30 ans sont les plus touchées par les accidents. Le diagramme des accidents par mois permet de visualiser les



variations des accidents au fil des périodes de l'année et de déterminer s'il existe des mois où les accidents sont plus fréquents. Cependant, nous ne pouvons pas observer de différence significative d'un mois à l'autre. Les accidents sont légèrement plus nombreux pendant l'été, mais cela peut être attribué à une augmentation des déplacements et des activités de loisirs liés aux vacances.

Afin de produire les visuels du nombre d'accident, et du taux d'accident, d'une part pour les regions et d'autre part pour les departements de France, nous avons du utiliser de nombreuses librairies ainsi que de trouver des bases de donnees supplémentaires afin de pouvoir representer ces données. Nous avons choisi de présenter des cartes de chaleures car elles nous permettent de bien représenter les avec des valeurs plus fortes ou plus nombreuses.

4.2 Visualisation 2

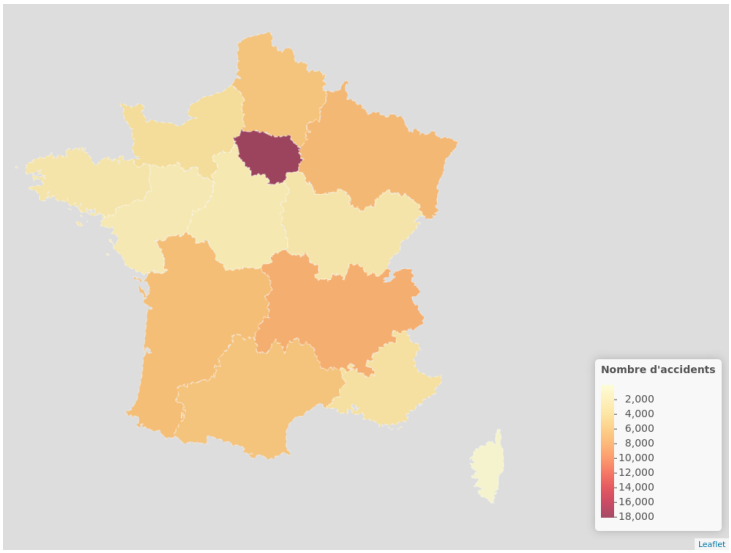


FIGURE 7 – Nombre d'accidents par régions (France

Sur cette cartes de chaleure, nous pouvons voir que le nombre d'accident sont assez répartis par régions sur le térritoire, sauf en Ile-de-France ou le nombre d'accident est très supérieur.

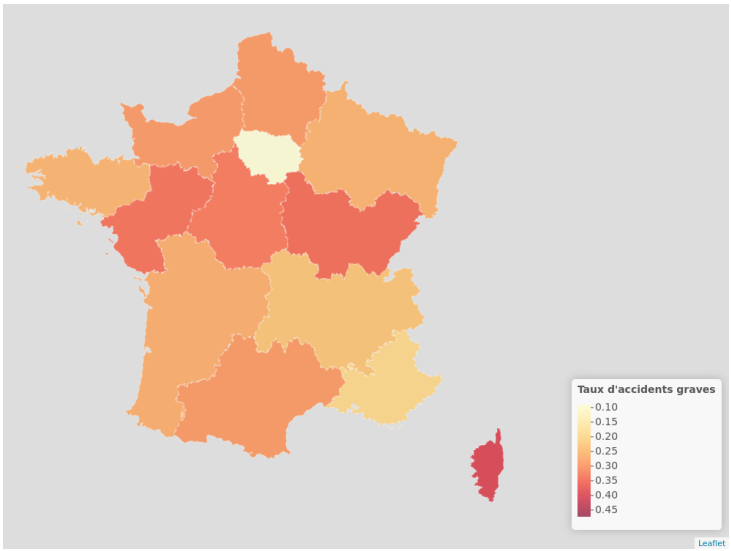


FIGURE 8 – Taux d'accidents graves par régions (France

Sur la carte du taux d'accident grave, nous pouvons constater que l'Ile-de-France n'est plus du tout mise en avant mais que des régions comme la Corse ont des taux d'accidents plus graves que la moyenne

Sur cette cartes,nous constqtons que certains departements n'ont constatés aucun accidents en 2009.

Sur la carte du taux d'accidents par départements, les taux sont plus répartis sur le térritoire. Le département Territoire de Belfort a un taux d'accident grave supérieur au reste.

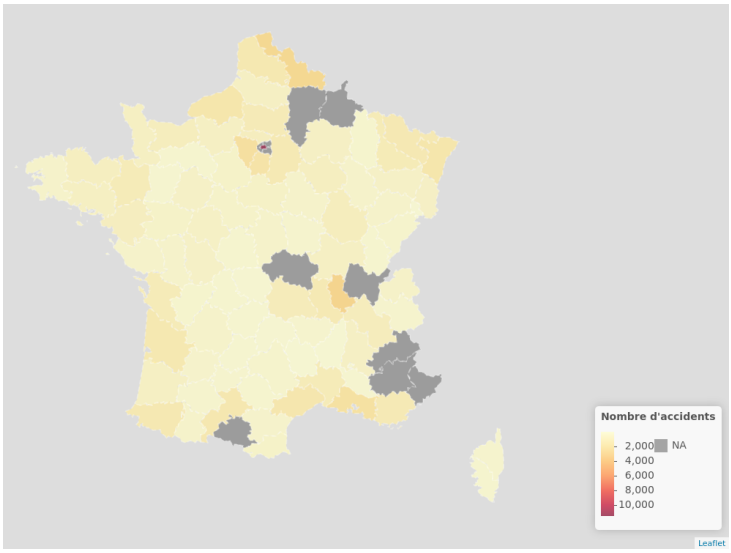


FIGURE 9 – Nombre d'accidents par départements (France

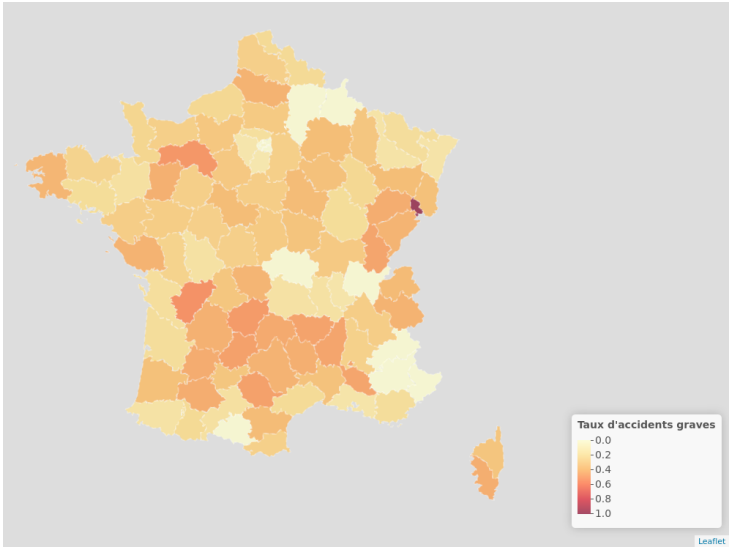


FIGURE 10 – Taux d'accidents graves par département (France

5 Analyse des données

5.1 Etude des relations entre variables qualitatives

Dans cette partie du projet, nous avons effectué une étude des relations entre variables qualitatives. Pour ce faire, nous avons réalisé des tableaux croisés et des tests d'indépendance du chi2 sur ces tableaux, en prenant en compte les différentes variables. Ensuite, nous avons représenté graphiquement ces tableaux à l'aide de la méthode du mosaicplot et les avons analysés en détail. Nous le verrons juste apres. D'abord commençons par expliquer les tests d'indépendance du chi2 qui sont utilisés pour évaluer la dépendance ou l'indépendance entre deux variables qualitatives. Ces tests permettent de déterminer si les différences observées dans les fréquences des catégories entre les variables sont statistiquement significatives.

Le test d'indépendance du chi2 repose sur l'hypothèse nulle selon laquelle les variables sont indépendantes les unes des autres dans la population sous-jacente. L'idée est de comparer les fréquences observées dans les cellules d'un tableau croisé avec les fréquences attendues si les variables étaient indépendantes.

La p-value (valeur de probabilité) est un résultat statistique obtenu lors du test d'indépendance du chi2. Elle indique la probabilité d'obtenir une statistique du chi2 aussi extrême que celle observée, en supposant que les variables sont réellement indépendantes.Lorsque la p-value est faible, cela signifie que les données observées sont peu probables sous l'hypothèse nulle d'indépendance, ce qui renforce l'idée qu'il existe une dépendance entre les variables.

Ensuite on a aussi créer un mosaicplot nous permetant de visualiser les tableaux.

Etude entre la gravité et les conditions athmosphériques

On obtient le tableau ci-dessous qui est le tableau croisé entre les conditions athmosphériques et la gravité des accidents ("Indemne" = 0, "Blessé léger" = 1, "Blessé hospitalisé" = 2, "Tué" = 3).

```
   Autre Brouillard - fumée Neige - grêle Normale Pluie forte Pluie légère
0   251               134          292   25244        668         2983
1   212               114          256   20544        525         2904
2   179               135          129   12114        300         1235
3    32                32           19    1630         39          174

   Temps couvert Temps éblouissant Vent fort - tempête
0           1044               342                  46
1            916               167                  34
2            580               184                  35
3            102                38                  10
[1] 7.881203e-52
```

FIGURE 11 – Tableau croisé entre la gravité et les conditions athmosphériques

On remarque deja a première vue que les accidents sont de plus grand nombre pour des conditions normales. On pourrait alors se poser la question d'une incohérence, mais en fait non. Ceci est lié au fait que les conditions athmosphériques "Normales" ont un taux d'apparition bien plus important que la neige ou la grêle par exemple qui ont un taux d'apparition bien plus faible. Cependant on remarque que dans des contionstel que les pluie on a proportionnellement beaucoup d'accidents sur les temps pluvieux et brouillardeux par exemple.

Dans ce cas la on obtient une p.value de $7{,}881203^{-52}$.

Ceci est très proche de 0. Ainsi on en déduit que les variables de gravité de l'accident et de conditions athmosphériques sont dépendantes l'une de l'autre.

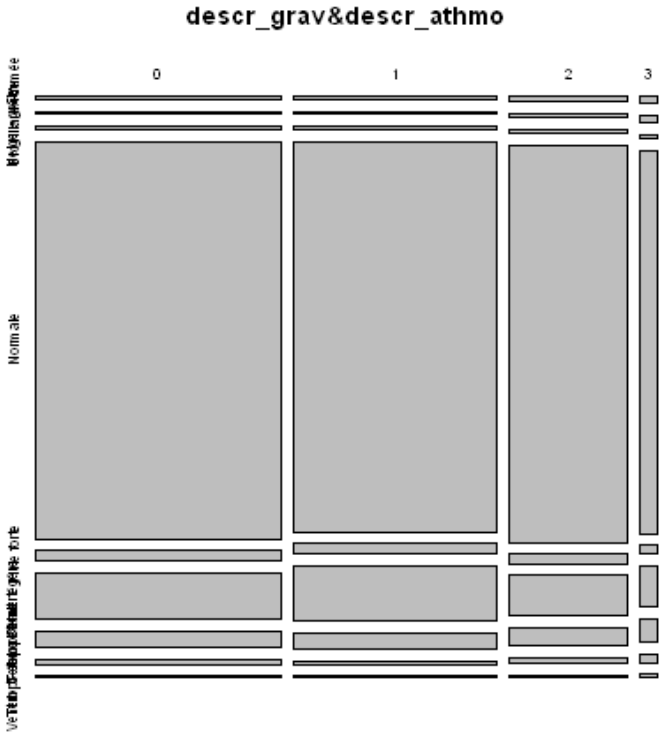


FIGURE 12 – Mosaicplot entre la gravité et les conditions athmosphériques

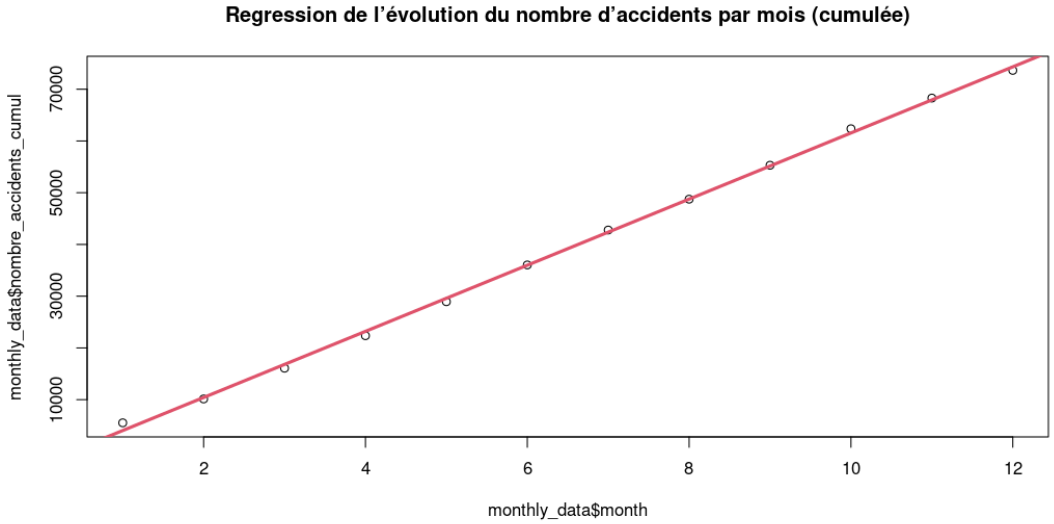
Les chiffres ne sont pas si facilement exploitables à première vue, on a du mal a visualiser ces chiffres. Pour palier ce soucis, nous avons construit un "mosaicplot" (visualisable ci dessous) qui nous permet d'avoir une vue globale sur ces chiffres.

Comme dit précédemment, les deux variables sont dépendantes. On peut aussi le remarquer sur le graphique en mosaïque par la présence de décallages au niveau de la taille des barres. En négligeant la valeure "Normale" qui a un taux de présence abérant par rapport aux autres conditions athmosphériques on peut voir que les valeurs du mosaicplot sont variables, ce qui traduit une relation entre les variables.

5.2 Régressions linéaires

Nous avons donc realises 2 regressions lineaires grace aux donnes de la table afin de verifier si les donnnes de la table sont coherentes. La premiere regression est entre le nombre d'accident par mois puis le nombre d'accident par semaine. Apres avoir fais la regression sur ces variables, nous nous sommes rendu compte qu'il n'y avait pas de correlation et qu'on ne pouvais pas obtenir de resultat exploitable pour les regressions. En effet comme les histogrammes nous ont montres plus tot, il y a tres peu de relation entre le mois de l'annee et les accidents, idem pour les semaines. Cependant, nous avons ensuite change notre modele pour prendre les valeurs des accidents par mois et par semaine cumules. A partir de ce moment la, les valeurs obtenues lors de la regressino ce sont trouves etre beaucoup plus interessante.

Nos resultats obtenus pour la regression des accidents par mois cumulee :



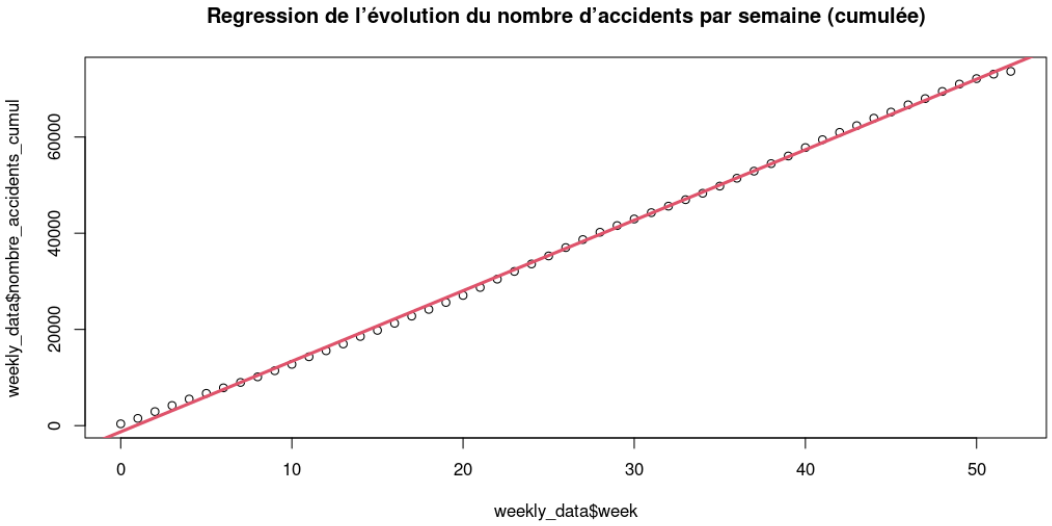
Nous avons traces en rouge, la ax+b line, de parametres a = -2351 et b = 6390 Nous pouvons donc reecrire le nombre d'accident comme etant : nombre d'accident = 6390 * (mois de l'annee) - 2351

```
Call:
lm(formula = monthly_data$nombre_accidents_cumul ~ monthly_data$month)

Residuals:
    Min      1Q  Median      3Q     Max
-848.05 -672.11   14.11  365.28 1486.28

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2351.39     454.37  -5.175 0.000416 ***
monthly_data$month  6390.11      61.74 103.507  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 738.3 on 10 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.999
F-statistic: 1.071e+04 on 1 and 10 DF,  p-value: < 2.2e-16
```

Lors de cette régression, nous avons obtenu une p-value de 2,2e-16. Étant donné que cette p-value est inférieure à 0,05, nous pouvons rejeter l'hypothèse nulle et conclure qu'il existe une relation significative entre les variables.

De plus, nous avons obtenu un coefficient de détermination R^2 de 0,9991 et un R^2 ajusté de 0,999 pour cette régression. Un coefficient de détermination proche de 1 indique que la ligne de régression linéaire correspond très bien aux données. Nous pouvons donc conclure que la variable indépendante explique une grande partie de la variation de la variable dépendante.

Regardons maintenant la regression du nombre d'accident par semaine cumulee :



Nous avons traces en rouge, la ax+b line, de parametres a = -1271 et b = 1465 Nous pouvons donc reecrire le nombre d'accident comme etant : nombre d'accident = 1465 * (semaine de l'annee) - 1271

```
Call:
lm(formula = weekly_data$nombre_accidents_cumul ~ weekly_data$week)

Residuals:
     Min       1Q   Median       3Q      Max
-1306.09  -524.89    -7.29   440.19  1648.13

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1271.128    182.663  -6.959 6.32e-09 ***
weekly_data$week  1465.773      6.055 242.068  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 674.3 on 51 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 5.86e+04 on 1 and 51 DF,  p-value: < 2.2e-16
```

Lors de cette régression, nous avons obtenu une p-value de 2,2e-16. Étant donné que cette p-value est inférieure à 0,05, nous pouvons rejeter l'hypothèse nulle et conclure qu'il existe une relation significative entre les variables.

De plus, nous avons obtenu un coefficient de détermination R^2 de 0,9991 et un R^2 ajusté de 0,9991 pour cette régression. Un coefficient de détermination proche de 1 indique que la ligne de régression linéaire correspond très bien aux données. Nous pouvons donc conclure que la variable indépendante explique une grande partie de la variation de la variable dépendante.

Étant donné que les valeurs de R^2 des deux régressions sont quasiment identiques, nous ne pouvons pas affirmer qu'un modèle est plus pertinent que l'autre. En ce qui concerne les prédictions, le modèle basé sur les mois prédit un nombre d'accidents de 74329 sur une année, tandis que le modèle basé sur les semaines prédit un nombre d'accidents de 74909. Étant donné qu'il y a eu 73643 accidents en 2009, on peut dire que la prédiction basée sur les mois est légèrement plus précise que celle basée sur les semaines.

5.3 Analyse 3

Dans cette partie nous avons fait une ACP. D'abor commencons par expliquer de quoi il s'agit. L'analyse en composantes principales est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées « composantes principales » ou axes principaux. Elle permet au statisticien de résumer l'information en réduisant le nombre de variables.

On obtient donc l'ACP suivant :

```
PCA(X = data_acp, scale.unit = TRUE, ncp = 5, graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
Variance               3.927   1.165   0.539   0.353   0.016   0.000
% of var.             65.456  19.414   8.982   5.875   0.268   0.005
Cumulative % of var.  65.456  84.870  93.852  99.728  99.995 100.000

Individuals (the 10 first)
                  Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1              |  2.480 |  2.154  6.948  0.754 |  0.969  4.738  0.153 |  0.256  0.717  0.011 |
2              |  0.845 | -0.108  0.017  0.016 |  0.048  0.012  0.003 | -0.698  5.322  0.683 |
3              |  0.691 | -0.486  0.354  0.495 |  0.342  0.590  0.245 |  0.321  1.126  0.216 |
4              |  0.924 | -0.648  0.629  0.492 | -0.082  0.034  0.008 | -0.570  3.547  0.381 |
5              |  2.620 | -1.217  2.219  0.216 |  1.339  9.060  0.261 |  1.468 23.526  0.314 |
6              |  2.369 |  2.217  7.361  0.876 |  0.218  0.240  0.008 | -0.712  5.536  0.090 |
7              |  2.550 | -2.379  8.474  0.870 | -0.854  3.682  0.112 | -0.279  0.849  0.012 |
8              |  3.622 | -3.268 15.997  0.814 | -0.853  3.675  0.055 | -0.093  0.095  0.001 |
9              |  1.871 |  1.517  3.446  0.657 |  0.113  0.064  0.004 | -1.045 11.924  0.312 |
10             |  5.294 |  4.065 24.747  0.589 | -3.185 51.219  0.362 |  1.167 14.867  0.049 |

Variables
                 Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
region_name   |  0.353  3.166  0.124 |  0.814 56.899  0.663 |  0.460 39.185  0.211 |
0             |  0.886 20.007  0.786 | -0.393 13.291  0.155 |  0.209  8.128  0.044 |
1             |  0.845 18.169  0.714 | -0.455 17.802  0.207 |  0.258 12.354  0.067 |
2             |  0.942 22.598  0.888 |  0.181  2.822  0.033 | -0.221  9.042  0.049 |
3             |  0.840 17.945  0.705 |  0.327  9.186  0.107 | -0.410 31.246  0.168 |
taux_accident |  0.843 18.115  0.711 |  0.001  0.000  0.000 | -0.015  0.044  0.000 |
```